Syrian Arab Republic
Ministry of Higher Education
Syrian Virtual University

الجمهورية العربية السورية
وزارة التعليم العالـي
الجامعة الافتراضية السورية

خطة بحث لرسالة ماجستير في علوم الوب

اسم المشرف:	د. وسيم صافي
بريد المشرف:	t_wsafi@svuonline.org
اسم الطالب:	هشام جمعة معاون
بريد الطالب:	Hisham_194930@svuonline.org
عنوان الرسالة بالعربية:	التعرف على الأحداث في الفيديوهات باستخدام تقنيات التعلم العميق
عنوان الرسالة بالإنكليزية:	Deep Learning based Action recognition in Videos
الكلمات المفتاحية:	Artificial Intelligence, Machine learning, Deep learning, Artificial Neural Networks, MobileNet, VGG16, Residual Network-50 , python, Django , CNN-LSTM, 3D CNN
فصل التسجيل على الرسالة:	**F23**
تاريخ إعداد الرسالة:	

الفهرس

1 المصطلحات والتعاريف

- مفهوم التعلم الآلي Machine Learning

يعني التعلم الآلي مسألة كيفية بناء حواسب قادرة على التعلم من خلال التجربة، أي قادرة على أن تتحسن تلقائياً من خلال الخبرة. يعتبر هذا المجال نتيجةً لتقاطع علوم الكمبيوتر مع الإحصاء بالإضافة إلى الرياضيات وعلوم البيانات. تشمل تطبيقاته العديد من المجالات مثل الرعاية الصحية، والتصنيع، والتعليم، والاقتصاد، والأمن، والتسويق.

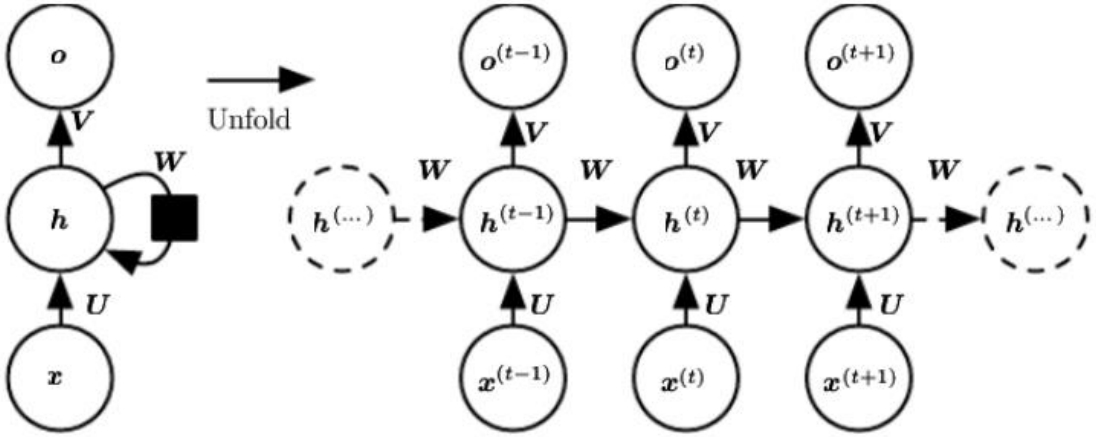
مثال على استخدام خوارزميات التعلم الآلي يعتمد عادة على تمثيل البيانات التي تتعلق بمسألة محددة، مثل مسألة الكشف المبكر عن السرطان، أن تساعد خوازرميات التعلم الآلي الأطباء في تحديد وجود الأورام بشكل مبكر، مما يزيد من فرص العلاج الناجح ويحسن من نتائج العلاج للمرضى. تتعلم خوارزميات التعلم الآلي كيفية استخدام مجموعة متنوعة من الواصفات (مثل السمات) والتي تمثل خصائص فريدة قابلة للقياس ومميزة لظاهرة معينة، لتحقيق نتائج مختلفة.
من الممكن حل العديد من مهام الذكاء الاصطناعي من خلال اختيار الجملة الصحيحة من الوصفات. ومع ذلك، في بعض الأحيان يكون من الصعب تحديد الوصفات المهمة لهذه المهام، على سبيل المثال، في إدخال هذه الوصفات إلى خوارزمية التعلم الآلي.
ومع ذلك، بالنسبة للعديد من المهام، يكون من الأفضل استخراج الواصفات بدلاً من تقديمها. تحل هذه المشكلة بشكل حقيقي من خلال اتباع نهج يهدف ليس فقط إلى اكتشاف الطريقة المثلى لاستخدام الواصفات، ولكن أيضًا لاكتشاف الواصفات ذاتها.
أحد التحديات الرئيسية في العديد من تطبيقات الذكاء الاصطناعي في الواقع العملي تتمثل في وجود العديد من العوامل التي تؤثر على كل جزء من البيانات ويمكن ملاحظتها. على سبيل المثال، قد يكون للبكسلات في صورة سيارة لون أحمر قريب جدًا من الأسود في الظلام، أو قد يتغير شكل ظل السيارة استنادًا إلى زاوية الرؤية. يتطلب معظم التطبيقات فصل الوصفات الهامة التي تزودنا بالمعلومات المميزة، بغض النظر عن الظروف المتغيرة، وتجاهل الوصفات الأخرى. قد يكون من الصعب جدًا استخراج هذه الوصفات ذات المستوى العالي من البيانات الأولية، ويمكن فهمها فقط من خلال فهم معقد للبيانات يتفوق على قدرات الإنسان.
عندما يكون من الصعب الحصول على تمثيل كافٍ للمشكلة الأصلية، يبدو أن التعلم التمثيلي يمكن أن يكون مفيدًا. يحل التعلم العميق هذه المشكلة في تعلم التمثيل من خلال تقديم تمثيلات أكثر تقدمًا من خلال تمثيلات أخرى أكثر بساطة. سنشرح هذا المفهوم ونوضح كيفية تطبيقه في الفقرة التالية.
يعرف هذا النهج بتعلم التوصيف (Representation Learning) [4] .

- الشبكات العصبونية الصنعية Artificial Neural Networks (ANN)

الشبكات العصبونية الصناعية (ANN) هي تقنية تعلم آلي مشهورة تحاكي آلية التعلم في الجهاز العصبي البيولوجي. تتضمن الشبكة العصبونية الصناعية وحدات حسابية تُسمى العصبونات، حيث يتوضع كل مجموعة من العصبونات ضمن طبقة. يرتبط الدخل مع العصبونات في الطبقة الأولى عن طريق وصلات لكل منها وزن، وتتصل كذلك العصبونات مع بعضها عن طريق وصلات موزونة. يتم حساب الإخراج عن طريق عملية تدعى الانتشار التقدمي، حيث يكون دخل كل عصبون هو جمع الإخراج من الوصلات التي تدخل إليه، كل منها بوزن معين. يتم معالجة الدخل من قبل العصبون عن طريق تطبيق وظيفة التنشيط عليه، وبعد ذلك يتم توجيه الإخراج لعصبون آخر وهكذا حتى الوصول إلى الطبقة النهائية. يحدث التعلم عن طريق تغيير أوزان الوصلات التي تصل العصبونات بالاعتماد على أمثلة معطيات التدريب. تُعطي معطيات التدريب تغذية راجعة حول صحة الأوزان التي يتم اختيارها بالاعتماد على مقدار صحة الإخراج الذي توقعته الشبكة العصبونية مقارنة بالقيمة الحقيقية للإخراج المتوقع لنفس هذا الدخل في معطيات التدريب.
ولها نوعان:

- **شبكات العصبونية ذات التغذية الأمامية Fee-Forward Neural Networks :** تتضمن عدة طبقات للشبكة العصبونية، حيث يتم استقبال الدخل من خلال الطبقة الأولى (طبقة الدخل) ويتم معالجته عبر الطبقات الوسيطة (الطبقات الخفية) حتى الوصول إلى الإخراج. في هذا النوع من الشبكات، لا توجد معلومات تنتقل من طبقة معينة إلى الطبقات السابقة، ويُعرف عدد الطبقات الخفية في الشبكة بعمق الشبكة العصبونية.

- **الشبكات العصبونية العودية Recurrent Neural Networks (RNNs)** : تم تصميم هذا النوع من الشبكات العصبونية للتعامل مع البيانات التي تكون على شكل سلسلة، وذلك عن طريق السماح للخرج في خطوة سابقة بأن يُستخدم كإحدى معطيات الدخل في الخطوة التالية. لفهم بنيتها بشكل أفضل، دعونا نلقي نظرة على الشكل (1).



الشكل (1) البنية العامة للشبكة العصبونية العودية

- **عمليات التعلم Learning في الشبكات العصبونية الصنعية:**

عملية التعلم في الشبكات العصبونية تتم من خلال عدة خطوات:

1- **معالجة المعطيات المدخلة (توابع التنشيط)** : تتمثل هذه الخطوة في تطبيق وظيفة التنشيط على المعطيات المدخلة إلى كل عصب في الشبكة. توابع التنشيط تعمل على تحويل المدخلات إلى إشارات خرجية محسوبة تمثل الإشارات الكهربائية التي تنتقل بين العصبونات في الشبكة.

2- **تقييم أداء النموذج (وظائف الخسارة):** بعد معالجة المعطيات، يتم تقييم أداء النموذج عن طريق مقارنة الإخراج الحالي للنموذج بالإخراج المتوقع. تُستخدم وظائف الخسارة (Loss Functions) لقياس الفارق بين الإخراج الفعلي والمتوقع. يهدف التعلم إلى تقليل هذه الخسارة عبر ضبط معاملات النموذج.

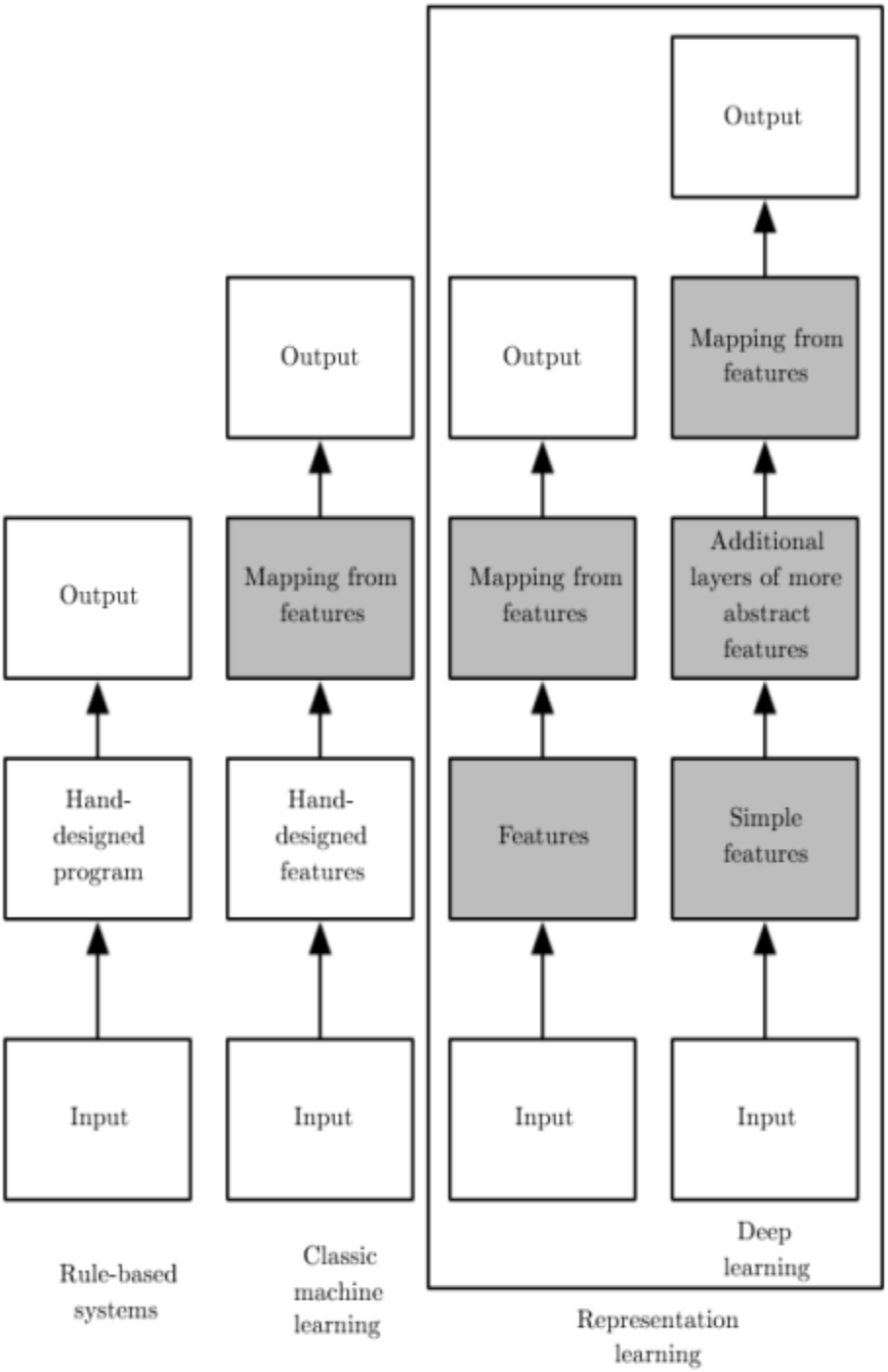
3- **ضبط الأوزان بناءً على التقييم:** بعد حساب الخسارة، يتم استخدام تقنيات مثل الانحدار العكسي (Backpropagation) لتحديث الأوزان والانحيازات في الشبكة بهدف تقليل الخسارة. يتم ذلك عن طريق حساب مشتقات الخسارة بالنسبة لكل معلمة في النموذج واستخدامها لتحديث قيم هذه المعلمات بشكل مناسب.

عملية التعلم تتكرر عدة مرات، حيث يتم تقدير الأداء، وتحديث الأوزان، ومعالجة المعطيات مرارًا وتكرارًا حتى يتم تحقيق أداء مرضٍ من قبل النموذج. استخدام تقنيات التعلم العميق، مثل الشبكات العصبونية العميقة، يمكن أن يتطلب تعيين معلمات كثيرة، وتدريباً طويلاً ومعقداً، ولكنه يمكن أن يؤدي إلى نماذج قادرة على تحقيق أداء متميز في مجموعة متنوعة من المهام الذكاء الصناعي. [5]

مفهوم التعلم العميق Deep Learning

التعلم العميق هو نوع خاص من التعلم الآلي الذي يحقق قوة كبيرة ومرونة من خلال تمثيل البيانات على أنها مجموعة من النماذج المتداخلة للمفاهيم، حيث أن كل مفهوم يرتبط بمفاهيم أبسط، وحيث أن الواصفات الأكثر تجريدية مشتقة من وصفات أقل تجريدية للبيانات. في حالة كانت المدخلات للحاسوب مجردة وحسية مثل قيم بكسلات صورة على سبيل المثال، يبدو أن تعلم أو تقييم هذا الارتباط يصعب تحديده مباشرة إذا كان الغرض هو التعامل معه بشكل مباشر ومعقد للغاية. يحل التعلم العميق

هذه الصعوبة عن طريق تقسيم هذا الارتباط المعقد إلى سلسلة من الارتباطات البسيطة التي تمتد عبر طبقات متعددة من النموذج. تقوم هذه الطبقات المتداخلة، والتي تُدخل كمدخلات في الطبقة المرئية، بمساعدة الموديل على استخراج المزيد من الوصفات التجريدية بشكل متزايد من البيانات المدخلة. تُسمى هذه الطبقات "المخفية" لأن قيمها غير معروفة في البيانات المدخلة، بل يتعين على النموذج تحديدها كمفاهيم مفيدة لشرح العلاقات في البيانات المرصودة. بشكل عام، سوف نوضح أكثر عن هذه الطبقات في الفقرة القادمة عن الشبكات العصبونية الصنعية[4]، كما في الشكل (2).



الشكل (2) اختصصات جمال الذكاء الصنعي وآلية عمل كل منها

2 فهم الحدث البشري وتحليله: أساسيات التعرّف وتصنيف الأحداث

تعريف الحدث في سياق التعرف على الحدث البشري يعتبر تحديًا، حيث يمكن للحدث أن يشمل مجموعة متنوعة من الحركات، من البسيطة إلى الأكثر تعقيدًا، والتي قد تكون جزءًا من سلسلة من الأحداث المتسلسلة والمترابطة. ولذلك، يبدو مصطلح "حدث" صعب التعريف، تقدم بعض التعاريف لمصطلح "الحدث" و"النشاط" في هذا السياق التالي:

1. تعريف الحدث:

- يُصف الحدث عادةً على أنه حركة بسيطة يقوم بها شخص واحد عادةً لفترة قصيرة جدًا، غالبًا ما يكون في نطاق عشرات الثواني.

- يُمكن أن يكون للحدث معنى محددًا وغالبًا ما يكون ذو مغزى للفرد الذي يقوم به.

2. تعريف النشاط:

- يُعرف النشاط على أنه سلسلة معقدة من الأحداث، والتي يمكن أن تتضمن العديد من الحركات المختلفة، وتكون منفذة من قبل عدة أشخاص يتفاعلون مع بعضهم.

- يمكن أن يكون للنشاط مجموعة متنوعة من الأهداف والمعاني، ويتأثر بالبيئة المحيطة والظروف المحيطة به.

استنادًا إلى هذه التعاريف، يُمكن تحديد **الحدث** كأصغر حركة يقوم بها الإنسان، ويكون لها معنى محدد ومعروف. في حين يُمكن تصنيف النشاط كسلسلة من الأحداث المترابطة التي يتفاعل فيها عدة أشخاص، وتكون لها أهداف ومعانٍ متعددة.

تحقيق التعرف على الحدث البشري يهدف إلى فهم سلوك الإنسان وتسمية أو تصنيف كل حدث. يشمل هذا مجموعة واسعة من التطبيقات التي لها تأثير هائل على المجتمع. هذه التطبيقات تتنوع من التفاعل بين الإنسان والحاسوب، إلى بناء أنظمة المراقبة والحماية، والتحليلات وتصنيف مقاطع الفيديو حسب المحتوى، بالإضافة إلى العديد من التطبيقات الأخرى.

تشمل هذه التطبيقات مجالات متنوعة مثل:

1. **الطب الرياضي:** تتضمن مراقبة وتتبع حركات الرياضيين لتحسين الأداء البدني والوقاية من الإصابات.
2. **التعليم والتدريب**: استخدام تقنيات التعرف على الحدث لتقديم تدريبات فعالة وشخصية في مجالات مثل التدريب الرياضي والتعليم عن بعد.
3. **الرعاية الصحية:** متابعة حالة المرضى وتقديم الرعاية الصحية المخصصة باستخدام مستشعرات الحركة وتقنيات التعرف على الحدث.
4. **الأمن والمراقبة:** استخدام التعرف على الحدث للكشف عن الأنشطة غير المرغوب فيها مثل السلوك العدواني أو الجريمة.
5. **تقديم المساعدة للأشخاص ذوي الاحتياجات الخاصة:** تصميم تطبيقات تستجيب لحركات الأفراد وتساعدهم على القيام بالمهام اليومية بشكل أفضل.

3 المشكلة العلمية ومبررات مشروع البحث

تكمن المشكلة التي يعالجها التعرف على الفعل البشري في الحاجة إلى إنشاء واجهات أكثر بديهية وطبيعية بين البشر وأجهزة الكمبيوتر. يمكن أن تكون طرق الإدخال التقليدية مثل لوحات المفاتيح وأجهزة الماوس مرهقة، خاصة في السيناريوهات التي تتطلب التفاعل بدون استخدام اليدين، أو للمستخدمين ذوي الإعاقة. وبالتالي، فإن التحدي يكمن في تطوير أنظمة يمكنها فهم وتفسير التصرفات والإيماءات البشرية من مصادر البيانات المختلفة، مثل تدفقات الفيديو، أو أجهزة استشعار العمق، أو الأجهزة القابلة للارتداء، وترجمتها إلى أوامر أو تفاعلات ذات معنى مع أجهزة الكمبيوتر.

1. **Enhanced User Experience:** من خلال تمكين أجهزة الكمبيوتر من التعرف على الإجراءات والإيماءات البشرية والاستجابة لها، يمكننا إنشاء تجارب مستخدم أكثر جاذبية. يمكن أن يؤدي ذلك إلى زيادة رضا المستخدم وإنتاجيته، خاصة في تطبيقات مثل الألعاب والواقع الافتراضي والواقع المعزز.

2. **Accessibility**: يمكن لأنظمة التعرف على الأفعال البشرية أن تجعل التكنولوجيا في متناول نطاق أوسع من المستخدمين، بما في ذلك الأشخاص ذوي الإعاقة أو ذوي الإعاقة الحركية. على سبيل المثال، يمكن للواجهات القائمة على الإيماءات أن توفر طرق إدخال بديلة للأفراد الذين يجدون صعوبة في استخدام أجهزة الإدخال التقليدية.

3. **Natural Interaction:** يمكن أن يؤدي التفاعل الطبيعي مع أجهزة الكمبيوتر إلى تحسين التواصل والتعاون بين البشر والآلات. ومن خلال فهم التصرفات والإيماءات البشرية، يمكن لأجهزة الكمبيوتر تكييف استجاباتها في الوقت الفعلي، مما يجعل التفاعلات أكثر مرونة وبديهية.

4. **Efficiency**: في سياقات معينة، مثل البيئات الصناعية أو بيئات الرعاية الصحية، يمكن أن يؤدي التفاعل مع أجهزة الكمبيوتر بدون استخدام اليدين إلى تحسين الكفاءة والسلامة بشكل كبير. على سبيل المثال، يمكن للجراحين الذين يقومون بإجراء العمليات الاستفادة من الواجهات القائمة على الإيماءات التي تسمح لهم بالوصول إلى المعلومات دون الحاجة إلى لمس لوحة المفاتيح أو الماوس.

4 أهداف البحث

1. **تطوير خوارزميات فعالة للتعرف على الإجراءات:** يمكن أن يكون الهدف الأول هو تطوير وتحسين خوارزميات قوية للتعرف على الإجراءات والإيماءات البشرية من مصادر البيانات المختلفة، مثل تدفقات الفيديو أو أجهزة استشعار العمق أو الأجهزة القابلة للارتداء. يتضمن هذا الهدف استكشاف مختلف بنيات التعلم العميق واستراتيجيات التدريب وتقنيات زيادة البيانات لتحسين دقة النماذج وقدرات تعميمها.
2. **تقييم الأداء عبر سيناريوهات متنوعة:** يمكن أن يركز الهدف الثاني على تقييم أداء الخوارزميات المطورة عبر سيناريوهات وبيئات متنوعة. يتضمن ذلك جمع مجموعات بيانات مشروحة تمثل مجموعة واسعة من الإجراءات والإيماءات البشرية، بما في ذلك السلوكيات الشائعة والمحددة السياق. الهدف هو تقييم قوة الخوارزميات ودقتها وقابلية التوسع في تطبيقات العالم الحقيقي.
3. **تحسين تصميم التفاعل وتجربة المستخدم**: قد يكون الهدف الآخر هو تحسين تصميم أنظمة التفاعل بين الإنسان والحاسوب (HCI) بناءً على قدرات خوارزميات التعرف على الإجراءات المطورة. يتضمن ذلك التصميم والتحسين المتكرر لواجهات المستخدم وآليات التغذية الراجعة وطرق التفاعل لتحقيق أقصى قدر من سهولة الاستخدام والحدس ورضا المستخدم. يمكن إجراء دراسات المستخدم وجلسات التغذية الراجعة feedback sessions لتقييم فعالية تصميمات التفاعل المختلفة.
4. **استكشاف التطبيقات الجديدة وحالات الاستخدام:** يهدف المشروع إلى استكشاف التطبيقات الجديدة وحالات الاستخدام التي تتيحها تقنية التعرف على الفعل البشري. يتضمن هذا الهدف تحديد المجالات والسيناريوهات التي يمكن أن يوفر فيها التفاعل القائم على الإيماءات مزايا كبيرة مقارنة بطرق الإدخال التقليدية. قد تشمل الأمثلة الرعاية الصحية والتعليم والألعاب والروبوتات والبيئات الذكية. الهدف هو إظهار الفائدة العملية وتعدد استخدامات التكنولوجيا المتقدمة في مواجهة تحديات العالم الحقيقي وتحسين التفاعل بين الإنسان والحاسوب.

5 النتائج التطبيقية المتوقعة من البحث والجهات المستفيدة منها

5.1- يمكن تلخيص النتائج المتوقعة من البحث بما يلي:

1- الاطلاع على مجموعة المعطيات المتعلقة بحدث معين .
2- التعرف على الأدوات والمكتبات البرمجية اللازمة للتدريب والتدرب عليها .
3- بناء وتدريب نموذجين دون استخدام تقنيات التعليم العميق
4- بناء وتدريب 11 نموذج باستخدام تقنيات التعليم العميق
5- إنشاء صفحة ويب لعرض النماذج

5.2 - الجهات التي يمكن أن تستفيد من هذا العمل:

1- **المستهلكون**: يمكن للمستهلكين يوميًا استخدام أنظمة التفاعل القائمة على الإيماءات للتحكم في الأجهزة المنزلية الذكية ووحدات تحكم الألعاب وأنظمة الترفيه وتطبيقات الواقع الافتراضي/الواقع المعزز (VR/AR). يمكن لتقنية التعرف على الإيماءات أن تعزز تجارب المستخدم من خلال توفير طرق أكثر سهولة وغامرة للتفاعل مع الأجهزة الرقمية والمحتوى.

2- **المتخصصون**: يمكن للمتخصصين في مجالات مثل الرعاية الصحية والتعليم والهندسة والتصميم الاستفادة من أنظمة التفاعل القائمة على الإيماءات لتعزيز الإنتاجية والكفاءة. على سبيل المثال، يمكن للجراحين استخدام عناصر التحكم بالإيماءات بدون استخدام اليدين للوصول إلى بيانات التصوير الطبي أثناء العمليات الجراحية، ويمكن للمعلمين استخدام الواجهات القائمة على الإيماءات لأنشطة التدريس والتعلم التفاعلية، ويمكن للمهندسين المعماريين/المهندسين استخدام عناصر التحكم بالإيماءات لمعالجة النماذج ثلاثية الأبعاد وتصميم النماذج الأولية.

3- **الأفراد ذوو الإعاقة**: يمكن للأشخاص ذوي الإعاقة أو ذوي الإعاقة الحركية الاستفادة بشكل كبير من أنظمة التفاعل القائمة على الإيماءات، والتي توفر طرق إدخال بديلة لا تتطلب معالجة مادية لأجهزة الإدخال التقليدية مثل لوحات المفاتيح أو الفئران. على سبيل المثال، يمكن للأفراد الذين يعانون من إعاقات حركية استخدام الإيماءات للتحكم في التقنيات المساعدة، والتنقل بين الواجهات الرقمية، والتواصل مع الآخرين.

4- **التطبيقات الصناعية والتجارية**: يمكن لصناعات مثل التصنيع والخدمات اللوجستية وتجارة التجزئة والسيارات الاستفادة من أنظمة التفاعل القائمة على الإيماءات لمختلف التطبيقات. على سبيل المثال، يمكن للعاملين في منشآت التصنيع استخدام عناصر التحكم بالإيماءات لتشغيل الآلات، ويمكن للفنيين في المستودعات استخدام الإيماءات لتتبع المخزون وإدارة الخدمات اللوجستية، ويمكن لموظفي التجزئة استخدام الواجهات القائمة على الإيماءات للشاشات التفاعلية ومشاركة العملاء.

5- **متخصصو الرعاية الصحية**: يمكن لمتخصصي الرعاية الصحية الاستفادة من أنظمة التفاعل القائمة على الإيماءات في الإعدادات السريرية لمهام مثل الوصول إلى سجلات المرضى، والتحكم في الأجهزة الطبية، وإجراء العمليات الجراحية. يمكن

أن تعمل عناصر التحكم بالإيماءات على تحسين النظافة من خلال تمكين التفاعل بدون استخدام اليدين مع الواجهات الرقمية، مما يقلل من خطر التلوث في البيئات المعقمة.

6- **صناعة الترفيه والإعلام**: يمكن لصناعة الترفيه والإعلام دمج أنظمة التفاعل القائمة على الإيماءات في المعارض التفاعلية، ومناطق الجذب في المتنزهات الترفيهية، والعروض الحية، والتجارب الغامرة. يمكن لتقنية التعرف على الإيماءات أن تعزز تفاعل الجمهور ومشاركته في مختلف الأماكن والفعاليات الترفيهية.[6]

6 الدراسة المرجعية

نقوم باستعراض تقنيات التعلم العميق المستخدمة في التعرف على الأحداث في الفيديوهات بشكل عام، ومن ثم سنتخصص التعرف على الفعل البشري، في النهاية سنتحدث عن تقنية نقل التعلم التي تم استخدامها في كثير من هذه الاعمال.

- **استخدام تقنيات التعلم العميق المستخدمة في التعرف على الأحداث في الفيديوهات:**

قبل استخدام التعلم العميق في مجال التعرف على الأحداث، كانت المحاولات تهدف لإيجاد طريقة للتمثيل ثلاثي الأبعاد للحدث، أي طريقة لإيجاد واصفات تمثل هذا الحدث والاستفادة منها في التعرف. تم تعريف طريقتين لتحقيق ذلك وهما التمثيل الكلي holistic representation والتمثيل المحلي local representation

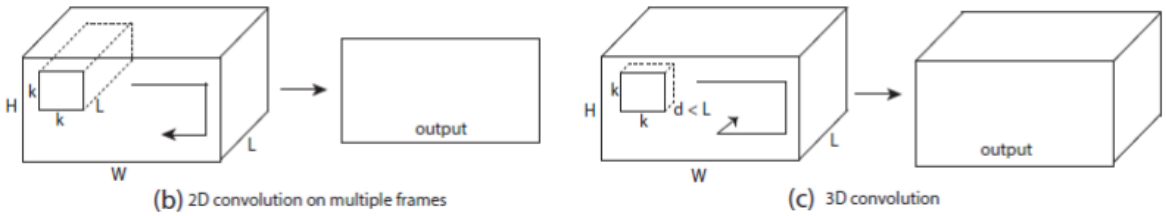
في التمثيل الكلي يعتمد التعرف على الحدث عن طريق استخراج تمثيل عام لبنية جسم الإنسان وشكله وحركاته مثل نموذج space-time volume.

أما في التمثيل المحلي يعتمد التعرف على الحدث على استخراج واصفات محلية أي استخراج نقاط اهتمام في البعد المكاني – الزمني باستخدام طرق مثل HOG أو FIT في كلا التمثيلين كانت تتم عملية استخراج الواصفات بشكل يدوي handcrafted دون تعلم. في بعض الأحيان، كان يتم استخدام التعلم الآلي للتعرف عن طريق إدخال الواصفات المستخرجة يدوياً إلى خوارزمية مثل آلة المتجهات الداعمة SVM (Support Vector Machine) أو الغابة العشوائية Random Forest. [3]

مع ظهور محاولات التعلم العميق والشبكات التلافيفية تم الاستغناء عن الطرق اليدوية، حيث أصبح النموذج يتعلم استخراج الواصفات التي تفيده في عملية التعرف بنفسه، مباشرةً من المعطيات الأولية. وبسبب طبيعة الشبكات العصبونية التي تعتمد على تمثيل الواصفات من خلال عدة طبقات متدرّجة التعقيد، استطاعت هذه الشبكات استخراج الواصفات من مجموعات المعطيات المعقدة والكبيرة وتقديم تمثيل عالي المستوى لها. هذا ما جعل تقنيات التعلم العميق تحقق نتائج مميزة جداً في مسائل التعرف على الأحداث.

توجد عدة تقنيات تم اتباعها في التعلم العميق للتعرف على الأحداث، تعتمد معظم هذه التقنيات بشكل أساسي على مرحلتين: أول مرحلة هي بناء نموذج يستخلص الواصفات المكانية الزمنية spatio-temporal features من إطارات الفيديو، والثانية هي بناء نموذج يستخدم هذه الواصفات من أجل القيام بعملية التعرف غالباً ما يكون النموذج في المرحلة الثانية هو شبكة كاملة الارتباط. تحدّد الواصفات المكانية طبيعة العناصر الموجودة ضمن إطار معين وأماكنها وارتباطها مع بعضها، بينما تحدد الواصفات الزمنية تغيّر هذه العناصر بين إطارات الفيديو سوف نذكر في الفقرات اللاحقة مجموعة من نماذج التعلم العميق التي تم استخدمها في عملية استخراج الواصفات المكانية الزمنية من الفيديوهات [3]

- **الشبكة العصبونية التلافيفية بثلاثة أبعاد (3D-CNN 3) D-Convolutional Network:** هي نسخة موسعة من D-CNN2، حيث أن لها بعد زمني إضافي كما رأينا فإن استعمال -CNN يتم لمعالجة الصور، بينما يتم استعمال D-CNN 3لمعالجة سلسلة من الإطارات ثنائية البعد. تتضمن هذه الشبكة طبقتين تلافيفية وتجميعية كل منهما بمرشح ثلاثي الأبعاد. عادةً ما يتم ترميز دخل هذه الشبكة برباعية تكون على الشكل W x H x L x C :حيث C هو عدد القنوات اللونية في الإطار، Lهو عدد الإطارات، HWهو طول وعرض كل إطار. في حال استعمال D-CNN 2لعملية معالجة سلسلة إطارات فيديو باعتبار كل إطار على قناة لونية مختلفة دون إجراء عملية تلافيفيّة على البعد الزمني فإن الناتج سيكون ثنائي البعد فقط وبالتالي فإن هذه الشبكة لن تملك القدرة على ربط الإطارات وستفقد المعلومات الزمنية. سيتم حل هذه المشكلة في حال استخدام D-CNN 3وسيكون الناتج ثلاثي البعد يتضمن واصفات مكانية وزمنية. يوضح الشكل 3 مقارنة بين الشبكة التلافيفية ثنائية وثلاثية البعد في مجال معالجة سلسلة من الصور. **[12]**

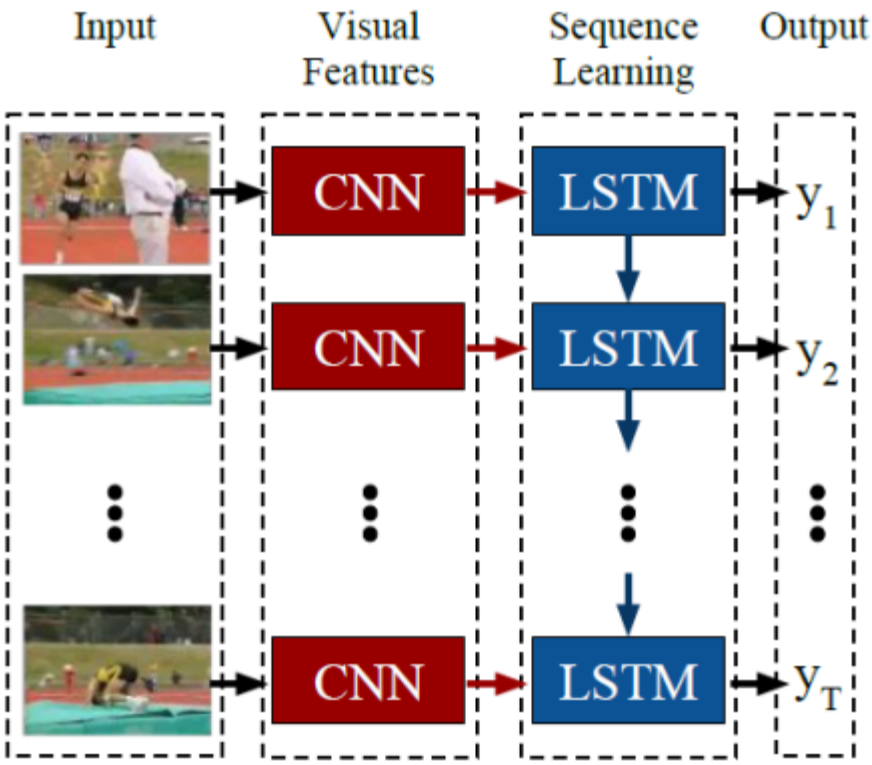


الشكل (3) الشبكات التلافيفية بثلاثة أبعاد (b) عند معالجة سلسلة إطارات فيديو باعتبار كل إطار على قناة لونية فإن الناتج سيكون ثنائي البعد فقط (c) ينتج من تطبيق التفاف ثلاثي الأبعاد على حجم فيديو حجم آخر، وبذلك يتم الاحتفاظ بالمعلومات الزمنية لإشارة الإدخال.

■ الشبكة التلافيفية المتبوعة بشبكة عودية CNN- RNN Architecture

يتم استخدام هذا النموذج لنمذجة المعطيات البصرية ذات التسلسل الزمني visual time-series modeling، أي يمكن استخدامه في مسألة التعرف على الأحداث في الفيديوهات. يتألف هذا النموذج بشكل أساسي من شبكة تلافيفية تقوم بعملية استخراج الواصفات الزمنية من كل إطار بشكل منفصل، ومن ثم يتم تجميع الواصفات المستخرجة من الإطارات على شكل سلسلة زمنية وإدخالها لشبكة عودية RNN تقوم باستخراج الواصفات الزمنية. يتم بعد ذلك إلحاق هذا النموذج بشبكة كاملة الارتباط لتتعلم باستخدام هذه الواصفات.تم في استخدام هذه الهيكلية حيث استعملت طبقة LSTM كطبقة عودية

وتم إطلاق اسم LRCN على النموذج. وتم إثبات أن هذه الطريقة ناجحة ليس فقط لعملية التعرف على الحدث وإنما أيضاً لعملية تفسير الصور والفيديوهات كما في الشكل (4). **[13]**

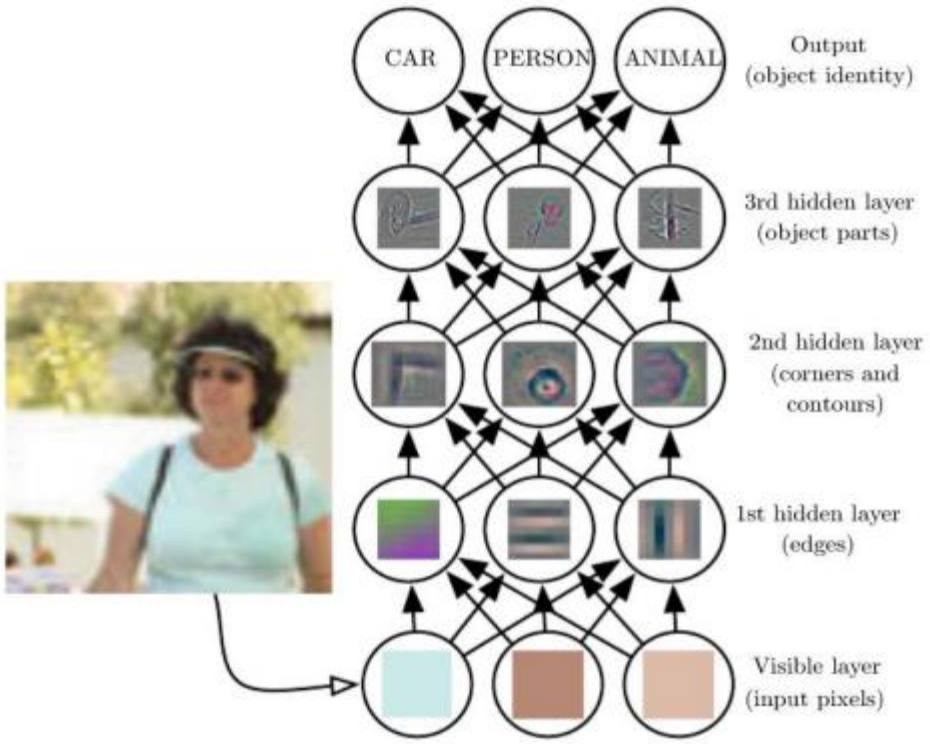


الشكل (4) نموذج LRCN (Long-term Recurrent Convolutional Network)

■ **الشبكات متعدّدة الدّفقةMultiple Stream Networks**

تم تصميم هذا الصنف من الشبكات لفصل المعلومات المتعلقة بالمظهر عن المعلومات المتعلقة بالحركة. تم تقديم أول شبكة تلافيفية متعددة الدفقة في، حيث تم استعمال شبكتين تلافيفيتين على التوازي من أجل عملية التعرف على الحدث. الشبكة التلافيفية على الدفقة المكانية spatial stream تأخذ كدخل إطار عشوائي من مجموعة إطارات معينة من الفيديو عددها L ، أي تأخذ دخل ثابت (صورة ثابتة).

بينما تأخذ الشبكة التلافيفية على الدفقة الزمنية temporal stream حقول التدفق البصري لهذه الإطارات. يتم تعريف التدفق البصري الكثيف dense optical flow لإطار معين على أنه مجموعة أشعة الإزاحة التي تنقل كل بكسل من موضعه في هذا الإطار إلى موضعه في الإطار التالي له. لكل شعاع من هذه الأشعة مركبتين أفقية وعمودية، ولذلك كان الدخل للشبكة على الدفقة الزمنية على شكل حجم ذو بعد WxHx2L باعتبار WH هما أبعاد الإطارات و 2 Lهو عدد القنوات حيث L هو عدد الإطارات التي يتم حساب التدفق البصري لها. نلاحظ أن هذا الدخل يصف بشكل صريح الحركة بين إطارات الفيديو، مما يجعل التعرف أسهل، حيث أنه لا تحتاج الشبكة بهذه الحالة إلى تقدير الحركة ضمنياً. أخيراً تم إلحاق كل من الدّفقتين بشبكة كاملة الارتباط للاستفادة من الواصفات المستخرجة من كل منهما، وتم ودمج الخرج النهائي بعدة طرق منها أخذ المتوسط. يوضح الشكل 5 هذه العملية.



الشكل (5) عملية استخراج المعطيات البصرية تدريجياً عبر الطبقات

- **تقنية نقل التعلّم مع ضبط دقيقTransfer Learning with Fine-tuning**

في تقنية نقل التعلم، يتم تطبيق معرفة نموذج تعلم آلي تم تدريبه على مشكلة مختلفة ذات صلة بالمشكلة التي نريد حلها. مع نقل التعلم، نحاول بشكل أساسي استغلال ما تم تعلمه في مهمة واحدة لتحسين التعميم في مهمة أخرى [4]. أي بشكل توضيحي أكثر نقوم بنقل الأوزان التي تعلمتها الشبكة في المهمة (أ) إلى المهمة (ب) الجديدة، وبدلاً من بدء عملية التعلم من البداية، نبدأ بالأنماط التي تم تعلمها من حل مهمة ذات صلة.

تستخدم هذه التقنية كثيراً في مجالات الرؤية الحاسوبية ومعالجة اللغات الطبيعية، وأحياناً يتم تطبيقها مع عملية ضبط دقيق Fine-tuningلملائمة النموذج الذي تم تدريبه مسبقاً ليحل مسألتنا، أي على سبيل المثال في مجال الرؤية الحاسوبية، تحاول الشبكات التلافيفيّة عادةً اكتشاف الحواف في الطبقات السابقة والأشكال في الطبقات الوسطى وبعض الميزات الخاصة بالمهمة في الطبقات اللاحقة باستخدام نقل التعلم مع ضبط دقيق، يتم استخدام الطبقات المبكرة والمتوسطة من النموذج المدرب مسبقاً ونقوم بإضافة طبقات جديدة وتدريب هذه الطبقات فقط

يعتبر نقل التعلم من التقنيات القوية جداً في مجال الرؤية الحاسوبية لأنه سمح بإنشاء تطبيقات كثيرة دون الحاجة لوقت تدريب طويل، وساهم في تحسين أداء النموذج بشكل كبير دون الحاجة إلى بيانات كثيرة أو قدرة حسابية هائلة واستعمال ذاكرة مفرط. سوف نتحدث في الفقرات الفرعية الآتية عن ثلاثة نماذج شهيرة جداً مدربة مسبقاً لاستعمالها في مجال الرؤية الحاسوبية. تم تدريب كل منها على قواعد معطيات هائلة من الصور وأعطت نتائج مميزة جداً.

- **نموذجMobileNet**

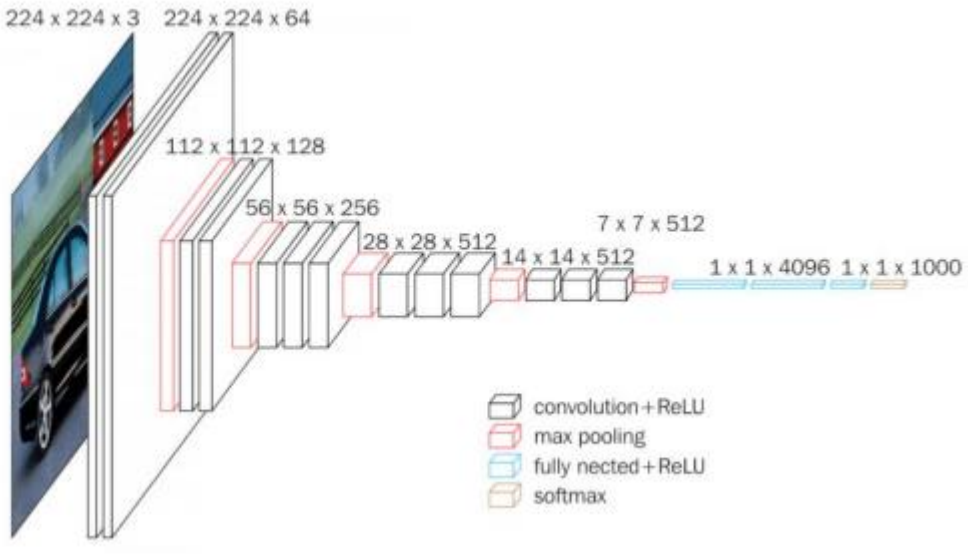
يتميز هذا النموذج بقدرته الكبيرة على تخفيف التعقيد الحسابي الهائل الموجود في الشبكات العميقة التلافيفية العادية عن طريق تعريف عمليات تلافيفية جديدة تدعىdepthwise separable convolution ، تتألف هذه العملية من جزأين:

- عملية تلافيفية واحدة من نمط depthwise يقوم المرشح فيها بمسح كل قناة لونية بمعزل عن الأخرى
- عملية (أو عدة يتم تنفيذها باستخدام مرشح من نمط 1 1 x لتجميع نتيجة خرج المرشح من عملية depthwise واستخراج الواصفات العملية التلافيفية التقليدية تقوم بالترشيح والتجميع بنفس الخطوة كما رأينا سابقاً.

عملية :pointwise convolution يتم تطبيق مرشح ذو حجم 1 1 x على خرج عملية depthwise لتجميعه واستخراج الواصفات. يتم تطبيق عدد N من هذه المرشحات على نفس الخرج وذلك بقدر عدد خرائط الواصفات المراد استخراجها. يحتاج هذه النموذج إلى 106 * 569 عملية ضرب جمع MADD(Multiply-Add) لأعداد ممثلة بالفاصلة العائمة من أجل عملية انتشار تقدّمي

■ **نموذجVGG16 (Visual Geometry Group16)**

تم طرح هذا النموذج لرؤية مدى تأثير عمق الشبكة التلافيفية على الدقة أثناء تدريبها للتعرف على الصور. قدم هذا النموذج مفهوماً جديداً، ألا وهو تجميع عدة طبقات تلافيفية بأحجام نواة صغيرة بدلاً من طبقة تلافيفية واحدة بحجم نواة كبير (التوسع بالعمق وليس بالعرض [26] ، حيث أن هذا النموذج اعتمد على استخدام حجم مرشح ثابت (3 (3 x واستغنى عن وضع طبقات بحجوم مرشحات كبيرة. تمكن هذا النموذج بالاعتماد على هذه الفكرة من تخفيض عدد المتغيرات القابلة للتدريب في الشبكة التلافيفية، وكذلك قدّم إثباتاً بأن العمق الكبير الشبكة التلافيفية مفيد لزيادة دقة التصنيف. كما في الشكل (6)

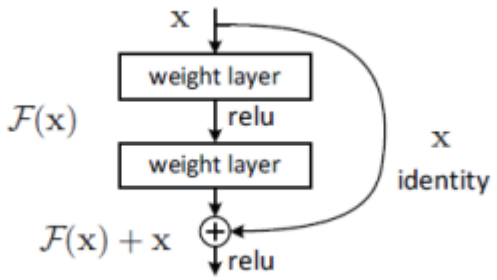


الشكل (6) بنية منوذج VGG16

يُعد vgg16 من أكثر النماذج المدربة مسبقاً المستخدمة حالياً وذلك بسبب بنيته المنتظمة التي تجعله مناسب جداً لمسائل الرؤية الحاسوبية. يتألف هذا النموذج من 16 طبقة تلافيفية ويحتاج 10 * 15.5 عملية MADD لأعداد ممثلة بالفاصلة العائمة من أجل عملية انتشار تقدّمي واحدة.

- **نموذج Residual Network-50 (ResNet50)-**

حقق إضافة مفهوم العمق للشبكات التلافيفية تحسيناً كبيراً على مستوى مسألة تصنيف الصور كما رأينا في نموذجvgg16 ، وهذا ما جعلنا نفترض بأن إضافة طبقات جديدة إلى النموذج إما سيحسن أداءه أو على الأقل سوف يبقيه كما هو، حيث أنه وبأسوأ الأحوال ستلعب الطبقات المضافة دور تابع مطابقة بينما الطبقات الأخرى ستكون مماثلة تماماً للنموذج الأول (قبل إضافة الطبقات الجديدة). ولكن أثبتت التجارب بأنه مع زيادة العمق تصل الدقة لحالة إشباع وتبدأ بالانخفاض وهذا ما يسمى مشكلة التراجع .degradation بشكل غير متوقع، هذا التراجع غير ناجم عن مشكلة فرط ملائمة وإنما يأتي من فكرة أن تعلم النموذج العميق ليس بسهولة بنائه، وأن الأنظمة تختلف فيما بينها من حيث سهولة أمثلتها [27]. استطاع نموذج ResNet حل هذه المشكلة بتعريف ما يسمى الوصلة المختصرة .shortcut connection لتوضيح ذلك دعونا ننظر للشكل7 .



الشكل (7) وحدة البناء الأساسية لنموذج ResNet

7 الدراسات المتشابهة

1- **التعرف على وضعية الإنسان في الوقت الفعلي في أجزاء من صور ذات عمق واحد:** بقلم جيمي شوتون وآخرون. (2011) حيث تقدم هذه الورقة طريقة للتعرف على وضعية الإنسان في الوقت الحقيقي باستخدام كاميرا عمق واحدة. ويعتمد هذا النهج على نموذج هرمي يمثل جسم الإنسان كمجموعة من الأجزاء مثل الرأس والجذع والذراعين والساقين.
ومن خلال اكتشاف أجزاء الجسم هذه وتتبعها في صور متعمقة، يمكن للنظام تقدير وضعية الشخص بدقة في الوقت الفعلي. هذه التكنولوجيا لها تطبيقات في مجالات مختلفة، بما في ذلك التعرف على الإيماءات، والتفاعل بين الإنسان والحاسوب، وتحليل الحركة. [7]

2- **التعلم العميق لاكتشاف الأجزاء البشرية في الصور:** بقلم بيوتر دولار وآخرون. (2014)حيث تقدم هذه الورقة أسلوب التعلم العميق لاكتشاف أجزاء جسم الإنسان في الصور، تستخدم الطريقة الشبكات العصبية التلافيفية (CNNs) لتتعلم تلقائيًا ميزات اكتشاف أجزاء الجسم، مثل الرأس والكتفين والمرفقين والركبتين.
من خلال تدريب الشبكة على مجموعات بيانات كبيرة من الصور المشروحة، يمكن للنظام تحديد وتصنيف أجزاء الجسم البشري بدقة في صور العالم الحقيقي. هذه التقنية لها تطبيقات في التعرف على الحركة، وتقدير الوضعية، والأنظمة التفاعلية. **[8]**

3- **الكشف عن إيماءات اليد والتعرف عليها في الوقت الفعلي باستخدام الشبكات العصبية التلافيفية**: بقلم Weicheng Xie et al. (2016) حيث : تقدم هذه الدراسة نظامًا للكشف عن إيماءات اليد والتعرف عليها في الوقت الفعلي استنادًا إلى الشبكات العصبية التلافيفية (CNNs).
يستخدم النظام بنية CNN لاستخراج الميزات من الصور اليدوية التي تلتقطها الكاميرا في الوقت الفعلي. يتم بعد ذلك استخدام هذه الميزات لتصنيف إيماءات اليد إلى فئات محددة مسبقًا، مما يتيح التفاعل الطبيعي والبديهي مع أجهزة الكمبيوتر والأجهزة. تتمتع هذه التقنية بتطبيقات في التفاعل بين الإنسان والحاسوب، والواقع الافتراضي، والألعاب، والبيئات الذكية. **[9]**

4- **التعلم العميق لتحليل الحركة البشرية:** التعرف على الأفعال البشرية ورسم الخرائط عبر الزمن" بقلم يورجن جال وآخرون. (2016) الوصف: تقدم هذه الورقة نظرة عامة على تقنيات التعلم العميق لتحليل الحركة البشرية، بما في ذلك التعرف على الحركة وتقدير الوضعية. يناقش المؤلفون مختلف بنيات CNN واستراتيجيات التدريب لنمذجة التبعيات الزمنية في التصرفات البشرية. من خلال التقاط المعلومات المكانية والزمانية من تسلسلات الفيديو، يمكن لهذه النماذج التعرف على الإجراءات المعقدة ورسم خريطة لوضعيات الإنسان عبر الزمن. لهذه التكنولوجيا تطبيقات في التفاعل بين الإنسان والحاسوب، والمراقبة، والروبوتات، والتعرف على الأنشطة.**[10]**

5- **التعرف على إيماءات اليد باستخدام الشبكات العصبية التلافيفية للتفاعل بين الإنسان والحاسوب**: بقلم S. Dharmawan وآخرون. (2018) حيث يركز هذا البحث على التعرف على إيماءات اليد باستخدام الشبكات العصبية التلافيفية (CNNs) لتطبيقات التفاعل بين الإنسان والحاسوب.
تستكشف الدراسة بنيات CNN المختلفة ومنهجيات التدريب للتعرف على الإيماءات بشكل قوي وفعال. من خلال تدريب الشبكة على مجموعات بيانات إيماءات اليد، يمكن للنظام تصنيف إيماءات اليد بدقة في الوقت الفعلي، مما يتيح التفاعل الطبيعي والبديهي مع أجهزة الكمبيوتر والأجهزة. لهذه التقنية تطبيقات في الواقع الافتراضي، والألعاب، والبيئات الذكية، والتعرف على لغة الإشارة. **[11]**

8 التقانات المقترح استخدامها

سوف نشرح عن الأدوات البرمجية المستخدمة، مع إيضاح المساهمة التي قدمتها كل أداة في تنفيذ القسم العملي.

1- لغة البرمجة python وبيئة التطوير Jupyter

- **لغة Python** هي لغة تفسيرية عالية المستوى تستخدم أسلوب البرمجة غرضية التوجه. تتميز لغة بايثون بسهولة تعلمها وبساطة نحوها syntax ولذلك يتم استخدامها في تطبيقات كثيرة. اعتمدنا هذه اللغة لأنها تقدم مكتبات عديدة مفتوحة المصدر تساعد على بناء نماذج التعلم العميق ومعالجة المعطيات الكبيرة.

بيئة Jupyter بيئة مفتوحة المصدر تدعم لغة بايثون تتميز هذه البيئة بإمكانية تقسيم الرماز البرمجي بواسطتها على شكل عدة أجزاء واستعراض نتائج تنفيذ كل جزء. استخدمنا هذه البيئة لتنفيذ عملنا عليها باعتبارها بيئة قوية لكتابة الرماز بشكل منظم وإجراء تعديلات وعمليات تجريب مرنة عن طريق إتاحة القدرة على تنفيذ جزء معين من الرماز دوناً عن الأجزاء البقية.

- **منصة الويب Django:**

Django هو إطار عمل framework عالي المستوى يستخدم في تطبيقات الويب، ويشجع فكرة التطوير السريع والتصميم النظيف. يوفر هذا الإطار الكثير من متاعب تطوير الويب بحيث يتم التركيز على كتابة التطبيق فقط. استخدمنا Django من أجل بناء صفحة ويب للتعرف على الأحداث البشرية العنيفة في مقطع فيديو يتم إدخاله.

2- المكتبات البرمجية المستخدمة:

- **مكتبة keras**

هي واجهة برمجة تطبيقات API للتعلم العميق مكتوبة بلغة Python، وتعمل فوق المكتبة الشهيرة Tensorflow.

تم تطوير هذه المكتبة لتوفير أداة قوية تتيح الإمكانية للتركيز على إجراء بحث جيد والانطلاق من الفكرة للتجريب العملي بسهولة ومرونة. استخدمنا الواجهات الآتية من هذه المكتبة أثناء عملنا:Layers API .

استخدمنا هذه الواجهة من أجل تنجيز الطبقات المختلفة التي تتكون منها النماذج، مثل الطبقات التلافيفية ببعدين وبثلاثة أبعاد، الطبقات التجميعية ببعد واحد وببعدين وبثلاثة أبعاد طبقات التجميع الشمولي ببعد واحد وببعدين، الطبقات العودية مثل GRU, STM BiLSTM الطبقة الكثيفة بالإضافة لطبقة التسرب Dropout وطبقة تنظيم الدفعة Batch Normalization.

- **Models API :-**

ا هذه الواجهة من أجل عملية بناء النماذج التي اطلعنا عليها في الدراسة المرجعية. استخدمنا الصف Sequential منها لبناء نموذج عن طريق تجميع عدة طبقات متسلسلة.

- Applications API

هذه الواجهة من أجل استخدام النماذج التلافيفية المدربة مسبقاً ,VGG16, MobileNet ResNet50 بأوزان تدريبها على مجموعة المعطيات ImageNet

- Optimizers API

هذه الواجهة من أجل تنجيز تابع الأمثلة للنموذج. استخدمنا من أجل كل نموذج الصف Adam لأمثلة تابع الخسارة أثناء عملية التدريب مع ضبط لقيم متغيراته beta_1, beta_2, learning rate.

Callbacks API

الصف EarlyStopping من هذه الواجهة من أجل إيقاف عملية التدريب في حال مرور عدة أدوار دون تحسن وذلك لتفادي مشكلة فرط الملائمة. استخدمنا أيضاً الصف ReduceLROnPlateau وذلك لتخفيض معدل التعلم في حال مرور عدة أدوار دون تحسن لأن معدل التعلم الكبير من الممكن أن يؤدي لتباعد خوارزمية الأمثلة أو تأرجحها حول الحل.

مكتبة OpenCV

توفر هذه المكتبة أدوات لبناء برمجيات الرؤية الحاسوبية والتعلم الآلي، وهي مفتوحة المصدر. تم إنشاؤها لتوفير بنية قوية وبيئة غنية تحوي أدوات متنوعة لمعالجة الصور والفيديوهات وذلك بغرض استخدامها في تطبيقات الرؤية الحاسوبية.

استخدمنا الصفVideoCapture

من هذه المكتبة من أجل قراءة إطارات ملف الفيديو عن طريق التابع read، ومن أجل معرفة معدل التأطير للفيديو framing rate، ومن أجل تغيير حجم الإطار عن طريق التابع resize.

- **مكتبة matplotlib ومكتبة seaborn**

استفدنا من أدوات هاتين المكتبتين في تمثيل واستعراض النتائج بيانياً. حيث استخدمناهما من أجل عملية رسم منحنيات التعلم الطيات الخمس، بالإضافة للمخططات الشريطية.

- **مكتبة sklearn**

توفر هذه المكتبة أدوات قوية من أجل بناء نماذج التعلم الآلي وإجراء الحسابات الإحصائية. استخدمنا هذه المكتبة من أجل

تقسيم المعطيات وفق طريقة تدريب / اختبار train test split ومن أجل تنفيذ التصديق المتقاطع باستخدام k طية.

9 الخطة الزمنية

المرحلة	الأعمال التنفيذية	أرقام الأشهر في العام												النتائج المتوقّعة
		1	2	3	4	5	6	7	8	9	10	11	12	
دراسة عامة ونظرية	– دراسة طبيعة المسالة – التعرف على الشبكات العصبونية والية العمل	X	X											تقرير عن الدراسة
دراسة مرجعية	دراسة التقنيات المستخدمة في مجال التعرف على الأحداث و دراسة هيكلية النماذج			X	X									تقرير عن الدراسة
التنفيذ العملي	اطلاع على مجموعة المعطيات - التعرف على الأداوات والمكتبات البرمجية اللازمة والتدرب عليها - بناء تدريب نموذجي دون استخدام تقنية نقل تعلم - بناء وتدريب 11 نموذج مع استخدام تقنية نقل تعلم - إنشاء موقع ويب الويب					X	X	X	X	X	X			تقرير عن الدراسة
التقرير	التقرير النهائي											X	X	تسليم نسخة نهائية

10 قائمة الاشكال

11 المراجع

1- Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video technology, 18(11), 1473-1488.

2- Wang, X., Farhadi, A., & Gupta, A. (2016). Actions~ transformations. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 2658-2667)

3- Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4-21

4- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

5- Aggarwal, C. C. (2018). Neural networks and deep learning. Springer, 10, 978-3.

6- Human-Computer Interaction, Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale.

7- Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116-124.

8- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2014). Deep learning for human part discovery in images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7), 1147-1162.

9- Xie, W., Noble, J. A., & Zisserman, A. (2016). Real-time hand gesture detection and recognition using convolutional neural networks. In European Conference on Computer Vision (pp. 103-117). Springer, Cham.

10- Gall, J., Zuffi, S., Stoll, C., Trimpe, S., & Theobalt, C. (2016). Deep learning for human motion analysis: Recognizing human actions and mapping poses across time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(3), 417-437.

11- Dharmawan, S., Indra, M. A., Hikmawan, K., Khotimah, C. N., & Satoto, B. D. (2018). Hand gesture recognition using convolutional neural networks for human-computer interaction. Journal of Physics: Conference Series, 1114(1), 012033.

12- Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221- 231.

13- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).