الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا

ماجستير في اختصاص المعلوماتية

التعرف على نشاط المجموعات باستخدام مقاربة معتمدة على الرؤية الحاسوبية

أُعدت هذه الأطروحة لنيل

شهادة الماجستير في نظم المعطيات الكبيرة

إعداد

غريس نعمة

إشراف

د. آصف جعفر د. ياسر رحال

دمشق 2023

Syrian Arabic Republic

Higher Institute for Applied Sciences and Technology

Computer Vision Based Approach For Group Activity Recognition

Submitted in Fulfillment of the Requirements for Master's Degree in Big Data Specialty

By

Grace Nemeh

Supervised By

Dr. Assef Jaafar Dr. Yasser rahal

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدثت بموجب المرسوم التشريعي رقم /24/ لعام 1983، وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي لدراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم والروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

إلى جانب النشاط التعليمي، يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات أطره العلمية ومختبراته. واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشابكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب 31983

Higher Institute for Applied Sciences and Technology – HIAST

P. O. Box 31983, Damascus, Syrian Arab Republic

هاتف 00963115123819 - فاكس 00963115140761

بريد إلكتروني contact@hiast.edu.sy

موقع إلكتروني www.hiast.edu.sy

تصريح

أنا الموقعة أدناه غريس نعمة معدّة أطروحة الماجستير التي تحمل العنوان: التعرف على نشاط المجموعات باستخدام مقاربة معتمدة على الرؤية الحاسوبية.

أصرح بأن:

- الأعمال والنتائج المعروضة في هذه الأطروحة هي نتيجة جهودي الشخصية وبتوجيه من المشرف، وأن ما عدا ذلك من معلومات ونتائج قد نُسبت إلى مصادرها ومؤلفيها، وأُشير إلى ذلك في متن النص وفي قائمة المراجع.
- المعطيات والمعلومات المستخدمة في هذه الأطروحة جرى تحصيلها بطرائق سليمة ومشروعة ونُسبت إلى مصادرها في المواضع الملائمة.
- كلّ مكوّن من مكونات هذه الأطروحة (مقطع نصّي، صورة، مخطط، ...) مقتبس من عمل آخر جرى تمييزه بوضوح ونُسب إلى مصدره.
- الأعمال والنتائج المعروضة في هذه الأطروحة لم تُستخدم سابقًا وليست قيد الاستخدام للحصول على أي شهادة أكاديمية أخرى.

التوقيع

دمشق 10/12 /2023

إهداء

إلى الجبل الذي يحميني من كل عواصف الحياة، السند الذي لا ينكسر، والقلب الذي لا يتهاون مع أحزاني

. . . . . **إلى أبي العزيز**

إلى من أخصّ الله الجنة تحت قدميها، إلى التي كانت تمسك بيدي كأنّني طفلة صغيرة وتؤمن بأنني سأكون يوماً ما أريد

. . . . . **إلى أمي الغالية**

إلى من وقف بجانبي طيلة سنوات الدراسة، إلى من تحمّل معي وتحمّلني وساعدني لتحقيق أهدافي، صديقي ورفيق دربي

وشريك أيامي . . . . . **مارسيل**

إلى من يسري حبهم في عروقي، إلى من تسعد عيني برؤية وجوههم، ويفرح فؤادي بسماع رنّات ضحكاتهم

أخوتي الأعزاء . . . . . **ماريا، كارلا**

إلى من لم يبخلن يوماً بدعواتهم الصادقة . . . . . **جدتيّ** . . . **إيلين، ليلى**

إلى من غمروني بنوايا هم الطيبة وشاركوني أجمل اللحظات . . . . . **عائلتي الثانية** . . . . **نزار، وداد، لؤي، شارلي، ماريو**

إلى من تمنّى لي الخير، وأرشدني إلى الدخول في هذا التّحدي، من كان نعم صديق . . . . . **مجد عيسى**

إلى من يهدأ ضجيج قلبي عند لقياه، من قدم لي الكثير دون أن يعلم . . . . . **ابن أختي** . . . . . **كريستيان عيّاش**

إلى كل من قدّم لي دعماً ولو كان مجرد كلمة . . أنا ممتن

شكر

يسرّني أن أوجه شكري لكل من نصحني أو أرشدني أو وجهني في إعداد هذا البحث بإيصالي للمراجع والمصادر المطلوبة في أي مرحلة من مراحله، وأقدم على وجه الخصوص خالص الشكر للدكتور آصف جعفر والدكتور ياسر رحّال اللذان لم يتهاونا عن تقديم النصائح والإرشاد لإنجاز هذا البحث، وإدارة المعهد العالي للعلوم التطبيقية والتكنولوجيا لتوفير البيئة المناسبة للبحث العلمي.

Abstract

The field of group activity recognition has witnessed a paradigm shift with the emergence of deep learning techniques. Modern methods often focus on individual analysis, such as tracking people or discovering key points of human body and studying people's relationships in different ways. These methods need an accurate description of all individual actions, bounding boxes and group activities, which makes the task of providing new domain-specific datasets demanding. Significantly, these approaches often ignore the context. This research presents a new approach based on the principle of adopting comprehensive scene analysis to recognize group activity. The goal is to interpret the entire scene as an integrated entity, based on our interest in bringing the solution closer to the human level of processing, rather than breaking it down into explicit components. In this approach we only need group activity labels.

Our methodology uses advanced deep learning architectures, specifically the video Swin transformer, which is introduced in this research for the first time to solve the problem of group activity recognition. This is done by building hierarchical relationships between the elements of the scene inherent in its structure in the temporal and spatial domains.

The evaluation was done using the accuracy metric according to what has been done in related works, on the widely used Volleyball dataset, and an accuracy of 85.3% was achieved for 8 classes and 93% for 6 classes of the mentioned data set.

Our results are promising and have positive implications, including the possibility of creating large, domain-specific datasets for research that is labeled only for the group's activity.

Keywords: Group Activity Recognition, Deep Learning, video Swin Transformer, hierarchical relationships, accuracy, Volleyball.

الملخص

شهد مجال التعرف على الأنشطة الجماعية تحولاً نموذجياً مع ظهور تقنيات التعلم العميق. الأساليب الحديثة غالباً ما تركز على تحليل الكيان الفردي مثل تتبع المسار للأشخاص أو اكتشاف النقاط الرئيسية في جسم الإنسان ودراسة علاقات الأشخاص بطرائق مختلفة. تحتاج هذه الأساليب إلى وصف دقيق لجميع الإجراءات الفردية ومربعات الإحاطة والإجراء الجماعي وهذا ما يجعل مهمة توفير مجموعات معطيات جديدة خاصة بالمجال تتطلب جهداً كبيراً، كما أن هذه الأساليب غالباً ما تهمل السياق. يقدم هذا البحث نهجاً جديداً ينطلق من مبدأ اعتماد تحليل المشهد الشامل للتعرف على نشاط المجموعة. الهدف هو تفسير المشهد بأكمله ككيان متكامل، انطلاقاً من اهتمامنا في تقريب الحل من المستوى البشري في المعالجة، بدلاً من تجزئته إلى مكونات صريحة. نحتاج في هذا النهج فقط إلى وسوم نشاط المجموعة.

تستخدم منهجيتنا بنى التعلم العميق المتقدمة، وبالتحديد محوّل video Swin، الذي تم تقديمه في هذا البحث لأول مرة في حل مسألة التّعرف على نشاط المجموعة، من خلال بناء العلاقات الهرمية بين عناصر المشهد المتأصلة في بنيته في المجالين الزماني والمكاني.

تم التقييم وفق معيار الدقة وفقاً لما تم اتباعه في الأعمال ذات الصلة، على مجموعة المعطيات المستخدمة بشكل واسع النطاق Volleyball وتم تحقيق دقة 85.3% من أجل 8 صفوف و93% من أجل 6 صفوف من مجموعة المعطيات المذكورة.

النتائج التي توصلنا إليها تعتبر واعدة ولها آثار إيجابية منها إمكانية تشكيل مجموعات معطيات كبيرة للبحث تستخدم فقط وسم لصنف نشاط المجموعة.

الكلمات المفتاحية: التعرف على نشاط المجموعة، التعلم العميق، محولvideo Swin، العلاقات الهرمية، الدقة، Volleyball.

فهرس المحتويات

الفصل الأول: مدخل إلى موضوع البحث وأهدافه..... 1

1.1. تمهيد:..... 1

2.1. الهدف من البحث:..... 1

3.1. لمحة عن الحل المقترح:..... 2

4.1. مساهمات البحث:..... 2

5.1. مخطط الأطروحة:..... 3

الفصل الثاني: مفاهيم نظرية..... 4

1.2. التعرف على النشاط البشري:..... 4

2.2. أنواع الأنشطة البشرية:..... 5

1.2.2. إجراءات المجموعة والرؤية الحاسوبية:..... 6

3.2. أحدث الأساليب الأساسية المتبعة في معالجة مسائل الرؤية الحاسوبية:..... 7

1.3.2. الشبكات العصبية الالتفافية:..... 7

2.3.2. الشبكات العصبية العوّدية: 8

3.3.2. المرمز ومفكك الترميز : 9

4.3.2. المحولات:..... 10

4.2. خاتمة..... 19

الفصل الثالث: الدراسة المرجعية..... 20

1.3. تمهيد:..... 20

2.3. المفاهيم الأساسية /الدراسات /والنظريات:..... 20

1.2.3. المناهج المعتمدة على الميزات اليدوية:..... 20

2.2.3. النهج المعتمد على التعلم العميق: 21

3.3. المحوّل في مجال الرؤية الحاسوبية:..... 25

1.3.3. لمحة تاريخية: 25

2.3.3. محوّل التصنيف ViT [32]: 25

3.3.3. المحوّل Swin [1] : 27
4.3.3. نمذجة الفيديو قبل Video SWIN:..... 34
4.3. المقاربات الحديثة في هذا المجال:..... 35
1.4.3. Actor-Transformers for Group Activity Recognition: 35
2.4.3. Social Adaptive Module for Weakly-supervised Group Activity Recognition : 40
3.4.3. Pose is all you need: The pose only group activity recognition system (POGARS): 42
5.3. خاتمة..... 45
الفصل الرابع: الحل المقترح..... 47
1.4. تمهيد:..... 47
2.4. نموذج Video Swin Transformer..... 48
1.2.4. مقدمة:..... 48
2.2.4. التهيئة من النموذج المدرب مسبقاً SWIN: 48
3.2.4. البنية العامة لنموذج Video Swin: 49
4.2.4. إصدارات النموذج:..... 52
5.2.4. نتائج نموذج Video Swin:..... 52
3.4. النموذج المقترح:..... 54
1.3.4. مراحل تنجيز النموذج:..... 54
2.3.4. تفاصيل التنفيذ: 58
4.4. أسباب اختيار النموذج وتوقع نجاحه:..... 58
5.4. خاتمة..... 59
الفصل الخامس: الاختبارات والنتائج..... 60
1.4. مقدمة..... 60
2.5. مجموعة المعطيات المستخدمة volleyball:..... 60
1.2.5. شرح توضيحي تفصيلي عام:..... 61
2.2.5. تفاصيل التطبيق:..... 62

3.2.5. أمثلة من مجموعة المعطيات: 62
3.5. معايير التقييم:..... 65
1.3.5. الدقة Accuracy:..... 65
2.3.5. مصفوفة الإرباك Confusion Matrix:..... 65
4.5. النتائج:..... 66
1.4.5. النتائج التي تم الحصول عليها بتدريب رأس النموذج فقط: 66
2.4.5. النتائج التي تم الحصول عليها بتدريب النموذج كاملاً:..... 71
5.5. خاتمة:..... 86
الفصل السادس: الخاتمة والآفاق المستقبلية..... 87
1.6. الخاتمة:..... 87
2.6. الآفاق المستقبلية:..... 88
1.2.6. المقترح 1:..... 88
2.2.6. المقترح 2:..... 89
المراجع..... 90

قائمة الأشكال

الشكل 1 الخطوات النموذجية للتعرف على النشاط البشري القائم على الرؤية [3]..... 4
الشكل 2 مستويات النشاط البشري [4]..... 6
الشكل 3 الفرق بين مرور المعلومات بين الشبكات العودية – اليمين، والشبكات التقليدية – اليسار [9]..... 8
الشكل 4 نموذج seq2seq مع بنية encoder–decoder [56] 9
الشكل 5 البنية الأساسية للمحول [14]..... 11
الشكل 6 البنية الأساسية للمرمز و مفكك الترميز داخل المحول [17] 11
الشكل 7 تطبيق التوابع داخل المحول لكل عنصر دخل بشكل تفرعي [17]..... 12
الشكل 8 الخطوة الأولى لحساب الانتباه الذاتي [17] 14
الشكل 9 الخطوة 2 إلى 5 من حساب الانتباه الذاتي [17]..... 15
الشكل 10 الحصول على مصفوفات Q و K و V [17] 16
الشكل 11 حساب الانتباه الذاتي في شكل مصفوفة [17]..... 16
الشكل 12 (يسار) الانتباه بالجداء السلمي المُقاس. (يمين) كتلة الانتباه متعدد الرؤوس المؤلفة من h طبقة انتباه تعمل على التوازي [14] 18
الشكل 13 تقسيم الصورة الى رقع في مرحلة Patch Partitioning [52] 29
الشكل 14 مقارنة آلية الانتباه الذاتي بين VIT و Swin [52] 30
الشكل 15 نهج النافذة المزاحة في بنية Swin [1] 31
الشكل 16 تمثيل النوافذ المزاحة باستخدام excel [53]..... 32
الشكل 17 كتلة ترميز المحولات بالمقارنة مع كتلة Swin Transformer [52] 32
الشكل 18 البنية الشاملة لـ Swin Transformer في الإصدار Tiny [1]..... 33
الشكل 19 بنية نموذج Actor–Transformers [40]..... 38
الشكل 20 نظرة عامة على النهج المقترح في SAM [41]..... 41
الشكل 21 ملخص لآلية عمل نظام POGARS [42]..... 45
الشكل 22 الهيكل العام لبنية نموذج Video Swin Transformer بنسخته Tiny [7] 50
الشكل 23 مثال توضيحي للنوافذ المزاحة ثلاثية الأبعاد [7]..... 51
الشكل 24 بنية نموذج نهجنا المقترح..... 54
الشكل 25 صور من الصفوف اليمنى من مجموعة البيانات المستخدمة 63
الشكل 26 صور من الصفوف اليسرى من مجموعة البيانات المستخدمة..... 64
الشكل 27 تغيرات دالّة الخسارة أثناء تدريب رأس النموذج المقترح فقط..... 67
الشكل 28 تغيرات دقة النموذج المقترح أثناء تدريب الرأس فقط..... 68
الشكل 29 تغيرات معدل التعلم للنموذج المقترح أثناء تدريب الرأس فقط..... 68

الشكل 30 تغيرات دقة اختبار النموذج المقترح بتدريب الرأس فقط 69
الشكل 31 مصفوفة الإرباك لاختبار النموذج بتدريب الرأس فقط..... 70
الشكل 32 مصفوفة إرباك النموذج المقترح بعد إجراء normalization على مستوى السطر..... 71
الشكل 33 تغييرات معدل التعلم للنموذج أثناء تدريب النموذج كاملاً 72
الشكل 34 تغيرات دقة تدريب النموذج المقترح بتدريب النموذج كامل 72
الشكل 35 تغيرات دالّة الخسارة أثناء تدريب النموذج المقترح كاملاً 73
الشكل 36 تغيرات دقة الاختبار للنموذج المقترح بتدريبه كاملاً 74
الشكل 37 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بتدريب النموذج كاملاً..... 75
الشكل 38 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بتدريب النموذج كاملاً بعد إجراء استنظام 76
الشكل 39 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بتدريب النموذج كاملاً 77
الشكل 40 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بتدريب النموذج كاملاً بعد إجراء استنظام..... 78
الشكل 41 تغيرات دالّة الخسارة أثناء تدريب النموذج بعد دمج الصفين set, pass 81
الشكل 42 تغيرات الدقة أثناء تدريب النموذج بعد دمج الصفين set,pass 81
الشكل 43 تغيرات معدّل التّعلم أثناء تدريب النموذج بعد دمج الصفين set,pass..... 82
الشكل 44 تغيرات الدقة أثناء اختبار النموذج بعد دمج الصفين set,pass 83
الشكل 45 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بعد دمج set وpass 83
الشكل 46 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بعد دمج set وpass وإجراء استنظام 84
الشكل 47 مصفوفة الإرباك لاختبار النموذج على بيانت الاختبار بعد دمج set وpass..... 85
الشكل 48 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بعد دمج set وpass وإجراء استنظام..... 85

قائمة الجداول

الجدول 1 مقارنة نتائج نموذج Video Swin Transformer على مجموعة معطيات kinetics-600 [7]..... 53
الجدول 2 عدد الأمثلة الموجودة من أجل كل صف في مجموعة المعطيات volleyball 62
الجدول 3 مقارنة النتيجة المحققة مع النموذج actor transformer وPOGARS..... 78
الجدول 4 مقارنة النتيجة مع النموذج SAM [42]..... 86

قائمة المصطلحات

الاختصار	المصطلح باللّغة الإنكليزية	المصطلح باللّغة العربية	الوصف
	Attention	انتباه	
	annotation	تنويط	وهي عملية وسم مجموعة البيانات بالوسوم التوضيحية الموافقة
HAR	Human Activity Recognition	التّعرف على النشاط البشري	مهمة تصنيف النشاط/الإجراء الذي يقوم به شخص ما أو التنبؤ به
	linear embedding layer	طبقة تضمين خطية	إسقاط شعاع من بعد إلى بعد آخر
CNN	Convolutional Neural Network	شبكة عصبونية التفافية	من أنواع الشبكات العصبونية
COCO	Common Objects in Context dataset		مجموعة معطيات خاصة بكشف وتجزئة الصور
	confusion matrix	مصفوفة الإرباك	مصفوفة تستخدم لتقييم أداء نموذج التصنيف
	concatenate	سَلسَلة	
	decoder	مفكك ترميز	
E	encoder	مُرمّز	مكوّن يقوم بترميز الدخل بشكل رقمي وهو مكون رئيسي من مكونات المحوّل
	embedding	التضمين	
FLOPS	floating point operations per second		عمليات الفاصلة العائمة في الثانية هي مقياس لأداء المعالج
FFN	Feedforward neural network		شبكة عصبونية ذات تغذية أمامية
	Feature Maps	خرائط الميزات	
	Fine-tuning	ضبط دقيق	
GPU	Graphics processing unit	وحدة معالجة رسومات	
GELU	Gaussian error linear unit		وحدة الخطأ الغاوسي الخطي

GCNs	Graph Convolutional Networks	الشبكات الالتفافية للرسم البياني	هي نوع من الشبكات العصبية المصممة للعمل مباشرة مع الرسوم البيانية ومعلوماتها الهيكلية، تُستخدم شبكات GCN لمهام مثل تصنيف العقدة والتنبؤ بالارتباط وتصنيف الرسم البياني.
	hidden state	الحالة المخفية	
	hierarchical	هرمي	
	kinetics		مجموعة بيانات للتعرف على الإجراء البشري في مقاطع الفيديو التي تغطي 700/600/400 فئة من فئات العمل البشري، حسب إصدار مجموعة البيانات
JFT-300M			مجموعة معطيات Google الداخلية المستخدمة لتدريب نماذج تصنيف الصور
LR	Learning Rate	معدّل التعلم	معلمة ضبط في خوارزمية التحسين التي تحدد حجم الخطوة عند كل تكرار أثناء التحرك نحو الحد الأدنى من دالّة الخسارة.
MHA	Multi-Head Attention	الانتباه متعدد الرؤوس	
MSA	Multi-Head Self Attention	الانتباه الذاتي متعدد الرؤوس	
	mask	قناع	
MLP	Multilayer Perceptron	شبكة عصبونية متعددة الطبقات	أحد أشكال الشبكات العصبونية
	Neural Network	الشبكات العصبونية	أحد خوارزميات التعلم الآلي
NLP	natural language processing	معالجة اللغات الطبيعية	
	Overfitting	المُلاءمة الزَّائدة	تحدث عندما يتعلم النموذج بيانات التدريب بشكل جيد للغاية، بما في ذلك الضوضاء والقيم المتطرفة، مما يقلل من قدرته على التعميم على البيانات الجديدة غير المرئية.
	Padding	حشو	
PE	Positional Encoding	ترميز مكاني	
	query	استعلام	
SW	shifted windows	نوافذ مزاحة	
	self-attention	الانتباه الذاتي	

	scale	مقياس	
	Softmax		دالة رياضية تستخدم في التعلم الآلي، وخاصة في سياق مشاكل التصنيف، تحوّل متجه الأعداد الحقيقية إلى توزيع احتمالي. يتم تطبيق Softmax بشكل شائع في الطبقة الأخيرة من الشبكة العصبية لتفسير المخرجات كاحتمالات لكل فئة.
	seq2seq models		نماذج دخلها أو خرجها سلاسل من العناصر
SSV2	SOMETHING-SOMETHING V2		مجموعة معطيات كبيرة من مقاطع الفيديو المُصنفة التي تُظهر البشر وهم يقومون بإجراءات أساسية محددة مسبقاً باستخدام الأغراض اليومية.
	Trade-off	مفاضلة	التنازل عن ميزة من أجل الحصول على أخرى
	Transformer	المحوّل	
W	window	نافذة	

الفصل الأول: مدخل إلى موضوع البحث وأهدافه

1.1. تمهيد:

أصبحت المراقبة بالفيديو حاجة حيوية في عصر المدينة الذكية لتحسين نوعية الحياة وتطوير المنطقة كمنطقة آمنة. عادة ما يتم تثبيت كاميرات المراقبة على مسافة معينة للتغطية المناسبة للمنطقة. لذلك، يلزم إجراء تحليل أفضل وفهم أكثر عمقاً لمقاطع الفيديو، مما يؤثر بشكل كبير على نظام الأمان.

اجتذب التعرف على نشاط المجموعة Group Activity Recognition (GAR) اهتماماً كبيراً من الباحثين في رؤية الكمبيوتر وهو مجموعة فرعية من مشكلة التعرف على النشاط البشري والتي تركّز على السلوك الجماعي لمجموعة من الناس. يعد التعرف على النشاط الجماعي مهمة أساسية للتحليل التلقائي للسلوك البشري في العديد من المجالات مثل المراقبة أو مقاطع الفيديو الرياضية.

2.1. الهدف من البحث:

يوجد العديد من الأبحاث التي قدّمت طرق متنوعة لحل مسألة التعرف على نشاط المجموعة، ولكن أغلبها اقتصر على دراسة المسألة من ناحية تجميع الميزات الفردية بشكل صريح لاستنتاج نشاط المجموعة مما يهمل معلومات السياق.

يختلف النشاط الجماعي عن العمل الذي يقوم به الفرد، فهو يحتاج إلى مراعاة التفاعلات المعقدة بين الأشخاص المختلفين. ومع ذلك، تتطلب معظم الأعمال السابقة شروحاً شاملة مثل معلومات التسمية الدقيقة للإجراءات الفردية، والتفاعلات الزوجية، ومربعات الإحاطة والوضعيات، والتي لا يمكن أن تكون متاحة بسهولة في الممارسة العملية.

يهدف هذا البحث إلى استكشاف القوة التنبؤية لنماذج المحولات في تحليل المشهد كوحدة كاملة متكاملة وذلك من خلال تطوير نموذج قائم على التعلم العميق يتعرف على نشاط المجموعة من الفيديو بشكل مباشر.

دون إضافة معلومات تفصيلية عن المشاركين في المشهد، أو استخدام استراتيجيات صريحة لتجميع معلومات الأفراد (سواء معلومات وضعية الجسم، معلومات مكانية والإجراءات الفردية وغيرها...)، ودون إهمال معلومات سياق المشهد.

ما دفعنا إلى البحث في هذا المجال لملائمة عمل النموذج مع السلوك البشري في المعالجة.

3.1. لمحة عن الحل المقترح:

نقدم في هذا البحث نموذج يهدف إلى التعرف على نشاط المجموعة في مشهد متعدد الممثلين، حيث يتم الافتراض أن آلية الانتباه الذاتي التي توفرها شبكات المحولات (وهي المكوّن الأساسي في نموذج المحول) التي أدخلت في السنوات الأخيرة في مجال الرؤية الحاسوبية، هي نماذج مرنة بدرجة كافية يمكن استخدامها بنجاح خارج الصندوق، دون حيل أو تعديلات إضافية، لاستدلال نشاط المجموعة بأكملها بالنظر إلى الفيديو المدخل كوحدة متكاملة. وباستخدام أحدث النماذج التي تم تطويرها يمكن تفادي مشكلة التعقيد الحسابي في محولات الصور القائمة على الانتباه الشامل، بحيث يتم استخدام الانتباه بشكل محلي وبذلك يصبح التعقيد متناسب بشكل خطي مع أبعاد الصورة.

هذا التخفيض في التعقيد الحسابي يجعل من محوّل SWIN [1] مناسباً أكثر لتطبيقات الصور ذات الحجوم الكبيرة ولتطبيقات الزمن الحقيقي، وذلك كما سنرى في الدراسة التفصيلية.

4.1. مساهمات البحث:

يمكن تلخيص مساهمات البحث بالنقاط التالية:

I. تطوير وتنجيز نموذج لتصنيف نشاط المجموعة البشرية.
II. تعقيد النموذج يتناسب مع أبعاد الصور خطياً وليس تربيعياً
III. تقييم النموذج المقترح عن طريق تطبيقه على مجموعة معطيات خاصة بمسألة التّعرف على نشاط المجموعة [2] Volleyball والحصول على نتائج مقبولة وواعدة.
IV. استخدام الوسوم على مستوى الفيديو فقط يمكّن من انخفاض كبير في الجهد اليدوي والوقت المستغرق في إعداد مجموعات التدريب وبالتالي زيادة القابلية لتوافر مجموعات جديدة لهذه المسألة.

5.1. مخطط الأطروحة:

قمنا في هذا الفصل بتحديد مسألة البحث وتقديم لمحة عن الحل المقترح والمساهمات الأساسية التي يقدمها. سنقوم في **الفصل الثاني** بعرض بعض من المفاهيم النظرية والتعريف بجوانبها المختلفة التي ستيم التطرق لها خلال البحث، ثم سنتابع في **الفصل الثالث** عرض أهم الأبحاث والدراسات المشابهة الحديثة ضمن نفس المجال؛ واستعراض نماذج المحول التي أدخلت إلى مجال الرؤية الحاسوبية وسنستعرض ضمن **الفصل الرابع** النموذج الأساسي المستخدم في الحل والمراحل الأساسية للحل مع شرح تفصيلي لها، ليتم في **الفصل الخامس** اختبار النهج المقترح على مجموعة معطيات خاصة بالتعرف على نشاط المجموعة ومقارنة أدائه بأداء الأنظمة المشابهة. ونختتم بحثنا في **الفصل السادس** بعرض الآفاق المستقبلية لهذا العمل.

الفصل الثاني: مفاهيم نظرية

1.2. التعرف على النشاط البشري:

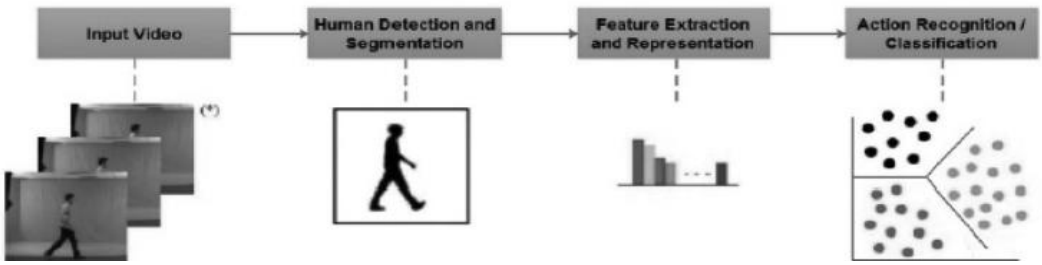
التّعرف على النشاط البشري (Human Activity Recognition (HAR هو مجال بحث في علوم الحاسوب والذكاء الصّنعي، هدفه التعرف على السلوكيات والأنشطة البشرية المختلفة وتصنيفها.

يعد التعرف على النشاط البشري أحد أنواع مسائل تصنيف التسلسل حيث تحتاج إلى بيانات من سلسلة من الخطوات الزمنية لتصنيف الإجراء الذي يتم تنفيذه بشكل صحيح.

حسب [3] تم تقسيم طرق التعرف على النشاط البشري بحسب الحسّاسات إلى الصنفين التاليين:

- **التعرف على النشاط البشري القائم على الرؤية:**

حيث تسجل الكاميرات المثبّتة في أماكن مختلفة لغرض المراقبة مقاطع الفيديو وتخزنها على الخوادم، ثم يتم استخدام موجزات الكاميرا أو مقاطع الفيديو المسجلة هذه لأغراض المراقبة. يستخدم هذا النوع من HAR للسلامة على الطرق، والأمن العام، وإدارة حركة المرور، ومراقبة الحشود، وما إلى ذلك. يوضح الشكل 1 الخطوات النموذجية للتعرف على النشاط البشري القائم على الرؤية:



الشكل 1 الخطوات النموذجية للتعرف على النشاط البشري القائم على الرؤية [3]

- **التعرف على النشاط البشري القائم على الاستشعار:**

أصبحت الهواتف الذكية أداة اتصال عالمية، ومؤخراً، أصبحت تقنية لدراسة البشر. يمكن لأجهزة الاستشعار المدمجة في الهواتف الذكية التقاط معلومات مستمرة حول الأنشطة البشرية. في هذا النهج، يتم استرداد البيانات من مقياس التسارع المدمج في الهاتف الذكي ومستشعرات الجيرسكوب، ثم يتم تطبيق تقنيات التعلم الآلي للتعرف على النشاط البشري.

هذا النوع من HAR مفيد لأنظمة مراقبة المريض، ومراقبة نشاط اللاعب الفردي أثناء ممارسة الرياضة، وما إلى ذلك.

2.2. أنواع الأنشطة البشرية:

تعتبر الأنشطة البشرية وسيلة للتواصل بين الأفراد والتفاعل مع الآلات ومع البيئة التي نعيش فيها. تشير الأنشطة إلى حركات أجزاء الجسم أو الجسم بالكامل وتتألف من عدة إجراءات أولية يتم تنفيذها بترتيب تسلسلي زمني. يمكن تحقيقها من قبل شخص واحد أو مجموعة من الأشخاص. يمكن إعطاء هذه لأنشطة البشرية تسلسلاً هرمياً اعتماداً على مدى تعقيدها من إجراء بسيط إلى أحداث أكثر تعقيداً كما يلي، بحسب [4]:

• **الإيماءة:** هي حركة بدائية لأي جزء من الجسم لنقل بعض المعلومات. يمكن أن تكون حركة يد صغيرة أو مجرد تعبير في الوجه، مدة هذه الأنشطة صغيرة جداً.

• **الفعل الذري atomic:** هو نشاط بسيط (يتضمن عدة إيماءات) يقوم به الإنسان. ومن الأمثلة على ذلك: الركض والجري والسباحة.

• **التفاعل:** هو نوع من النشاط يقوم به ممثلان. يجب أن يكون أحد الممثلين إنساناً والآخر قد يكون إنساناً أو أي كائن آخر. وبالتالي، يمكن أن يكون تفاعل إنسان–إنسان أو تفاعل إنسان–كائن ما. حيث "القتال بين شخصين" و "المصافحة" و "المعانقة" أمثلة على تفاعل إنسان–إنسان، في حين أن "شخصاً يستخدم آلة الصّراف الآلي" و "شخصاً يستخدم الحاسوب" و "شخصاً يسرق حقيبة" أمثلة على تفاعل بين إنسان–كائن.

• **إجراءات المجموعة:** هي نشاط معقد يتضمن أكثر من شخصين وكائن. من أمثلة الأنشطة الجماعية: لعبة في فريق مثل كرة القدم، ولعب الورق في مجموعة، والقتال الجماعي والاستعراض وغيرها.

من الشكل 2 يُفهم بوضوح، بينما نتحرك على المحور باتجاه السهم، يزداد مستوى تعقيد الأنشطة البشرية، ومن ثم، تزداد صعوبة أتمتة العملية.

الشكل 2 مستويات النشاط البشري [4]

1.2.2. إجراءات المجموعة والرؤية الحاسوبية:

نهتم في هذا البحث بمحور **إجراءات المجموعة** والذي كما رأينا أنه مجموعة فرعية من مشكلة التّعرف على النشاط البشري والتي تركز على السلوك الجماعي لمجموعة من الناس، الناتج عن التصرفات الفردية للأشخاص وتفاعلاتهم، الفرق الرئيسي بين التعرف على النشاط الجماعي وتصنيف العمل الفردي هو الحاجة إلى التفكير في وقت واحد حول عدة أشخاص.

يتمثل التحدي الذي يواجه هذه المهمة في كيفية تصميم شبكات مناسبة للسماح لخوارزمية التعلّم بالتركيز على التمييز بين فئات الأنشطة ذات المستوى الأعلى (من الإجراء الفردي) والتي تتعلق بالتطور المكاني والزمني لنشاط الأفراد في المجموعة.

تتمثل القدرات الرئيسية التي يجب أن يتمتع بها هكذا نموذج في نمذجة العلاقات الهرمية بكفاءة داخل المشهد واستخراج الميزات الزمانية المكانية المميزة من المجموعات بدقة [5]. ونظراً لقابلية تطبيق هذه التقنية، فقد حظي تحديد الأنشطة الجماعية باهتمام بحثي كبير. حيث يعد التعرف على النشاط الجماعي مهمة أساسية لتحليل السلوك البشري التلقائي في العديد من المجالات مثل المراقبة أو مقاطع الفيديو الرياضية. تتضمن الأمثلة في مجال الرياضة تحديد الأحداث مثل التسديدات على المرمى في مباريات كرة القدم، وأحداث الالتقاط والتدحرج في كرة السلة. تشمل الأمثلة في مقاطع فيديو المراقبة، الأشخاص الذين يتقاتلون، أو يسيرون معاً، أو تتم ملاحقتهم، أو أنشطة بشرية مماثلة أخرى.

كان التركيز في المراحل الأولى من أبحاث الرؤية الحاسوبية، على استخراج ميزات الصورة. ومع تقدم تقنيات أجهزة الاستشعار والكاميرا وتوسيع المعرفة، تحول التّركيز نحو دراسة طرق التعرف على الأنشطة بناءً على مقاطع الفيديو. حيث أدى التوافر المتزايد لبيانات الفيديو إلى تزايد الطلب على تقنية فهم الفيديو في تطبيقات الحياة الواقعية. على سبيل المثال، يعد الأمان أمراً بالغ الأهمية في تحديد الأنشطة غير الطبيعية وتوفير التحذيرات في الوقت الفعلي. وبالتالي، برزت مشكلة الفهم البصري في بيانات الفيديو كمدخلات، وتطوير تكنولوجيا الفهم المرئي كتوجهات بحثية بارزة [6].

تقدّم مقاطع الفيديو بعداً زمنياً، مما يستلزم نمذجة معلومات متعددة الإطارات لفهم التسلسلات الزمنية الديناميكية. ومع ذلك، تنشأ التحديات من التكرار غير ذي الصلة وضبابية الحركة، مما يؤدي إلى أعباء حسابية. وبالتالي، فإن معالجة وفهم بيانات الفيديو تشكّل تحديات كبيرة.

يركز هذا العمل على التعرف على النشاط الجماعي في مقاطع الفيديو، حيث يهدف النموذج إلى تصنيف الأنشطة المستمرة في الفيديو، سنتطرق تباعاً في الفقرات التالية إلى البنى المستخدمة في حل مسائل الرؤية الحاسوبية وكيف تدرجت وصولاً إلى آلية الانتباه الذاتي التي تعد أساس النموذج [7] الذي يقوم عليه هذا العمل.

3.2. أحدث الأساليب الأساسية المتبعة في معالجة مسائل الرؤية الحاسوبية:

1.3.2. الشبكات العصبية الالتفافية:

بدأت الشبكات العصبية الالتفافية Convolution Neural Network (CNN) بتحطيم الأرقام القياسية في مجال الرؤية الحاسوبية عاماً بعد عام. على الرغم من أن الجذور البرمجية قد تم وضعها في وقت مبكر جداً، إلا أن العتاديات الملائمة للتطبيق استغرقت بعض الوقت للحاق بها. تبدأ الطبقات السفلية من CNN باكتشاف الحواف والخطوط، وتتعلم الطبقات اللاحقة أنماطاً أكثر تعقيداً حتى تنظر الطبقات الأخيرة في النهاية إلى الصورة ككل. تتشابه بنية CNN مع القشرة البصرية البطنية للفقاريات إلى حد ما والتي هي عبارة عن مسار يتكون من طبقات متعددة لمعالجة المعلومات. ومع تدفق المعلومات عبر المسار البصري، تصبح الميزات التي يتم تعلمها أكثر تعقيداً، تماماً كما هو الحال في شبكة CNN. التشابه الأكثر إثارة للاهتمام هو حجم المجال البصري الاستقبالي. وفي كلتا الحالتين، فإن حجم المجال يزداد عبر الطبقات حيث يتم تجميع المزيد والمزيد من المعلومات حول الصورة. هذا أمر منطقي بشكل بديهي، حيث للتعرف على شيء مثل السيارة، يجب على الشبكة أولاً التعرف على الميزات البسيطة مثل الحواف ثم الزوايا والحواف ثم تجميعها لتكوين أشكال مثل العجلات والنوافذ وغطاء المحرك وما إلى ذلك قبل تجميع كل هذه العناصر والتوصل إلى أن ذلك من المرجح أن يكون سيارة [8].

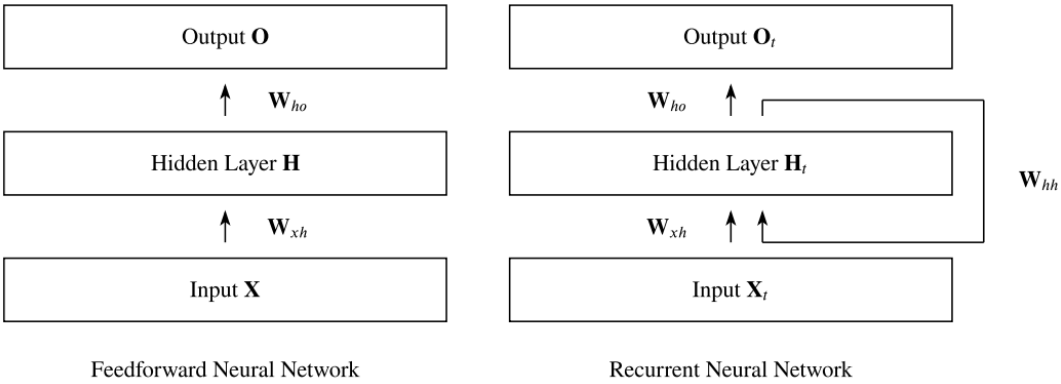
وبالتالي، بدا الأمر طبيعياً إلى حد ما عندما بدأت نماذج الرؤية الحاسوبية التي تستفيد من شبكة CNN في تحقيق أداء يقارب المستوى البشري في مهام مثل تصنيف الصور.

1.1.3.2. ميزات تصميم الشبكات العصبية الالتفافية:

- CNN ثابتة في الانسحاب. مثال: عندما تنظر إلى صورة لصديقك مبتسماً، فسوف تتعرف على صديقك بغض النظر عمّا إذا كان يقف إلى اليسار أو اليمين أو في النصف السفلي أو العلوي من الإطار. سوف تتعرف على صديقك حتى إذا كانت زاوية الصورة مختلفة، أو الإضاءة خافتة أو إذا كانت الصورة معكوسة، أو تم تكبيرها، وما إلى ذلك. هذا هو الثبات. مثل العين البشرية.
- CNN تحترم المحلّية المكانية. تأخذ كل عملية CNN في الاعتبار منطقة محلية من الصورة فقط. هذا منطقي بشكل حدسي. حيث تُفهم الصورة من خلال البدء بمجالات استقبالية صغيرة تركز على المعلومات المحلّية منخفضة المستوى ثم يتم تجميعها عبر الطبقات حتى التّمكن من فهم الصورة الكاملة.

2.3.2. الشبكات العصبية العوّدية:

تُستخدم RNN بشكل رئيسي للتعرف على الأنماط في المعطيات التسلسلية، كالنصوص، الفيديوهات أو أية متسلسلة زمنية وما يميزها عن الشبكات العصبونية التقليدية متعددة الطبقات Multilayer perceptron هو كيفية مرور المعلومات من خلالها[9] :



الشكل 3 الفرق بين مرور المعلومات بين الشبكات العودية – اليمين، والشبكات التقليدية – اليسار [9]

يوضح الشكل 3 الاختلاف بين MLP و RNN حيث تمرر شبكات MLP المعلومات دون وجود أية حلقات، على العكس من شبكات RNN، حيث يُحسب الخرج في شبكات MLP كما في المعادلات التالية:

$$H = \phi_h (XW_{xh} + b_h)$$
$$O = \phi_o (HW_{ho} + b_o)$$

حيث H خرج الطبقة المخفية، ϕ_h دالّة تفعيل الطبقة المخفية، O الخرج النهائي، ϕ_o دالّة تفعيل طبقة الخرج، W مصفوفة الأوزان و b شعاع الانحياز.

أما في RNN فمعادلة الخرج في كل خطوة زمنية t تحسب من خلال المعادلات التالية:

$$H_t = \phi_h (X_t W_{xh} + H_{t-1} W_{hh} + b_h)$$
$$O_t = \phi_o (H_t W_{ho} + b_o)$$

حيث H_t يسمى الحالة المخفية hidden state في اللحظة t.

نلاحظ أنه في كل خطوة زمنية لحساب الخرج نحتاج إلى الدخل X_t وإلى الحالة المخفية في الخطوة الزمنية السابقة H_{t-1}.

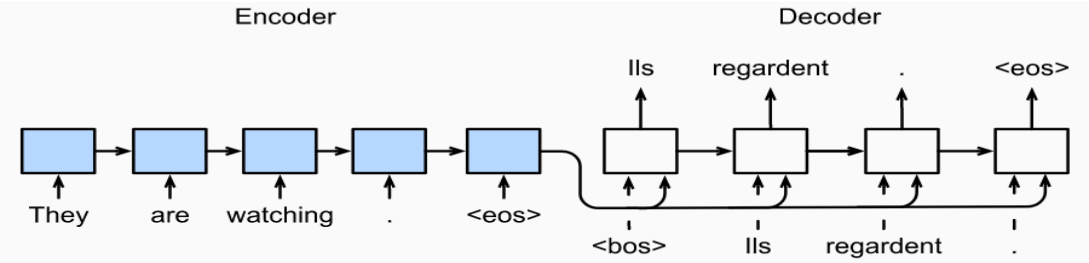
1.2.3.2. مشكلات RNN:

إحدى سلبيات شبكات RNN هي معالجة العناصر بشكل تسلسلي، فكما رأينا في الفقرة السابقة فإن الخرج في كل خطوة زمنية يحتاج إلى معالجة المداخل السابقة بشكل تسلسلي وهذا ما يجعل زمن التدريب طويلاً، وبسبب بنية الشبكة تصبح المعالجة بشكل تفرعي أمراً صعباً أو حتى مستحيل [10] . المشكلة الثانية هي تلاشي المشتقات vanishing gradient وهو ما يزيد من صعوبة التدريب وبخاصة في حالة السلاسل الطويلة، إحدى أشهر الحلول لتقليل أثر هذه المشكلة هي باستخدام LSTM (Long Short–Term Memory). من الممكن أن تختلف أبعاد سلسلتي الدخل والخرج ولحل هذا الاختلاف في الطول يمكن استخدام شبكتين عوّديتين منفصلتين الأولى نسميها مرمز encoder والثانية مفكك ترميز decoder.

3.3.2. المرمز ومفكك الترميز:

يعالج المرمز كل عنصر من عناصر سلسلة الدخل ويستخرج المعلومات ويضعها في شعاع يسمى شعاع السياق context vector، وبعد الانتهاء من معالجة كامل عناصر الدخل يرسل شعاع السياق إلى مفكك الترميز ليقوم بتوليد سلسلة الخرج بشكل تتابعي كما في الشكل 4 .

المشكلة الأساسية في هذه النماذج هي كيفية توليد شعاع السياق ليعبر بأفضل طريقة ممكنة عن معلومات سلسلة الدخل. فإذا كان شعاع السياق هو الحالة المخفية لآخر خطوة زمنية في سلسلة الدخل عندها لن يعبر بشكل فعال عن بدايات السلسلة وخاصةً في حالة سلسلة الدخل الطويلة. عندها يمكن أن يكون شعاع السياق عبارة عن كامل الحالات المخفية لكل خطوة زمنية في سلسلة الدخل، أو أن ينتج عن تعديل هذه الحالات المخفية بتطبيق دالّة معينة، ومن هنا ظهرت فكرة الانتباه attention وذلك بالتركيز فقط على الأجزاء المهمة وجعل الأجزاء غير المهمة أقل تأثيراً. بما أن المشكلة الأساسية هي كيفية توليد شعاع السياق كان هناك الكثير من الأبحاث لتحسين تمثيل هذا الشعاع من أبرزها [12] [11].



الشكل 4 نموذج seq2seq مع بنية encoder–decoder [56]

إذ كانت هذه الأعمال بدايات فكرة الانتباه، والذي أظهر نتائج جيدة ومبشرة في تطبيقات ترجمة الآلة، وخاصة في حالة الأبعاد الكبيرة لسلاسل الدخل [13].

4.3.2. المحولات:

ظهر نموذج المحول في عام 2017 [14] حيث تم تصميمه ليكون إحدى نماذج سلسلة إلى سلسلة seq2seq وهي نماذج يكون كل من دخلها وخرجها عبارة عن سلاسل من العناصر. من الممكن أن تكون السلسلة أي نوع من المعلومات ككلمات، صور، محارف أو غيرها.

ولأول مرة تم اختبار نموذج يعتمد بشكل كامل على توابع الانتباه ويستغني عن RNN، حيث كانت البنية الأساسية لنماذج seq2seq هي الشبكات العصبية العوّدية.

لذلك نجح المحول في التغلب على مشاكل هذا النوع من الشبكات، مثل عدم القدرة على المعالجة التفرعية لعناصر الدخل. حيث أن أحد ميزات المحول هي قابلية معالجة العمليات الحسابية بشكل تفرعي وهذا ما يسرّع عملية التدريب والاستدلال في حال وجود GPU.

إن التطبيق الأساسي لنموذج المحول الأصلي هو ترجمة الجمل من اللغة الإنكليزية إلى لغات أخرى، إذ أن دخل النموذج سلسلة من الكلمات (عدة جمل)، كل كلمة يتم التعبير عنها بشعاع من الأرقام نسميه token، فيكون دخل المرمز من أجل سلسلة كلمات طولها L وبعد نموذج d هو عبارة عن مصفوفة $X \in R^{Lxd}$.

لاقى نموذج المحول نجاحاً كبيراً في مجال معالجة اللغات الطبيعية [14]، وبسبب هذا النّجاح تم استخدامه في تطبيقات الرؤية الصنعية.

لتمييز مكان كل عنصر من السلسلة يتم إضافة قيم للترميز المكاني (PE(positional encoding. خرج كل طبقة من طبقات المرمّز وطبقات مفكك الترميز هو مصفوفة بأبعاد Lxd .

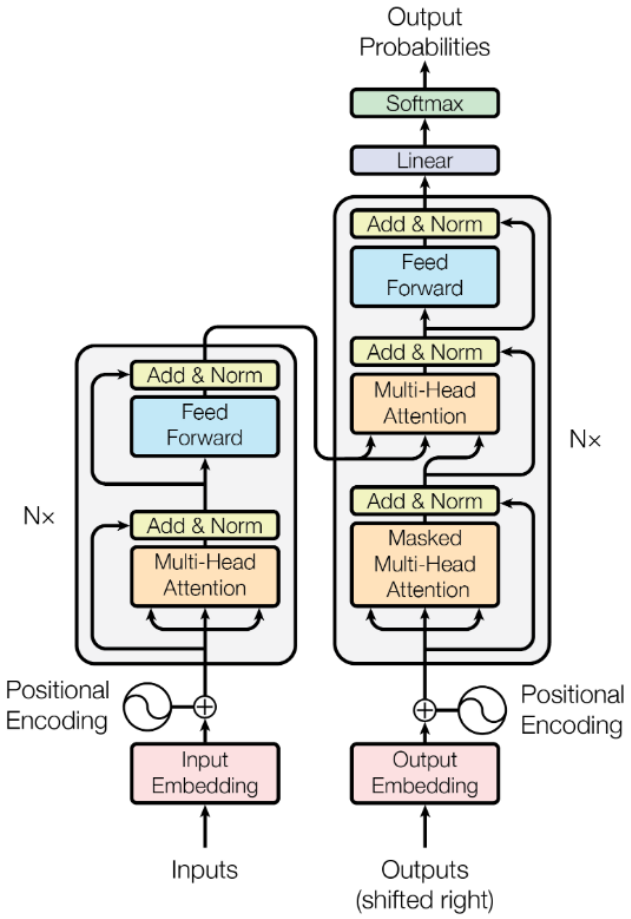
كما يوضح الشكل 5، يتكون كل من المرمّز ومفكك الترميز من عدة كتل متسلسلة ومتطابقة في البنية.

1.4.3.2. المرمّز:

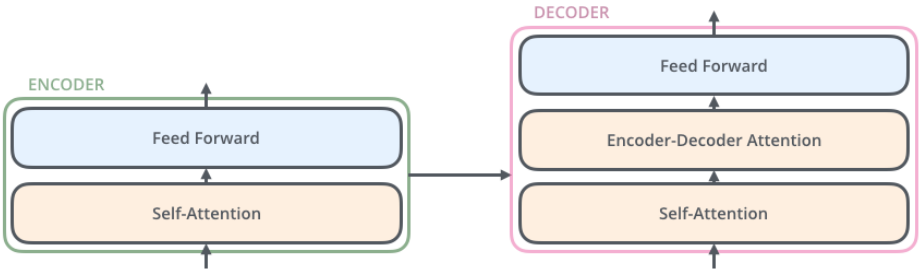
بداخل كل كتلة مرمّز كتلتان أساسيتان كما في الشكل 6، الأولى لحساب الانتباه الذاتي متعدد الرؤوس (Multi-Head Self Attention) MHA، تقوم بتعديل تمثيل عناصر سلسلة الدخل بحسب السياق. والكتلة الثانية هي شبكة عصبونية FFN تتكون من طبقتين مع تابع تفعيل ReLU بحسب المعادلة التالية:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

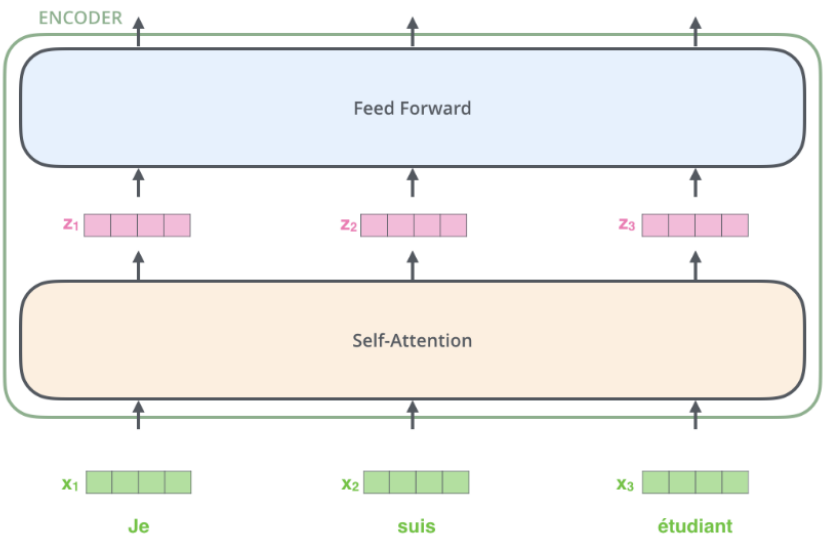
تطبق هذه الشبكة على كل token أو عنصر دخل بشكل منفصل ومتطابق، وهذا ما نسميه position–wise كما في الشكل 7، مما يجعل المحوّل قادراً على إجراء الحسابات بشكل تفرعي. من أجل كل من الكتلتين السابقتين يتم تطبيق residual connection [15] متبوعة ب layer normalization [16].



الشكل 5 البنية الأساسية للمحول [14]



الشكل 6 البنية الأساسية للمرمز و مفكك الترميز داخل المحول [17]



الشكل 7 تطبيق التوابع داخل المحول لكل عنصر دخل بشكل تفرعي [17]

2.4.3.2. مفكك الترميز:

يوضح الشكل 6 بنية مفكك الترميز، وكما في المرمّز فإنه يتكون بالإضافة إلى كتلتي الانتباه الذاتي وشبكة التغذية الأمامية فإنه أيضاً يحتوي على كتلة إضافية وهي الانتباه التقاطعي متعدد الرؤوس. في خوارزمية المحول الأساسية [14] يكون دخل مفكك الترميز هو خرج آخر طبقة في المرمز.

3.4.3.2. ترميز المعلومات المكانية:

كما ذكرنا سابقاً فإن معظم نماذج seq2seq كانت تستخدم الشبكات العوّدية قبل ظهور نموذج المحول [14] وهذا النوع من الشبكات يحافظ على المعلومات المكانية النسبية بين عناصر سلسلة الدخل، لكن الاستغناء عن شبكات RNN والاستعانة فقط بتوابع الانتباه لمعالجة السلاسل يفقد هذه المعلومات.

لمعالجة هذه المشكلة كان من اللازم إدخال هذه المعلومات للسلسلة بشكل ما، هنا تم طرح طريقة ترميز الموقع لإضافة المعلومات المكانية لكل عنصر من عناصر السلسلة وذلك بإضافة قيم مستخرجة من توابع بترددات مختلفة [14]، كما في المعادلات التالية:

$$PE_{(pos,2i)} = sin(pos/1000^{2i/dmodel})$$
$$PE_{(pos,2i+1)} = cos(pos/1000^{2i/dmodel})$$

حيث pos هو موقع عنصر الدخل من السلسلة، و i هو البعد dimension. ويتم إضافة قيم هذه التوابع إلى دخل كل من المرمّز ومفكك الترميز. اختبرت العديد من الأبحاث طرق الترميز المكاني، فأبسط الأشكال هي إسناد رقم طبيعي أو حقيقي مميز أو رقم حقيقي ضمن المجال [0 , 1] مميز إلى كل عنصر من عناصر السلسلة.

لكن هذه الطرق البسيطة لم تحسّن النتائج لأنها لم تستطع أن تجعل النموذج يلتقط معلومات المواقع بين العناصر. هنا اقترح البحث [14] أن يكون ترميز الموقع تابع جيبي كما في المعادلات السابقة، إذ يمكن اعتبار سلسلة الدخل سلسلة زمنية وكل عنصر هو خرج السلسلة عند خطوة زمنية معينة، هذه الطريقة في التفكير ساعدت النموذج على كشف معلومات الموقع النسبية بين العناصر.

4.4.3.2. الانتباه في المحولات:

كما يوحي الاسم، فإن الهدف من الانتباه هو السماح للنموذج بالتركيز على أجزاء مهمة من المدخلات. هذا منطقي حيث عندما ننظر إلى إدخال (مثل صورة أو نص)، فإن بعض الأجزاء أكثر أهمية لفهمنا من غيرها. يمكننا ربط أجزاء معينة من المدخلات ببعضها البعض وفهم السياق بعيد المدى.

تسمح آليات الانتباه لنماذج المحولات بالتعلم بطريقة مماثلة. في حين أثبتت آلية الانتباه أنها فعّالة للغاية، إلا أن لها مشكلة عملية فتعقيدها تربيعي بالنسبة لطول المدخلات. ولكن هناك الكثير من الأبحاث المكرّسة لجعل الانتباه أكثر فعالية.

5.4.3.2. الانتباه الذاتي:

هو نوع معين من الانتباه. الفرق بين الانتباه العادي والانتباه الذاتي هو أنه بدلاً من ربط المدخلات بتسلسل الإخراج، يركّز الانتباه الذاتي على تسلسل واحد. حيث يسمح للنموذج بجعل التسلسل يقوم بمعرفة معلومات عن نفسه.

6.4.3.2. المكونات الرئيسية المستخدمة من قبل انتباه المحولات:

- q و k تشير إلى متجهات ببعد d_k، والتي تحتوي على الاستعلامات (query) والمفاتيح (key)، على التوالي.
- v دلالة على متجه ببعد d_v، الذي يحتوي على القيم (value) .
- Q و K و V تشير إلى مصفوفات تجمع مجموعات من الاستعلامات والمفاتيح والقيم معاً، على التوالي.
- W^Q، W^K و W^V تشير إلى مصفوفات الإسقاط المستخدمة في إنشاء تمثيلات فضاء فرعية مختلفة للاستعلام ومصفوفات القيمة والمفتاح.
- W^O للدلالة على مصفوفة الإسقاط للمخرج متعدد الرؤوس.

إذا استعرضنا مثالاً في شرح الانتباه الذاتي في NLP:

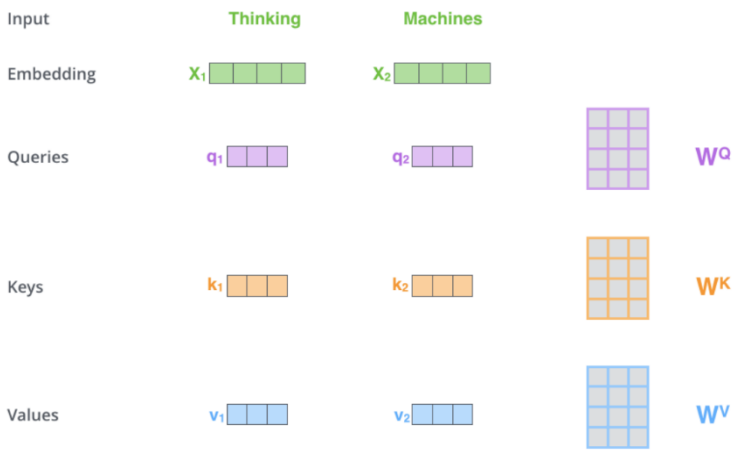
لنفترض أن الجملة التالية هي جملة إدخال نريد ترجمتها:

Thinking machines

الخطوة الأولى في حساب الانتباه الذاتي المبينة في الشكل 8، هي إنشاء ثلاث متجهات من كل من متجهات إدخال المرمّز (في هذه الحالة، تضمين كل كلمة) وهي: متجه الاستعلام q ، متجه المفتاح k ومتجه القيمة v .

كل متجه يتم حسابه عبر ضرب تضمين الكلمة المدخلة في ثلاث مصفوفات، يتم تدريبها أثناء عملية التدريب.

بفرض x_1 يشير إلى تضمين الكلمة الأولى فإن ضرب x_1 بالمصفوفة W^Q ينتج متجه الاستعلام q_1 المرتبط بهذه الكلمة. وهكذا بالنسبة لباقي المتجهات (المفتاح والقيمة) والتي ستستخدم لاحقاً لتمثيل كل كلمة في الجملة المدخلة.



الشكل 8 الخطوة الأولى لحساب الانتباه الذاتي [17]

الخطوة التالية في حساب الانتباه الذاتي تتمحور حول حساب نتيجة score.

مثال: لنفترض أننا نرغب في حساب الانتباه الذاتي للكلمة **(thinking)** في المثال المذكور.

نحتاج في هذه المرحلة أن نحسب score لكل كلمة في الجملة المدخلة على حسب ارتباطها بكلمة **(thinking)** وذلك لتحديد مناطق التركيز الأخرى في الجملة المدخلة عند ترميز الكلمة **(thinking)** .

يتم حساب score من حاصل الجداء السلمي dot product لمتجه الاستعلام للكلمة المراد معالجتها ومتجه المفتاح للكلمة المعنية التي نسجل score لها حيث إن الجداء السلمي يقيس مدى التشابه بين المتجهات المضمنة. لذلك إذا كنا نعالج

الانتباه الذاتي للكلمة الموجودة في الموضع رقم 1، فستكون النتيجة الأولى هي حاصل الجداء السلمي لـ q_1 وk_1. ستكون النتيجة الثانية هي حاصل الجداء السلمي لـ q_1 وk_2.

الخطوة الثالثة والرابعة هي قسمة النتائج على الجذر التربيعي ل d_k (تم اختيار هذا الرقم للحصول على انحدار gradient مستقر) ومن ثم يتم حساب ال softmax الذي يضمن أن تكون القيم مقيسة بين [0,1].

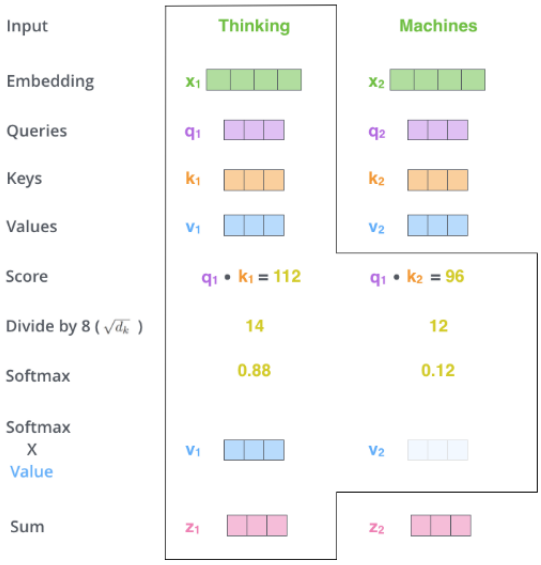
عملية SoftMax ستحدد مدى ارتباط كل كلمة في الجملة المدخلة مع الكلمة المراد معالجتها، مثلا مدى ارتباط كلمة (machines) مع كلمة (thinking).

الخطوة الخامسة، وهي التركيز على الكلمات ذات العلاقة بالكلمة المراد معالجتها، ويتم ذلك عبر ضرب كل متجه v بنتيجة SoftMax الخاصة به. وبالتالي ستتضخم قيم الكلمات ذات الصلة (عبر قيمة SoftMax العالية) وبالتالي سيتم التركيز عليها، وستتضاءل قيمة الكلمات غير المرتبطة ذات قيمة SoftMax منخفضة وبالتالي يتم تجاهلها.

يمثل المخرج من هذه العملية النتيجة النهائية لحساب الانتباه الذاتي للكلمة المراد معالجتها في تلك المرحلة.

وهكذا يتم حساب الانتباه الذاتي لكل كلمة حيث يكون الناتج النهائي عبارة عن متجه يُمرر بعد ذلك إلى شبكة التغذية الأمامية كما يوّضح الشكل 9.

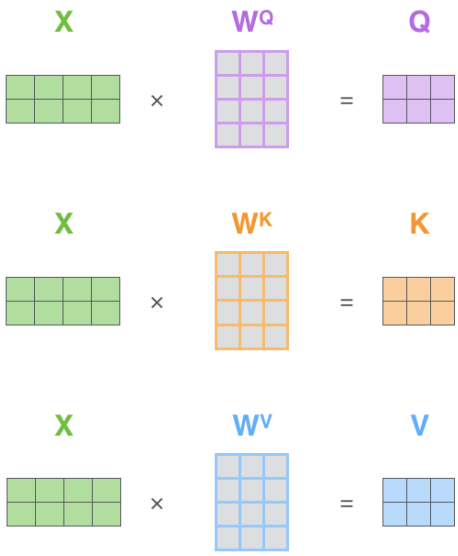
عملياً، لتسريع العمليات الحسابية، يتم استخدام المصفوفات لإتمام حساب الانتباه الذاتي لجملة الكلمات. الجزء التالي يشرح ذلك بالتفصيل.



الشكل 9 الخطوة 2 إلى 5 من حساب الانتباه الذاتي [17]

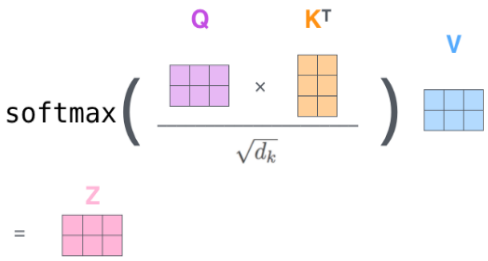
7.4.3.2. حساب الانتباه الذاتي باستخدام المصفوفات:

في الخطوة الأولى يتم حساب مصفوفات الاستعلام، المفتاح والقيمة. وذلك يتم ببساطة عبر ترتيب تضمينات الكلمات في مصفوفة X ثم ضربها في مصفوفات الأوزان (W^Q ، W^K، W^V) كما في الشكل 10.



الشكل 10 الحصول على مصفوفات Q و K و V [17]

حيث كل صف في المصفوفة X يمثل كلمة في الجملة المدخلة. وبهذه الطريقة يمكن دمج الخطوات السابق ذكرها (2 الى 5) في عملية حساب واحدة لحساب الانتباه الذاتي في كل طبقة. وبالتالي نحصل على مصفوفة حساب الانتباه الذاتي التالية مفصلة في الخطوات اللاحقة ومبينة في الشكل 11:



الشكل 11 حساب الانتباه الذاتي في شكل مصفوفة [17]

يمكن وصف الإجراء التدريجي لحساب الانتباه بالجداء السلمي المُقاس scaled–dot product attention بالخطوات التالية:

1. ضرب مجموعة الاستعلامات التي تم تجميعها في المصفوفة **Q**، بالمفاتيح الموجودة في المصفوفة **K** إذا كانت المصفوفة **Q** من الحجم $m \times d_k$ والمصفوفة **K**، من الحجم $n \times d_k$، فإن المصفوفة الناتجة ستكون بالحجم $m \times n$:

$$\mathbf{QK}^T = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ e_{21} & e_{22} & \dots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{m1} & e_{m2} & \dots & e_{mn} \end{bmatrix}$$

2. تقييس scale كل من القيم الناتجة بضربها ب

$$\frac{1}{\sqrt{d_k}}$$

فينتج:

$$\frac{\mathbf{QK}^T}{\sqrt{d_k}} = \begin{bmatrix} \frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \dots & \frac{e_{1n}}{\sqrt{d_k}} \\ \frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \dots & \frac{e_{2n}}{\sqrt{d_k}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \dots & \frac{e_{mn}}{\sqrt{d_k}} \end{bmatrix}$$

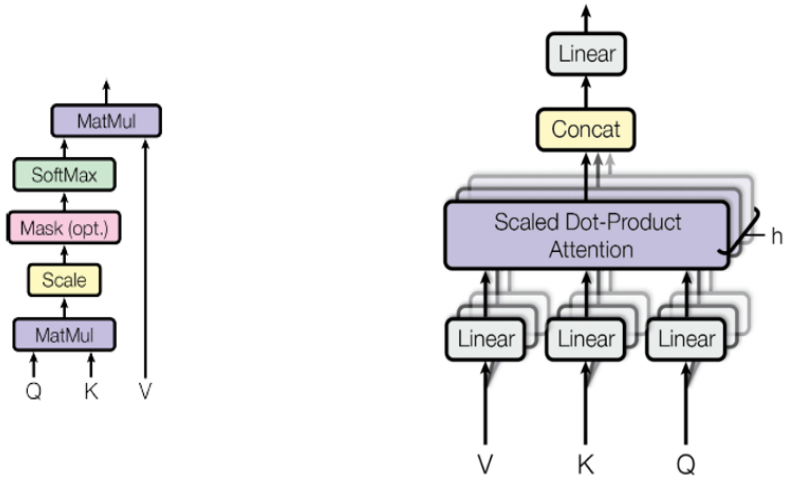
3. تطبيق عملية softmax للحصول على مجموعة من الأوزان والذي يضمن أن تكون القيم ضمن المجال[0,1]:

$$\text{softmax}(\frac{\mathbf{QK}^T}{\sqrt{d_k}}) = \begin{bmatrix} \text{softmax}(\frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \dots & \frac{e_{1n}}{\sqrt{d_k}}) \\ \text{softmax}(\frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \dots & \frac{e_{2n}}{\sqrt{d_k}}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{softmax}(\frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \dots & \frac{e_{mn}}{\sqrt{d_k}}) \end{bmatrix}$$

4. أخيراً، يتم تطبيق الأوزان الناتجة على القيم الموجودة في المصفوفة $\mathbf{V}$ من الحجم $n \times d_v$:

$$\text{SoftMax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} = \begin{bmatrix} \text{softmax}\left(\frac{e_{11}}{\sqrt{d_k}} & \frac{e_{12}}{\sqrt{d_k}} & \dots & \frac{e_{1n}}{\sqrt{d_k}}\right) \\ \text{softmax}\left(\frac{e_{21}}{\sqrt{d_k}} & \frac{e_{22}}{\sqrt{d_k}} & \dots & \frac{e_{2n}}{\sqrt{d_k}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{softmax}\left(\frac{e_{m1}}{\sqrt{d_k}} & \frac{e_{m2}}{\sqrt{d_k}} & \dots & \frac{e_{mn}}{\sqrt{d_k}}\right) \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1d_v} \\ v_{21} & v_{22} & \dots & v_{2d_v} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \dots & v_{nd_v} \end{bmatrix}$$

وهذا ما يتم تطبيقه من أجل كل رأس من رؤوس الانتباه المتعدد الرؤوس الذي له الشكل 12 التالي (اليمين) :



الشكل 12 (يسار) الانتباه بالجداء السلمي المُقاس. (يمين) كتلة الانتباه متعدد الرؤوس المؤلفة من h طبقة انتباه تعمل على التوازي [14]

تحسن هذه الآلية أداء طبقة الانتباه عبر طريقتين:

1. تطوير قدرة النموذج على التركيز على أكثر من كلمة في نفس الوقت.
2. تعدد مساحات التمثيل في طبقة الانتباه حيث يصبح هناك مصفوفات متعددة للاستعلام، المفتاح والقيمة. يستخدم المحوّل في الورقة [14] رؤوس انتباه عددها 8 وفي كل رأس يتم إسناد قيم أولية بشكل عشوائي ثم استخدامها لتمثيل قيم التضمين للكلمات المدخلة في مساحة تمثيل خاصة.

يتم تكرار حساب مصفوفة الانتباه بالطريقة التي تم شرحها سابقاً لكل من الثمان رؤوس، يؤدي إلى الحصول على 8 مصفوفات Z .

لكن طبقة التغذية الأمامية تستقبل فقط مصفوفة واحدة مخرجة، كل صف فيها يمثل حساب الانتباه لكل كلمة مدخلة، لذلك سيتم إعادة تمثيل المصفوفات المخرجة من الرؤوس المتعددة بحيث يتم تكديسها في مصفوفة واحدة فقط عبر عملية ضم المصفوفات Concatenation ثم ضربها بمصفوفة الوزن W^O حتى يتم عكس التأثير الحقيقي لكل كلمة كما تبيّن المعادلة التالية:

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_H)W_O$$
$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

وبحسب [14]، فإن معالجة سمات الدخل في عدة رؤوس يسمح لكل رأس أو نموذج انتباه بالتركيز على مجموعة معينة من السمات. في النموذج الأساسي [14] استخدمت توابع انتباه بعدد الرؤوس $h = 8$.

وهكذا نلاحظ وجود نوعين من الحسابات المتوازية مخبأة داخل الانتباه الذاتي من خلال تجميع متجهات التضمين في مصفوفة الاستعلام، ومن خلال تقديم الانتباه متعدد الرؤوس.

4.2. خاتمة

تم في هذا الفصل شرح معنى التعّرف على نشاط المجموعة ومايهم من مفاهيم للدخول في تفاصيل البحث والدراسات.
وبما أن العمل الحالي يعتمد على بنية قائمة على المحولات للتعرف على الفيديو كان لا بد من شرح العمليات التي تقوم عليها المحوّلات بشكل أساسي، والتعريف بآلية الانتباه وكيفية عملها بشكل مفصل فهي أساس عمل نموذجنا المقترح.

الفصل الثالث: الدراسة المرجعية

1.3. تمهيد:

سيتم التطرق إلى محورين أساسيين في هذا الفصل: في المحور الأول سنقدم الأساليب العامة التي تقوم بحل مسألة التعرف على النشاط في المجموعة، ونقدم بعضاً من أبرز هذه الدراسات بشكل مفصّل، والمحور الثاني وانطلاقاً من أن عملنا قائم بشكل كامل على بنية المحول، سيتم التطرق إلى موضوع إدخال المحولات إلى مجال الرؤية ومزاياها التي دفعتنا لاستعمالها كنموذج مناسب للحل المقترح في الفصل الرابع.

2.3. المفاهيم الأساسية /الدراسات /والنظريات:

التعرف على نشاط المجموعة له أهمية نظرية وعملية. ومع ذلك، تركز معظم المراجعات السابقة للتعرف على النشاط البشري على التعرف على العمل الفردي، تم نشر مسح متعلق بالتعرف على نشاط المجموعة في عام 2017 [18] يركز بشكل أساسي على الأساليب القائمة على الميزات اليدوية بينما لا تتم مناقشة الأساليب القائمة على التعلم العميق بعمق. علاوة على ذلك، تم إحراز تقدم ملحوظ في هذا المجال في السنوات الأخيرة بسبب تقنيات التعلم العميق القوية.

يتم التمييز في المراجعة [19] بين الأساليب التقليدية القائمة على الميزات اليدوية وتلك القائمة على التعلم العميق.

1.2.3. المناهج المعتمدة على الميزات اليدوية:

يمكن تصنيف الأساليب التقليدية للتعرف على نشاط المجموعة في فئتين: المناهج التنازلية [[20]، [21]، [22]] التي تحلل الأنشطة من حيث الحركة والتفاعل على مستوى المجموعة. تتمثل عيوب هذه الأساليب في الافتقار إلى وصف تفصيلي للنشاط الذي لا يمكّنهم من استغلال الميزات بالكامل على المستوى الفردي.

والمناهج التصاعدية [[23]، [24]، [25]]. تركز هذه المناهج على التعرف على كل فرد ووصف النشاط بناءً على مجموعة من السمات الفردية وإحصاءاتها. لذلك، فهي حساسة لفشل استخراج الميزات الفردية بسبب الحجب occlusion [19].

تأتي الأساليب التقليدية القائمة على هندسة الميزات اليدوية للتعرف على النشاط الجماعي بشكل عام مع قيود مختلفة مثل تبعية المجال والتكاليف الحسابية الكبيرة والدقة الناتجة الأقل، وقد تحوّل تركيز نماذج التعرف على النشاط الجماعي الحديثة نحو استخدام الشبكات العصبية العميقة.

2.2.3. النهج المعتمد على التعلم العميق:

في الآونة الأخيرة، أظهرت الشبكات العصبية الالتفافية العميقة أداءً رائعاً في مجموعة متنوعة من مهام الرؤية الحاسوبية بما في ذلك تصنيف الصور، والتجزئة الدلالية، والتعرف على الفيديو. تم اقتراح العديد من مناهج التعلم العميق للتعرف على نشاط المجموعة وحققت نتائج متفوقة على الأساليب اليدوية.

سيتم في هذا القسم استعراض الأساليب القائمة على التعلم العميق للتعرف على نشاط المجموعة. حيث يتم تلخيص ثلاث مشاكل رئيسية للتعرف على نشاط المجموعة: النمذجة الزمنية الهرمية، ونمذجة العلاقة، ونمذجة الانتباه. يتم التقسيم إلى هذه الأساليب على أساس المشكلة الحاسمة التي يركزون عليها.

1.2.2.3. النمذجة الزمنية الهرمية:

حققت الشبكة العصبية ذات الذاكرة الطويلة قصيرة المدى (Long Short Term Memory) LSTM، وهي نوع خاص من الشبكات العصبية العوّدية، نجاحاً كبيراً في المهام المتسلسلة بما في ذلك التعرف على الكلام وتوليد التسميات التوضيحية للصور.

في التعرف على نشاط المجموعة، يحاول بعض الباحثين تطبيق LSTM لبناء تمثيل هيكلي هرمي لاستنتاج الإجراءات الفردية وأنشطة المجموعة.

تم اقتراح نموذج هرمي عميق [2] من مرحلتين (Hierarchical Deep Temporal Model) HDTM. يتطلب النموذج كدخل مسارات الأشخاص في المشهد. في المرحلة الأولى يتم استخدام شبكة CNN لاستخراج الميزات من المربع المحيط بالشخص في كل خطوة زمنية على مسار الشخص. يتم اعتبار مخرجات CNN، بمثابة ميزات معقدة قائمة على الصور تصف المنطقة المكانية المحيطة بالشخص. تشكّل هذه الميزات مدخل خلية LSTM في اللحظة t.

بسبب النشر من البوابات على دفق المعلومات ستتشكل الحالة المخفية بناءً على ذاكرة قصيرة الأمد للسلوك السابق للشخص. لذلك، يمكن تمرير إخراج خلية LSTM في كل مرة إلى طبقة تصنيف للتنبؤ بالإجراء الفردي على مستوى الشخص لكل مسار. تُشكل طبقة LSTM المرحلة الأولى من النموذج الهرمي. وهذا يمثل المرحلة الأولى من النموذج التي تم تصميمها لنمذجة الإجراءات على مستوى الشخص وتطورها الزمني.

يتم تدريب النموذج بطريقة مرحلية، حيث يتم التدريب أولاً على التنبؤ بالإجراءات على مستوى الشخص، ثم تمرير الحالات المخفية لطبقة LSTM إلى المرحلة الثانية للتعرف على نشاط المجموعة. حيث يحتوي محتوى الذاكرة لطبقة LSTM الأولى في كل خطوة زمنية على معلومات تمييزية تصف تصرفات الفرد بالإضافة إلى التغييرات السابقة في تصرفاته. إذا تم جمع محتوى الذاكرة بشكل صحيح من جميع الأشخاص في المشهد، فيمكن استخدامه لوصف نشاط المجموعة في المشهد بأكمله.

تتم سَلسَلة كل من الميزات المستخرجة من شبكة CNN والميزات المستخرجة من طبقة LSTM للحصول على تمثيل الميزة الزمنية للشخص. يمكن استخدام استراتيجيات التجميع المختلفة لتجميع هذه الميزات على جميع الأشخاص الموجودين في المشهد في كل خطوة زمنية لتمثيل الميزات على مستوى الإطار. تشكل مخرجات طبقة التجميع تمثيل نشاط المجموعة. وتُستخدم شبكة LSTM الثانية، التي تعمل فوق التمثيل الزمني، لنمذجة الديناميكيات الزمنية لنشاط المجموعة بشكل مباشر. ترتبط طبقة LSTM الخاصة بالشبكة الثانية مباشرة بطبقة تصنيف من أجل اكتشاف فئات نشاط المجموعة في تسلسل الفيديو.

تعّد هذه الطريقة أول عمل يتضمن إطار عمل LSTM عميق لمعالجة التعرف على نشاط المجموعة.

ولا بدّ من الإشارة إلى أنه جرى في هذه الورقة تقديم مجموعة المعطيات المستخدمة في بحثنا Volleyball والتي تم جمعها من مقاطع فيديو الكرة الطائرة المتاحة للجمهور على YouTube.

2.2.2.3. نمذجة العلاقات العميقة:

يعد بناء العلاقات بين الأشخاص وأداء الاستنتاج العلائقي أمراً ضرورياً للتعرف على الأنشطة عالية المستوى. ومع ذلك، فإن نمذجة العلاقات بين الأشخاص يمثل تحدّياً في التعرف على نشاط المجموعة لسبب أنه لا يمكن الوصول إلّا إلى تسميات الإجراءات الفردية وتسميات النشاط الجماعي، دون معرفة إضافية بمعلومات التّفاعل.

في هذا السياق تم تطبيق شبكات (Graph Convolutional Networks) GCNs في العديد من مجالات الرؤية الحاسوبية مثل التتبع البصري والتعرف على العمل الفردي البشري.

تعد GCN نموذجاً مناسباً لمعالجة التعرف على نشاط المجموعة الذي يمكن من خلاله اعتبار كل شخص عقدة.

في [26] عام 2019 تم إدخال GCN في التّعرف على نشاط المجموعة، حيث يتم استخراج الميزات على مستوى الشخص من خلال الشبكات العصبية الالتفافية ويتم إنشاء بيان لعلاقة الممثل (الفرد في المشهد) على أساس التشابه البصري ومسافة الموقع المكاني مع بقية الممثلين.

تم اعتماد GCN لأداء التفكير العلائقي على بيان علاقة الممثل (Actor Relation Graph) ARG لاكتساب الميزات العلائقية لكل شخص. وذلك كما يلي:

يتم استخراج أشعة الميزات للممثلين من إطارات فيديو الدخل التي تم أخذ عينات منها حيث يتم اتباع استراتيجية استخراج الميزات والتي تعتمد Inception−v3 لاستخراج خريطة ميزات متعددة المقاييس لكل إطار وتطبيق RoIAlign [27] لاستخراج

الميزات لكل مربع محيط بالممثل من خريطة ميزات الإطار. بعد ذلك، يتم تنفيذ طبقة متصلة بالكامل على هذه الميزات للحصول على شعاع ميزات ببعد d من أجل كل ممثل. يُشار إلى العدد الإجمالي للمربعات المحيطة في إطارات عددها K بالرمز N. ويتم استخدام مصفوفة بعدها $N \times d$ لتمثيل أشعة ميزات الممثلين.

بعد ذلك، بناءً على هذه الميزات الأصلية للممثلين، يتم بناء رسوم بيانية لعلاقات الممثلين المشار إليها ب ARGs، حيث تشير كل عقدة إلى ممثل. كل حافة في الرسوم البيانية عبارة عن وزن عددي، والذي يتم حسابه وفقاً لميزات مظهر اثنين من الممثلين وموقعهما النسبي. وبهدف تمثيل معلومات العلاقات المتنوعة، يتم إنشاء رسوم بيانية متعددة للعلاقات من نفس مجموعة ميزات الممثلين.

بعد ذلك، تُستخدم GCNs لإجراء الاستدلال العلائقي على الرسوم البيانية. يتم بعد ذلك دمج مخرجات جميع الرسوم البيانية لإنشاء التمثيل العلائقي للممثلين، والذي يكون أيضاً ببعد $N \times d$. وأخيراً، يتم تجميع الميزات الأصلية والميزات العلائقية وإدخالها في مصنفات النشاط الجماعي والعمل الفردي.

نلاحظ أنه في هذا العمل تتم نمذجة معلومات التفاعل بين ممثلي المشهد بشكل صريح.

3.2.2.3. نمذجة الانتباه:

للتعرف على نشاط المجموعة، عادة ما يكون هناك العديد من الأشخاص النشطين في المشهد بينما يساهم عدد قليل فقط من الأشخاص الرئيسيين في أنشطة المجموعة، والآخرون قد يجلبون معلومات محيرة لاستنتاج أنشطة المجموعة ليست ذات صلة بالنشاط القائم. نظراً لعدم وجود الوسوم التوضيحية للشخص الرئيسي لمجموعات بيانات التعرف على النشاط الجماعي، يمكن تعريف هذه المشكلة على أنها اكتشاف الأشخاص المهمين الخاضعة للإشراف الضعيف. لمعالجة هذه المشكلة، صممت عدة طرق لنمذجة الانتباه.

في [28] تم اقتراح آلية انتباه من مستويين للتعرف على نشاط المجموعة. الأول انتباه على المستوى الفردي يتم توجيهه بميزات الوضعية للتحكم في الحالة المخفية في كل خطوة زمنية. والثاني انتباه على مستوى المشهد يربط الأفراد بأوزان مختلفة لبناء تمثيل مشهد مميّز. تعتمد هذه الطريقة على تقدير الوضعية.

يمكن مقارنة الأساليب السابقة كما يلي:

تُظهر الأساليب الحديثة القائمة على التعلم العميق للتعرف على نشاط المجموعة تحسينات واعدة في الأداء على الطرق التقليدية.

تستخدم الأساليب القائمة على النمذجة الزمنية الهرمية نموذج LSTM من مرحلتين لتعلم التمثيل الزمني للإجراءات على المستوى الفردي وتطبيق وظائف التجميع على الميزات الفردية لإنشاء تمثيل على مستوى المجموعة. ألهَم إطار عمل LSTM المكون من مرحلتين الكثير من الأعمال اللاحقة. قصورها يكون بمعاملة جميع الأفراد بنفس الأهمية، على الرغم من أنه عادة ما يتم تحديد نشاط المجموعة من قبل عدد قليل من الأشخاص الرئيسيين في بعض السيناريوهات مثل مقاطع الفيديو الرياضية.

تحاول الأساليب القائمة على نمذجة الانتباه حل هذه المشكلة. هذا الأسلوب له أداء أعلى من الأساليب القائمة على النمذجة الزمنية الهرمية. ومع ذلك، نظراً لعدم وجود وسوم توضيحية للأشخاص الرئيسيين، فإن كيفية تعلّم نموذج مستقر يمكنه العثور بدقة على الفرد الرئيسي لا يزال يمثل مشكلة صعبة.

حالياً، تم استخدام العلاقات بين الكيانات على نطاق واسع في مهام الرؤية الحاسوبية المختلفة. يتم تقديم طرق مختلفة لاستدلال العلاقة في التعرف على نشاط المجموعة، مثل GCN والمحولات. تمتاز الأساليب القائمة على نمذجة العلاقات العميقة بتمكنها من التقاط التفاعلات المحتملة والعلاقات بين الأشخاص التي يمكن أن تميز بشكل فعّال نشاط الفرد والجماعة. تحقق هذه الفئة من الأساليب أفضل النتائج في مجموعة بيانات CAD [29] و volleyball [2] .

ممّا جاء في المراجعة [19] وكان محفّزاً لعملنا هذا، هو أنه تُقر معظم الأساليب الحالية بشكل مباشر المربع المحيط من الوسوم التوضيحية التي يتعذر الوصول إليها في التطبيقات العملية. حيث أنّ البحث في هذا الموضوع محدود. يمكن أن تكون المهام الخاضعة للإشراف الضعيف للتعرّف على نشاط المجموعة حيث يمكن الوصول إلى نشاط المجموعة على مستوى الفيديو فقط اتجاهاً آخر للتعرف على نشاط المجموعة.

والذي كان ملهماً لهذا العمل أيضاً، هو ظهور بنية قوية مثل بنية المحولات ودخولها عالم الرؤيا الحاسوبية والتطوير عليها بشكل كبير ومحفّز. لذلك وبعد أن قمنا بشرح العمليات التي تقوم على أساسها المحولات بأساسها النظري في الفصل السابق، سنقدم في الفقرة التالية كيف أُدخلت المحولات في مجال الرؤية وصولاً للمحوّل الذي يعدّ الأساس القائم عليه عملنا هذا.

3.3. المحوّل في مجال الرؤية الحاسوبية:

1.3.3. لمحة تاريخية:

في السنوات الأخيرة، سيطرت المحولات على بنى التعلم العميق في مهام معالجة اللغة الطبيعية التي قُدمت في الورقة البحثية [14]. تستخدم المحولات آلية الانتباه الذاتي المصممة لنمذجة التسلسل والمهام التحويلية للترجمات وتلخيص النص. حققت المحولات ذات الانتباه الذاتي نجاحاً هائلاً نظراً لقدرتها وسهولة نمذجة التبعيات بعيدة المدى في البيانات.

حفّز النجاح الهائل للمحولات في مجال معالجة اللغات الطبيعية جهود البحث لتكييف المحولات لمهام الرؤية. وقد أعطى نتائج تفوّقت على أحدث وأفضل الخوارزميات في هذا المجال وذلك بمقارنة النتائج على مجموعات معطيات مثل ImageNet، COCO وغيرها [30]. فقد تم استخدام المحوّل في مجال التوليد [31]، التصنيف [32] ، الكشف [33]، الملاحقة [34] ، وتصنيف الفيديو [7] والذي هو موضوع بحثنا.
سنتكلم بإيجاز في هذه الفقرة عن النموذج الأول الذي استخدم المحول في مجال الرؤية الحاسوبية وهو نموذج ViT [32].

2.3.3. محوّل التصنيف ViT [32]:

ViT [32] هو محوّل الرؤية الحاسوبية الذي تم تقديمه في الورقة البحثية: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

يقوم المحوّل بتحويل الصورة إلى سلسلة عن طريق تقطيع الصورة إلى رقع patches مؤلفة من 16×16 بكسل وتغذيتها في طبقة خطية تحوّل كل رقعة إلى تضمين. كل رقعة تعامل معاملة token (كلمة) في المحول الأصلي [14].

عند تدريب هذا النموذج على معطيات تدريب ذات حجم متوسط مثل ImageNet، كان أداؤه متواضعاً وأقل من أداء نموذج ResNet [15] وبعدد أوزان متقارب. لكن النتائج تغيرت عند تدريبه على معطيات تدريب بحجوم كبيرة (14 مليون – 300 مليون عينة) مثل JFT-300M،ImageNet- 21K حيث أعطى النموذج نتائج تفوقت على الخوارزميات السابقة.

دخل النموذج الأصلي للمحول هو سلسلة ببعد واحد. للتعامل مع الصورة وللمحافظة على البنية الأساسية للمحول فقد تم تعديل الصورة $x \in R^{HxWxC}$ وتحويلها إلى سلسلة من الأجزاء المسطحة $x_p \in R^{Nx(P^2.C)}$ ذات بعد واحد، وذلك لمحاكاة دخل المحول الأصلي.

حيث HxW أبعاد الصورة، C عدد القنوات والتي هي $RGB = 3$، PxP أبعاد كل رقعة من الصورة (patch)، عدد الرقع أو الأجزاء $N = \frac{HW}{P^2}$.

يتم إدخال الشعاع x_p إلى طبقة إسقاط خطي، وذلك لتحويل أبعاده إلى (N,D) كما في المعادلة التالية الأولى [32] :

$$z_0 = [x_{class}; X_P^1 E; X_P^2 E; \ldots\ldots X_P^N E] + E_{Pos}\,, \qquad E \in R^{(P^2.C)\mathrm{xD}}\ , E_{Pos} \in R^{(N+1)xD}$$

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}\ , \qquad l = 1\ldots L$$

$$z_l = MLP\left(LN(z_l')\right) + z_l'\ , \qquad l = 1\ldots L$$

$$y = LN(z_l^0)$$

حيث D هو hyperparameter ويمكن تسميته ببعد التضمين.

E هي طبقة الإسقاط الخطي، بعدد بارامترات $E \in R^{(P^2.C)\mathrm{xD}}$ قابلة للتدريب. وكما في خوارزمية BERT [35] يعتمد المؤلفون على تضمين قابل للتعلم في تسلسل الرقع حيث يتم إضافة شعاع من البارامترات يدعى class token قابل للتدريب بأبعاد $(1\mathrm{x}D)$، وهو x_{class} في المعادلة الأولى السابقة.

حيث يتم اعتبار أن حالة هذا الشعاع في الخرج النهائي للمحول z_l^0 يمكن تدريبها لتعبر عن صنف الصورة، وذلك بعد تعديل هذا الشعاع بإدخاله إلى رأس التصنيف classification head (رأس التصنيف عبارة عن MLP بسيط يتألف من طبقتين مخفيتين مع تابع تفعيل GELU).

$\boldsymbol{L}$ هي عدد طبقات المرمز. تتم إضافة شعاع ترميز الموقع E_{Pos} إلى هذا الشعاع، وذلك لإدخال المعلومات المكانية النسبية بين أجزاء الصورة، حيث استخدموا في ViTعمليات تضمين الموضع أحادية الأبعاد القياسية القابلة للتعلم.

z_0 هو دخل المحول كما في المعادلة الأولى السابقة. يستخدم ViT بنية مرمز مطابقة لبنية المرمز في المحول الأصلي من حيث توابع انتباه متعددة الرؤوس كما في المعادلة الثانية وشبكة MLP كما في المعادلة الثالثة، ويتم تطبيق طبقة تقييس قبل كل كتلة، بالإضافة إلى residual connections [15] بعد كل كتلة. هذا هو جوهر ViT [32].

بالرغم أن خوارزمية ViT [32] لم تعدّل على بنية المرمز الأصلي للمحول [14] ، إذ أنها فقط جزأت الصورة إلى 16x16 جزء إلا أنها تفوقت على أفضل المصنفات حين تم تدريبها على معطيات تدريب كافية.

ولكن على الرغم من أن ViT Transformers [32] جيدة في التصنيف، إلّا أنها تعاني من العديد من أوجه القصور، حيث أن العدد الكبير للبكسلات في الصورة مقارنة بعدد الكلمات في الجملة، يجعل المحول الأصلي ذا تعقيد حسابي كبير من أجل المهمات التي تتطلب معالجة على مستوى البكسل مثل التجزئة الدلالية، أو من أجل التعامل مع الصور ذات الحجوم الكبيرة حيث يكون التعقيد الحسابي للانتباه الذاتي متناسباً بشكل تربيعي مع أبعاد الصورة إذ يتم حساب تابع الانتباه بين جميع tokens.

كما أن tokens في ViT [32] لها حجم ومقياس ثابت، وهذا غير مناسب لأغراض الرؤية الحاسوبية حيث تكون العناصر المرئية ذات نطاق متغير.

تبعت سلسلة من الأعمال البحثية ViT [32]، وأجرى معظمهم تحسينات على بنية المحولات القياسية من أجل معالجة أوجه القصور المذكورة أعلاه. من بين هذه الأعمال عام 2021، نشر باحثوا Microsoft المحوّل Swin Transformer [1] والذي يمكن القول إنه واحد من أكثر الأبحاث إثارة التي تبعت ViT الأصلي **وهو أساس المحول الذي يستند إليه هذا العمل.**

3.3.3. المحوّل Swin [1] :

1.3.3.3. مقدمة:

بعد الدخول المزدهر لمحول الرؤية في عام 2021، أصبح مجتمع الأبحاث مفرط النشاط لتحسين ViT [32]، نظراً لأن ViTs الأصلية كانت متعطشة للغاية للبيانات وتم تدريبها على مجموعة بيانات Google JFT-300M الخاصة الداخلية.

لذا، توصل فريق Facebook AI إلى DeiT [36]، وهو محول كفؤ تجاه البيانات وكان قادراً على التفوق على الشبكات العصبية الالتفافية و ViTs، من حيث مفاضلة الدقة/FLOPS. تم تدريب DeiT [36] على بيانات ImageNet21 فقط. لكنه لم يكن حلّاً خالياً من الالتفاف تماماً.

جاءت Microsoft Research Asia بمحوّل Swin [1] وهو محوّل هرمي للرؤية الحاسوبية يستخدم النوافذ غير المتداخلة، مع فكرة جعل محوّل الرؤية الحاسوبية أكثر عمومية وقابلية للتكيف مع مهام الرؤية الأخرى.

في جميع حلول التعلم العميق، لدينا شبكة عصبية عميقة تُستخدم لاستخراج الميزات، وتسمى هذه الشبكة العمود الفقري. ثم يتم تمرير الميزات المستخرجة إلى رأس مناسب بناءً على الهدف النهائي، مثل التصنيف، الكشف، تحليل المشاعر والترجمة.

تم تقديم SWIN Transformer [1] باعتباره العمود الفقري للأغراض العامة لمهام الرؤية الحاسوبية، والتي يمكن استخدامها لإجراء التصنيف والكشف والتجزئة، بشكل أفضل من الشبكات الالتفافية وViT [32] و DeiT [36].

حازت هذه الورقة البحثية على جائزة David Marr المشرّفة للغاية في مؤتمر ICML'21، نظراً لتصميمها وأدائها الذكيين.

2.3.3.3. تعريفات عامة في نموذج Swin:

- تشير "Win" في "SWIN" إلى النوافذ "windows" حيث يتم تقسيم الصورة في هذا المحوّل إلى نوافذ غير متداخلة.
- تشير "S" في "SWIN" إلى "Shifted" تسمح باتصال النوافذ ببعضها.
- هو نموذج هرمي hierarchical يسمح بالنمذجة من أجل مقاييس متنوعة، حيث يبني "خرائط ميزات هرمية" عبر طبقاته.
- الرقعة: مفهوم تقسيم الصورة إلى رقع Patches هو نفسه في ViT.

إذا كان حجم الصورة الواردة $H \times W \times 3$ (تشير 3 إلى RGB) مقسمة إلى رقع بحجم $4 \times 4 \times 3$، أي أن كل رقعة ستحتوي بالتالي على قيم 48 بكسل بما في ذلك ألوان RGB. يتم تحويل هذه الميزات الـ 48 إلى تضمين خطي أحادي ببعد C.

- حجم القناة (C): يعبر C عن بعد أشعة التضمين عندما يتم تحويل رقع الصورة مبدئياً إلى tokens أحادية البعد. فمثلاً من أجل الطبقة المخفية للمرحلة الأولى يكون C=192 (في النسخة الكبيرة من النموذج). وهكذا يتم تمثيل كل رقعة من خلال تضمين بطول 192.
- النافذة: في Swin، لدينا نوافذ مرتبة لتقسيم الصورة بالتساوي بطريقة غير متداخلة. حيث يتم تقسيم الصورة بحيث تحتوي كل نافذة على M × M رقعة حيث تم اختيار M=7 في الورقة البحثية، أي لدينا $7 \times 7 = 49$ رقعة في كل نافذة.
- الانتباه المحلي: يحدث الانتباه بين الرقع الـ 49 من كل نافذة من نوافذ الصورة. وبالتالي يكون الانتباه محليّاً، أي أن كل رقعة تهتم فقط بـ 49 رقعة في نافذتها (بما في ذلك نفسها). **وهذا ما يخفض التعقيد الحسابي**، مقارنة بالمحول الأصلي الذي يحسب الانتباه الذاتي بين كل العناصر في الصورة فإن التعقيد الحسابي من أجل سلسلة دخل بأبعاد $h \times w \times C$ يكون كما في المعادلة التالية [1]:

$$\Omega\,(MHA) = 4hwC^2 + 2(hw)^2C$$

نلاحظ أن التعقيد يزداد بشكل تربيعي مع زيادة أبعاد الصورة بينما في حال استخدام طريقة النوافذ وحساب الانتباه داخل العناصر ضمن النافذة الواحدة فقط فيكون التعقيد الحسابي من أجل كل نافذة تحوي رقعاً بعدد M × M كما المعادلة التالية [1]:

$$\Omega\,(W-MHA) = 4hwC^2 + 2M^2hwC$$

نلاحظ بأن التعقيد الحسابي متناسب بشكل خطي مع أبعاد الصورة في حال كانت M ثابتة
إن هذا التحفيض في التعقيد الحسابي يجعل من محوّل SWIN مناسب أكثر لتطبيقات الصور ذات الحجوم الكبيرة ولتطبيقات الزمن الحقيقي.

3.3.3.3. بنيان المحول Swin:

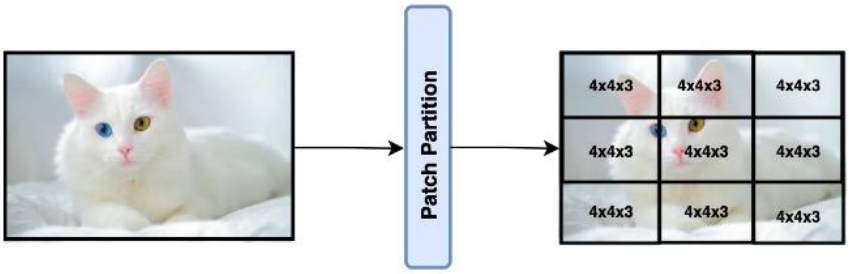
سنتطرق إلى شرح المكونات الرئيسية لـ Swin أولاً، ثم سنجمع كل الأجزاء معاً للحصول على النموذج النهائي:

المكونات الرئيسية لـ Swin الموضح ببنيته الشاملة في الشكل 18 :

- **Patch Partitioning:**

تشير "Patch" إلى أصغر وحدة في خريطة الميزات. يتم تمرير الصورة عبر محول Swin عن طريق تقسيمها إلى رقع غير متداخلة، مثل ViT. يُطلق على كل رقعة اسم token بحجم $4 \times 4 \times 3 = 48$ بكسل، حيث يعود الرقم 3 إلى القنوات RGB والرقم 4 إلى طول وعرض الرقعة المربعة كما في الشكل 13.

.



الشكل 13 تقسيم الصورة الى رقع في مرحلة Patch Partitioning [52]

- **Patch Merging:**

يعد دمج الرقع من أهم الطبقات في بنية محول Swin حيث يعدّ القسم المسؤول عن هرمية النموذج، حيث في الشبكات العصبية الالتفافية مثل ResNet، يتم اختزال خرائط الميزات باستخدام عملية الالتفاف. لذلك لاختزال خرائط الميزات في شبكة محولات خالصة دون استخدام الالتفاف يستخدم Swin Transformer طبقة دمج الرقع.

إذ أن هدف هذه الطبقة الخطية هو تخفيض عدد عناصر سلسلة الدخل أو tokens. ويزداد هذا التخفيض بازدياد عمق النموذج كما نلاحظ من أبعاد دخل كل كتلة المبينة في الشكل 18.

تقوم طبقة دمج الرقع الأولى بسَلسَلة ميزات كل مجموعة مؤلفة من n×n من الرقع المتجاورة باتجاه العمق، حيث تم اختيار n=2 (أي يتم سلسلة كل 2 × 2 من الرقع بالعمق) .

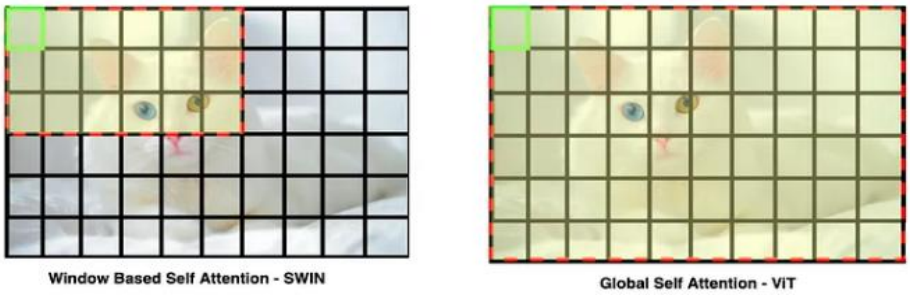
يؤدي هذا إلى تقليل عدد tokens بمقدار 2×2=4 (2× اختزال الدقة). وبالتالي يصبح العمق 4 C بدل C ويتم تطبيق طبقة خطية بحيث يتم ضبط بُعد الإخراج على 2C. وبما أن الجزء الأول من كل مرحلة من المراحل 2 و3 و4 من النموذج هو طبقة patch merging تكون دقة خرج كل مرحلة من المراحل المذكورة على التوالي:

$\frac{H}{8} \times \frac{W}{8} \times 2C$ خرج المرحلة 2 ، $\frac{H}{16} \times \frac{W}{16} \times 4C$ خرج المرحلة 3، $\frac{H}{32} \times \frac{W}{32} \times 8C$ خرج المرحلة 4.

نلاحظ أن الدقة المكانية لخرائط الميزات الهرمية هذه مطابقة لتلك الموجودة في ResNet، وذلك يمكّن Swin Transformers من استبدال شبكات ResNet [15] بشكل ملائم في الأساليب الحالية لمهام الرؤية وبالتالي استخدامه كعمود فقري بديل لتطبيقات الرؤية الحاسوبية المختلفة.

- **Window-Multi Head Self Attention (W-MSA):**

يستخدم محول Swin كتل المرمّز (Encoder blocks) من بنية المحولات الأصلية. بحيث تتكون كل كتلة ترميز من وحدة انتباه ذاتي متعددة الرؤوس وشبكة تغذية أمامية. في ViT *[32]* ، يتم استخدام الانتباه الذاتي متعدد الرؤوس القائم على الجداء السلمي dot product لحساب ترميزات الانتباه لكل رقعة بالنسبة إلى جميع الرقع الأخرى في صورة الإدخال. لذلك في الشكل 14 بالنسبة إلى ViT *[32]*، على اليمين، إذا أردنا حساب الانتباه للرقعة الخضراء العلوية اليسرى، فإننا نهتم بجميع الرموز الأخرى، والتي تصبح تربيعية في الحساب.

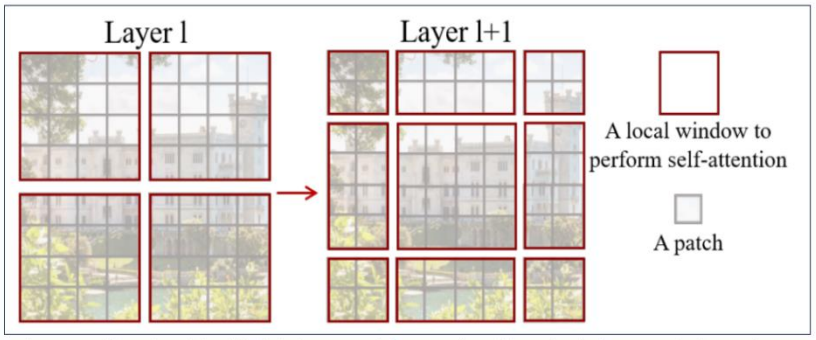


الشكل 14 مقارنة آلية الانتباه الذاتي بين VIT و Swin [52]

في Swin [1] ، يتم تقسيم الصورة إلى نوافذ، بحيث تحتوي كل نافذة على عدد ثابت من الرقع. النافذة في المثال في الصورة اليسرى من الشكل 14، فيها رقع 3x6، في الورقة البحثية [1] ، يتم أخذ نوافذ مربعة وكل نافذة فيها MxM رقعة حيث تم اختيار M=7. الآن من أجل حساب ترميز الانتباه للرقعة اليسرى العلوية، يتم الاهتمام فقط بالرقع في النافذة نفسها. هذا النهج أكثر كفاءة وقابلية للتطوير من الانتباه لجميع الرموز في الصورة.

- **Shifted Window-Multi Head Self Attention (SW-MSA):**

تفتقر وحدة الانتباه الذاتي القائمة على النوافذ إلى الاتصالات عبر النوافذ، مما يحد من قدرتها على النمذجة. لإدخال اتصالات عبر النوافذ مع الحفاظ على الحساب الفعّال للنوافذ غير المتداخلة، تم اقتراح نهج لتقسيم النوافذ المزاحة والذي يتناوب بين تكوينين للتقسيم في كتل Swin Transformer المتتالية.



الشكل 15 نهج النافذة المزاحة في بنية Swin [1]

في الشكل 15 على اليسار، لدينا خريطة ميزات 8 × 8 لنسميها خريطة الميزات "A" في الشكل 16 . مقسمة بالتساوي إلى 4 نوافذ بحجم 4 × 4. هنا، حجم النافذة M = 4. الآن في الجزء الأول من الكتلتين المتتاليتين من كتل المحوّل، يتم حساب الانتباه داخل هذه النوافذ. لكن هناك حاجة أيضاً إلى الانتباه عبر النوافذ حتى تتعلم الشبكة بشكل أفضل وذلك (لأنه لم يعد يتم استخدام سياق انتباه شامل).

لذلك، في الجزء الثاني من كتلة محول Swin، تتم إزاحة النوافذ بمقدار $(\lfloor\frac{M}{2}\rfloor،\lfloor\frac{M}{2}\rfloor)$ رقعة نحو الزاوية اليمنى السفلية من النوافذ المقسمة بانتظام يؤدي هذا للحصول على خريطة الميزات "B" حيث تم تمييز كل نافذة مرة أخرى بلون مختلف في الشكل 16 المرفق، وإجراء الانتباه بين هذه النوافذ الجديدة، هذا يؤدي إلى اتصالات عبر النوافذ.

في هذه الحالة، نظراً لأن M = 4، تتم إزاحة النوافذ بمقدار (2،2). الآن، يتم إجراء الانتباه الذاتي داخل النوافذ المحلية المزاحة.

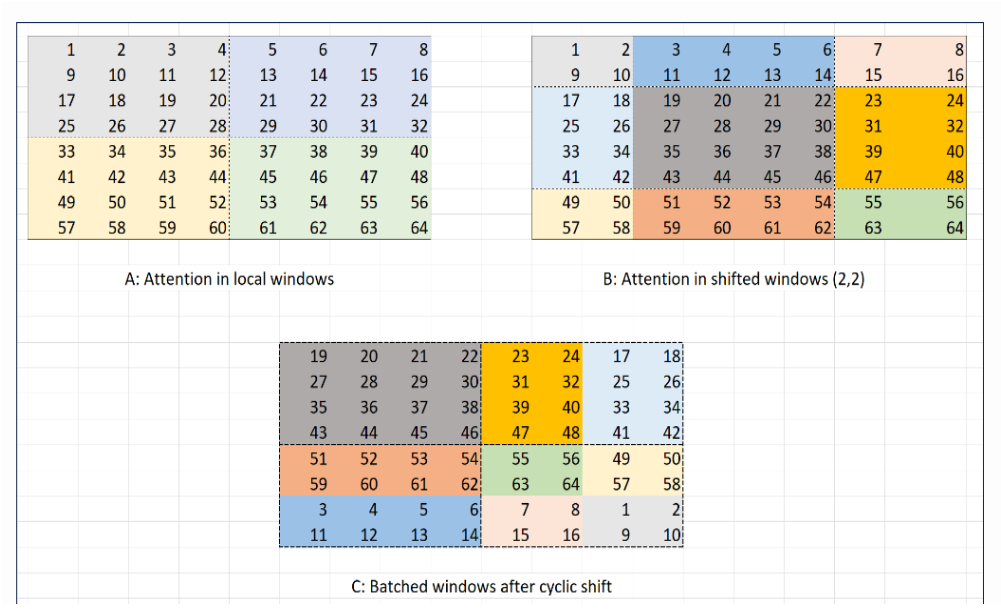
ومع ذلك، ينتج عن هذا التحول رقعاً "معزولة" لا تنتمي إلى أية نافذة، كما ستكون بعض النوافذ ذات حجم أصغر من M × M لذلك يطبق Swin Transformer تقنية "Cyclic Shift"، والتي تنقل الرقع "المعزولة" إلى نوافذ عدد الرقع فيها أصغر من M × M.

حيث يوجد الآن 9 نوافذ. كان من الممكن أن يكون الحل الساذج هو حشو النوافذ التي يقل حجمها عن 4 x 4، ولكن هذا من شأنه أن يؤدي إلى حسابات مكلفة.

لذا تم الاقتراح في الورقة البحثية [1] إجراء "حساب دفعي فعّال" الفكرة هي إنشاء خريطة الميزات "C" كما في الشكل 16 عن طريق التحويل الدوري لخريطة الميزات "A". ولكن، يؤدي هذا للحصول على خريطة الميزات تتكون فيها جميع النوافذ من رقع فرعية لا تكون متجاورة حقاً في الصورة الأصلية، باستثناء النافذة العلوية اليسرى.

يجب أن يكون الانتباه المحلي وفقاً للألوان المميزة في خريطة الميزات "C" وليس وفق الحدود المتقطعة في "C". لذلك فإن الخطوة التالية هي استخدام التقنيع عند إجراء الانتباه المحلي. يتطابق القناع تماماً مع الألوان المميزة في خريطة الميزات "C". وبالتالي يمكن تنفيذ الانتباه مع التقنيع أثناء استخدام تقنية "Cyclic Shift"، سيكون هذا معادلاً لتنفيذ الانتباه المحلي داخل النوافذ في خريطة الميزات "B".

يقدم نهج النافذة المزاحة هذا اتصالات متقاطعة مهمة بين النوافذ، وقد وجد أنه يحسن أداء الشبكة.



1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

A: Attention in local windows

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64

B: Attention in shifted windows (2,2)

19	20	21	22	23	24	17	18
27	28	29	30	31	32	25	26
35	36	37	38	39	40	33	34
43	44	45	46	47	48	41	42
51	52	53	54	55	56	49	50
59	60	61	62	63	64	57	58
3	4	5	6	7	8	1	2
11	12	13	14	15	16	9	10

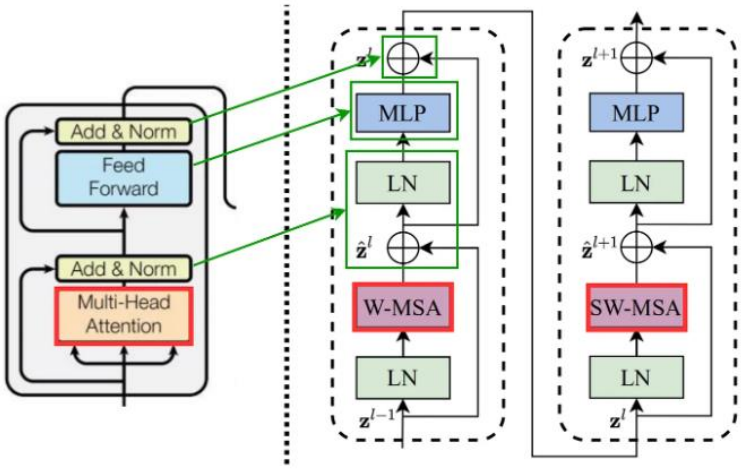
C: Batched windows after cyclic shift

الشكل 16 تمثيل النوافذ المزاحة باستخدام excel [53]

- **SWIN Transformer Block**

في كتلة محول Swin، يتم وضع مرمّزين على التسلسل، ويتم تغذية خرج المرمّز الأول إلى الثاني.

يحسب المرمّز الأول W–MSA والثاني يحسب SW–MSA على خرج المرمّز الأول. تشبه البنية العامة لكتلة محولات Swin كتلة ترميز المحولات الأصلية، باستثناء آلية حساب الانتباه. حيث بشكل مختلف عن MSA الشامل، تحتوي كتلة محول Swin على W–MSA وSW–MSA كما في الشكل التالي:



الشكل 17 كتلة ترميز المحولات بالمقارنة مع كتلة Swin Transformer [52]

4.3.3.3. المعمارية الشاملة:

تم تقسيم تدفق صورة المدخلات عبر محول SWIN إلى 4 مراحل مبينة في الشكل 18، وهي كالتالي:

I. المرحلة 1:

1. يتم تمرير صورة الإدخال عبر طبقة patch partition، لتقسيمها إلى رقع ذات حجم ثابت. إذا كانت الصورة بحجم $H \times W$، وكان حجم الرقعة 4x4، فإن خرج طبقة التقسيم إلى رقع يعطينا رقع بعدد $\frac{H}{4} \times \frac{W}{4}$.
كل رقعة لها أبعاد بكل قناة، لذا فإن كل رقعة تكون بحجم 4 × 4 × 3 = 48 بكسل. لتحويل كل رقعة من 48 بكسل إلى بعد أفضل C يتم تمرير كل رقعة عبر طبقة خطية، والتي تقوم بإسقاط كل رقعة على بعد C، والآن لدينا رقع كل منها ببعد C وبالتالي تكون خريطة الميزات بحجم $\frac{H}{4} \times \frac{W}{4} \times C$.

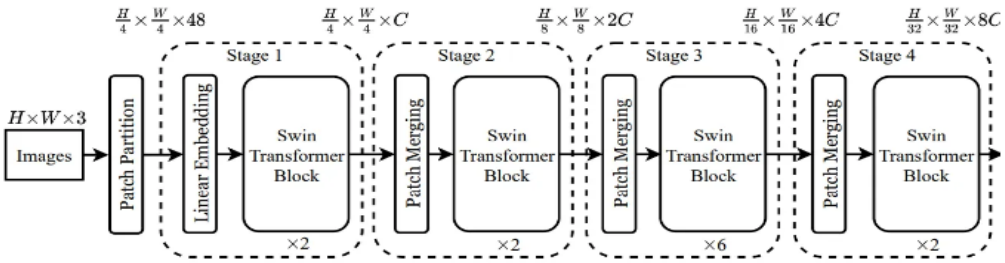
2. يتم تمرير خريطة الميزات هذه عبر كتلة محول Swin، ونظراً لأن كتلة محول Swin تتكون من كتل "transformer encoder"، فإن حجم المدخلات والمخرجات يبقى كما هو. وبالتالي، يبقى خرج كتلة محول Swin في المرحلة 1 مماثلاً لحجم خريطة ميزات الإدخال، أي $\frac{H}{4} \times \frac{W}{4} \times C$

II. المرحلة 2:

1. يتم الآن تمرير خريطة الميزات ذات الحجم $\frac{H}{4} \times \frac{W}{4} \times C$ عبر طبقة patch merging، التي تقوم باختزال الدقة بمقدار x2 وزيادة عمق خريطة الميزات بمقدار 2 كما شرحنا.
وبالتالي $\frac{H}{4} \times \frac{W}{4} \times C$ تصبح $\frac{H}{8} \times \frac{W}{8} \times 2C$.
2. يتم تمرير خريطة الميزات عبر كتلة محول Swin أخرى، والتي تحافظ على أبعادها سليمة.

III. المرحلة 3 والمرحلة 4:

1. يتم تكرار نفس الإجراء الخاص بالمرحلة 2، وستقل دقة خريطة الميزات بمقدار النصف بعد المرور على كل طبقة patch merging في كل مرحلة كما شرحنا سابقاً.



الشكل 18 البنية الشاملة لـ Swin Transformer في الإصدار Tiny [1]

يوضح الشكل 18 نسخة النموذج Tiny وفق عدد كتل المحول في كل مرحلة وحجم خرائط الميزات، مما يعكس أنه بينما نتعمق في الشبكة، فإن دقة خريطة الميزات تتناقص وعمقها يتزايد.

4.3.3. نمذجة الفيديو قبل Video SWIN:

جميع نماذج الفيديو التي تعتمد على المحولات قبل video Swin [7] (النموذج الذي سيتم شرحه في قسم النموذج المقترح الذي يقوم عليه هذا العمل) بُنيت على طبقات المحولات التي تربط الرقع بشكل شامل عبر الأبعاد المكانية والزمانية.

ونظراً لأن النمذجة الزمانية المكانية المشتركة ليست اقتصادية أو سهلة التحسين، فقد تم اقتراح تحليل النطاقات المكانية والزمانية إلى عوامل لتحقيق مفاضلة أفضل بين السرعة والدقة. حيث في المحاولات الأولية للتعرف على الفيديو المستند إلى المحولات، يتم أيضاً استخدام نهج التحليل إلى العوامل الزمانية والمكانية، وقد ثبت أن هذا يقلل بشكل كبير من حجم النموذج دون انخفاض كبير في الأداء من الأمثلة على ذلك:

- تقترح VTN [37] إضافة مُرمّز انتباه زمني فوق ViT [32] المدرب مسبقاً، والذي ينتج عنه أداء جيد في التعرف على الإجراء في الفيديو.
- ViViT [38] يقوم النموذج باستخراج الرموز الزمانية المكانية من فيديو الإدخال، والتي يتم ترميزها بعد ذلك بواسطة سلسلة من طبقات المحولات. من أجل التعامل مع التسلسلات الطويلة من الرموز المميزة التي تتم مواجهتها في الفيديو، يتم اقتراح العديد من المتغيرات الفعّالة للنموذج الذي يقوم بتحليل الأبعاد المكانية والزمانية للمدخلات باستخدام تصميمات مختلفة لنموذج ViT [32] المُدرّب مسبقاً.
- MViT [39] عبارة عن محول رؤية متعدد المقاييس للتعرف على الفيديو تم تدريبه من نقطة الصفر ويقلل من الحساب عن طريق تجميع الانتباه للنمذجة الزمانية المكانية.

أما في [7]، تم أولاً دراسة المنطقة الزمانية المكانية ثم أُظهر تجريبياً أن محول الفيديو ذو الانحياز المكاني الزماني يتفوق على أداء جميع محولات الرؤية الأخرى في مهام التعرف على الفيديو المختلفة كما سنرى في الفصل القادم.

4.3. المقاربات الحديثة في هذا المجال:

سيتم في هذا القسم شرح تفصيلي ل 3 مقاربات في حل مسألة التعرف على نشاط المجموعة، عام 2020 [40] و [41]، وعام 2022 [42]على التوالي.

1.4.3. Actor-Transformers for Group Activity Recognition:

تسعى هذه الورقة [40] إلى التّعرف على الإجراءات الفردية والأنشطة الجماعية من مقاطع الفيديو، حيث يتم اقتراح نموذج قادر على التعلم واستخراج المعلومات ذات الصلة بشكل انتقائي للتعرف على نشاط المجموعة.

ظهرت شبكة المحولات [14] كطريقة متفوقة لمهام معالجة اللغات الطبيعية لأنها تعتمد على آلية الانتباه الذاتي التي مكّنتها من صياغة التبعيات بشكل أفضل بين الكلمات بمرور الوقت دون مكوّن عودي تسمح هذه الآلية للشبكة باستخراج المعلومات والعلاقات الأكثر صلة بشكل انتقائي، وانطلاقاً من ذلك تم الافتراض في هذا البحث أن شبكة المحولات يمكنها أيضاً أن تكون نموذجاً أفضل للعلاقات بين الممثلين في المشهد وتجمع المعلومات على مستوى الممثل للتعرف على نشاط المجموعة مقارنة بالنماذج التي تتطلب قيوداً مكانية وزمانية واضحة.

1.1.4.3. النهج المتّبع في هذا البحث:

يتكوّن النهج المتّبع من ثلاث مراحل رئيسية مبينة في الشكل 19: **استخراج ميزات الممثل وتجميع النشاط الجماعي والاندماج.**

باختصار، الدخل إلى النموذج عبارة عن سلسلة من إطارات الفيديو F_t بالشكل التالي حيث $t=1,2....,T$ مع التزويد بـN مربع إحاطة حول الممثلين في كل إطار، حيث N عدد الممثلين في المشهد و T هو عدد الإطارات.

في المرحلة الأولى يتم الحصول على التمثيل الثابت والديناميكي لكل ممثل من خلال تطبيق شبكة pose ثنائية الأبعاد على إطار واحد وشبكة CNN ثلاثية الأبعاد على جميع إطارات الدخل. بحيث يتم بناء التمثيل الديناميكي من إطارات RGB أو إطارات التدفق البصري، والتي تتم معالجتها بواسطة شبكة CNN ثلاثية الأبعاد متبوعة بطبقة RoIAlign [27].

بعد أن تم الحصول على هذه التمثيلات يتم تضمينها في فضاء جزئي بحيث يتم تمثيل كل ممثل بمتجه أحادي البعد.

في المرحلة الثانية، يتم تطبيق شبكة محولات على هذه التمثيلات للحصول على الميزات على مستوى الإجراء. ويتم تجميع هذه الميزات وفق القيمة الكُبرى MaxPooling لالتقاط الميزات على مستوى نشاط المجموعة.

يتم استخدام المصنف الخطي للتنبؤ بالإجراءات الفردية ونشاط المجموعة باستخدام الميزات على مستوى الإجراء وعلى مستوى نشاط المجموعة، على التوالي.

كما يتم تقديم استراتيجيات الاندماج قبل وبعد شبكة المحولات لاستكشاف فائدة دمج المعلومات عبر تمثيلات مختلفة.

2.1.4.3. استعراض المراحل:

استخراج ميزات الممثل:

تتضمن جميع أفعال الإنسان حركة مفاصل الجسم، مثل اليدين والساقين. ينطبق هذا على الإجراءات الدقيقة التي يتم إجراؤها في الأنشطة الرياضية وعلى الإجراءات اليومية مثل المشي والتحدث. هذا يعني أنه من المهم التقاط ليس فقط موقع المفاصل ولكن أيضاً ديناميكياتها الزمنية. لهذا الغرض، يتم استخدام نموذجين أساسيين مختلفين لالتقاط كل من الموضع والحركة للمفاصل والممثلين أنفسهم.

للحصول على مواضع المفاصل، يتم تطبيق نموذج تقدير الوضعية. يتلقى كدخل صندوقاً يحيط بالممثل ويتنبأ بموقع المفاصل الرئيسية. تم اختيار HRNet المنشورة مؤخراً [43] كشبكة وضعيات نظراً لتصميمها البسيط نسبياً، وتحقيقها أحدث النتائج في معايير تقدير الوضعية. يتم استخدام الميزات من الطبقة الأخيرة من الشبكة، مباشرة قبل طبقة التصنيف النهائية، في جميع التجارب. على وجه التحديد، يتم استخدام نسخة الشبكة الأصغر وهي pose_hrnet_w32 المدربة على نقاط COCO الرئيسية [44]، والتي تُظهر أداء جيد بما يكفي للمهمة الحالية أيضاً.

الشبكة الأساسية الثانية مسؤولة عن نمذجة الديناميكيات الزمنية. حيث أظهرت العديد من الدراسات أن شبكات CNN ثلاثية الأبعاد، مع ما يكفي من البيانات المتاحة للتدريب [[45]، [46]]، يمكنها بناء تمثيلات مكانية–زمانية قوية للتعرف على الإجراءات. وفقاً لذلك، يتم استخدام شبكة I3D [46] في إطار هذا العمل، نظراً لأن شبكة الوضعية وحدها لا يمكنها التقاط حركة المفاصل من إطار واحد.

تعالج شبكة I3D الإطارات المكدسة F_t، حيث $t=1,2....,T$ بوساطة inflated 3d convolutions. يتم الأخذ في الاعتبار تمثيلات التدفق الضوئي و RGB لأنها يمكن أن تلتقط جوانب مختلفة من الحركة. نظراً لأن شبكات CNN ثلاثية الأبعاد مكلفة من الناحية الحسابية، يتم استخدام طبقة RoIAlign [27] لاستخراج ميزات لكل ممثل باستخدام N مربع إحاطة حول الممثلين ومعالجة إطارات الدخل بالكامل بواسطة الشبكة مرة واحدة فقط.

المحوّل:

تتكون شبكة المحوّلات من جزأين: ترميز وفك ترميز. في هذا البحث، تم تطبيق فقط جزء الترميز من بنية المحولات تاركين جزء وحدة فك الترميز للعمل في المستقبل.

حيث كما ذكرنا أن الانتباه A في المحولات هو دالة تمثل مجموعاً موزون للقيم V. يتم حساب الأوزان عن طريق مطابقة استعلام Q مع مجموعة من المفاتيح K. يمكن أن يكون لوظيفة المطابقة أشكالاً مختلفة، والأكثر شيوعًا هو الجداء السلّمي المقاس. يمكن كتابة الانتباه باستخدام وظيفة مطابقة الجداء السلمي المقاسة على النحو التالي:

$$A(Q,K,V) = \text{softmax}\left(\frac{Q\,K^T}{\sqrt{d}}\right) V$$

حيث d هو بُعد كل من الاستعلامات والمفاتيح. في وحدة الانتباه الذاتي، يتم حساب التمثيلات الثلاثة (Q، K، V) من تسلسل الإدخال S عبر الإسقاطات الخطية:

$$A(S) = A(Q(S), K(S), V(S))$$

في نمذجة سلسلة إلى سلسلة، تعطي هذه الآلية أهمية أكبر للكلمات الأكثر صلة في سلسلة المصدر، هذه خاصية مرغوبة للتعرف على نشاط المجموعة أيضاً لأنه يمكن تحسين المعلومات الخاصة بميزات كل ممثل بناءً على الممثلين الآخرين في المشهد دون أي قيود مكانية.

تتكون طبقة الترميز في المحول المستخدم E من الانتباه متعدد الرؤوس جنباً إلى جنب مع الشبكة العصبية للتغذية الأمامية، حيث أن الانتباه متعدد الرؤوس A_h هو امتداد للانتباه مع العديد من وظائف الانتباه المتوازية باستخدام إسقاطات خطية منفصلة h_i ل (Q, K, V).

يمكن أن يحتوي مرمز المحول على العديد من هذه الطبقات التي تعالج بالتتابع الإدخال S.

S هنا عبارة عن مجموعة من ميزات الممثلين $S = \{si|i = 1, .., N\}$ التي تم الحصول عليها بواسطة مستخرجات ميزات الممثل. نظراً لأن الميزات s_i لا تتبع أي ترتيب معين، فإن آلية الانتباه الذاتي هي نموذج ملائم لتحسين وتجميع هذه الميزات.

يتم إظهار أن مُرمّز المحوّل يمكن أن يستفيد ضمنياً من استخدام العلاقات المكانية بين الممثلين عبر الترميز الموضعي لـ s_i، من خلال تمثيل كل مربع محيط ثنائي لميزات الممثل المعني s_i بنقطة المركز (x_i, y_i) وترميز النقطة المركزية بنفس الوظيفة PE في المحول الأصلي [14].

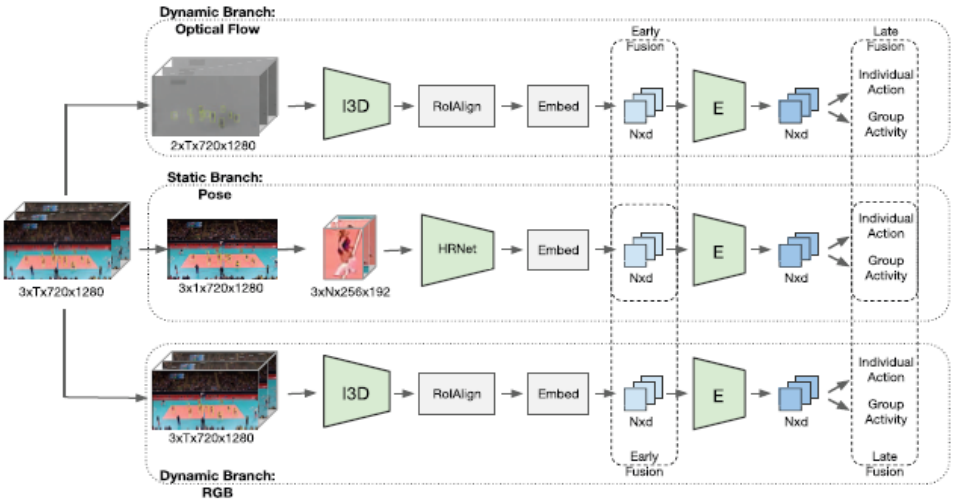
الاندماج:

يتم تمثيل الفرع الثابت من خلال pose network التي تلتقط الوضعية الثابتة لمفاصل الجسم، بينما يتم تمثيل الفرع الديناميكي بواسطة I3D وهو مسؤول عن الميزات الزمنية لكل ممثل في المشهد. نظراً لأن RGB والتدفق البصري يمكنهما التقاط جوانب مختلفة من الحركة، يتم دراسة الفروع الديناميكية مع كل من تمثيلات فيديو الإدخال.

لدمج الفروع الثابتة والديناميكية، يتم استكشاف استراتيجيتين للاندماج:

الاندماج المبكر لخصائص الممثلين قبل شبكة المحولات والاندماج المتأخر الذي يجمع تنبؤات المصنفات، على غرار [47]. يتيح الاندماج المبكر الوصول إلى كل من الميزات الثابتة والديناميكية قبل استنتاج نشاط المجموعة. يعالج الاندماج المتأخر بشكل منفصل الميزات الثابتة والديناميكية للتعرف على نشاط المجموعة ويمكن أن يركز على الميزات الثابتة أو الديناميكية، بشكل منفصل. ولكن أوضحت الاختبارات أن الدمج المبكر ليس مفيداً للنموذج، حيث يؤدي أداءاً مشابهاً أو حتى أسوأ من نماذج الفرع الواحد.

تتطلب استراتيجيات الاندماج المبكرة من المحول أن يفكر في كل من الميزات الثابتة والديناميكية. ونظراً لصغر حجم مجموعة بيانات الكرة الطائرة، لا يستطيع النموذج استغلال هذا النوع من الدمج بشكل كامل. يساعد التركيز على كل من التمثيلين بشكل منفصل النموذج على استخدام إمكانات الميزات الثابتة والديناميكية بشكل أفضل.



الشكل 19 بنية نموذج Actor-Transformers [40]

بنظرة عامة على النموذج يمكن تلخيصه وفق ما يلي:

الدخل: فيديو مؤلف من T من الإطارات و N مربع إحاطة ل N ممثل.

تتم معالجة الدخل بواسطة فرعين: ثابت وديناميكي.

الفرع الثابت: يقوم بإخراج تمثيلات الوضعية لكل مربع محيط بكل ممثل عبر شبكة HRNet.

الفرع الديناميكي: يعتمد على I3D، والذي يتلقى كإدخال إما إطارات RGB أو إطارات تدفّق بصري مكدّسة لاستخراج ميزات على مستوى الممثل، بعد I3D يتم تطبيق طبقة RoIAlign [27] .

يعمل مرمّز المحوّل E على تنقيح وتجميع الميزات على مستوى الممثل متبوعة بمصنفات النشاط الفردي والنشاط الجماعي.

يدعم النموذج استراتيجيتين للاندماج. من أجل الاندماج المبكر، يتم جمع الميزات على مستوى الممثل من الفرعين قبل (E)، وفي الاندماج المتأخر يتم الجمع بين درجات تنبؤ المصنّف.

3.1.4.3. اختبارات Actor-Transformers والنتائج:

من أجل التبسيط، تم التعبير عن الفرع الثابت باسم "Pose"، والفرع الديناميكي بإطارات RGB باسم "RGB" والفرع الديناميكي بإطارات التدفق البصري باسم "Flow".

غالباً ما يكون أداء نماذج المحولات أفضل عندما يكون لديها المزيد من الطبقات والرؤوس. وذلك لأن المزيد من الطبقات والرؤوس تسمح للنموذج بالتقاط أنماط وعلاقات أكثر تعقيداً في البيانات. بالإضافة إلى ذلك، تستفيد هذه النماذج عادةً من مجموعات البيانات الكبيرة، حيث يساعد المزيد من البيانات في تدريب نماذج أكثر تعقيداً دون الإفراط في المُلاءمة.

ونظراً لصغر حجم مجموعة بيانات الكرة الطائرة، يعترف المؤلفون بأن نموذج المحوّلات ذو الطبقات والرؤوس المتعددة لن يصل إلى إمكاناته الكاملة. وذلك لأنه لا توجد بيانات كافية لتدريب مثل هذا النموذج المعقّد بشكل صحيح. ولمعالجة هذه المشكلة، اختاروا استخدام نسخة أبسط من نموذج المحولات في تجاربهم، نسخة تحتوي على طبقة واحدة فقط ورأس واحد. من المحتمل أن يتم اتخاذ هذا القرار لمنع المُلاءمة الزَّائدة وللتأكد من أن النموذج لا يزال قادراً على تعلم أنماط ذات معنى من مجموعة البيانات الأصغر.

تم تقييم هذا النهج على مجموعتي بيانات للنشاط الجماعي المتاحة للعامة وهي مجموعة بيانات الكرة الطائرة (التي نستخدمها في عملنا والتي سنشرحها مفصلاً) [2] ومجموعة البيانات الجماعية Collective Activity Dataset [29] التي تتكوّن من 44 مقطعاً بأطوال مختلفة تبدأ من 193 إطار إلى حوالي 1800 إطار في كل مقطع. يحتوي كل إطار عاشر على تعليق توضيحي للمربعات المحيطة بالأشخاص مع أحد الإجراءات الفردية الخمسة: (العبور، والانتظار، والوقوف في الطابور، والمشي، والتحدث)، يتم تحديد نشاط المجموعة من خلال الإجراء الذي يقوم به معظم الأشخاص في المقطع.

تم استخدام دقة التعرّف على نشاط المجموعة كمقياس وذلك باتباع نفس النهج المتّبع في الأعمال ذات الصلة.

حيث باستخدام 2 backbones هما HRNet + I3D تم تحقيق أعلى دقة على مجموعة المعطيات volleyball باستخدام الفرعين Pose+Flow وهي 94.4% للتعرف على نشاط المجموعة.

وباستخدام 2 backbones هما I3D + I3D تم تحقيق أعلى دقة على مجموعة المعطيات Collective باستخدام الفرعين RGB+Flow أي الفرعين الديناميكيين دون استخدام معلومات ال pose وهي 92.8% للتعرف على نشاط المجموعة.

3.4.2. Social Adaptive Module for Weakly-supervised Group Activity Recognition:

الهدف في هذه الورقة البحثية [41] تقديم مهمة جديدة تسمى التعرف على نشاط المجموعة الخاضع للإشراف الضعيف والتي تختلف عن مهام GAR التقليدية من حيث توفر تسميات مستوى الفيديو فقط (كما في عملنا المقترح في هذا البحث) حيث عادةً ما تستخرج هذه الأساليب الأخرى ميزات كل شخص وفقاً للمربعات المحيطة المقابلة حيث يكون كل شخص في المشهد موسوم بالإجراء الفردي الذي يقوم به، ثم تقوم بدمج ميزات مستوى الشخص في تمثيل واحد لكل إطار. وبذلك فإن الأساليب السابقة حساسة للعدد المتغير للأشخاص في كل إطار وتتطلب مواقع صريحة لهم، وهي محدودة في التطبيقات العملية.

يُقترح استخدام وحدة التكيف الاجتماعي (Social Adaptive Module) SAM للعثور على التمثيل الفعال على مستوى الشخص وعلى مستوى الإطار لمواجهة تحديات المدخلات غير المؤكدة والبيانات غير ذات الصلة في إطارات الفيديو وذلك بناءً على الافتراض الاجتماعي بأن التمثيلات الرئيسية (الأشخاص/الإطارات) عادةً ما ترتبط ارتباطاً وثيقاً ببعضها البعض، حيث تم تصميم SAM لتحديد مربعات الإحاطة والإطارات التمييزية بشكل انتقائي من مقاطع الفيديو مما يتيح التّعرف الأكثر دقة بأنشطة المجموعة في بيئة تعليمية ضعيفة الإشراف.

3.4.2.1. النهج المتّبع:

بالنسبة لمهمة التعرف على نشاط المجموعة، فإن التمثيل الوسيط المكوّن من السمات الفردية والعلاقات الأساسية فيما بينها، يُشار إليه بالتمثيل الاجتماعي في هذه الورقة [41].

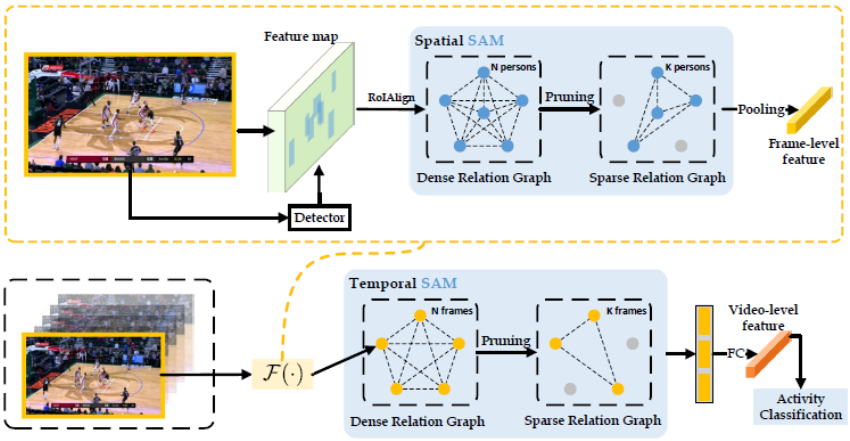
الفكرة الأساسية هي أولاً بناء كل التمثيلات الاجتماعية الممكنة ثم العثور على التمثيلات الفعالة بناءً على الافتراض الاجتماعي بأن الأمثلة الرئيسية (الأشخاص/الإطارات) ترتبط ارتباطًا وثيقًا ببعضها البعض. حيث يتم استخراج التمثيلات الرئيسية من خلال العلاقات الاجتماعية، وذلك باستخدام وحدة التكيف الاجتماعي (SAM). تم تصميم SAM لتحديد تمثيلات على مستوى الشخص ومستوى الإطار بشكل تكيفي من خلال الاستفادة من الافتراض الاجتماعي بأن الحالات الرئيسية مترابطة بشكل وثيق. يقوم هذا النهج ببناء وتقليم الرسوم البيانية للعلاقات الكثيفة حيث يحتوي الرسم البياني الكثيف على العديد من العلاقات، ليست جميعها على نفس القدر من الأهمية لفهم نشاط المجموعة. ولذلك، يتم تطبيق تقليم على هذا الرسم البياني الكثيف لتصفية الاتصالات الأقل صلة والتركيز على العقد الأكثر صلة، وتحويلها إلى رسوم بيانية للعلاقات المتفرقة للتضمين العلائقي الفعال. تتناول هذه الطريقة تحديات المدخلات غير المؤكدة في البيئات ضعيفة الإشراف، مما يتيح استخلاص السمات التمييزية من البيانات الحاوية على ضجيج.

2.2.4.3. مراحل التنفيذ في هذه الورقة:

اكتشاف الأشخاص واستخراج الميزات: بالنسبة لكل إطار، تم اعتماد أولاً Faster-RCNN [33] المدرب مسبقًا على MS-COCO [44] لاكتشاف الأشخاص المحتملين في المشهد. بعد ذلك، يجري تتبعها على جميع الإطارات بواسطة متتبع الارتباط correlation tracker. بعد ذلك، تم اعتماد ResNet-18 [15] كعمود فقري لاستخراج خريطة الميزات الالتفافية لكل إطار، ثم يتم الحصول على الميزات المحاذية لكل مربع إحاطة من خريطة الميزات بواسطة RoIAlign [27] بحجم اقتصاص قدره 5 × 5 يتم بتضمينها في متجه ميزات ذو 1024 بعد بواسطة طبقة متصلة بالكامل.

وحدة التكيف الاجتماعي: يأتي بعدها دور هذه الوحدة لاختيار الميزات الفعالة K من بين الميزات المدخلة N.

في الشكل 20، من أجل GAR الخاضع للإشراف الضعيف، تكون المدخلات عبارة عن مجموعة من الإطارات والمربعات المحيطة المكتشفة مسبقاً للأشخاص، حيث يتم تطبيق SAM لتحديد الميزات التمييزية على مستوى الشخص بشكل متزامن في المجال المكاني والتمثيلات الفعالة على مستوى الإطار في المجال الزمني.



الشكل 20 نظرة عامة على النهج المقترح في SAM [41]

3.2.4.3. اختبارات SAM والنتائج:

تم تقديم مجموعة بيانات جديدة قائمة على الفيديو في هذه الورقة [41]، وهي مجموعة بيانات NBA. تصف الأنشطة الجماعية الشائعة في ألعاب كرة السلة. حيث تم تنميطها بوسم النشاط الجماعي لكل مقطع. تحوي مجموعة البيانات 9172 مقطع فيديو، ينتمي كل منها إلى أحد تسعة أنشطة.

تم استخدام دقة التعرّف على نشاط المجموعة في تقييم أداء النموذج على مجموعتي البيانات NBA و Volleyball.

تم تحقيق دقة 50.3% على مجموعة بيانات NBA. وبالنسبة إلى مجموعة بيانات Volleyball تم دمج الصفين pass و set وبالتالي أصبح عدد الصفوف الإجمالي 6، كما تم استخدام 3 إطارات من كل مقطع، وحصلوا على دقة 93.1%.

3.4.3. Pose is all you need: The pose only group activity recognition system (POGARS):

الهدف في هذه الورقة البحثية [42] عام 2022 هو استكشاف القوة التنبؤية لبيانات الوضعية (pose) من خلال تطوير نموذج تعلّم عميق يتنبأ بتسمية نشاط مجموعة الكرة الطائرة من تعقب الوضعيه فقط، دون استخدام بيانات RGB بشكل مباشر.

انطلقوا في هذه الورقة البحثية من مبدأ أن استخدام الوضع لتمثيل البشر له مزايا عديدة، تتجاهل تمثيلات الوضعية العوامل الأقل صلة في بيانات الإدخال مثل ميزات مظهر الشخص والخلفية التي تسمح للنماذج المدربة باستخدام الوضعية للتعميم بشكل أفضل على المواقف الجديدة مثل الأحداث الرياضية التي تتضمن فرقاً وأماكن غير مُشاهدة من قبل. يسمح تجاهل ميزات المظهر التي يمكن تحديدها على مستوى الشخص لبيانات الوضع بالحفاظ على خصوصية الأشخاص الذين يتم التقاط بياناتهم.

بالإضافة إلى معلومات الوضعية، في [48] تم استكشاف استخدام مسارات الكائنات ذات الصلة (مثل الكرة) لتحديد أنشطة المجموعة في مقاطع الفيديو الرياضية، ووجدوا أن معلومات الإدخال الإضافية مثل مسارات الكرة في لعبة الكرة الطائرة يمكن أن تحسن بشكل كبير دقة التصنيف. تدعم التجارب في هذه الورقة [42] هذه النتيجة بشكل جزئي، مع إدراج تعقب الكرة تظهر نتائج تجريبية محسّنة قليلاً. ومع ذلك، فإن النهج المتّبع لا يعتمد على مسارات الكرة لتحقيق دقة تصنيف عالية.

تستخدم POGARS شبكات التفافية أحادية البعد لتعلم الديناميكيات الزمانية المكانية للأفراد المشاركين في النشاط الجماعي وذلك باستخدام تقديرات النقاط الرئيسية للوضعية الخاص بممثلي المشهد وتغيرات مواضع هذه النقاط.

يستخدم النموذج المقترح آلية الانتباه المكاني والزماني حيث أن آلية الانتباه الزمني قادرة على تحديد الأهمية المحدّدة لكل إطار في مقطع الفيديو بينما توفر آلية الانتباه المكاني درجة الأهمية لكل فرد بناءً على مشاركته في نشاط جماعي معين وذلك انطلاقاً من فكرة أنه في معظم النشاطات الجماعية، هناك أشخاص معينون يلعبون أدواراً مركزية أكثر من غيرهم، وبالتالي يكونون أكثر تنبؤاً بمهمة التعرف على نشاط المجموعة.

فمثلاً عند التعرف على نشاط spike في الكرة الطائرة، يتوقع أنه ينبغي إيلاء المزيد من الاهتمام للاعبين المشاركين بشكل مباشر في الرمي السريع وصد الكرة. يستخدم POGARS آلية الانتباه الذاتي المكاني لتحديد أهمية كل فرد في نشاط جماعي معين.

بالإضافة إلى ذلك، الانتباه الزمني مهم أيضاً. على سبيل المثال، يمكن اعتبار إطارات الفيديو التي تُمرَّر فيها الكرة الطائرة بمثابة لحظة زمنية مهمة في مقطع الفيديو. يستخدم النهج المقترح آلية الانتباه الذاتي الزمنية لتعيين أوزان أهمية مختلفة لإطارات الفيديو المختلفة.

1.3.4.3. مجموعة البيانات المستخدمة وبعض المقارنات:

على الرغم من أن التعرف على نشاط المجموعة هو مجال بحث مهم في الرؤية الحاسوبية، إلا أن هناك القليل من مجموعات بيانات الفيديو المتاحة لتقييم النموذج، أُجريت التجارب في هذه الورقة على مجموعة بيانات الكرة الطائرة وفق ما يلي:

تحقق POGARS دقة تصل إلى 93.2٪ للتعرف على نشاط المجموعة من خلال استخدام معلومات الوضعية المتعقبة فقط في مجموعة بيانات الكرة الطائرة بينما نموذج actor–transformer الذي تم التحدث عنه في الورقة السابقة [40] يحقق 92.3٪ عند استخدام الوضعية فقط (تتطلب الدقة البالغة 94.4٪ بيانات التدفق الضوئي كطريقة إدخال إضافية).

تم التنويه في هذا البحث عن فائدة أخرى لاستخدام الوضعيات المتعقبة فقط هي أن النموذج المدرب يمكن أن يعمم بشكل أفضل لاختبار البيانات ذات الخصائص المختلفة من خلال تجاهل عوامل مثل ظروف الإضاءة، ولون الفريق الموحد، ولون الملعب، ومظهر الجمهور، وما إلى ذلك.

ولاختبار هذه الفرضية، أنشؤوا تقسيماً منحرفاً للتدريب/الاختبار لبيانات الكرة الطائرة حيث تكونت بيانات التدريب من المباريات التي تم لعبها في مكان واحد فقط في أولمبياد لندن 2012 وتألفت مجموعة الاختبار من المباريات من جميع الملاعب الأخرى. تظهر النتائج أن دقة POGARS انخفضت بشكل طفيف فقط من 93.2٪ إلى 89.7٪ عند استخدام البيانات المنحرفة. في المقابل، أظهر نموذج I3D [46] الذي تم تدريبه باستخدام صور RGB كمدخلات انخفاضاً كبيراً في الدقة، من 84.6٪ إلى 73.9٪.

2.3.4.3. نظرة عامة على المقاربة المقترحة في POGARS:

يتمثل أحد أكبر التحديات التي تواجه نمذجة البيانات في الفيديو في العثور على التماسك الزمني (تتبع كيف تنتقل نقطة واحدة في إحدى الأطر إلى الإطار التالي). لنمذجة التماسك الزمني في تسلسل الوضعيات، يتم تعيين كل نقطة رئيسية keypoint لبعد مختلف في متجه الوضعية. ومن ثم يمكن لشبكات أحادية البعد التفافية أن تصمّم بشكل فعّال التغييرات في موضع كل نقطة رئيسية عبر الزمن. حيث أن النقاط الأساسية المستخدمة عددها 16 نقطة أساسية (الكاحلين، الركبتين، الوركين، الحوض، العمود الفقري، الرقبة، الرأس، الرسغين، المرفقين والكتفين) لكل لاعب في إطارات الفيديو.

بالإضافة إلى تسليط الضوء على أهم البيانات للتعرف على النشاط، يتم أيضاً استخدام آلية الانتباه الذاتي لتعيين وزن أعلى للأشخاص والإطارات الأكثر أهمية لتصنيف نشاط المجموعة بدقة.

بالنظر إلى مقطع فيديو إدخال V يحتوي على N من الأشخاص وإطارات عددها (البعد الزمني) T

أولاً، يتم الحصول على إحداثيات الموقع والمربعات المحيطة المرتبطة بكل شخص عبر الوقت في فيديو الإدخال من التعليقات التوضيحية اليدوية المستخدمة في [49].

بعد ذلك، يتم إنشاء 16 تقديراً ثنائي الأبعاد للنقاط الرئيسية لكل فرد عن طريق تغذية خوارزمية الساعة الرملية المكدسة لتقدير الوضع البشري[50] بمسارات الصندوق المحيط.

كما يتم تغذية مسارات إحداثيات الموضع إلى طبقة متصلة بالكامل لإنشاء تضمين ميزة وصفية لموضع كل فرد في كل إطار. تتم سلسلة كل من تضمين الموقع وتمثيلات النقاط الأساسية للوضعية لإنتاج تضمين مركب لكل فرد تحتوي على معلومات الوضعية والموقع F .

تساهم تمثيلات الميزات للإطارات المتتالية في مقطع الفيديو بدرجات مختلفة في توقُّع الصّف. لذلك يتم إدخال الميزات التي تمت سلسلتها F لكل شخص في كل إطار في آلية الانتباه الزمني لحساب أوزان الانتباه حسب الإطار.

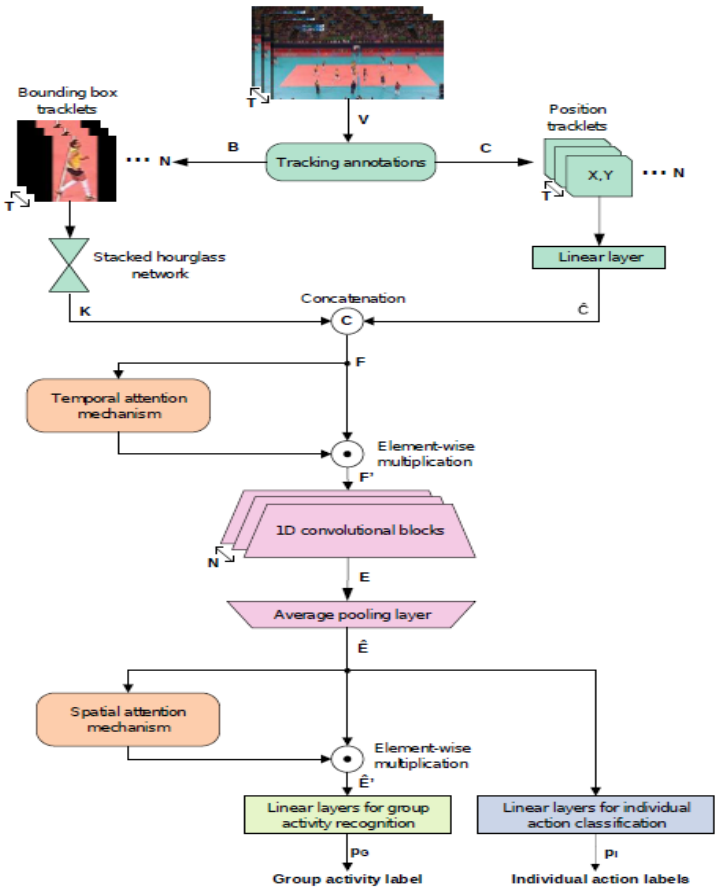
يتم تنفيذ عملية الضرب حسب العنصر بين تضمين الميزات وأوزان الانتباه المحسوبة حسب كل عنصر عبر البعد الزمني وينتج عن ذلك ميزات تضمين مقطع الفيديو الموزونة وفقاً للأهمية الزمنية للإطار F' . قبل تغذية الميزات في مجموعة من كتل CNN أحادية البعد.

كتل CNN المتتابعة ذات الوصلات المتبقية "residual " قادرة على تعلم الديناميكيات الزمنية للأفراد باستخدام تضمينات الموقع والوضعية. حيث كما ذكرنا في الطريقة المقترحة، يتم تمثيل الأشخاص في مقطع الفيديو من خلال تضمينات ميزات أوضاعهم المتتبعة. حيث تم استخدام شبكات 1D CNN لنمذجة التطورات الزمنية للأفراد . حيث كل كتلة التفافية في الشكل 21 تتكون من ثلاث طبقات CNN أحادية الأبعاد مع اتصال تخطي، ويحتوي POGARS على أربع كتل من هذه الطبقات مكدسة معاً. تأخذ المكدس تضمينات ميزات ما بعد الاهتمام F' كمدخلات وتحلل حركة الأفراد من خلال إنتاج تمثيل E يتم تجميع هذه التضمينات في المتوسط في البعد الزمني لإنتاج تضمين الميزات لكل شخص في مقطع الفيديو $\hat{E}$.

تؤدي POGARS التعلم متعدد المهام من خلال التنبؤ بكل من تسمية النشاط الجماعي p_G وتسميات الإجراءات الفردية p_I.

يتم تحقيق ذلك من خلال إرفاق رأسين منفصلين (أحدهما للتنبؤ بالنشاط الجماعي والآخر للتنبؤ بالنشاط الفردي) بمجموعة الميزات المخرجة من كتل 1D CNN.

لتنفيذ الانتباه المكاني، يتم ضرب أوزان الانتباه لكل فرد مع تضمين الميزات للشخص في مقطع الفيديو $\hat{E}$.لضمان حصول تضمينات الميزات للأشخاص الرئيسيين في الفيديو على مزيد من الاهتمام في عملية التنبؤ بنشاط المجموعة ينتج عن ذلك تضمين لجميع الأفراد في مقطع الفيديو الموزونة وفقًا لأهمية كل شخص $\widehat{E'}$، حيث يوضّح الشكل 21 ملخص آلية العمل.



الشكل 21 ملخص لآلية عمل نظام POGARS [42]

5.3. خاتمة

في هذا الفصل تمت مراجعة الطرق المتبعة لحل مشكلة التعرف على نشاط المجموعة من الطرق التقليدية إلى أبرز وأحدث الأبحاث التي وجدنا أنها بدأت بإدخال آلية الانتباه المستخدمة في المحولات في حل هذه المشكلة.

وجرى استعراض كل حل مع ما استعمل من مجموعة معطيات والنتائج التي حصل عليها.

نستنتج من خلال الدراسة المرجعية بأن الاتجاه في تطوير نظم التصنيف هو الاعتماد على التعلم العميق وخاصة نماذج المحوّل.

في POGARS يرى المؤلفون أنه يمكن أن يؤدي استخدام ميزات الوضع المستخرجة بدلاً من ميزات RGB إلى إنتاج نموذج أكثر قدرة على التعميم عبر مجموعات البيانات نظراً لأن الوضع يستبعد عوامل من المدخلات، مثل ظروف الإضاءة ولون زيّ الفريق ولون الملعب ومظهر الجمهور وعوامل الإزعاج الأخرى. ولكن من وجهة نظر أخرى ممكن أن نرى أنه لو تم في

POGARS اختبار النموذج على مجموعة بيانات Collective التي تحوي فئاتها المختلفة على وضعيات بشرية متشابهة عندها ستفشل الخوارزميات المعتمدة على الوضع فقط دون أخذ سياق المشهد كاملاً أو العناصر الأخرى الموجودة في فيديو معين، مثلا وضعية "عبور الشارع" مشابهة أو مطابقة لوضعية "المشي" ولكن الخطوط في الشارع تلعب دور في التصنيف يتم إهماله في مثل هذه المنهجيات في الحل.

كما وجدنا أن النتائج تُظهر أن المحولات تبدو وكأنها آلات عامة جداً، مثل العثور على بنية قادرة على تعلم أية مهمة على أي نوع من بيانات الإدخال "ذكاء اصطناعي حقيقي". يمكن الآن استخدام بنية واحدة في أي مجموعة من المهام سواء كانت صورة أو نص أو سلسلة زمنية أو مقاطع فيديو، وما إلى ذلك. يمكن تدريبهم الآن بمزيد من البيانات أكثر من أي وقت مضى.

أصبح من الممكن الآن إنشاء بنية موحدة في التعلم الآلي ويمكننا أن نرى تطورات مثيرة في هذا المجال في السنوات القليلة المقبلة.

وبناءً على ذلك تم اقتراح الحل الذي سنرى تفاصيله في الفصل القادم كما سيتم الاستفاضة في شرح نسخة المحوّل المستخدمة التي بدأت تحل محل الشبكات الالتفافية في مسائل الرؤية الحاسوبية.

الفصل الرابع: الحل المقترح

1.4. تمهيد:

سنتناول في هذا الفصل المخطط العام للحل المقترح حيث سنستعرض كل جزئية في المخطط العام للحل بشكل مفصّل.

سنتكلم في هذا الفصل عن الورقة البحثية التي اعتمدنا عليها بشكل أساسي في بحثنا وهي Video Swin Transformer [7] التي تعتمد على توسيع المحوّل Swin [1] إلى المجال الزمني، وذلك باستغلال بنية المحول Swin التي تحدثنا عنها في الفصل السابق والتي تم تصميمها لنمذجة الصّور، وتكييفها من أجل نمذجة الفيديو مع الاستمرار في الاستفادة من قوّة نماذج الصور المدرّبة مسبقاً.

حيث قمنا باستخدام المحوّل Video Swin Transformer [7] كعمود فقري لاستخراج الميزات من مقاطع الفيديو في مجموعة البيانات المستخدمة لدينا Volleyball [2] ثم تمريرها إلى رأس تصنيف I3d head [46] للتنبؤ بالصف الصحيح للنشاط، حيث ينقسم النموذج في غالب معماريات الشبكات العصبية إلى جزأين: العمود الفقري، وهو مجموعة من الطبقات التي تستخرج الميزات من المدخلات، والرأس، وهو مجموعة من الطبقات التي تقوم بالتنبؤات بناءً على الميزات المستخرجة من العمود الفقري.

تم اختيار هذا النموذج كنموذج أساس لحل مشكلة التعرف على نشاط المجموعة في الفيديو انطلاقاً من فكرة أن Video Swin Transformer [7] ممكن اعتباره خط أساس بسيط ولكنه قوي للدراسة المستقبلية. توفر هندسته المعمارية المبتكرة، التي تستفيد من النوافذ المزاحة للتمثيل الزماني والمكاني، منظوراً جديداً في فهم الفيديو. إن كفاءة النموذج في التعامل مع أطوال الفيديو المختلفة وقدرته على التكيف مع مهام تحليل الفيديو المختلفة تجعله أداة جيدة جداً لتطوير البحث في مجالات مثل التعرف على الإجراءات وتصنيف الفيديو.

2.4. نموذج Video Swin Transformer

1.2.4. مقدمة:

أدى النجاح الكبير الذي حققته محولات الصور إلى دراسة البنى القائمة على المحولات لمهام التعرف على الفيديو حيث شهد مجتمع الرؤية تحولاً في النمذجة من CNNs إلى Transformers، إذ حققت بنية Transformer الخالصة دقة قصوى في معايير التعرف على الفيديو الرئيسية، ولكن تم بناء جميع نماذج الفيديو هذه على طبقات المحولات التي تربط الرقع بشكل شامل عبر الأبعاد المكانية والزمانية.

أمّا في الورقة البحثية [7] ، فقد دعى الباحثون إلى التحيّز الاستقرائي للمحلية في محولات الفيديو من خلال الاستفادة من المحلية الزمانية المكانية المتأصلة في مقاطع الفيديو، حيث من المرجّح أن تكون البكسلات الأقرب لبعضها البعض في المسافة الزمانية المكانية أكثر ارتباطاً، بسبب هذه الخاصية، يمكن تقريب الانتباه الذاتي الزماني المكاني الشامل جيداً عن طريق الانتباه الذاتي المحسوب محلياً، **مع توفير كبير في الحساب وحجم النموذج**. مما يؤدي إلى مقايضة دقة – سرعة بشكل أفضل مقارنة بالنهُج السابقة التي تحسب الانتباه الذاتي الشامل حتى مع التحليل للعوامل المكانية والزمانية. يتم تحقيق المحليّة في بنية الفيديو المقترحة من خلال تكيّيف Swin Transformer [1] المصمم لمجال الصورة، مع الاستمرار في الاستفادة من قوة نماذج الصور المدربة مسبقاً. حقق هذا النهج دقة متطورة في مجموعة واسعة من معايير التعرف على الفيديو.

2.2.4. التهيئة من النموذج المدرب مسبقاً SWIN:

نظراً لأن البنية مقتبسة من Swin Transformer [1]، يمكن تهيئتها بسهولة باستخدام النموذج الذي تم تدريبه مسبقاً على مجموعة بيانات صور واسعة النطاق ImageNet-21K.

بالمقارنة مع Swin Transformer الأصلي في [1]، فإن كتلتين فقط من كتل البنية في Video Swin Transformers لهما أشكال مختلفة، طبقة التضمين الخطيّة في المرحلة الأولى وانحيازات الموضع النسبي.

في إصدار الفيديو من المحوّل Swin، يتم تعديل رمز الإدخال token ليشمل المعلومات الزمنية (البعد الزمني) الموجودة في بيانات الفيديو. حيث يتم توسيع رمز الإدخال إلى بُعد زمني قدره 2، وبالتالي يصبح شكل طبقة التضمين الخطية 96×C من 48×C في Swin [1] الأصلي. وذلك لأنه يتم اعتبار إطارين متتاليين معاً، لذلك سيكون حجم كل رمز إدخال 4x4x3x2 (96). وهذا يعني أن طبقة التضمين الخطية يجب أن تستوعب رموز الإدخال الكبيرة هذه، ويصبح شكلها 96×C بدلاً من 48×C الأصلي. وهذا، يتم بتكرار الأوزان مباشرة في النموذج المُدرّب مسبقاً مرتين ثم يتم ضرب المصفوفة بأكملها بمقدار 0.5 للحفاظ على متوسط وتباين الإخراج دون تغيير.

شكل مصفوفة انحياز الموضع النسبي هو (2M−1, 2M−1) في Swin [1] الأصلي (حيث يعبر M عن حجم النافذة التي تحوي M × M رقعة)، لكن البعد الجديد لهذه المصفوفة نظراً لوجود بعد زمني في محول الفيديو يجب أن يكون (2P−1, 2M−1, 2M−1) حيث P هو البعد الزمني للنافذة ثلاثية الأبعاد حيث تصبح النوافذ ثلاثية الأبعاد بحجم P × M × M في نسخة الفيديو من المحول (كما سيتم شرحها في الفقرات التالية).

لجعل انحياز الموضع النسبي هو نفسه داخل كل إطار، يتم ضرب المصفوفة في النموذج المُدرب مسبقاً بمقدار 2P − 1 مرة للحصول على الشكل (2P−1, 2M−1, 2M−1) لمصفوفة انحياز الموضع النسبي من أجل التهيئة.

باستخدام نموذج تم تدريبه مسبقًا على ImageNet−21K، تم إيجاد أن معدل التعلّم للعمود الفقري للنموذج يجب أن يكون أصغر (على سبيل المثال 0.1×) من معدل التعلّم للرأس، الذي تتم تهيئته بشكل عشوائي.

ونتيجة لذلك، ينسى العمود الفقري المعلّمات والبيانات المدربة مسبقاً ببطء أثناء ملاءمة مدخلات الفيديو الجديد، مما يؤدي إلى تعميم أفضل. تشير هذه الملاحظة إلى اتجاه لمزيد من الدراسة حول كيفية الاستفادة بشكل أفضل من الأوزان المدرّبة مسبقاً.

3.2.4. البنية العامة لنموذج Video Swin:

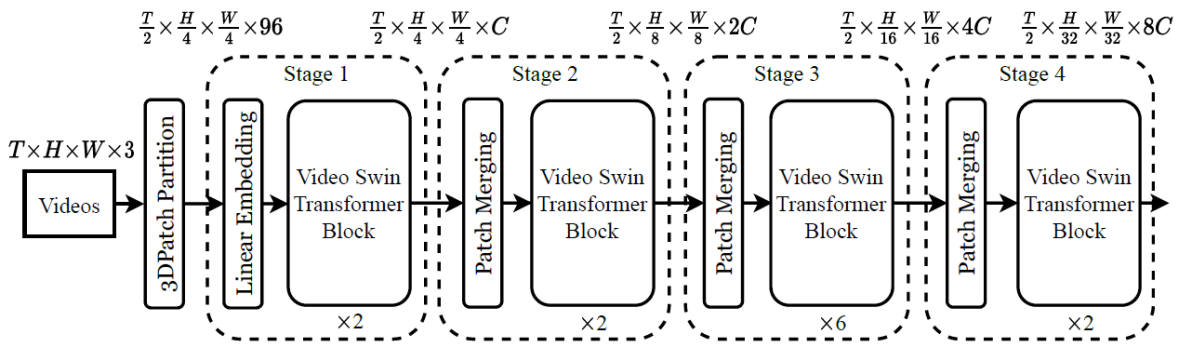
يظهر الهيكل العام المقترح في الشكل 22 ، والذي يوضح نسخته الصغيرة (Swin−T). يتم تعريف فيديو الإدخال على أنه بحجم T × H × W ×3، أي يتألف من T إطار يحتوي كل منها على H × W × 3 بكسل.

في Video Swin Transformer [7] ، يتم التعامل مع كل رقعة ثلاثية الأبعاد بحجم 2 × 4 × 4 × 3 كأنها token.

وبالتالي، فإن طبقة 3D patch partitioning تُنتج $\left(\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}\right)$ tokens ثلاثي الأبعاد، حيث كل token ببعد 96. يتم بعد ذلك تطبيق طبقة تضمين خطية لإسقاط ميزات كل token إلى بُعد يُشار إليه ب C.

إن عدم الاختزال على طول البعد الزمني يسمح باتباع البنية الهرمية لمحول Swin [1] الأصلي بصرامة والتي تتكون من أربع مراحل وتقوم بالاختزال المكاني بمقدار (2×) في طبقة patch merging (كما شرحنا في فقرات المحول Swin) من كل مرحلة. تُسلسل طبقة patch merging ميزات كل مجموعة مؤلفة من (2 × 2) من الرقع المتجاورة مكانياً وتطبق طبقة خطية لإسقاط الميزات إلى بعد يساوي نصف أبعادها.

المكون الرئيسي في هذه البنية هو كتلة Video Swin Transformer، والذي تم إنشاؤه عن طريق استبدال وحدة الانتباه الذاتي متعددة الرؤوس (MSA) في طبقة Transformer القياسية بوحدة الانتباه الذاتي متعددة الرؤوس المستندة إلى النافذة ثلاثية الأبعاد وتسمى (3D(S)W−MSA) مع الحفاظ على المكونات الأخرى دون تغيير.



الشكل 22 الهيكل العام لبنية نموذج Video Swin Transformer بنسخته Tiny [7]

3D Shifted Window based MSA Module .1.3.2.4

تم اتباع Swin Transformer [1] من خلال تقديم التحيز الاستقرائي المحلي إلى وحدة الانتباه الذاتي، والتي تبين لاحقاً أنها فعّالة في التعرف على الفيديو.

- **الانتباه الذاتي المتعدد الرؤوس القائم على النوافذ ثلاثية الأبعاد غير المتداخلة:**

أثبتت آليات الانتباه الذاتي متعدد الرؤوس في النوافذ ثنائية الأبعاد غير المتداخلة أنها ذات كفاءة في التعرّف على الصور. هنا يتم توسيع هذا التصميم بشكل مباشر لمعالجة مدخلات الفيديو.

بالنظر إلى فيديو مؤلف من $T' \times H' \times W'$ من tokens الثلاثية الأبعاد وحجم $P \times M \times M$ للنافذة ثلاثية الأبعاد، يتم ترتيب النوافذ لتقسيم الفيديو المدخل بالتساوي بطريقة غير متداخلة. أي يتم تقسيم tokens الدخل إلى $\lceil\frac{T'}{P}\rceil \times \lceil\frac{H'}{M}\rceil \times \lceil\frac{W'}{M}\rceil$ من النوافذ الثلاثية الأبعاد غير المتداخلة.

- **النوافذ ثلاثية الأبعاد المُزاحة:**

نظراً لتطبيق آلية الانتباه الذاتي متعددة الرؤوس داخل كل نافذة ثلاثية الأبعاد غير متداخلة، هناك نقص في الاتصالات عبر النوافذ المختلفة، مما قد يحد من قوة التمثيل للبنية.

وبالتالي، تم توسيع آلية النوافذ ثنائية الأبعاد المزاحة لـ Swin Transformer [1] إلى نوافذ ثلاثية الأبعاد بغرض تقديم اتصالات عبر النوافذ مع الحفاظ على الحساب الفعال للانتباه الذاتي المستند إلى النوافذ غير المتداخلة.

بالنظر إلى فيديو مؤلف من $T' \times H' \times W'$ من ال tokens الثلاثية الأبعاد وحجم $P \times M \times M$ للنافذة ثلاثية الأبعاد، لطبقتين متتاليتين، تستخدم وحدة الانتباه الذاتي في الطبقة الأولى استراتيجية تقسيم النافذة العادية التي تم ذكرها في الفقرة السابقة.

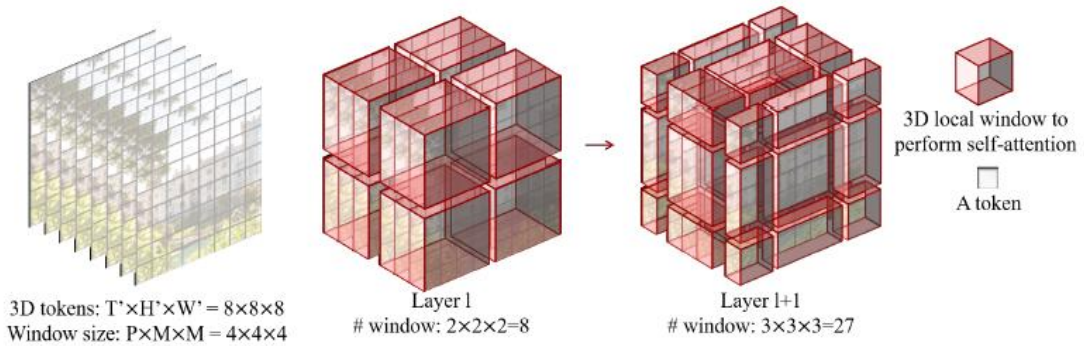
وبالنسبة لوحدة الانتباه الذاتي في الطبقة الثانية، يتم إزاحة ترتيب تقسيم النافذة على طول المحاور الزمنية والارتفاع والعرض بمقدار tokens $(\frac{P}{2}, \frac{M}{2}, \frac{M}{2})$ من وحدة الانتباه الذاتي للطبقة السابقة.

في الشكل 23 مثال توضيحي، حجم الإدخال هو 8×8×8، وحجم النافذة في المثال هو 4×4×4. بما أن الطبقة l تتبنى تقسيم النوافذ بشكل منتظم، فإن عدد النوافذ في الطبقة l هو 2×2×2=8. بالنسبة للطبقة $l + 1$ ، حيث يتم إزاحة النوافذ بمقدار $(\frac{P}{2}, \frac{M}{2}, \frac{M}{2}) = (2,2,2)$ يصبح عدد النوافذ 3×3 ×3= 27. على الرغم من زيادة عدد النوافذ، يمكن اتباع حساب الدُفعات الفعال في محوّل Swin [1] لإعدادات الإزاحة بحيث يبقى العدد النهائي من النوافذ للحساب 8.

باستخدام نهج تقسيم النافذة المزاحة، يتم حساب كتلتين متتاليتين من محول Video Swin كالتالي:

$$\hat{z}^l = 3DW - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1}$$
$$z^l = FFN\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l$$
$$\hat{z}^{l+1} = 3DSW - MSA\left(LN\left(z^l\right)\right) + z^l$$
$$z^{l+1} = FFN\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1}$$

حيث تشير $\hat{z}^l$ و z^l في المعادلات السابقة إلى ميزات إخراج الوحدة 3DW–MSA والوحدة FFN للكتلة l ، على التوالي، تشير 3DW–MSA و 3DSW–MSA إلى الانتباه الذاتي متعدد الرؤوس المستند إلى النوافذ ثلاثية الأبعاد باستخدام تقسيم النوافذ العادية والمزاحة، على التوالي، وبالتالي تشير $\hat{z}^{l+1}$ و z^{l+1} في المعادلات السابقة إلى ميزات إخراج الوحدة 3DSW–MSA والوحدة FFN للكتلة $l + 1$ ، على التوالي.



الشكل 23 مثال توضيحي للنوافذ المزاحة ثلاثية الأبعاد [7]

- **انحياز الموضع النسبي ثلاثي الأبعاد:**

أظهرت العديد من الأعمال السابقة أنه قد يكون من المفيد تضمين انحياز موضعي نسبي $B \in R^{P^2 \times M^2 \times M^2}$ لكل رأس في حساب الانتباه الذاتي:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q\,K^T}{\sqrt{d}} + B\right) V$$

حيث $Q, K, V \in R^{PM^2 \times d}$ هي مصفوفات الاستعلام والمفتاح والقيمة، d هو بُعد ميزات الاستعلام والمفتاح، و PM^2 هو عدد الرموز المميزة في نافذة ثلاثية الأبعاد.

4.2.4. إصدارات النموذج:

على غرار Swin transformer [1]، تم تقديم أربعة إصدارات مختلفة من النموذج Video Swin Transformer هي كالتالي:

Swin-T: C = 96, layer numbers= {2, 2, 6, 2}
Swin-S: C = 96, layer numbers= {2, 2, 18, 2}
Swin-B: C = 128, layer numbers= {2, 2, 18, 2}
Swin-L: C = 192, layer numbers= {2, 2, 18, 2}

حيث تشير C إلى بعد شعاع التضمين في المرحلة الأولى من النموذج.
هذه الإصدارات الأربعة هي (0.25 × ، 0.5 × ، 1 × و 2 ×) من حجم النموذج الأساسي والتعقيد الحسابي، على التوالي.
يتم ضبط حجم النافذة على P = 8 و M= 7 افتراضياً.
بُعد الاستعلام q لكل رأس هو d= 32، ويتم تعيين طبقة التوسيع لكل MLP على α=4

5.2.4. نتائج نموذج Video Swin:

يُظهر النهج المقترح في هذه الورقة البحثية [7] أداءً قوياً في مهام التّعرف على الفيديو الخاصة بالتّعرف على الإجراءات الفردية، وذلك على مجموعات بيانات Kinetics-400/Kinetics-600 والنمذجة الزمنية على Something-Something v2 (التي يشار لها بـ SSv2).
بالنسبة للتعرف على الإجراء في الفيديو، تم تحقيق 84.9% top-1 accuracy على بيانات Kinetics-400 و 85.9% top-1 accuracy على بيانات Kinetics-600 حيث تفوق بذلك قليلاً على النتائج الحديثة السابقة في ViViT [38] بمقدار 0.1 ، وذلك بحجم نموذج أصغر (200.0 مليون معلّمة في نسخة النموذج Swin-L مقابل 647.5 مليون معلّمة لـ ViViT-H) ومجموعة بيانات أصغر للتدريب المسبق (ImageNet-21K مقابل JFT-300M).
بالنسبة للنمذجة الزمنية على SSv2، فقد حصل النموذج على 69.6% top-1 accuracy وهو تحسن بمقدار 0.9 نقطة مقارنة ب MViT [39].

نستعرض في الجدول 1 مقارنة نتائج نموذج Video Swin Transformer مع أحدث ما توصلت إليه الدراسات على مجموعة معطيات Kinetics-600. حيث في الجدول يشير "384↑" إلى أن النموذج يستخدم دقة مكانية أكبر تبلغ 384×384، وتشير "views" إلى عدد المقاطع الزمنية × عدد الاقتصاصات المكانية، الوحدات المستخدمة هي Giga (10^9)، وMega (10^6) لـ FLOPs و Param على التوالي:

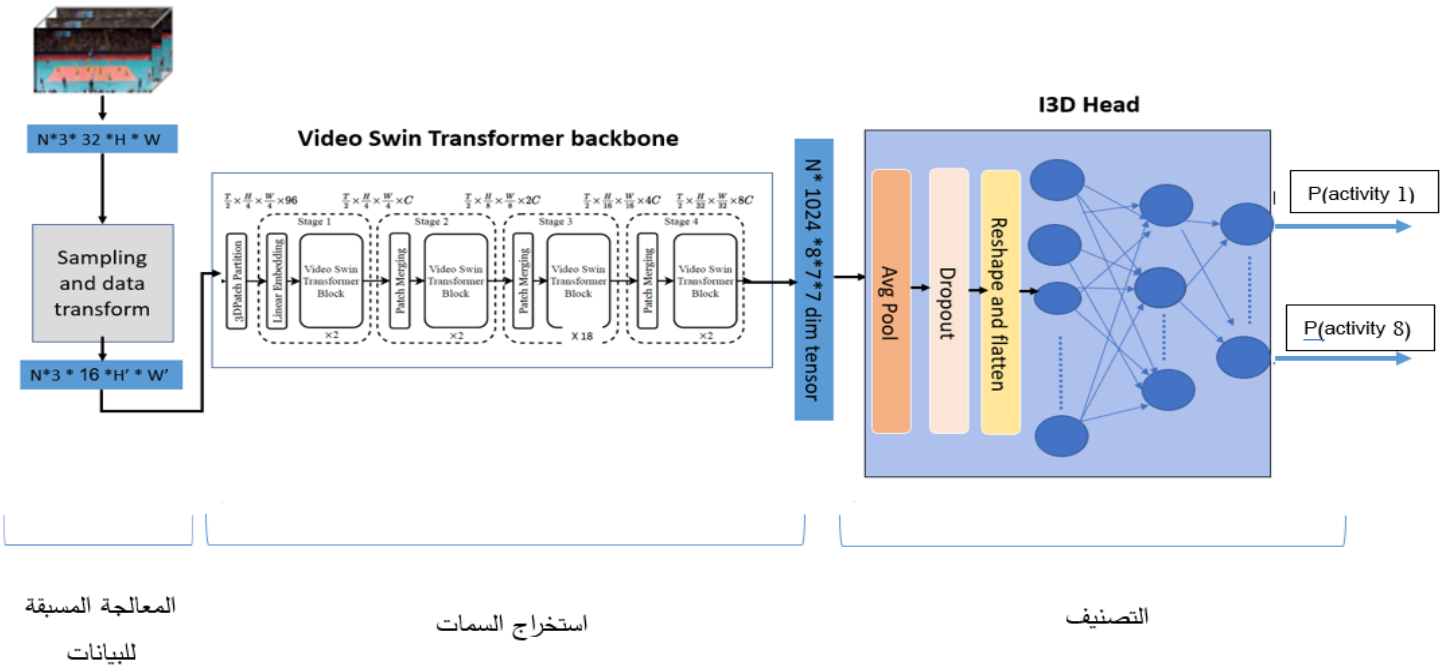
Method	Pretrain	Top−1	Top−5	Views	FLOPs	Param
SlowFast R101+NL	−	81.8	95.1	10 × 3	234	59.9
X3D XL	−	81.9	95.5	3 × 10	48	11.0
MViT−B−24, 32×3	−	83.8	96.3	5 × 1	236	52.9
TimeSformer−HR	ImageNet−21K	82.4	96	1 × 3	1703	121.4
ViViT−L/16x2 320	ImageNet−21K	83.0	95.7	4 × 3	3992	310.8
ViViT−H/16x2	JFT−300M	85.8	96.5	3 × 4	8316	647.5
Swin−B	**ImageNet−21K**	**83.8**	**96.4**	**4 × 3**	**282**	**88.1**
Swin−L (384↑)	**ImageNet−21K**	**85.9**	**97.1**	**4 × 3**	**2107**	**200.0**

الجدول 1 مقارنة نتائج نموذج Video Swin Transformer على مجموعة معطيات kinetics−600 [7]

بعد أن تم استعراض بنية النموذج الأساسي الذي نستند إليه في هذا العمل بشكل معمق والنتائج التي حققها في مجموعات معطيات التعرّف على الإجراء والتي كانت حافزاً لاستخدامه لحل مشكلة التعرف على النشاط الجماعي في الفيديو، سيتم استعراض وشرح بنية النموذج المقترح لحل مسألة بحثنا.

3.4. النموذج المقترح:

يبين الشكل 24 بنية النموذج الذي تم تصميمه بالشكل الموافق لحل مشكلة التّعرف على نشاط المجموعة في الفيديو:



الشكل 24 بنية نموذج نهجنا المقترح

1.3.4. مراحل تنجيز النموذج:

في هذا البحث تم استخدام مجموعة المعطيات Volleyball [2] التي تم تدريب النموذج المقترح عليها وإجراء الاختبارات.

1. **المعالجة المسبقة للبيانات:**

- تجهيز مجموعة المعطيات Volleyball [2] التي تم تحميلها إلى Google Drive (سنتطرق إلى شرحها بالتفصيل في الفصل القادم)، وذلك بإعادة هيكلة ملفات التنويط الموافقة للمعطيات بشكل يناسب دخل النموذج.

يتم استخدام حجم الدفعة 64 (أي أن N في الشكل 24 تساوي 64)، عدد إطارات الفيديو الكلّية المستخدمة عند الإدخال إلى النموذج هو 32، H,W تشير إلى أبعاد الصورة المكانية، 3 تشير إلى العمق المعبّر عن RGB.

- من أجل كل فيديو يتم تمرير معطيات التدريب عبر مسار البيانات data pipeline وهو عبارة عن سلسلة من خطوات المعالجة المسبقة المطبقة على بيانات الفيديو قبل تغذيتها في الشبكة العصبية للتدريب. وفيما يلي شرح كل خطوة:

أخذ العينات Sample Frames:

تتعلق هذه الخطوة باختيار الإطارات من الفيديو. بحيث اخترنا clip_len=16 يعني أن كل مقطع سيحتوي على 16 إطاراً. وframe_interval=2 لضبط التباعد بين الإطارات، وهو شكل من أشكال الاختزال الزمني لتقليل الحمل الحسابي. و num_clips=1 أي أنه يتم أخذ مقطع واحد من كل فيديو للمجموعة التدريبية.

ترميز الإطار الخام Raw Frame Decode:

يتم ترميز الإطارات الأولية المستخرجة من مقاطع الفيديو إلى تنسيق يمكن معالجته بواسطة النموذج.

تغيير الحجم Resize:

تضمن خطوة تغيير الحجم هذه أن يتم تغيير حجم الجانب الأقصر من الإطار إلى 256 بكسل مع الحفاظ على نسبة العرض إلى الارتفاع للإطارات وذلك كخطوة تحضيرية قبل الاقتصاص.

Random Resize Crop:

بعد تغيير الحجم، يتم أخذ جزء من الإطار بشكل عشوائي. يخدم هذا غرضين: يساعد النموذج على التعميم بشكل أفضل، ويقلل الأبعاد المكانية إلى حجم يمكن التحكم فيه للشبكة.

تغيير الحجم Resize:

يتم تغيير حجم الإطارات التي تم اقتصاصها عشوائياً إلى حجم ثابت يبلغ 224 × 224 بكسل.

تنظيم Normalize:

يتم تطبيق التنظيم باستخدام القيم المتوسطة والقياسية المتوفرة لكل قناة ألوان (RGB). تقوم هذه الخطوة بضبط قيم البكسل بحيث يكون لمجموعة البيانات متوسط 0 وانحراف معياري 1. ويساعد ذلك في تسريع التدريب من خلال توفير الاستقرار الرقمي.

تنسيق شكل الإدخال Format Shape:

تقوم هذه الخطوة بتنسيق شكل tensor الإدخال ليكون مناسباً للشبكة العصبية. 'input_format='NCTHW يعني أنه سيتم تشكيل البيانات في tensor بأبعاد تقابل (عدد المقاطع، القنوات، البعد الزمني (الإطارات)، الارتفاع، العرض).

تشكل هذه الخطوات معاً مساراً شاملاً للمعالجة المسبقة للبيانات، يحول بيانات الفيديو الأولية إلى تنسيق موحد وجاهز لتدريب نموذج التعلم العميق. ويضمن المسار هذا أن تكون بيانات الإدخال بتنسيق متّسق، ومعزز لتحسين التعميم، وموحد لتسهيل التدريب الفعال.

- ومما سبق يتضح أن دخل النموذج الموضح في الشكل 24 هو:

$$N \times 3 \times 16 \times H' \times W' = 64 \times 3 \times 16 \times 224 \times 224$$

2. استخراج الميزات:

تطرقنا إلى شرح النموذج Video Swin transformer [7] القائم على توسيع نموذج Swin [1] إلى البعد الزمني، وهو العمود الفقري المستخدم في عملنا والمبيّن في الشكل 24 بنسخته الأساسية Swin Base المكوّن من 4 مراحل أساسية حيث يكون عدد كتل محول Video Swin في كل مرحلة وبعد أشعة التضمين C في المرحلة الأولى كما يلي:

$$\text{Swin-B: } C = 128, \text{ layer numbers} = \{2, 2, 18, 2\}$$

تعمل كل كتلة على خرائط الميزات بدقة محددة وبعد القنوات (يشار إليها بـ HxWxC). وهذا ما يسمح لها بالنمذجة على مستويات مختلفة وتكون فعّالة في التعامل مع الأبعاد العالية لبيانات الفيديو. حيث يتم استخراج الميزات البصرية بالشكل التالي:

يصبح عدد رموز الإدخال بعد مرحلة التقسيم إلى رقع ثلاثية الأبعاد 56 × 56 × 8 رمزاً ثلاثي الأبعاد ومن خلال المرور بجميع مراحل النموذج يصبح tensor الإخراج من أجل عينة فيديو واحدة ممثلاً للميزات المستخرجة من المقطع المدخل، حيث تصبح خريطة المعالم المكانية ببعد 7 × 7 والبعد الزمني لها 8، وكل عنصر في تلك الخريطة له عمق 1024.

تتناقص الأبعاد المكانية خلال المراحل الأربعة (56 × 56، 28 × 28، 14 × 14، 7 × 7) على الترتيب، ويزداد العمق عند المرور بمراحل النموذج.

3. التصنيف:

يتم إدخال الميزات المستخرجة من المرحلة السابقة إلى رأس التصنيف بحيث يكون الخرج النهائي tensor أبعاده (N, 8) يحوي على class score لكلّ عينة من الدخل من أجل كل من الصفوف الثمانية لمجموعة المعطيات.

I3DHead في النموذج هو رأس تصنيف مصمم لبيانات الإدخال ثلاثية الأبعاد. يستخدم طبقة تسرب للتنظيم، وطبقة متصلة بالكامل للتصنيف، ويستخدم وظيفة CrossEntropyLoss لتحسين النموذج أثناء التدريب.

تتألف شبكة التصنيف المستخدمة من الطبقات التالية المبينة في الشكل 24:

avg_pool: يؤدي التجميع وفق القيمة المتوسطة إلى تقليل كل خريطة ميزات (7، 7، 8) إلى قيمة واحدة (متوسط جميع القيم في خريطة المعالم)، مما يقلل بشكل فعال الأبعاد إلى (1، 1، 1).

Dropout: عبارة عن تقنية تنظيم تقوم بشكل عشوائي بتعيين جزءٍ من وحدات الإدخال إلى 0 عند كل تحديث أثناء وقت التدريب، مما يساعد على منع الإفراط في الملاءمة من خلال ضمان أن النموذج لا يعتمد بشكل كبير على أي خلية عصبية واحدة ويمكنه تعميمه بشكل أفضل.

إعادة التشكيل: يتم بعد ذلك إعادة تشكيل tensor من [N, in_channels, 1, 1, 1] إلى [N, in_channels]. يؤدي هذا إلى تسطيح tensor بحيث يتم تمثيل كل مثال في الدفعة بمتجه واحد بطول in_channels والتي تساوي 1024، وذلك من أجل التحضير للطبقة المتصلة بالكامل. الآن، كل مقطع فيديو في الدفعة يعبر عنه بمتجه ببعد 1024 يمثل الميزات المستخرجة.

الطبقة المتصلة بالكامل fc_cls: يتم تمرير tensor المسطح عبر طبقة متصلة بالكامل (fc_cls)، وهو تحويل خطي. تقوم الطبقة المتصلة بالكامل بتحويل tensor من [N, in_channels] إلى [N, num_classes]، حيث num_classes هو عدد الصفوف في مهمة التصنيف (8 في عملنا هذا). وبذلك تقوم الطبقة الأخيرة بإخراج احتمالات صفوف الإجراءات الثمانية، والتي تمثّل تنبؤ الشبكة بالإجراء الذي يحدث في الفيديو بناءً على الميزات المكانية والزمانية التي تم تعلّمها. **هذه الخطوة هي المكان الذي يحدث فيه التصنيف الفعلي.**

2.3.4. تفاصيل التنفيذ:

- تم استخدام مُحسِّن AdamW هو أحد أشكال مُحسِّن Adam الذي يطبق تناقص الوزن بطريقة أكثر اتساقًا مع الهدف الأصلي المتمثل في L2 regularization.
- معدل التعلم lr=0.0003 وهو معدل التعلم الأولي للمُحسّن. يتحكم في حجم الخطوة في تحديث الأوزان.
- تناقص الوزن Weight Decay: 0.05، يضيف هذا عقوبة L2، مما لا يشجع الأوزان الكبيرة في النموذج لمنع المُلاءمة الزَّائدة.
- تم استخدام MMAction والتي هي عبارة عن مجموعة أدوات مفتوحة المصدر تعتمد على PyTorch، وتدعم العديد من نماذج فهم الفيديو، بما في ذلك المهام التالية: action recognition، skeleton based action recognition، spatial–temporal action detection و action localization.
- إعدادات نموذج العمود الفقري الأساسية:
  - ✓ حجم الرقع الثلاثية الأبعاد: (2, 4, 4)
  - ✓ بعد التضمين في المرحلة الأولى للمحول بنسخته الأساسية: 128
  - ✓ عدد طبقات المحوّل في كل مرحلة من مراحل النموذج من اليسار إلى اليمين بنسخته الأساسية:

    [2, 2, 18, 2]

  - ✓ عدد رؤوس كتلة الانتباه في كل مرحلة من مراحل النموذج من اليسار إلى اليمين بنسخته الأساسية: [4, 8, 16, 32]
  - ✓ حجم النافذة الثلاثية الأبعاد : (8, 7, 7)
  - ✓ mlp_ratio=4.0
- إعدادات نموذج رأس التصنيف الأساسية:
  - ✓ عدد قنوات الإدخال لهذه الطبقة : 1024
  - ✓ عدد الصفوف: 8
  - ✓ dropout_ratio: 0.5

4.4. أسباب اختيار النموذج وتوقع نجاحه:

يمكن أن يؤدي استخدام محول فيديو قوي مثل Video Swin Transformer [7] لاستخراج الميزات الزمانية المكانية من بيانات الفيديو إلى التقاط علاقات هرمية في كل من المكان والزمان، وهو أمر مهم لأنشطة المجموعة المعقدة. حيث يمكنه تعلم التقاط كل من التفاعلات داخل المجموعة وبين المجموعات باستخدام مستويات مختلفة من الانتباه الذاتي وذلك لأن النموذج يتكون من أربع مراحل، كل منها بحجم نافذة ودقة patch مختلفة. فالمرحلة الأولى لديها أصغر حجم نافذة وأعلى

دقة للرقعة، بينما المرحلة الأخيرة لديها أكبر حجم نافذة وأدنى دقة للرقعة، يسمح الحجم الأصغر للنافذة والدقة الأعلى للرقع في المرحلة الأولى للنموذج بالتركيز على التفاصيل المحلية داخل كل مجموعة، مثل الإجراءات والوضعيات الفردية، كما يسمح الحجم الأكبر للنافذة والدقة الأدنى للرقع في المرحلة الأخيرة بالتقاط سياق عام عبر مجموعات مختلفة، مثل أنشطة المجموعات والعلاقات. توفر المراحل المتوسطة توازناً بين المعلومات المحلية والشامِلة.

5.4. خاتمة

تحدثنا في هذا الفصل عن النموذج الأساسي المستخدم في بحثنا بالتفصيل Video Swin Transformer [7] وقمنا ببيان سرعته وتفوقه في الأداء على كثير من نماذج التعرف في الفيديو. لذلك اقترحنا نموذجاً يستفيد من قدرة هذا المحوّل على نمذجة المعلومات المكانية الزمانية.

وبينّا كيفية استخدامه في عملنا على مجموعة المعطيات المستخدمة وكيف تمت ملاءمة الحل للحصول على الخرج المطلوب.

سيتم في الفصل القادم شرح مجموعة البيانات والتجارب التي قمنا بها لاختبار قدرة النموذج المقترح في حل مشكلة التعرف على النشاط في المجموعة.

الفصل الخامس: الاختبارات والنتائج

1.5. مقدمة

في هذا الفصل سنتحدث عن مجموعة المعطيات المستخدمة في هذا البحث، حيث في جميع الأعمال التي قدمت في موضوع التّعرف على نشاط المجموعة تم التطرق إلى مجموعتي بيانات للأنشطة الجماعية المتاحة للجمهور في هذا السياق، هما مجموعة بيانات الكرة الطائرة Volleyball [2] ومجموعة البيانات الجماعية Collective Activity Dataset [29].

نستخدم في نهجنا المقترح مجموعة بيانات الكرة الطائرة Volleyball [2] لتدريب النموذج والاختبار والمقارنة مع الأعمال المشابهة. وسيتم التطرق إلى معايير التقييم المستخدمة في هذا النوع من المسائل، ومن ثم سنستعرض نتائج التدريب والاختبار التي تم الحصول عليها.

2.5. مجموعة المعطيات المستخدمة volleyball:

بالمقارنة مع المعايير المتاحة لفهم الإجراءات الفردية، هناك عدد قليل من الموارد المُشاركة في أنشطة المجموعة البشرية المعقدة. تنتمي جميع مجموعات البيانات للتعرّف على الأنشطة الجماعية إلى مقاطع فيديو المراقبة أو مقاطع الفيديو الرياضية التي تحفّزها المتطلبات العملية لبناء أنظمة السلامة أو أنظمة التحليل الرياضي.

في هذا العمل، تم استخدام مجموعة معطيات الكرة الطائرة قدمها إبراهيم وآخرون في [2] ، حيث تم جمع مجموعة المعطيات باستخدام مقاطع فيديو YouTube للكرة الطائرة المتاحة للجمهور. تم تنويط 4830 إطاراً تم انتقاؤها يدوياً من 55 مقطع فيديو مع وسوم 9 لاعبين و 8 أنشطة جماعية.

يبلغ طول كل مقطع 41 إطاراً. تتضمن الوسوم المتوفرة **وسم النشاط الجماعي** و**المربعات المحيطة للاعبين الفرديين** و**الإجراءات الفردية الخاصة بهم**، والتي يتم توفيرها فقط للإطار الأوسط للمقطع.

في الورقة [51] تم توسيع مجموعة البيانات بإضافة **المربعات المحيطة للاعبين الفرديين لبقية الإطارات** في المقاطع.

تحتوي قائمة وسوم نشاط المجموعة على أربعة أنشطة رئيسية خاصة بلعبة الكرة الطائرة (Set، Spike، Pass، Winpoint) والتي تنقسم إلى مجموعتين فرعيتين، اليسار واليمين، أي تحتوي مجموعة البيانات على ثمان وسوم لأنشطة المجموعة في المجموع، ويشار لها كما يلي:

l-set, l-spike, l-pass, l-winpoint, r-set, r-spike, r-pass, r-winpoint

ويمكن لكل لاعب تنفيذ أحد الإجراءات الفردية التسعة:

الحجب، والحفر، والسقوط، والقفز، والتحرك، والإعداد، والارتفاع، والوقوف، والانتظار، ويشار لها بالأسماء التالية:

blocking, digging, falling, jumping, moving, setting, spiking, standing and waiting.

1.2.5. شرح توضيحي تفصيلي عام:

- تحتوي مجموعة المعطيات على 55 مقطع فيديو. لكل فيديو مجلد خاص به، بمعرفات فريدة (0 ، 1 ... 54)
- حجم مجموعة المعطيات 58 Giga byte.
- مقاطع فيديو التدريب training: 1 3 6 7 10 13 15 16 18 22 23 31 32 36 38 39 40 41 42 48 50 52 53 54
- مقاطع فيديو التحقق validation: 0 2 8 12 17 19 24 26 27 28 30 33 46 49 51
- مقاطع فيديو الاختبار testing: 4 5 9 11 14 20 21 25 29 34 35 37 43 44 45 47
- داخل كل مجلد للفيديو، هناك مجموعة من المجلدات التي تتوافق مع الإطارات المشروحة (مثلاً 39/29885) تعني فيديو رقم 39، مجلد الإطار 29885 (الإطار المنوّط). يحتوي كل مجلد إطار frame directory على 41 صورة (20 إطار قبل الإطار المستهدف، الإطار المستهدف، 20 إطاراً بعد الإطار الهدف).
- يحتوي كل مجلد للفيديو على ملف annotations.txt يحتوي على وسوم ومعلومات توضيحية للإطارات المحددة حيث كل سطر في الملف هو من الشكل:

{Frame ID} {Frame Activity Class} {Player Annotation} {Player Annotation}...

يتوافق Player Annotation مع مربع محيط ضيق يحيط بكل لاعب حيث أن كل Player Annotationهي بالشكل التالي:

{Action Class} X Y W H

- مقاطع الفيديو بدقة 1920 × 1080 هي: 2 37 38 39 40 41 44 45 (8 في المجموع). كل المقاطع الأخرى هي بدقة 1280 × 720.

2.2.5. تفاصيل التطبيق:

في هذا العمل:

- استخدمنا 32 إطاراً بدلاً من 41 كاملة وذلك لتفادي بعض مشاكل الحجوم التي تمت مواجهتها حيث 32 إطار كافي للمعالجة حيث في الورقة البحثية [2] التي قدمت مجموعة البيانات Volleyball تم استخدام 10 إطارات.
- استخدمنا الوسوم التوضيحية الخاصة بنشاط المجموعة **فقط**، أي تم إهمال معلومات مربعات الإحاطة ووسوم النشاطات الفردية للاعبين، حيث الهدف من النموذج هو اختبار الحل المقترح في فهم الفيديو ككل من حيث النشاط الجماعي، كالأسلوب البشري (دون تفصيلات فردية وتجميع معلومات فردية بشكل صريح، لاستنتاج النشاط الجماعي).
- تم تحويل شكل ملفات التنويط annotations إلى شكل يتوافق مع الشكل المتوقع من النموذج للقيام بعمليات التدريب والاختبار، كما تم تجميع ملفات التنويط التي هي بالأساس على مستوى الفيديو إلى أن تصبح على مستوى تقسيم المعطيات (تدريب، تحقق، اختبار)، أي من 55 ملف إلى 3 ملفات.

3.2.5. أمثلة من مجموعة المعطيات:

يعرض الجدول 2 عدد الأمثلة الموجودة في كل صف في مجموعة المعطيات:

Group Activity Class	No. of Instances	Action Classes	No. of Instances
Right set	644	Waiting	3601
Right spike	623	Setting	1332
Right pass	801	Digging	2333
Right winpoint	295	Falling	1241
Left winpoint	367	Spiking	1216
Left pass	826	Blocking	2458
Left spike	642	Jumping	341
Left set	633	Moving	5121
		Standing	38696

الجدول 2 عدد الأمثلة الموجودة من أجل كل صف في مجموعة المعطيات volleyball

يعرض الشكل 25 والشكل 26 صوراً من مجموعة البيانات لكل من صفوف set, spike, pass, winpoint في الجهة اليمنى وفي الجهة اليسرى:

r-set

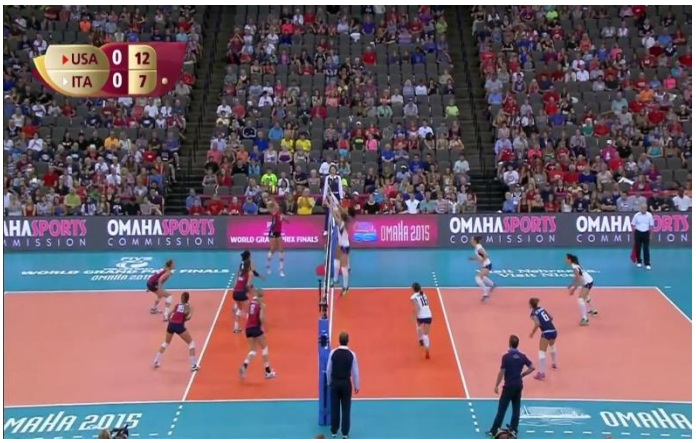
r-spike

r-pass

r-winpoint

الشكل 25 صور من الصفوف اليمنى من مجموعة البيانات المستخدمة

l−set



l−spike

l−pass

l−winpoint

الشكل 26 صور من الصفوف اليسرى من مجموعة البيانات المستخدمة

3.5. معايير التقييم:

عند تقييم مهام التّعرف على نشاط المجموعة، والتي تصنّف تحت مسائل التصنيف، هناك العديد من المعايير والاعتبارات الرئيسية التي تستخدم عادة لتقييم فعالية ومتانة مثل هذه الأنظمة. وفيما يلي توضيح لما تم اتباعه في التقييم وذلك وفقاً لما ورد أيضا في الأبحاث المشابهة:

1.3.5. الدقة Accuracy:

وهو المقياس الأكثر وضوحاً، حيث يقيس عدد المرات التي يحدد فيها النظام نشاط المجموعة بشكل صحيح. ويتم تقديمها كنسبة مئوية للأنشطة المحددة بشكل صحيح من بين جميع الأنشطة.

2.3.5. مصفوفة الإرباك Confusion Matrix:

مصفوفة الإرباك متعددة الفئات لمسألة التصنيف ب N فئة، هي مصفوفة M ببعد N×N حيث يمثل كل عنصر $M_{i,j}$ عدد الحالات التي هي في الفئة i ولكن تم التوقع أن تكون في الفئة j.

العناصر القطرية ($M_{i,i}$): تمثل عدد الحالات التي تكون فيها الفئة المتوقعة مساوية للفئة الحقيقية، أي الإيجابيات الحقيقية لكل فئة.

العناصر خارج القطر ($M_{i,j}$ حيث $i \neq j$): تمثل هذه الحالات المصنفة بشكل خاطئ، مما يشير إلى عدد مثيلات الفئة i التي تم التنبؤ بها كفئة j.

الغرض من مصفوفة الإرباك:

تقييم أداء النموذج: توفر حساباً تفصيلياً لأداء النموذج، مما يسمح بتقييم ليس فقط الدقة الإجمالية ولكن أيضاً مدى جودة أداء النموذج لكل فئة على حدى.

تحديد التصنيفات الخاطئة: من خلال تحليل الصفوف التي غالباً ما يتم الخلط بينها وبين بعضها البعض، يساعد ذلك في فهم نقاط الضعف في النموذج ويمكن أن يوجه التحسينات في تدريب النماذج وهندسة الميزات.

معالجة عدم توازن الفئة: إذا كانت بعض الصفوف تحتوي على أمثلة أكثر بكثير من غيرها، فسيكون ذلك واضحاً في مصفوفة الإرباك ويمكن أن يوضح لنا كيف قد تحتاج إلى تعديل النموذج أو بيانات التدريب الخاصة به.

4.5. النتائج:

في هذه الفقرة سنعرض نتائج التجارب التي قمنا بها، باستخدام النموذج المقترح الذي تم تقديمه في الفصل السابق على مجموعة المعطيات التي تم شرحها Volleyball، يتكون النموذج من كتلتين رئيسيتين هما :

Backbone: Video Swin Transformer

Head : I3d

كما ذكرنا فإن النموذج الأساسي في بحثنا هو Video Swin [7]، وجميع اختباراتنا في الوقت الحالي هي من أجل نسخة النموذج Video Swin Base، الذي تمت تهيئته من نموذج الصور المدرب مسبقاً Swin-Transformer Base [1] على ImageNet-21k.

1.4.5. النتائج التي تم الحصول عليها بتدريب رأس النموذج فقط:

في التجارب الأولى على النموذج المقترح تم تدريب رأس النموذج فقط واستخدام العمود الفقري المدرّب مسبقاً كمستخرج سمات فقط.

كما ذكرنا في شرح النموذج video Swin [7] في فصل النموذج المقترح، في فقرة التهيئة من النموذج المدرب مسبقاً Swin [1] (الخاص بالصور)، فإن نماذج الفيديو تم تهيئتها من نماذج الصور القوية المدربة مسبقاً على مجموعة معطيات كبيرة.

نسخة النموذج المذكورة التي تم استخدامها مهيئة مسبقاً من نسخة نموذج الصور المدرب على ImageNet-22K وفق معدل التعلّم والنتيجة المحققة على مجموعة بيانات Kinetics 600، وفق الجدول التالي:

(حيث تحوي مجموعة معطيات ImageNet-21k على 14,197,122 صورة مقسمة إلى 21,841 فئة. تقوم بعض الأوراق البحثية بتجميع هذا وتسميته ImageNet-22k)

Backbone	Pretrain	Lr Schd	spatial crop	acc@1	acc@5	#params	FLOPs
Swin-B	ImageNet-22K	30ep	224	84.0	96.5	88M	281.6G

استخدمنا أوزان النموذج المدرب مسبقاً على مجموعة المعطيات kinetics600 (مجموعة معطيات للتعرّف على الإجراء الفردي). وتم التدريب بتجميد هذه الأوزان وتعديل أوزان الرأس فقط.

تم التدريب في بيئة colab باستخدام وحدة معالجة الرسومات T4 GPU.

تم اختيار Batch size =64 و Epochs=45، من أجل كل عصر تدريب يتم التدريب على 2146 عينة.

تم التوقف عند 45 عصر تدريب حيث أنه تم التجريب من أجل 50 لكن النموذج بعد 45 عصر تدريب لم يتعلم.

1.1.4.5. نتائج التدريب على مجموعة بيانات التدريب:

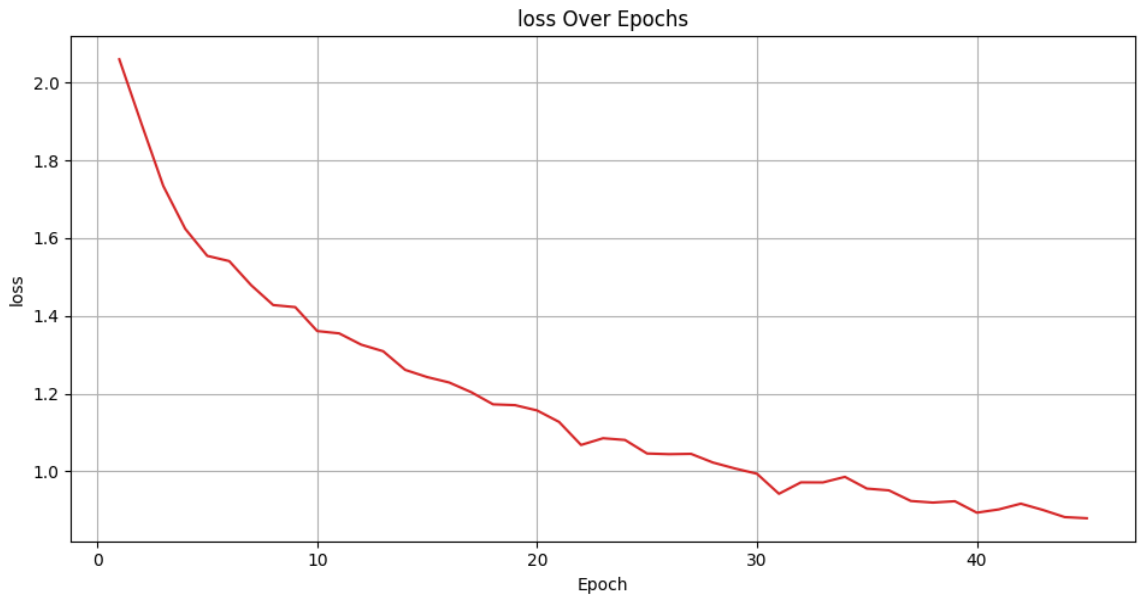
لتوليد منحنيات الخطأ وتتبع أداء النظام أثناء التدريب تم استخدام مكتبة matplotlib. حيث يبين في الأشكال عدد عصور التدريب على المحور الأفقي

يبين الشكل 27 تغير دالّة الخسارة إذ نلاحظ هبوط منحنيات الخطأ، وذلك بسبب تغيير أوزان النموذج أثناء تدريبه على مجموعة المعطيات.

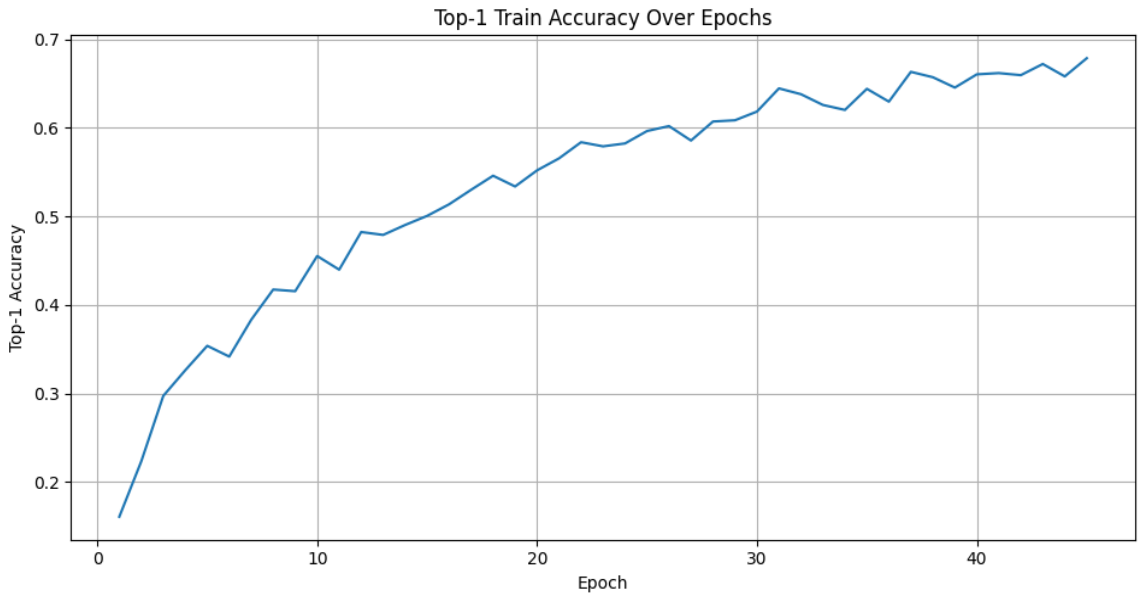
يبين الشكل 28، تغيرات الدقة التي حققها النموذج أثناء التدريب والتي وصلت إلى أعلى قيمة لها وهي 68%.

يبين الشكل 29 تغيرات معدل التعلم أثناء التدريب، والذي تم ضبطه في بداية العملية على القيمة:

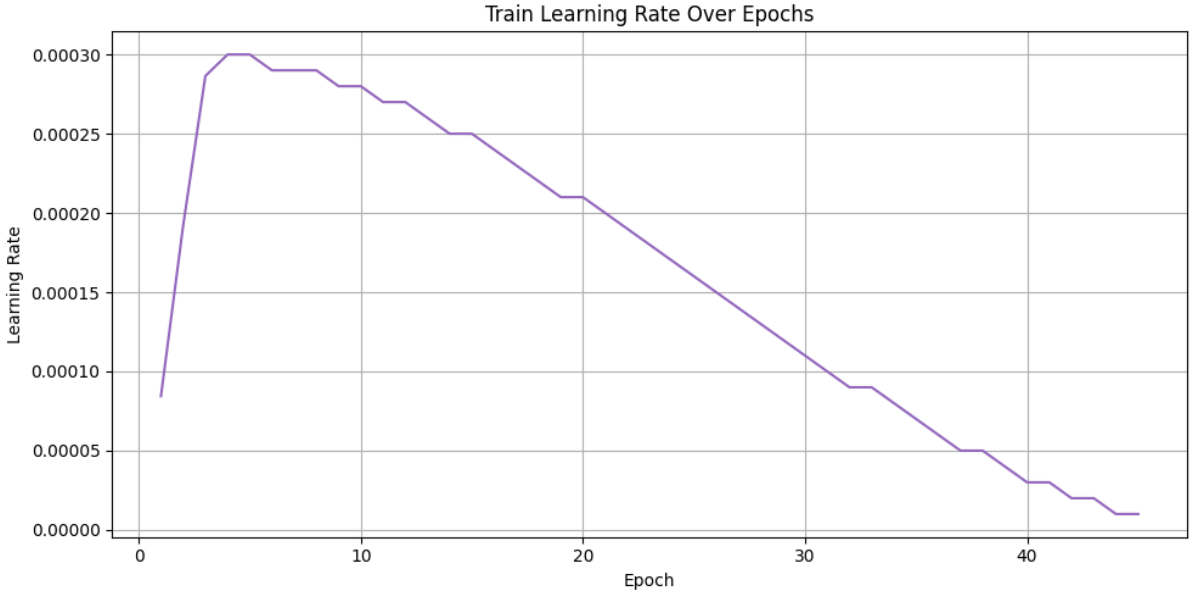
lr= 0.0003.



الشكل 27 تغيرات دالّة الخسارة أثناء تدريب رأس النموذج المقترح فقط



الشكل 28 تغيرات دقة النموذج المقترح أثناء تدريب الرأس فقط

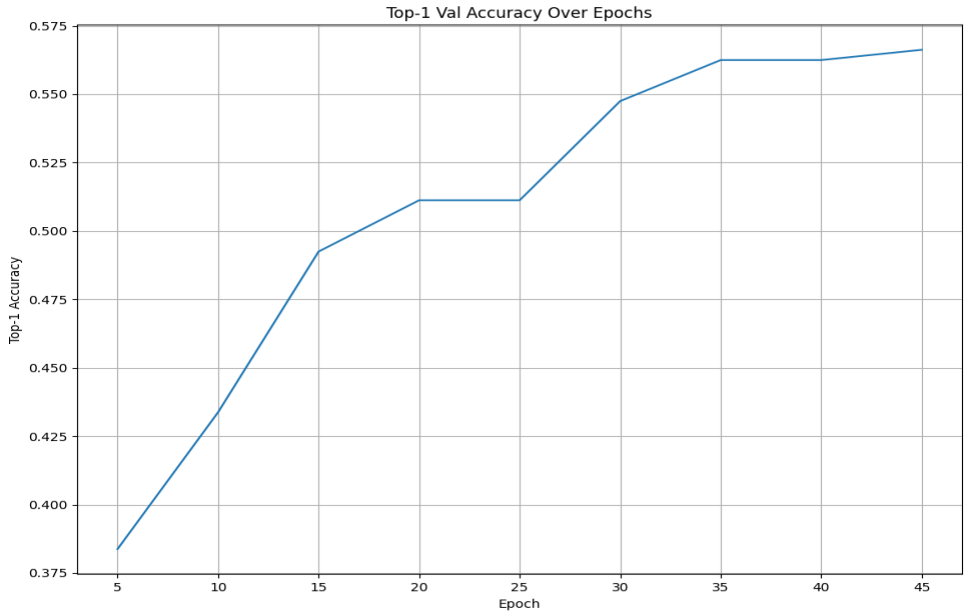


الشكل 29 تغيرات معدل التعلم للنموذج المقترح أثناء تدريب الرأس فقط

2.1.4.5. نتائج التجربة على مجموعة بيانات التحقق:

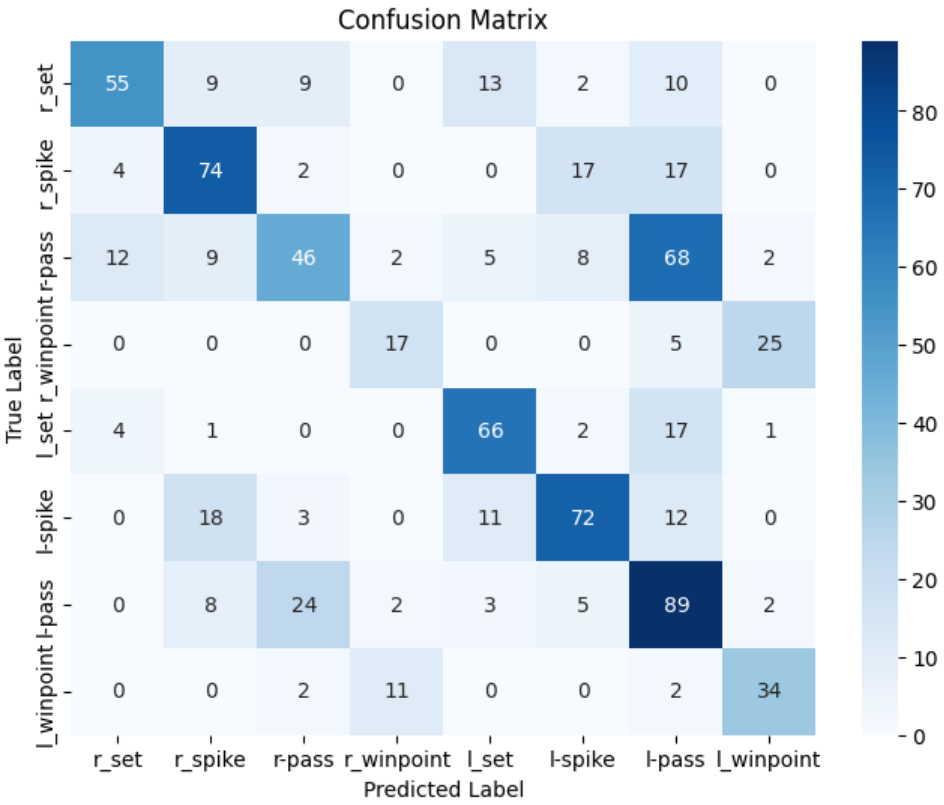
يتم إجراء اختبار للنموذج على مجموعة بيانات التحقق بعد كل 5 عصور أثناء عملية التدريب، حيث لم يتم إجراء الاختبار بعد كل عصر وذلك بسبب قيود يفرضها استخدام colab دعت بنا إلى إيجاد حل وسطي للحصول على النتيجة وفق زمن الاستخدام المتاح. وجرى ذلك من أجل 800 عينة.

يبين الشكل 30 ، تغيرات الدقة التي حققها النموذج والتي وصلت إلى 56.6%.



الشكل 30 تغيرات دقة اختبار النموذج المقترح بتدريب الرأس فقط

نبين أيضاً في الشكل 31 مصفوفة الإرباك الموافقة للعصر 45، حيث يتم بيان عدد التنبؤات الصحيحة من أجل كل صف على حد سواء وما الصفوف التي يتم الخلط بينها.



الشكل 31 مصفوفة الإرباك لاختبار النموذج بتدريب الرأس فقط

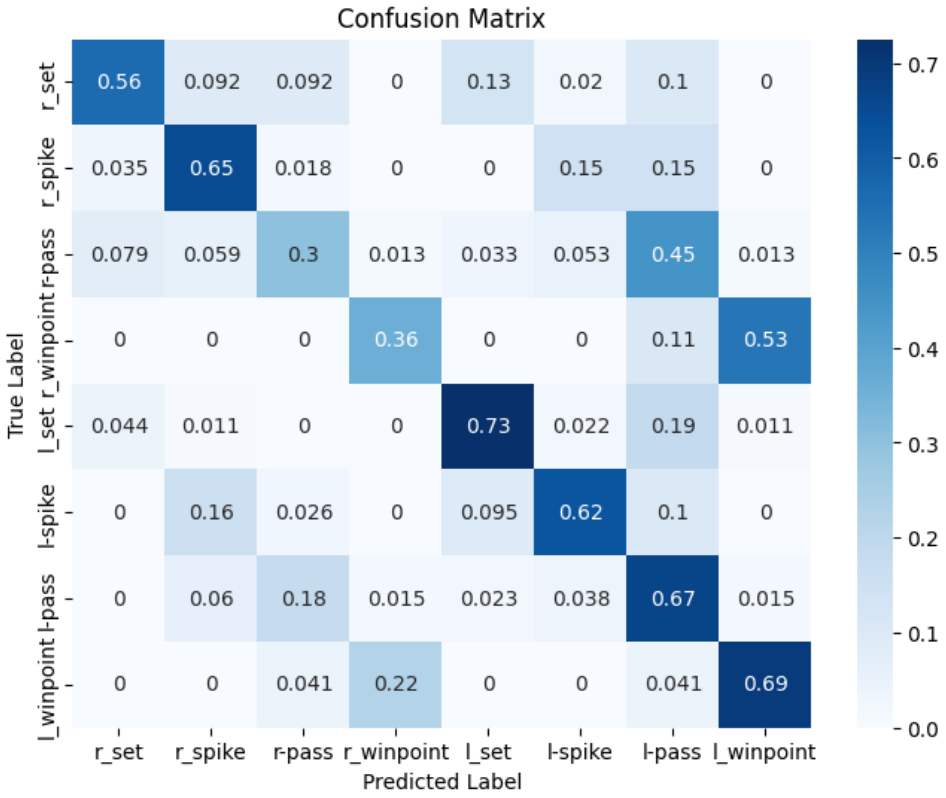
وبعد إجراء عملية استنظام على مستوى السطر للمصفوفة السابقة نحصل على الدقة المبينة في الشكل 32، على مستوى كل صف من الصفوف، حيث يحقق الصف l–set أعلى دقة وتصل إلى 73%.

نلاحظ في هذه التجربة من الشكل 32 أن صفوف الأنشطة اليسرى حققت دقة أعلى من صفوف الأنشطة اليمنى، حتى أن الصف r–pass تم تصنيفه على أنه l–pass بنسبة 45% و r–winpoint تم تصنيفه على أنه l–winpoint بنسبة 53% أي بنسب أعلى من نسبة صفّهم الفعلي.

بالنسبة لباقي الأنشطة فإن ثان أعلى نسبة من أجل كل سطر هي لنظير الصف من الجهة الأخرى، مثال على ذلك r–set صنّف بشكل صحيح بنسبة 56% وثان أعلى نسبة في سطره هي 13% تعود إلى l–set.

يعود ذلك على الأغلب إلى أن عدد أمثلة الفئات اليسرى في بيانات التدريب أعلى من نظيرتها اليمنى من أجل كل نشاط حيث يؤدي ذلك إلى معدلات تنبؤ أعلى لتلك الفئة لأن النموذج شاهد المزيد من الأمثلة عليها.

وبذلك نرى أنه من المفيد تجربة نموذج أكثر تعقيدًا لمعرفة ما إذا كان بإمكانه التقاط الفروق الدقيقة بين الفئات التي يفتقدها هذا النموذج، لذلك سيتم تجريب تدريب النموذج كاملاً بهدف استخراج ميزات أفضل.



الشكل 32 مصفوفة إرباك النموذج المقترح بعد إجراء normalization على مستوى السطر

2.4.5. النتائج التي تم الحصول عليها بتدريب النموذج كاملاً:

1.2.4.5. نتائج التدريب على مجموعة بيانات التدريب:

تم التدريب في بيئة Colab Pro+ باستخدام وحدة معالجة الرسومات A100 GPU.

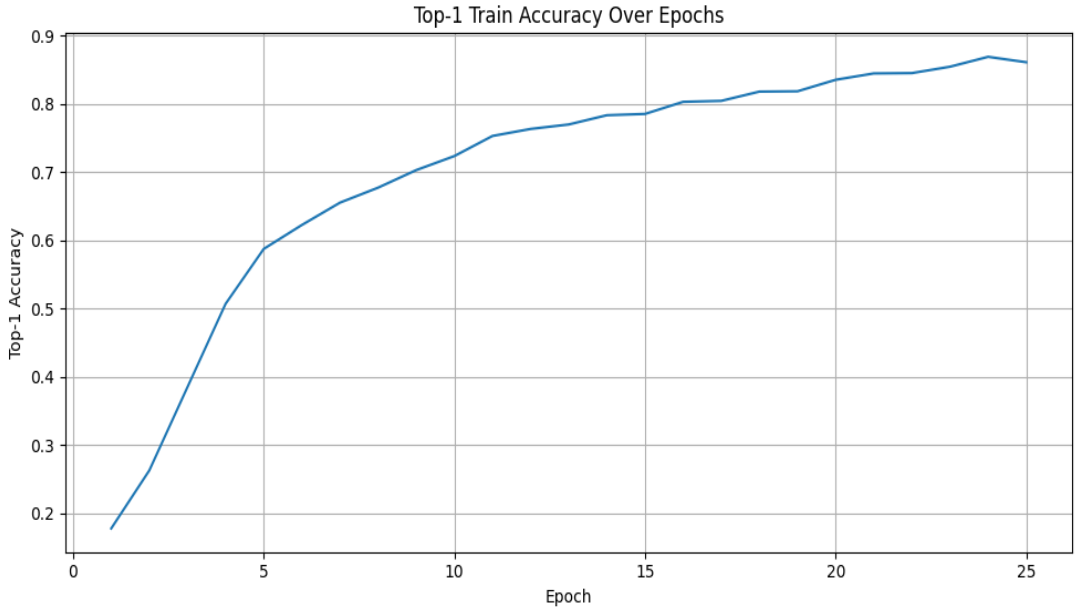
تم اختيار Batch size =64 و Epochs=25، من أجل كل عصر يتم التدريب على 2146عينة.

يبين الشكل 34 تغيرات دقة تدريب النموذج المقترح بتدريب النموذج كاملاً.

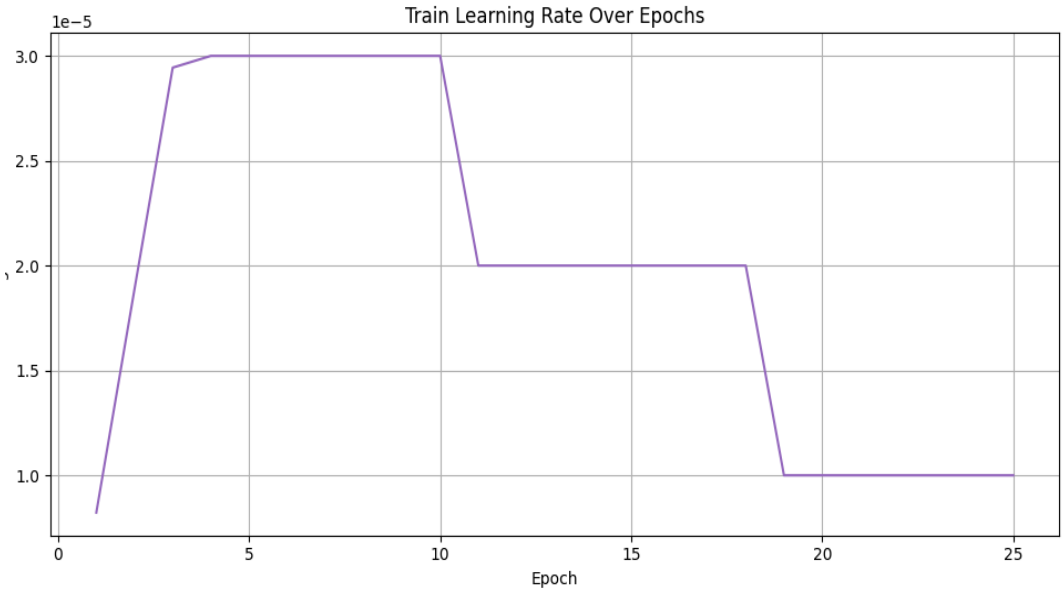
يبين *الشكل 33* تغيرات معدل التعلم للنموذج أثناء تدريب النموذج كاملاً، والذي تم ضبطه في بداية العملية على القيمة:

0.00003 =lr من أجل العمود الفقري و lr=0.0003 من أجل رأس التصنيف.

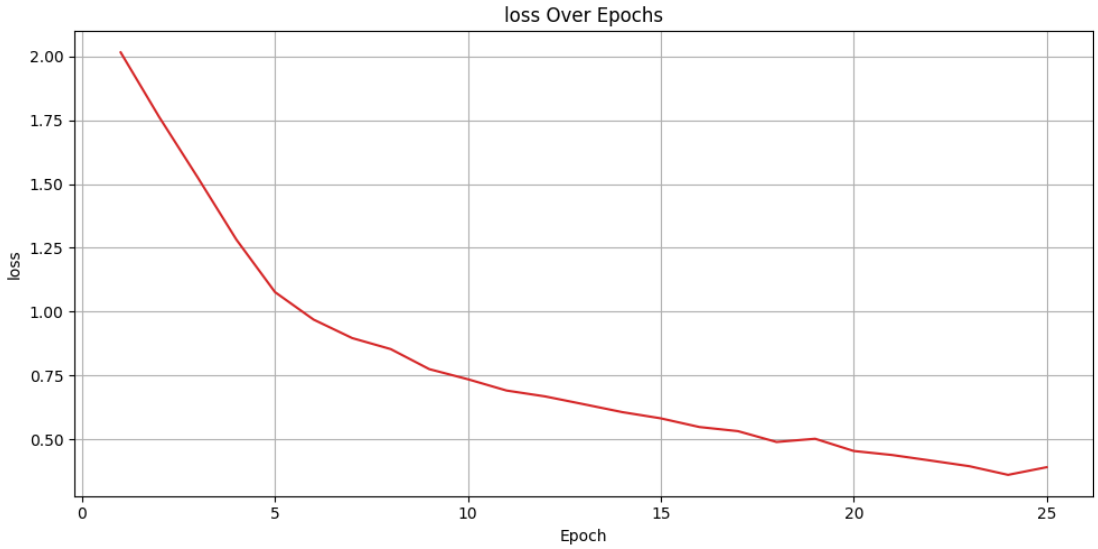
يبين الشكل 35 تغيرات دالّة الخسارة أثناء تدريب النموذج المقترح كاملاً، إذ نلاحظ هبوط منحنيات الخطأ، وذلك بسبب تغيير أوزان النموذج أثناء تدريبه على مجموعة المعطيات.



الشكل 34 تغيرات دقة تدريب النموذج المقترح بتدريب النموذج كامل



الشكل 33 تغييرات معدل التعلم للنموذج أثناء تدريب النموذج كاملاً

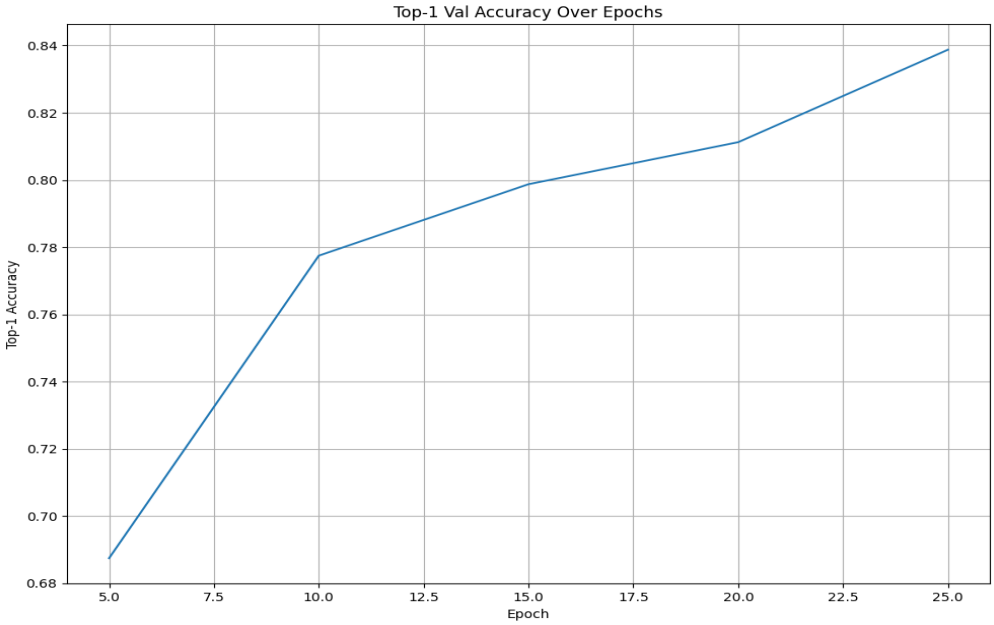


الشكل 35 تغيرات دالّة الخسارة أثناء تدريب النموذج المقترح كاملاً

2.2.4.5. نتائج التجربة على مجموعة بيانات التحقق:

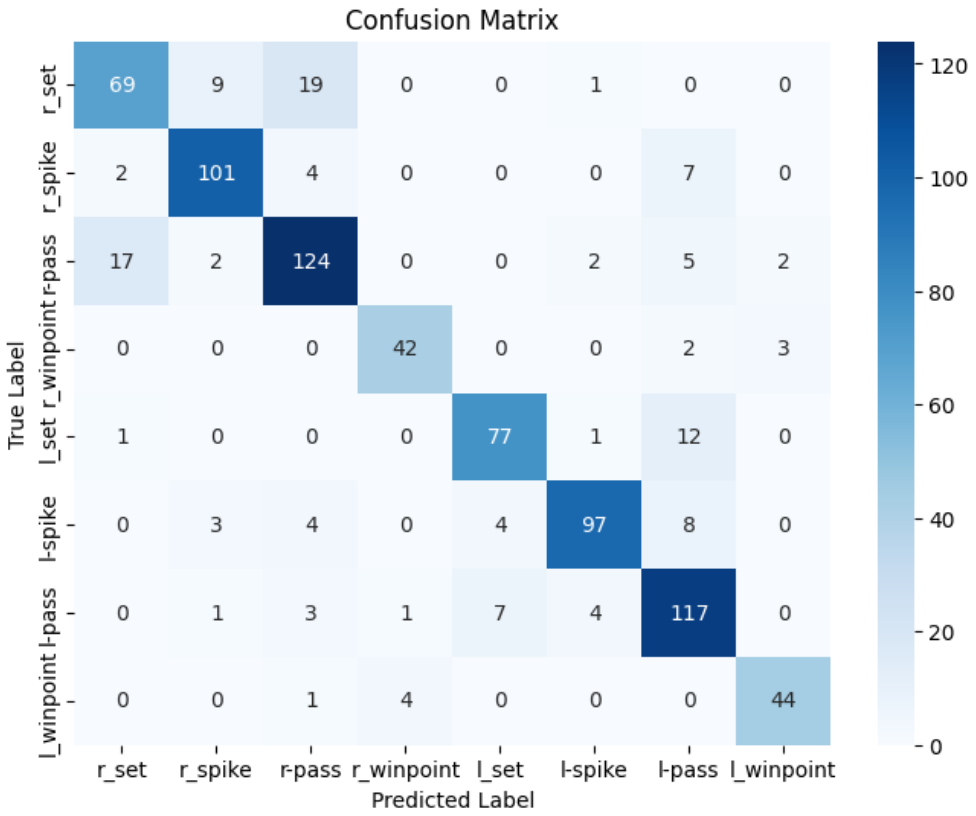
يتم إجراء اختبار للنموذج على مجموعة التحقق بعد كل 5 عصور أثناء عملية التدريب.

يبين الشكل 36 تغيرات الدقة التي حققها النموذج والتي وصلت إلى 83.8%.



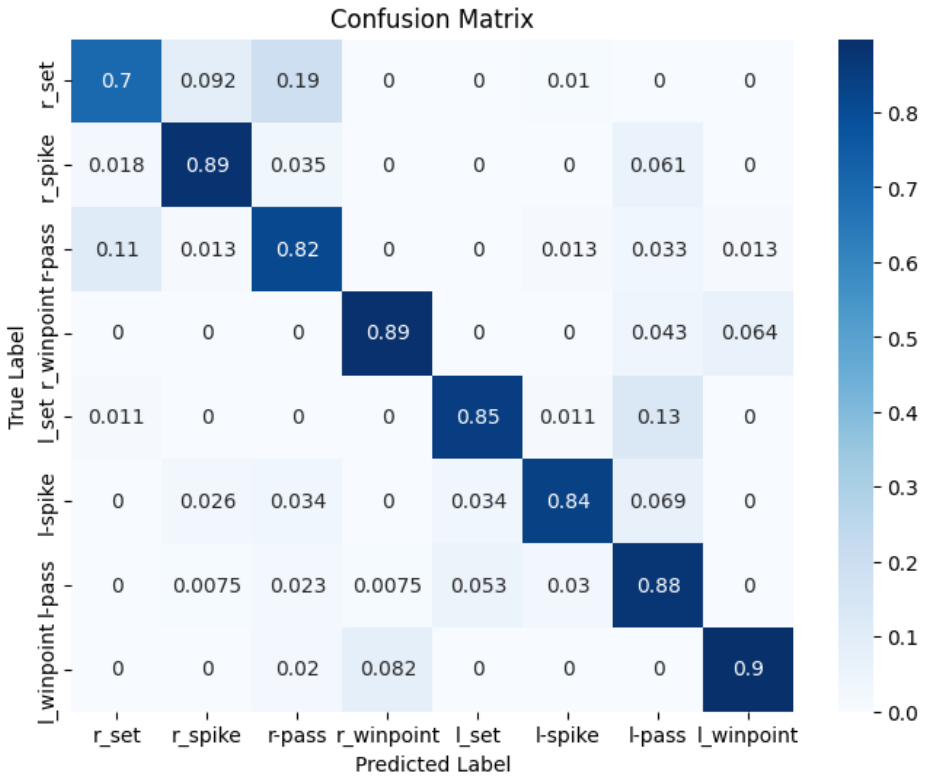
الشكل 36 تغيرات دقة الاختبار للنموذج المقترح بتدريبه كاملاً

نعرض أيضاً مصفوفة الإرباك الموافقة في الشكل 37، حيث يظهر عدد التنبؤات الصحيحة من أجل كل صف على حد سواء، وما الصفوف التي يتم الخلط بينها.



الشكل 37 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بتدريب النموذج كاملاً

وبعد إجراء عملية استنظام على مستوى السطر للمصفوفة السابقة نحصل على الدقة المبينة في الشكل 38 على مستوى كل صف من الصفوف حيث حقق الصف l–winpoint أعلى دقة وصلت إلى 90%.

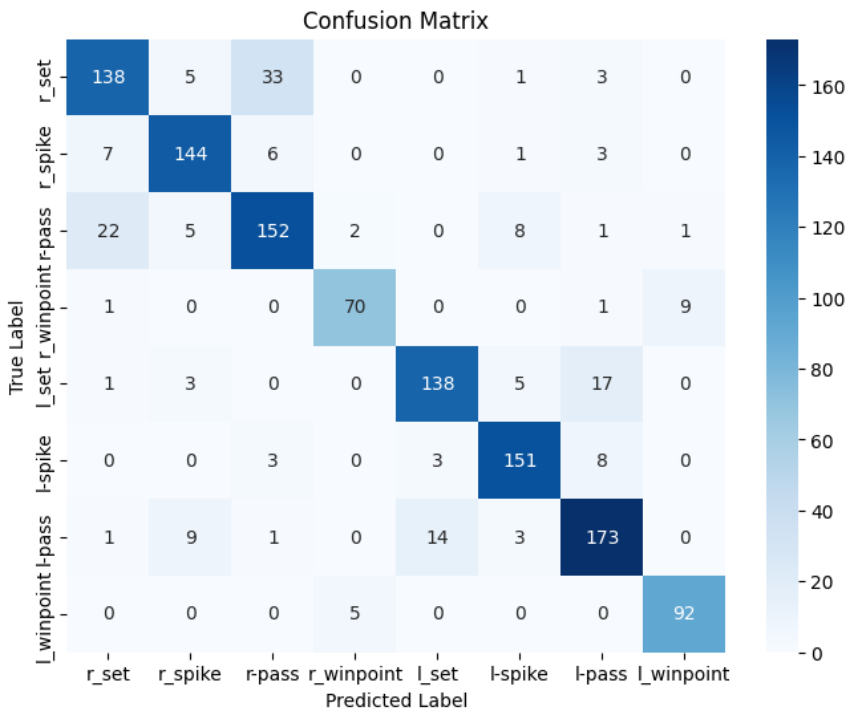


الشكل 38 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بتدريب النموذج كاملاً بعد إجراء استنظام

3.2.4.5. نتائج التجربة على مجموعة بيانات الاختبار:

تم إجراء اختبار للنموذج على مجموعة الاختبار المؤلفة من 1240 عينة، الدقة التي حققها النموذج وصلت إلى 85.3%.

نبين أيضا في الشكل 39 مصفوفة الإرباك لبيانات الاختبار، حيث يتم بيان عدد التنبؤات الصحيحة من أجل كل صف على حد سواء وما الصفوف التي يتم الخلط بينها.



الشكل 39 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بتدريب النموذج كاملاً

وبعد إجراء عملية استنظام على مستوى السطر للمصفوفة السابقة نحصل على الدقة المبينة في الشكل 40 على مستوى كل صف من الصفوف.

نلاحظ من المصفوفة في الشكل 40 تحسّن الدقة بشكل ملحوظ عند تدريب النموذج كاملاً وذلك بسبب استخراج ميزات بشكل أفضل من مستخرج الميزات.

الصف l–winpoint يحقق أعلى دقة وتصل إلى 95%.

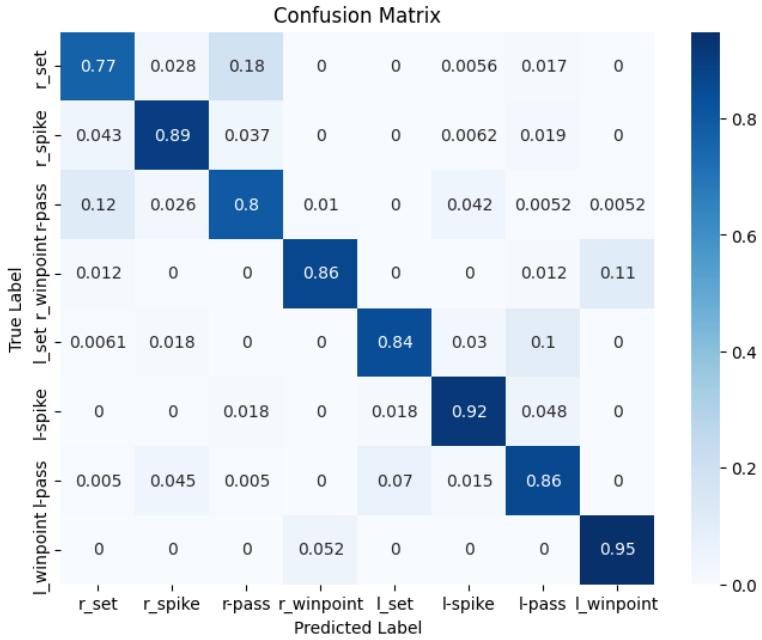
بالنسبة للصفين l–winpoint و r–winpoint فإن ثان أعلى نسبة في سطر كل منهما هي للجهة الأخرى من النشاط و ذلك يعود إلى أن بعض مشاهد هذين الصفين تكون الكاميرا قريبة وبالتالي يحصل الخطأ في تمييز الجهة وليس في الخلط مع غير أنشطة نظراً لتميُّز هذا النشاط بطبيعته. أما باقي الأنشطة فلم تعد ثان أعلى دقة في كل سطر هي للنظير من الجهة الأخرى أي تم تمييز الجهات بشكل أفضل في هذه التجربة.

النشاطين winpoint و spike يحققان دقة جيدة والخطأ في أسطر هذه الصفوف موجود بنسب ضئيلة جداً وذلك لتميُّز حركات هذه الأنشطة.

بالنسبة للنشاطين set و pass فإن الأخطاء بينهما هي بالشكل التالي: r–set يخطئ مع r–pass، r–pass يخطئ مع r–set، l–set يخطئ مع l–pass و l–pass يخطئ مع l–set حيث ذلك بالنسبة لأعلى دقة تصنيف خاطئ في سطر

كل صف من الصفوف المذكورة، ويعود ذلك لصعوبة تمييز النشاطين حيث في جميع الدراسات المشابهة ذكر أن هذين الصفين متشابهان في طبيعتهما، كما أن SAM [41] قامت بدمجهما كما قمنا بذلك في التجربة التالية.

ما زالت صفوف الأنشطة اليسرى تحقق دقة أعلى من نظيرتها اليمنى وذلك يعود إلى أن عدد أمثلتها أكثر في بيانات التدريب.



الشكل 40 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بتدريب النموذج كاملاً بعد إجراء استنظام

يعرض الجدول 3 مقارنة النتيجة المحققة مع نتائج [40] و [42] على مجموعة معطيات Volleyball:

عدد مقاطع التدريب	النموذج	الدقة %
3493	Actor Transformer [40]	94.4
3493 + augmentation	POGARS [42]	93.2
2146	OURS	85.3

الجدول 3 مقارنة النتيجة المحققة مع النموذج actor transformer وPOGARS

في [40] استخدموا عمودين فقريين هما HRNet , I3D وذلك من أجل الفرعين pose, flow. في عملنا هذا نستخدم مستخرج ميزات وحيد، كما أن البحث [40] يستخدم مربعات الإحاطة المزودة من مجموعة البيانات إضافة إلى النسخة الموسّعة بمربعات إحاطة لجميع الإطارات

في [42] اعتمدوا على معلومات الوضعية وأهملوا المعلومات السياقية في المشهد، ورغم تحقيقهم لنتيجة جيدة فإن هذا الأسلوب في الحل ليس فعّالاً في جميع الحالات، حيث يكون لعناصر المشهد الأخرى دور أساسي في معرفة النشاط. فمثلاً من الأمثلة في مجموعة معطيات Collective Activity Dataset [29] فإن نشاطيّ عبور الشارع والمشي لهم وضعيات متماثلة للأشخاص ولكن الذي يفصلهما هو خطوط الطريق مثلاً التي تهملها هكذا نماذج. ومن الأمثلة في الحياة اليومية: نجد أنه في المسيرات لدى الأشخاص الذين يرفعون لافتة وضعيات مماثلة لأشخاص يرفعون يديهم بغرض آخر، فتكون اللافتات في هذا السياق مهمة لتحديد النشاط وكذلك فإن وضعيات جلوس مجموعة من الأشخاص تكون متشابهة ولكن ممكن أن تختلف الأنشطة باختلاف المكان أو بوجود عناصر أخرى في المشهد، مثلاً يشاهدون تلفاز أو يتحدثون أو يجتمعون في مكتب وتطول الأمثلة المشابهة. ولذلك، يمكن القول أن هذه الأساليب ممكن أن تكون منحازة للتعرف على النشاط الرياضي كون الوضعيات البشرية تلعب دوراً مهم في معرفة النشاط الحاصل ولكن لا يمكن تعميمها على جميع مسائل الحياة، لذلك من المحتمل أن يكون عدم قيامهم باختبارات على Collective Activity Dataset [29] بسبب عدم فعالية النموذج في هذه الحالات.

حيث أن من أهداف حلّنا عدم تخصيص المسألة بشكل أو بآخر واكتشاف قدرة هكذا نموذج على تحقيق نتائج جيدة في موضوع أنشطة المجموعة كون بنيته الهرمية تسمح ببناء العلاقات بين عناصر المشهد الزمانية المكانية التي يجب تحقيقها في مسائل التّعرف على النشاط الجماعي، ومن خلال هذه النتائج المقدمة نجد أن النموذج تعلّم بشكل جيّد وأعطى نتيجة تعتبر واعدة وبالتحسينات التي سنقترحها ممكن أن يحقق نتائج ممتازة بطرق غير معقدة وذلك بفضل بنيته الملائمة لمسألتنا.

في كلا المنهجين [40] و [42] تم الاستفادة من معلومات مربعات الإحاطة والمعلومات المكانية والإجراء الفردي لكل ممثل في المشهد إضافة إلى النشاط الجماعي والتي هي المعلومة الوحيدة المستخدمة في نهجنا. حيث من الطبيعي أن يجعل إثراء النموذج بمعلومات كثيرة محددة المسألة أسهل على النموذج لكن نسعى في هذا العمل لاكتشاف القوة التنبؤية للنموذج الهرمي بإجراءات المجموعة فقط، مما يتيح للباحثين في هذا الموضوع البحث باتجاه جديد أقل تكلفة وتوفير مجموعات معطيات أكثر لإجراء تجارب أكثر عمومية بأقل جهد من حيث وضع الوسوم اللازمة عليها.

4.2.4.5. نتائج التدريب على مجموعة بيانات التدريب بعد دمج الصفين pass و set:

ورد في الورقة البحثية [41] إبراز مشكلة في مجموعة المعطيات، حيث تم التنويه في هذه الورقة إلى وجود وسوم توضيحية غير صحيحة بين إجراءين محددين في مجموعة معطيات Volleyball هما "pass" و"set". حيث تم العثور على أن هذه الوسوم الخاطئة تؤثر بشكل كبير على دقة تقييم النموذج.

التكيف المنهجي الذي تم اقتراحه لمعالجة هذه المشكلة، هو دمج صفّي "pass " و"set " في صف واحد، وذلك لضمان تدريب وتقييم أكثر موثوقية لنموذجهم وبما أن النهج المقترح في هذه الورقة يعتمد على وسوم المجموعة فقط من ملفات التنويط، دفعنا هذا إلى اتباع نهج مماثل من أجل عدالة المقارنة مع نهجهم المقترح.

حيث تم في عملنا هذا دمج الصّفين المذكورين كما يلي، وتعديل ملفات التنويط وفقاً لهذا التعديل:

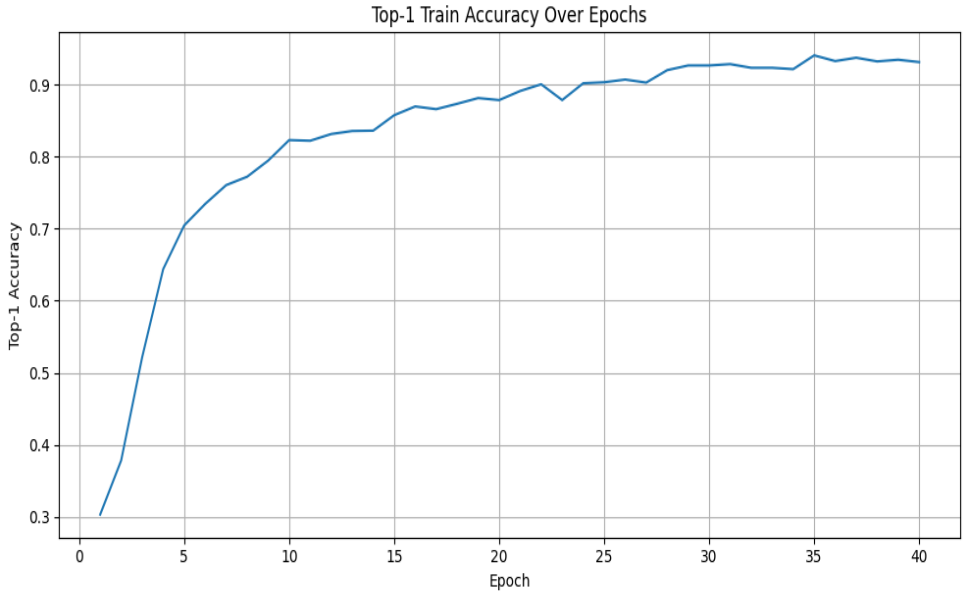
- الصف الجديد r_set_pass ناتج عن دمج الصفين r_set و r_pass.
- الصف الجديد l_set_pass ناتج عن دمج الصفين l_set و l_pass.

وبالتالي أصبح عدد صفوف النشاط الجماعي بعد الدمج هو 6 بدل 8.

تم التدريب في بيئة Colab Pro+ باستخدام وحدة معالجة الرسومات T4 GPU.

تم اختيار Batch size =64 و Epochs=40، من أجل كل عصر يتم التدريب على 2146 عينة.

يبين الشكل 41 تغيرات دالّة الخسارة أثناء تدريب النموذج بعد دمج الصفين set, pass.

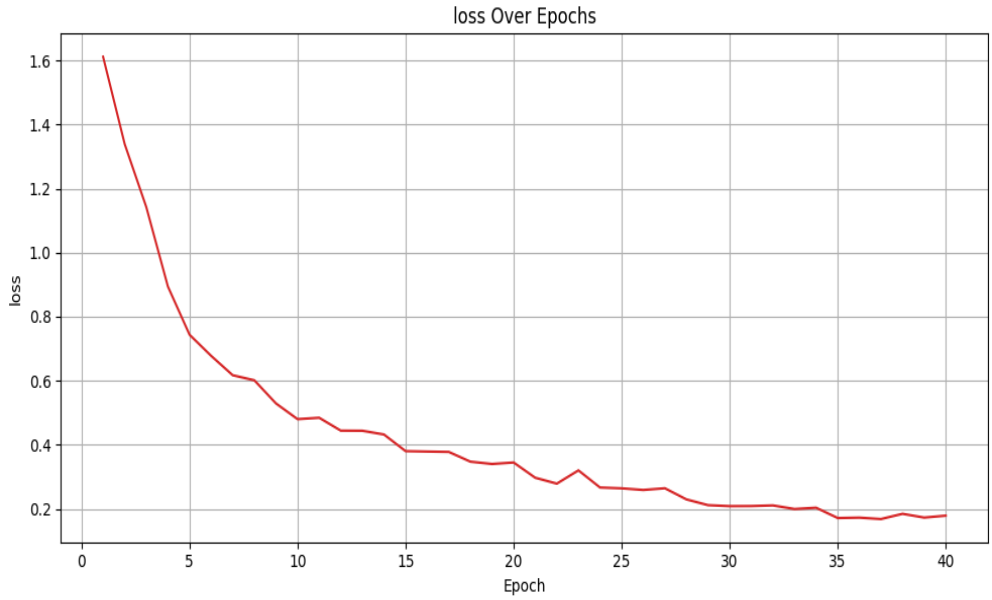


يبين

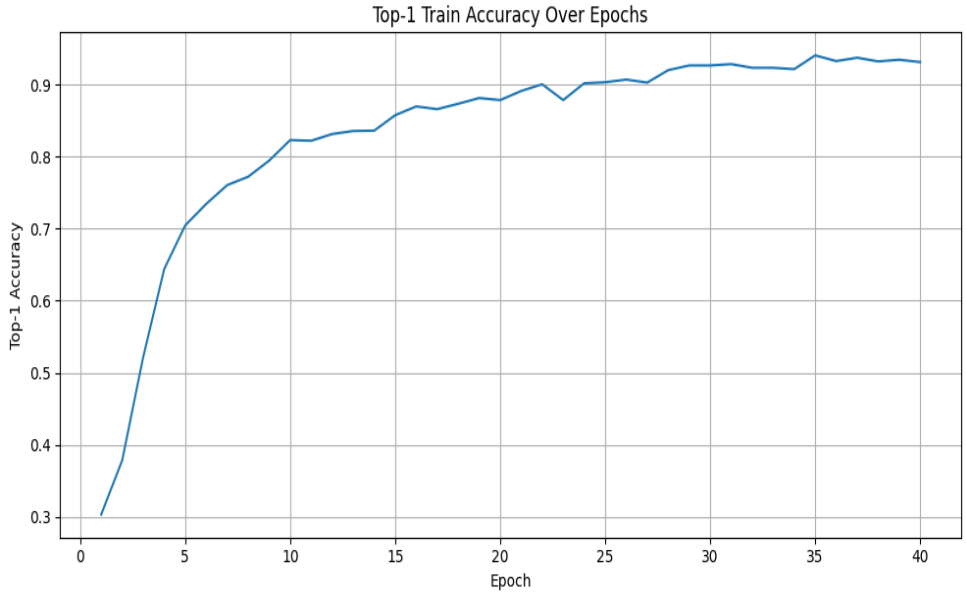
الشكل 42 تغيرات الدقة أثناء تدريب النموذج بعد دمج الصفين set,pass حيث وصلت أعلى دقة حققها النموذج 93%.

يبين الشكل 43 تغيّر معدل التعلم أثناء التدريب والذي تم ضبطه في بداية العملية على القيمة:

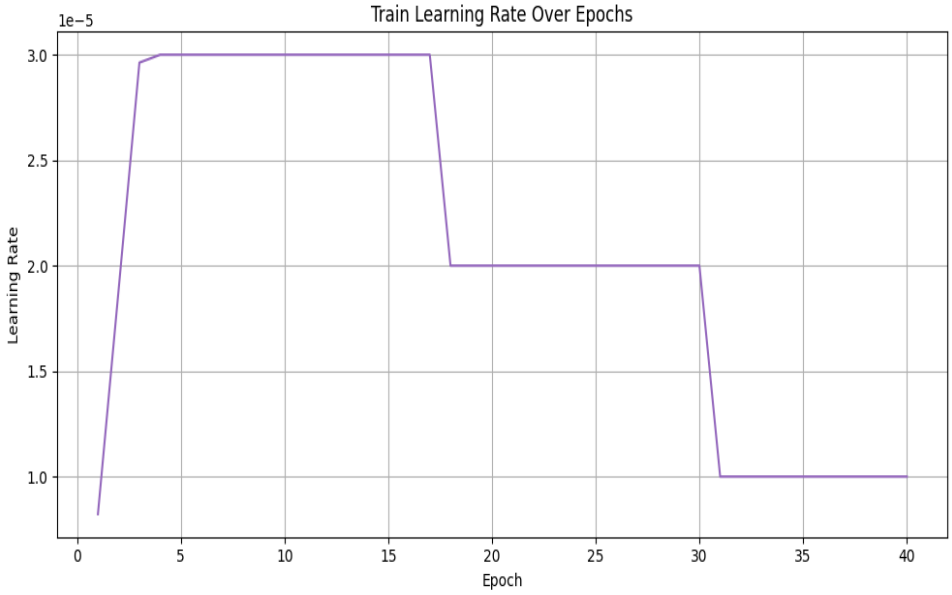
lr= 0.00003 من أجل العمود الفقري و lr=0.0003 من أجل رأس التصنيف.



الشكل 41 تغيرات دالّة الخسارة أثناء تدريب النموذج بعد دمج الصفين set, pass



الشكل 42 تغيرات الدقة أثناء تدريب النموذج بعد دمج الصفين set,pass

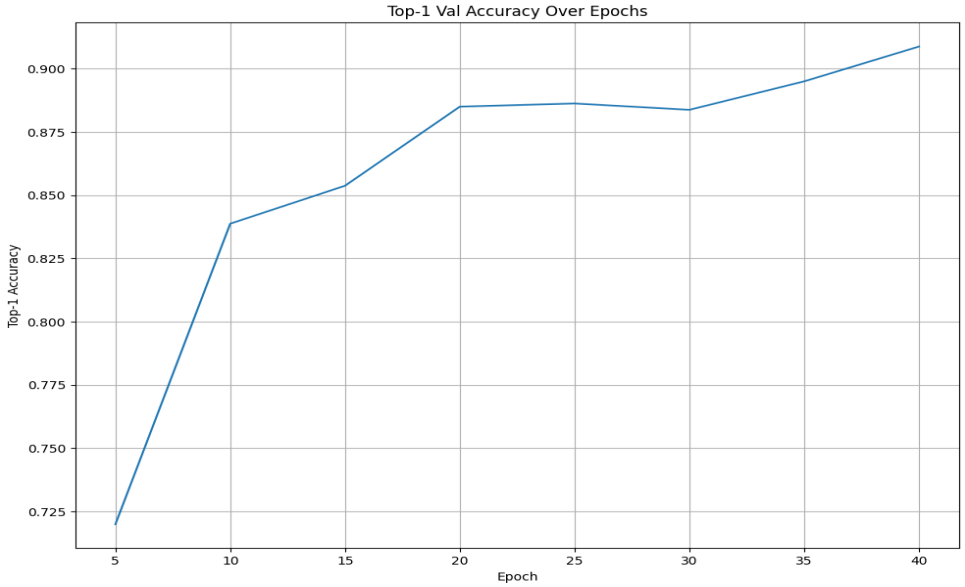


الشكل 43 تغيرات معدّل التّعلم أثناء تدريب النموذج بعد دمج الصفين set,pass

5.2.4.5. نتائج التجربة على مجموعة بيانات التحقق بعد دمج الصفين pass و set:

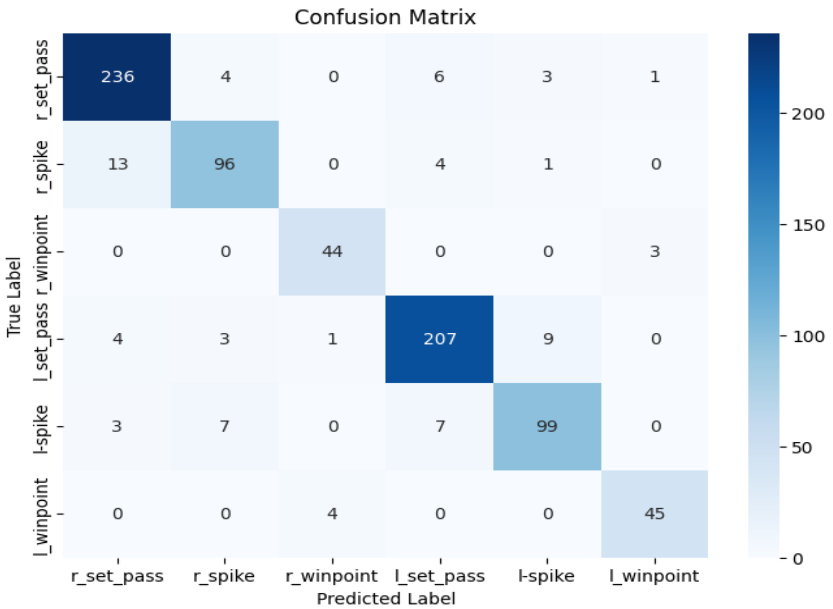
يتم إجراء اختبار للنموذج على مجموعة ال validation بعد كل 5 عصور أثناء عملية التدريب.

يبين الشكل 44، تغيرات الدقة التي حققها النموذج والتي وصلت إلى 91%.



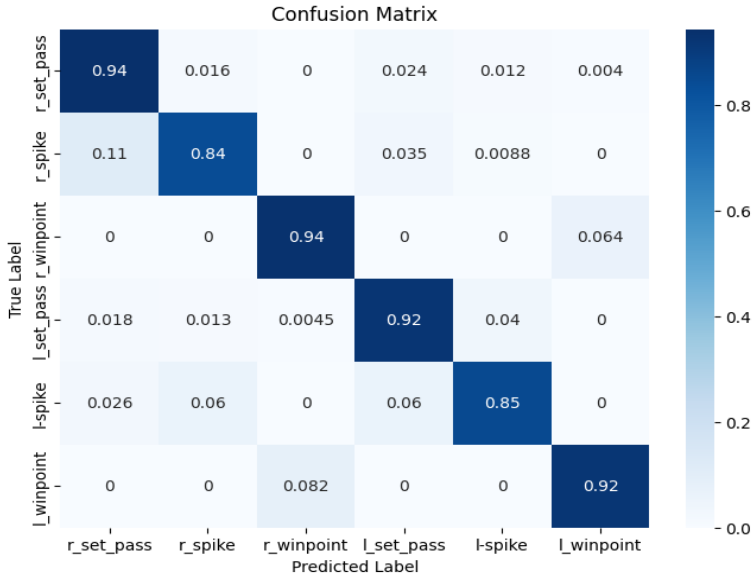
الشكل 44 تغيرات الدقة أثناء اختبار النموذج بعد دمج الصفين set,pass

نبين أيضا في الشكل 45 مصفوفة الإرباك الموافقة للعصر 40، حيث يتم بيان عدد التنبؤات الصحيحة من أجل كل صف على حد سواء وما الصفوف التي يتم الخلط بينها.



الشكل 45 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بعد دمج set و pass

وبعد إجراء عملية استنظام على مستوى السطر للمصفوفة السابقة نحصل على الدقة المبينة في الشكل 46 على مستوى كل صف من الصفوف.

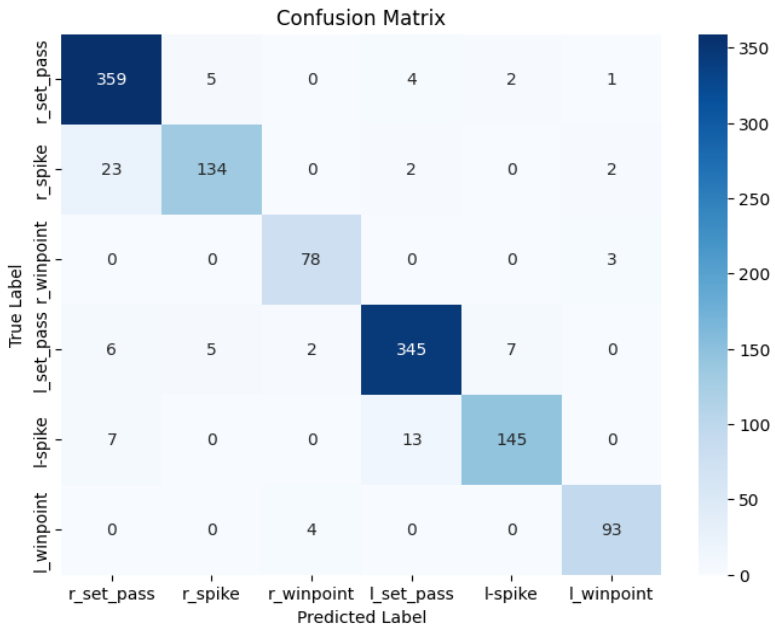


الشكل 46 مصفوفة الإرباك لاختبار النموذج على بيانات التحقق بعد دمج set وpass وإجراء استنظام

6.2.4.5. نتائج التجربة على مجموعة بيانات الاختبار:

تم إجراء اختبار للنموذج على مجموعة الاختبار المؤلفة من 1240 عينة، الدقة التي حققها النموذج وصلت إلى 93%.

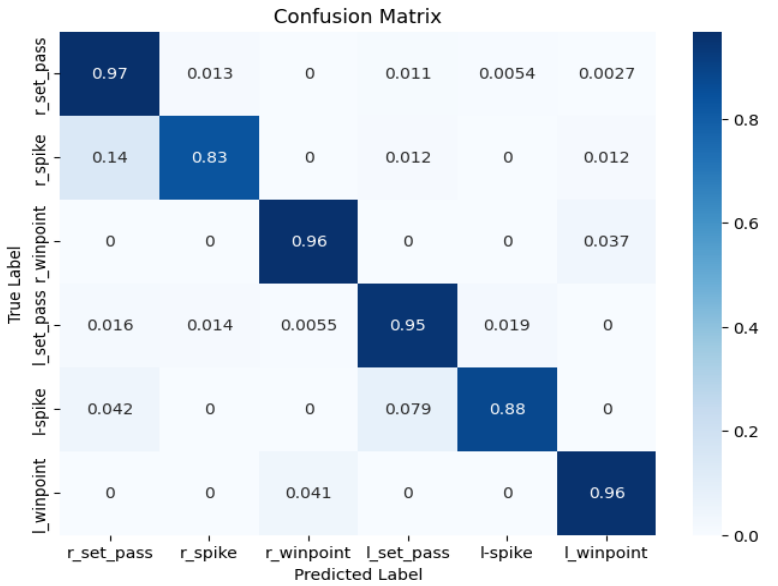
نبين أيضا في الشكل 47 مصفوفة الإرباك لبيانات الاختبار، حيث يتم بيان عدد التنبؤات الصحيحة من أجل كل صف على حد سواء وما الصفوف التي يتم الخلط بينها.



الشكل 47 مصفوفة الإرباك لاختبار النموذج على بيانت الاختبار بعد دمج set وpass

وبعد إجراء عملية استنظام على مستوى السطر للمصفوفة السابقة نحصل على الدقة المبينة في الشكل 48 على مستوى كل صف من الصفوف.

نلاحظ تحسّن دقة الصفوف التي تم دمجها وبالتالي تحسّن دقة التصنيف الإجمالية والحصول على نتيجة واعدة من النموذج.



الشكل 48 مصفوفة الإرباك لاختبار النموذج على بيانات الاختبار بعد دمج set وpass وإجراء استنظام

يعرض الجدول 4 مقارنة النتيجة المحققة مع نتيجة نموذج SAM [41] على مجموعة المعطيات Volleyball:

الدقة %	النموذج	عدد مقاطع التدريب
93.1	SAM [41]	3493
93	OURS	2146

الجدول 4 مقارنة النتيجة مع النموذج SAM [42]

نجد تفوق النهج المقترح في [41] بفارق 0.1% أي فرق بسيط جداً رغم استخدامنا بيانات تدريب أقل بسبب قيود الموارد، وفي هذه الورقة نوهوا أنهم استخدموا 3 إطارات من فيديو الدخل وهذا يجعل النمذجة الزمنية ليست حقيقية بشكل كامل، أما في عملنا فقد تم استخدام 16 إطار كما أن هذه الدراسة ذكرت أنها لا تستخدم مربعات الإحاطة التي يتم تزويدها من قبل ملفات التنويط ولكنهم في المرحلة الأولى من عملهم قاموا باستخدام خوارزمية كشف جاهزة للتزويد بمربعات الإحاطة وإدخالها إلى الكتلة التي تم اقتراحها SAM بشكل أساسي، أي بالنتيجة نهجنا أبسط بكثير من النهج المقترح في هذه الورقة ومعتمد على إطارات الفيديو كدخل أي يأخذ معلومات المشهد كاملة وليس فقط مربعات الإحاطة وقادر على تحقيق نتيجة تعتبر منافسة وواعدة.

5.5. خاتمة:

وجدنا في هذا الفصل نتائج نموذج التصنيف المقترح من أجل التنبؤ بالفئة الصحيحة للنشاط الجماعي حيث تبين أن قدرة هذه البنية تعتمد إلى حد كبير على جودة ميزات الفيديو المستخرجة. إذا كان أداء مستخرج الميزات ضعيف، يؤدي ذلك إلى الإضرار بالنتائج النهائية. حيث أن تدريب النموذج بشكل كامل أي استخراج ميزات أقوى حسن من جودة التصنيف.

يمكن للعوامل التالية من زيادة الدقة المكانية من 224 x 224 إلى 384 x 384، أو زيادة عدد الإطارات (استخدام 41 إطاراً المتوفرة) أو زيادة المعطيات كونها نسبةً لتدريب محوّل تعتبر صغيرة الحجم، أو المتابعة بتغيير ال hyperparameters للنموذج واستخدام إصدارات مختلفة من video swin أن تزيد من الدقة التي حققها النموذج، لكن محدودية الموارد المتاحة، وذلك وفقاً لمحدودية الحجوم وكذلك محدودية الوقت الذي يتيحه colab حتى مع استخدام النسخ المدفوعة منه، آلت إلى صعوبة اختبار النموذج وفق جميع المقترحات.

سنبين في الفصل القادم، أساليب مقترحة يمكن أن تؤدي إلى تحسين الميزات المستخرجة إذا تم تطبيقها بشكل دقيق، وبالتالي تحسين أداء النموذج في عملية التصنيف.

الفصل السادس: الخاتمة والآفاق المستقبلية

1.6. الخاتمة:

قمنا في هذا البحث بالتعرف على أنواع الأنشطة البشرية، وبالأخص نشاط المجموعة البشرية، وأساليب التّعرف عليها باستخدام تقنيات التعلم العميق، وقمنا بتقديم الشروحات عن نموذج المحول وعلى التابع الأساسي فيه وهو تابع الانتباه، كما أننا ذكرنا تطبيقات المحول في مجال الرؤية الحاسوبية، وركزنا على محوّل Swin [1] كونه الأساس الذي يقوم عليه العمل وتعرفنا على سبب ملائمته لتطبيقات الرؤية الصنعية. ثم شرحنا عن النموذج الأساسي المستخدم الذي يعتبر نموذج ملاءمة Swin في الصور إلى الفيديو، وهو Video Swin Transformer [7] الذي تبيّن أنه يعد خياراً واعداً لمهام التعرف على نشاط المجموعة نظراً لعدة أسباب وهي:

تعلم التمثيل الهرمي، حيث تم تصميمه لالتقاط التفاعلات المحلية والشاملة بطريقة هرمية، وهو أمر مفيد بشكل خاص للتعرف على نشاط المجموعة، والنمذجة الزمانية المكانية: يمكن لمحول الفيديو التقاط المعلومات المكانية والزمانية للفيديو عن طريق معالجة إطارات الإدخال بطريقة متسلسلة مما يسمح له بنمذجة الديناميكيات الزمنية بشكل أكثر فعالية مقارنة بالنماذج التي تعالج كل إطار بشكل مستقل.

ثم شرحنا النموذج المقترح وكيفية استخدامه مع مجموعة المعطيات Volleyball [2] وبيننا النتائج التي تم التوصل إليها، باستخدام فقط معلومة وسم نشاط المجموعة من مجموعة المعطيات المستخدمة.

2.6. الآفاق المستقبلية:

وفقاً للطرق المقترحة يوجد عدة آفاق للتحسينات التي يمكن العمل عليها مستقبلاً، نذكر منها المقترحين الآتيين:

1.2.6. المقترح 1:

استخدام نهج ثنائي الدفق، وذلك بحيث يتألف من عمودين فقريين لاستخراج الميزات، أحدهما هو video Swin transformer المستخدم في عملنا لاستخراج الميزات البصرية من مجموعة المعطيات، والآخر يقوم باستخراج التعليقات التوضيحية captionsمن مجموعة المعطيات، وبعد ذلك يتم تجميع هذه الميزات باستخدام آلية الانتباه لتصنيف المجموعة النهائي.

فيما يلي مخطط تفصيلي للنهج:

1. Feature extraction:

أ. استخراج الميزات البصرية: استخدام video Swin Transformer لاستخراج الميزات المكانية والزمانية من إطارات الفيديو. أي استخدام جزء مستخرج الميزات من هذا العمل.

ب. استخراج ميزات التعليقات التوضيحية: استخدام نموذج لغة مدرّب لاستخراج ميزات النص من التعليقات التوضيحية.

2. combining Features:

استخدام الانتباه التقاطعي cross attention لدمج الميزات البصرية والتعليق، من أجل الدمج متعدد الوسائط باستخدام آلية الانتباه متعدد الرؤوس من بنية المحولات، حيث تسمح هذه الآلية للنموذج بمعرفة العلاقات المختلفة بين الميزات البصرية والتعليق.

يتكون الإدخال لهذه الطبقة من تمثيل الفيديو كمجموعة من الميزات، بينما يتم تمثيل التعليقات التوضيحية على شكل سلسلة من الكلمات.

لدمج التمثيلات المرئية والنصية، يمكن استخدام الانتباه التقاطعي لمحاذاة الميزات المرئية مع الإدخال النصي، والعكس. على وجه التحديد، يمكن استخدام الميزات المرئية كـ "مفاتيح" و "قيم" في آلية الانتباه، بينما يمكن استخدام الإدخال النصي كـ "استعلامات". ستقوم آلية الانتباه بعد ذلك بتقييم أهمية كل ميزة مرئية بناءً على ملاءمتها للإدخال النصي، والعكس. ثم يتم دمج الميزات الموزونة لتشكيل تمثيل مشترك يلتقط العلاقة بين المدخلات المرئية والنصية.

وبعد ذلك يتم استخدام التمثيل المشترك الناتج لمهام التصنيف النهائية.

حيث بشكل عام، يعد الانتباه التقاطعي أسلوباً قوياً للجمع بين متجهين للتمثيل في المهام متعددة الوسائط، ويمكن أن يحسن بشكل كبير دقة وأداء نماذج التعلم العميق لهذه المهام.

3. Classification:

بمجرد دمج الميزات باستخدام آلية الانتباه، يمكن تدريب رأس التصنيف المستخدم في هذا العمل للتنبؤ بنشاط المجموعة النهائي من شعاع الميزات الجديد.

يوفر ربط السمات المرئية واللغوية تزويد نموذج التصنيف النهائي بمجموعة أكثر ثراءً من الميزات التي تلتقط كلاً من المظهر المرئي والمعنى الدلالي للأنشطة. يمكن أن يؤدي ذلك إلى تحسين قدرة النموذج على التمييز بين الاختلافات الدقيقة بين الإجراءات المتشابهة.

2.2.6. المقترح 2:

يمكن باستخدام النموذج الأساسي من هذا العمل، عدم الاعتماد فقط على خرج الطبقة الأخيرة من مستخرج الميزات video Swin transformer، ولكن يمكن دمج مخرجات مراحل مختلفة من النموذج (المكوّن من 4 مراحل) باستخدام كتلة انتباه تقاطعي بهدف إغناء التمثيل النهائي قبل إدخاله إلى رأس التصنيف.

المراجع

[1] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin Transformer: Hierarchical Vision Transformer using ShiftedWindows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[2] Mostafa S. Ibrahim and Srikanth Muralidharan and Zhiwei Deng and Arash Vahdat and Greg Mori, "Hierarchical Deep Temporal Models for Group Activity Recognition.," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016.

[3] Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra & Ajai, "A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets," *Applied Artificial Intelligence,* 2022.

[4] Palak Girdhar, Prashant Johri, Deepali Virmani, "Vision Based Human Activity Recognition: A Comprehensive Review of Methods & Techniques," *Turkish Journal of Computer and Mathematics Education,* 2021.

[5] Wang, Chuanchuan, and Ahmad Sufril Azlan Mohamed, "Group Activity Recognition in Computer Vision: A Comprehensive Review, Challenges, and Future Perspectives," *arXiv:2307.13541,* 2023.

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He, "SlowFast Networks for Video Recognition," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 2019.

[7] Liu, Ze, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[8] Allohvk, "Swin/Vision Transformers — Hacking the Human Eye," 17 1 2022. [Online]. Available: https://towardsdatascience.com/swin-vision-transformers-hacking-the-human-eye-4223ba9764c3.

[9] R. M. Schmidt, "Recurrent Neural Networks (RNNs): A gentle Introduction and Overview," *arXiv:1912.05911v1,* 2019.

[10] V. Perez, "Transformers in Computer Vision: Farewell Convolutions!," 23 11 2020. [Online]. Available: https://towardsdatascience.com/transformers-in-computer-vision-farewell-convolutions-f083da6ef8ab.

[11] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," *arXiv preprint arXiv:1508.04025,* 2015.

[12] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473,* 2014.

[13] R. Karim, "Attn: Illustrated Attention," 2022. [Online]. Available: URL https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3.

[14] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *dvances in neural information processing systems 30*, 2017.

[15] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition.," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[16] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton, "Layer Normalization," *arXiv preprint arXiv:1607.06450 ,* 2016.

[17] J. Alammar, "illustrated-transformer," 2018. [Online]. Available: https://jalammar.github.io/illustrated-transformer/.

[18] S. A. Vahora, N. C. Chauhan, "A comprehensive study of group activity recognition methods in video," *Indian Journal of Science and Technology,* vol. 10, no. 23, pp. 1-11, 2017.

[19] Wu, L. F., Wang, Q., Jian, M., Qiao, Y., & Zhao, B. X., "A Comprehensive Review of Group Activity in Video," *nternational Journal of Automation and Computing 18 ,* pp. 334-350, 2021.

[20] Y. M. Zhang, W. N. Ge, M. C. Chang, X. M. Liu, "Group context learning for event recognition," in *IEEE Workshop on the Applications of Computer Vision,IEEE, Breckenridge, USA,* , USA,, 2012.

[21] K. N. Tran, A. Gala, I. A. Kakadiaris, S. K. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," in *Pattern Recognition Letters*, 2014, pp. 49-57.

[22] Zha, Z.-J., Zhang, H., Wang, M., Luan, H., Chua, T.-S, "Detecting group activities with multi-camera context," *IEEE Transactions on Circuits and Systems for Video Technology,* 2013.

[23] Peng Dai, Huijun Di, Ligeng Dong, Linmi Tao, Guangyou Xu , "Group interaction analysis in dynamic context," *IEEE Trans Syst Man Cybern B Cybern,* 2008.

[24] M. R. Amer, S. Todorovic., "A chains model for localizing participants of group activities in videos," in *In Proceedings of International Conference on Computer Vision*, Barcelona, Spain,, 2011.

[25] X. B. Chang, W. S. Zheng, J. G. Zhang, "Learning person-person interaction in collective activity recognition," 2015, p. 1905-1918.

[26] J. C. Wu, L. M. Wang, L. Wang, J. Guo, G. S. Wu, "learning actor relation graphs for group activity recognition," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019.

[27] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[28] Lu, Lihua, Huijun Di, Yao Lu, Lin Zhang, and Shunzhou Wang, "A two level attention-based interaction model for multi-person activity recognition," *Neurocomputing,* no. 322 , pp. 195-205, 2018.

[29] Wongun Choi, Khuram Shahid, and Silvio Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," *ICCV Workshops,* 2009.

[30] Liu, Yang, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He, "A Survey of Visual Transformers," *IEEE Transactions on Neural Networks and Learning Systems,* 2023.

[31] Jiang, Y., Chang, S. and Wang, Z., "Transgan: Two pure transformers can make one strong gan, and that can scale up," in *Advances in Neural Information Processing Systems*, 2021, pp. 14745-14758.

[32] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929,* 2020.

[33] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A.,Zagoruyko, S., "End-to-End Object Detection with Transformers," *arXiv:2005.12872,* no. cs, 2020.

[34] Chen, Xin, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu, "Transformer Tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[35] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[36] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021.

[37] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[38] Arnab, Anurag, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[39] Fan, Haoqi, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[40] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek, "Actor-transformers for group activity recogntion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2020.

[41] Yan, R., Xie, L., Tang, J., Shu, X., & Tian, Q, "Social adaptive module for weakly-supervised group activity recognition," in *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, 2020.

[42] Thilakarathne, Haritha, Aiden Nibali, Zhen He, and Stuart Morgan, "Pose is all you need: The pose only group activity recognition system (POGARS)," *Machine Vision and Applications,* vol. 33, no. 6, 2022.

[43] Sun, Ke, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[44] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Doll´ar, "Microsoft coco: Common objects in context," *ECCV,* 2014.

[45] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani,and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[46] Jo˜ao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *CVPR,* 2017.

[47] Simonyan, Karen, and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, 2014.

[48] Perez, Mauricio, Jun Liu, and Alex C. Kot, "Skeleton-based relational reasoning for group activity analysis," *Pattern Recognition,* vol. 122, 2022.

[49] Sendo, Kohei, and Norimichi Ukita, "Heatmapping of people involved in group activities," in *2019 16th International Conference on machine vision applications* , 2019.

[50] Newell, A., Yang, K., Deng, J., "Stacked hourglass networks for human pose estimation," *Lecture Notes in Computer Science,* 2016.

[51] Timur M. Bagautdinov, Alexandre Alahi, Franc¸ois Fleuret, Pascal Fua, and Silvio Savarese, "Social scene understanding End-to-end multi-person action localization and collective activity recognition," *CVPR,* 2017.

[52] m. Ljaz, "Swin Transformer: Hierarchical Vision Transformer using Shifted Window — Part I," 13 2 2022. [Online]. Available: https://medium.com/aiguys/swin-transformer-hierarchical-vision-transformer-using-shifted-window-part-i-5dc3fe7ae774.

[53] A. Arora, "Swin Transformer Model Architecture explained with PyTorch implementation line-by-line," 4 7 2022. [Online]. Available: https://amaarora.github.io/posts/2022-07-04-swintransformerv1.html.

[54] W. Choi, S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision*, Florence, Italy,, 2012.

[55] Chuanchuan Wang, Ahmad Sufril Azlan Mohamed, "Group Activity Recognition in Computer Vision: A Comprehensive Review, Challenges, and Future Perspectives," *arXiv:2307.13541,* 2023.

[56] "Encoder-Decoder Seq2Seq for Machine Translation - Dive into Deep Learning," [Online]. Available: https://d2l.ai/chapter_recurrent-modern/seq2seq.html.

[57] Azar, S.M., Atigh, M.G., Nickabadi, A., "A multi-stream convolutional neural network framework for group activity recognition.," *URL: http://arxiv.org/abs/1812.10328,* 2018.

[58] Azar, Sina Mokhtarzadeh, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi, "Convolutional relational machine for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[59] Shu, T., Todorovic, S., Zhu, S.C.,, "Cern: Confidenceenergy recurrent network for group activity recognition," *IEEE Conference on Computer Vision and Pattern Recognition,* 2017.

[60] Mehrasa, Nazanin, Yatao Zhong, Frederick Tung, Luke Bornn, and Greg Mori, "Learning person trajectory representations for team activity analysis," *arXiv preprint arXiv:1706.00893,* 2017.