الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم النظم المعلوماتية

أعدت هذه الأطروحة لنيل
درجة الماجستير في نظم المعطيات الكبيرة بعنوان

توصيف الصور نصياً باستخدام تقنيات التعلم العميق

Image Captioning Using Deep Learning Techniques

إعداد

م. محمد عبد الهادي الملا

إشراف

د. آصف جعفر **د. ندى غنيم**

كانون الأول 2021

تعريف بالمعهد العالي للعلوم التطبيقية والتكنولوجيا

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

كلمة شكر

أتقدم بجزيل الشكر إلى كل من الدكتور آصف جعفر والدكتورة ندى غنيم، اللذان شرفاني بقبولهما الإشراف على رسالتي، على كل ما قدماه من جهد ووقت ثمينين وأفكار ونصائح أدت إلى نجاح بحثي هذا.

أتوجه أيضاً بالشكر إلى أساتذتي في الماجستير وفي المرحلة الجامعية، الذين لم يوفروا جهداً وكان همهم نجاحنا.

م. محمد عبد الهادي الملا

الملخص

يعد توصيف الصور الآلي من المسائل الشائعة في الذكاء الصنعي الحديث، حيث يهتم هذا المجال بإنشاء نص كخرج يصف صورة ما في الدخل، ويمكن لهذا الوصف أن يكون جملة واحدة أو أكثر. يمكن استخدام توصيف الصور لفهرسة الصور تِلْقَائِيًّا وهو أمر مهم في العديد من تطبيقات معالجة الصور.

تعالج مسألة توصيف الصور مؤخراً باستخدام تقنيات التعلم العميق، وخاصة نماذج الترميز–فك الترميز Encoder–Decoder. في هذه الأطروحة، نقدم نموذجاً معتمداً على تقنية الانتباه attention من نمط Encoder–Decoder يستفيد من السمات من الطبقات التلافيفية convolutional المستخرجة من نموذج Xception المدرب مسبقاً على مجموعة بيانات ImageNet، وسمات الأغراض المستخرجة من نموذج YOLOv4 الذي درب مسبقاً على مجموعة بيانات MSCOCO. نقدم أيضاً مخطط ترميز موضعي Positional Encoding Scheme جديداً لسمات الأغراض نسميه "عامل الأهمية"، ونبين تأثيره على معايير التقييم. نختبر في هذا البحث نموذجنا على مجموعة بيانات MSCOCO ونقارنها بأعمال مماثلة.

نقدم أيضاً دراسة تجريبية شاملة حول استخراج السمات باستخدام شبكات CNN لمسألة توصيف الصورة في سياق التعلم العميق. أجرينا مجموعة من 72 تجربة باستخدام 3 مجموعات بيانات على 12 شبكة لتصنيف الصور مدربة مسبقاً على مجموعة بيانات ImageNet. ندرس تأثير تغيير مستخرج السمات CNN على جودة التوصيف للصور، ونجد علاقة قوية بين بنية النموذج ومجموعة بيانات الصور المستخدمة. للاستفادة من هذه النتائج، نوصي بمجموعة من شبكات CNN المدربة مسبقاً لكل من مقاييس تقييم توصيف الصور التي نريد تحسينها.

تحسب درجات التقييم باستخدام المقاييس الثمانية القياسية في مجال توصيف الصور الآلي. يساهم هذا العمل في تطوير مجال توصيف الصور الآلي من خلال تقديم طرق تمثيل أفضل للصور.

الكلمات المفتاحية: توصيف الصور آلياً؛ سمات الأغراض؛ الشبكة العصبونية CNN؛ التعلم العميق؛ استخراج السمات

Abstract

Image captioning is one of the trending problems in modern Artificial Intelligence (AI). It is concerned with generating an output text describing an input image, where the output can be one or more sentences. Image captioning is important for many reasons. For example, it can be used for automatic image indexing, which is important for many applications.

The problem of image captioning has been solved recently by deep learning techniques, especially Encoder-Decoder methods. In this thesis, we present an Encoder-Decoder attention-based architecture that makes use of convolutional features extracted from the Xception model pre-trained on ImageNet, and object features extracted from the YOLOv4 model, pre-trained on MSCOCO. We also introduce a new positional encoding scheme for object features, "the importance factor", and show its effect on evaluation scores. We test our model on the MSCOCO dataset and compare it to similar works.

We also present a thorough experimental study about feature extraction using Convolutional Neural Networks (CNNs) for the task of image captioning in the context of deep learning. We perform a set of 72 experiments using 3 datasets on 12 image classification CNNs pre-trained on the ImageNet dataset. We study the effect of changing the CNN feature extractor on image captioning quality, and find a strong relationship between the model structure and the image captioning dataset. To benefit from these results, we recommend a set of pre-trained CNNs for each of the image captioning evaluation metrics we want to optimise.

The evaluation scores are calculated using the eight standard metrics in the image captioning field. Our work contributes to image captioning by introducing better representation schemes for images.

Keywords: Image Captioning; Object Features; Convolutional Neural Network; Deep Learning; Feature Extraction

قائمة المحتويات

تعريف بالمعهد العالي للعلوم التطبيقية والتكنولوجيا.....2
كلمة شكر.....4
الملخص.....5
Abstract.....6
قائمة المحتويات.....8
قائمة الأشكال.....11
قائمة الجداول.....12
قائمة الاختصارات والمصطلحات.....13
الفصل الأول: مقدمة عامة.....15
1.1 تمهيد.....15
2.1 إشكالية ودوافع البحث.....17
3.1 فكرة الحل المقترح.....17
4.1 المساهمات الأساسية في البحث.....17
5.1 مخطط البحث.....18
الفصل الثاني: الدراسة المرجعية.....19
1.2 مقارنة بين المنهجيات المتعاكسة.....20
1.1.2 الفضاء الصوري Visual Space والفضاء متعدد الوسائط Multimodal Space.....20
1.1.1.2 الفضاء الصوري Visual Space.....20
2.1.1.2 الفضاء متعدد الوسائط Multimodal Space.....21
2.1.2 التعلم بالإشراف Supervised Learning وطرق التعلم العميق الأخرى Other Deep Learning.....21
1.2.1.2 منهجيات التعلم بالإشراف Supervised Learning.....22
2.2.1.2 طرق التعلم العميق الأخرى Other Deep Learning.....22
3.1.2 التوصيف الكثيف Dense Captioning والتوصيف للمشهد بأكمله Captions for the Whole Scene.....23
1.3.1.2 التوصيف الكثيف Dense Captioning.....23
2.3.1.2 التوصيف للمشهد بأكمله Captions for the Whole Scene.....24

4.1.2 بنية الترميز وفك الترميز Encoder–Decoder Architecture والبنية التركيبية Compositional Architecture.....24
1.4.1.2 بنية الترميز وفك الترميز Encoder–Decoder Architecture.....24
2.4.1.2 توصيف الصور المعتمد على البنية التركيبية Compositional Architecture.....25
5.1.2 تصنيفات أخرى.....26
1.5.1.2 توصيف الصور المعتمد على الانتباه Attention–based Image Captioning.....26
2.5.1.2 توصيف الصور المعتمد على المفهوم الدلالي Semantic Concept–Based Image Captioning.....30
3.5.1.2 توصيف الصور القادر على التعرف على الأغراض الجديدة Novel Object–based Image Captioning.....31
4.5.1.2 التوصيف ذو النص المنمط Stylized Captioning.....32
2.2 دراسات لمنهجيات استخراج السمات في توصيف الصور الآلي.....33
1.2.2 التوصيف الآلي للصور بالاستعانة بمعلومات أغراض الصورة.....33
2.2.2 مقارنة نماذج CNN لاستخدامها في استخراج السمات.....35
3.2 مقارنة المنهجية المستخدمة في هذا البحث مع منهجيات الأبحاث السابقة.....37
4.2 مجموعات البيانات.....38
1.4.2 مجموعة بيانات MSCOCO.....38
2.4.2 مجموعة بيانات Flickr30K.....38
3.4.2 مجموعة بيانات Flickr8K.....38
4.4.2 مجموعة بيانات Visual Genome.....39
5.4.2 مجموعة بيانات Instagram.....39
6.4.2 مجموعة بيانات IAPR TC–12.....39
7.4.2 مجموعة بيانات Stock3M.....40
8.4.2 مجموعة بيانات MIT–Adobe FiveK.....40
9.4.2 مجموعة البيانات FlickrStyle10k.....40
10.4.2 مجموعة بيانات Google Conceptual Captions.....40
5.2 معايير التقييم.....41
الفصل الثالث: النماذج المقترحة في هذه الأطروحة.....43

1.3 النموذج الأولي (Baseline Model).....43
2.3 التعديل 1 (Double Word Embedding).....44
3.3 التعديل 2 (YOLO Bounding Boxes).....45
4.3 التعديل 3 (YOLO Raw Features v1).....48
5.3 التعديل 4 (YOLO Raw Features v2).....49
6.3 التعديل 2.1 (YOLO Bounding Boxes).....51
الفصل الرابع: الاختبارات والنتائج.....53
1.4 المرحلة الأولى من التجارب.....53
2.4 المرحلة الثانية من التجارب.....54
3.4 المرحلة الثالثة من التجارب.....59
4.4 المرحلة الرابعة من التجارب.....65
5.4 المرحلة الخامسة من التجارب.....69
6.4 المرحلة السادسة من التجارب.....71
الفصل الخامس: الخاتمة والآفاق المستقبلية.....81
الملاحق.....83
الملحق أ: التقنيات المستخدمة.....83
أ.1 مكتبة TensorFlow 2.....83
أ.2 مكتبة Keras.....83
أ.3 MSCOCO Evaluation Toolkit.....84
أ.4 مكتبة Numpy.....84
أ.5 مكتبة Matplotlib.....85
أ.6 مكتبة yolov4.....85
أ.7 Hardware.....85
المراجع.....86
Abstract.....93

قائمة الأشكال

الشكل 1.1.....16
الشكل 1.2.....20
الشكل 1.3.....45
الشكل 2.3.....46
الشكل 3.3.....48
الشكل 4.3.....50
الشكل 5.3.....51
الشكل 6.3.....53
الشكل 1.4.....57
الشكل 2.4.....58
الشكل 3.4.....59
الشكل 4.4.....65
الشكل 5.4.....66
الشكل 6.4.....72
الشكل 7.4.....74
الشكل 8.4.....76

قائمة الجداول

الجدول-1.2.....37
الجدول-1.3.....48
الجدول-1.4.....54
الجدول-2.4.....56
الجدول-3.4.....57
الجدول-4.4.....58
الجدول-5.4.....59
الجدول-6.4.....60
الجدول-7.4.....61
الجدول-8.4.....62
الجدول-9.4.....68
الجدول-10.4.....69
الجدول-11.4.....70
الجدول-12.4.....72
الجدول-13.4.....74

قائمة الاختصارات والمصطلحات

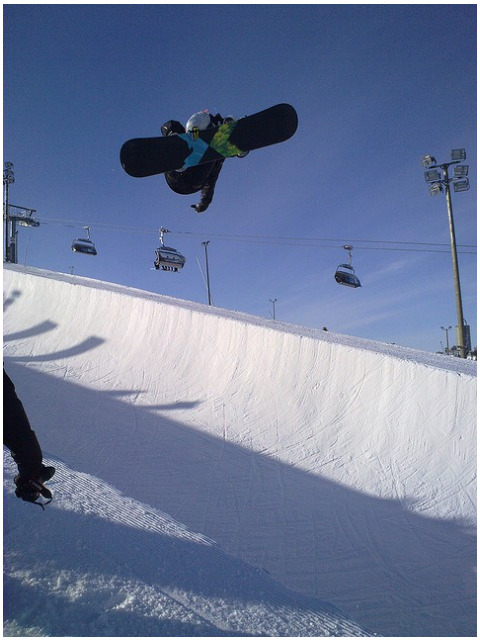
المعنى باللغة العربية	المعنى باللغة الإنكليزية	الاختصار
الذكاء الصنعي	Artificial Intelligence	AI
استرجاع الصور المستند إلى المحتوى	Content-Based Image Retrieval	CBIR
الذاكرة الطويلة قصيرة الأمد	Long Short-Term Memory	LSTM
وحدة التكرار ذات البوابات	Gated Recurrent Unit	GRU
شبكة توليدية متقابلة	Generative Adversarial Network	GAN
حجز Dirichlet الكامن	Latent Dirichlet Allocation	LDA
شبكة ذاكرة تسلسلية سياقية	Context Sequence Memory Network	CSMN
علاقة شجرة التبعية	Dependency Tree Relation	DTR
دراسة التقييم ثنائية اللغة	Bilingual Evaluation Understudy	BLEU
تقييم الانحراف الجوهري	Recall-Oriented Understudy for Gisting Evaluation	ROUGE
مقياس تقييم الترجمة بالترتيب الصريح	Metric for Evaluation of Translation with Explicit ORdering	METEOR
تقييم توصيف الصور بالإجماع	Consensus-based Image Description Evaluation	CIDEr

TF-IDF	Term Frequency-Inverse Document Frequency	تكرار الكلمة ومعكوس تكرار الوثيقة
SPICE	Semantic Propositional Image Caption Evaluation	تقييم توصيف الصور بالاقتراح الدلالي
VQA	Visual Question Answering	الإجابة على الأسئلة الصورية
RPN	Region Proposal Network	شبكة اقتراح المساحات
IoU	Intersection over Union	تقاطع على الاجتماع
CNN	Convolutional Neural Network	شبكة عصبونية تلفيفية

الفصل الأول: مقدمة عامة

1.1 تمهيد

يعد التوصيف الآلي للصور إحدى المسائل الشائعة في الذكاء الصنعي الحديث (AI)، حيث يهتم بتوليد نص يصف صورة دخل ويمكن أن يكون الناتج جملة واحدة أو أكثر (الشكل 1.1).. يتبع توصيف الصور آلياً لمجالي الرؤية الحاسوبية ومعالجة اللغات الطبيعية. فيما سبق، كانت مسألة توصيف الصور تعالج تقليدياً باستخدام تقنيات التعلم الآلي التقليدية، أما مؤخراً فقد اكتسبت تقنيات التعلم العميق شعبية أكبر لمثل هذه التطبيقات.



A skier performing a jump against some snow

الشكل 1.1. مثال عن صورة مع توصيفها

يعد توصيف الصور آلياً مكوناً هاماً في العديد من التطبيقات، حيث يمكن استخدامه على سبيل المثال للفهرسة التلقائية للصور وهي عملية هامة في استرجاع الصور باعتماد المحتوى Content-Based Image Retrieval (CBIR)، ومن ثم يمكن تطبيقه في العديد من المجالات، بما في ذلك الطب الحيوي، والتجارة، والتطبيقات العسكرية، والتعليم، والمكتبات الرقمية، والبحث في الوب. يمكن لمنصات التواصل الاجتماعي مثل Facebook و Twitter إنشاء توصيفات مباشرة من الصور، ويمكن أن تشمل التوصيفات مكان وجودنا (على سبيل المثال: شاطئ، مقهى) وما نرتديه، والأهم من ذلك ما نفعله في هذه الصور.

يعدُّ استخدام بنية Encoder-Decoder من أكثر الطرق نجاحاً لحل مسألة توصيف الصور [1]، حيث يقوم هذا الأسلوب على ترميز الصور وفق تمثيل عالي المستوى ثم يقوم بفك ترميز هذا التمثيل باستخدام نموذج توليدٍ لغوي، مثل وحدة Long Short-Term Memory (LSTM) أو وحدة Gated Recurrent Unit (GRU) أو أحد أنواعهما.

أثبتت آلية الانتباه attention فعاليتها في تطبيقات ترجمة السلاسل إلى سلاسل أخرى، وخصوصاً في توصيف الصور والترجمة الآلية، حيث أثبتت أنها قادرة على زيادة الدقة من خلال جعل النموذج يركز على الأجزاء المهمة من الدخل عند كل خطوة من إنشاء سلسلة الخرج.

لفهم صورة ما، استخدمت العديد من نماذج التعلم العميق الحديثة شبكات CNN مدربة مسبقاً لاستخراج مصفوفات السمات من الطبقات التلافيفية الأخيرة. ساعد هذا في فهم العديد من جوانب الأغراض في الصورة والعلاقات فيما بينها وتمثيل الصورة على مستوى أعلى.

حاول بعض الباحثين في الأبحاث الأخيرة استخدام سمات الأغراض في توصيف الصور [1] [2] [3]، فكان من بين نماذج كشف الأغراض المستخدمة نماذج YOLOv3 و YOLOv4 و YOLO9000 المعروفة بسرعتها ودقتها وفعاليتها لتطبيقات الزمن الحقيقي. عادةً ما تكون سمات الأغراض مجموعة توصيفات للأغراض في الصورة، حيث يتضمن كل تمثيل لغرض معلومات المربع المحيط به وصنف الغرض ومستوى الثقة بصنف الغرض المتنبأ به.

2.1 إشكالية ودوافع البحث

تتناول معظم الأبحاث الأخيرة في نظم توصيف الصور التي تعتمد على التعلم العميق بنىً كثيرٌ منها فعال جداً في توصيف الصور وتوليد النصوص لكن الأعمال قليلة في الجزء من النماذج الذي يعنى بفهم الصورة واستخراج السمات التي تفيد حقاً في توصيفها التوصيف القريب من توصيف البشر. يهدف هذا البحث إلى تطوير نظم توصيف الصور، وخصوصاً في جزئية فهم الصورة واستخلاص سمات فعالة وذات معنى لنظم توصيف الصور الآلي. من هذا الهدف، استخلصنا الأسئلة الآتية لتوجيه البحث:

- ما هي الأساليب المتبعة في نظم التعلم العميق من أجل فهم الصورة وتوصيفها؟
- ما أهمية مكون فهم الصورة واستخراج السمات منها في نظام التوصيف؟ وكيف يمكن تطويره؟
- هل تعمل أساليب فهم الصورة بالكفاءة ذاتها على جميع مجموعات الصور؟

3.1 فكرة الحل المقترح

نبني في هذا البحث نموذجاً لتوصيف الصور معتمداً على التعلم العميق وآلية الانتباه attention، ونهتم فيه بجزئية فهم الصورة وتحليلها. يتكون النموذج من جزأين أساسيين: جزء تحليل الصورة واستخراج السمات منها Image Encoder، وجزء التوليد اللغوي لتوصيف الصورة بالاعتماد على تلك السمات Language Decoder. نطور في جزء فهم الصورة طريقة جديدة لاستخراج السمات تعتمد على استخراج معلومات الأغراض من الصورة بنموذج مستقل مدرب مسبقاً، واستخراج سمات بمستوىً عالٍ من التجريد بنموذج CNN مدرب مسبقاً أيضاً واستخدام نوعي السمات لتحقيق زيادة في جودة التوصيف.

4.1 المساهمات الأساسية في البحث

نلخص المساهمات الأساسية المقدمة في هذا البحث بالنقاط الآتية:

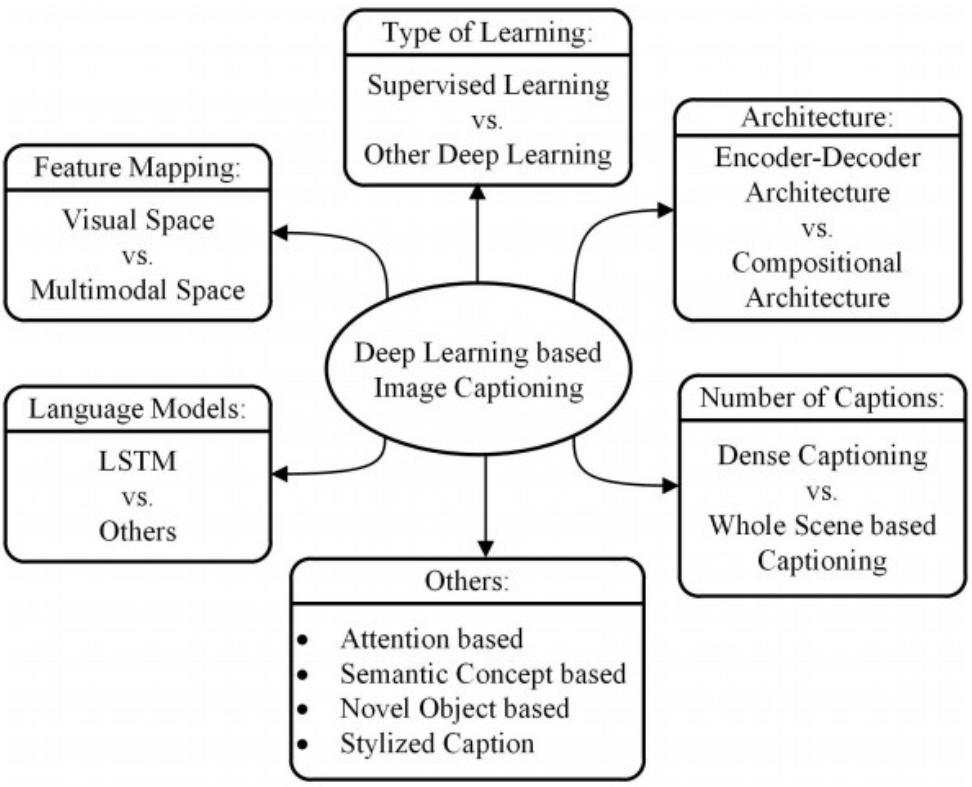
- تطوير أسلوب جديد لفهم الصورة واستخراج السمات المناسبة لتوصيفها.
- إجراء تجارب موسعة غير مسبوقة في الأبحاث السابقة تبين الترابط بين اختيار بنية نموذج CNN وطبيعة الصور عند استخراج سماتها في نظام توصيف الصور.
- ابتكار أسلوب جديد لترتيب سمات الأغراض المكتشفة في الصورة Positional Encoding، وإظهار تأثيره على معايير التقييم.

5.1 مخطط البحث

من أجل الإجابة على أسئلة البحث السابقة، قسمنا محتوى الأطروحة على خمسة فصول. يقدم الفصل الثاني دراسة مرجعية حول نظم التوصيف الآلي للصور باستخدام تقنيات التعلم العميق، حيث نجيب على السؤال الأول للبحث. يعرض الفصل الثالث النماذج المقترحة في هذا البحث بالتفصيل، حيث نجيب على السؤال الثاني للبحث. أما الفصل الرابع فيعرض التجارب التي قمنا بها، التي تجيب على السؤال الثالث للبحث. أخيراً، يسلط الفصل الخامس الضوء على الآفاق المستقبلية لمسار البحث.

الفصل الثاني: الدراسة المرجعية

يعد ظهور مفهوم توصيف الصور آلياً حديثاً نسبياً، وقد ظهرت العديد من المنهجيات المتبعة لأتمتة عملية توصيف الصور التي سايرت تطور منهجيات الذكاء الصنعي من تعلم آلي ثم تعلم عميق. ركز البحث في السنوات الأخيرة في مجال توصيف الصور الآلي على تقنيات التعلم العميق التي تستفيد من الشبكات العصبونية التلافيفية CNNs لاستخراج السمات، واستخدام وحدات توليد لغوي مثل GRU أو LSTM في مرحلة فك الترميز decoding. لقد درس حسين وآخرون [4] هذه المنهجيات ووصفوها وفقاً للمخطط الآتي (الشكل 1.2):



الشكل 1.2. تصنيف شامل لمنهجيات توصيف الصور المعتمد على التعلم العميق.

بما أن البحث ركز في السنوات الأخيرة في مجال توصيف الصور الآلي على أساليب التعلم العميق فقط، (بسبب نتائجها الجيدة نسبياً وتوفر القوة الحسابية)، ستركز هذه الدراسة المرجعية على التعلم العميق.

1.2 مقارنة بين المنهجيات المتعاكسة

فيما يأتي مقارنة لبعض المنهجيات المتعاكسة في التعلم العميق.

1.1.2 الفضاء الصوري Visual Space والفضاء متعدد الوسائط Multimodal Space

يمكن أن تقسم أساليب توصيف الصور المعتمدة على التعلم العميق إلى طرق visual space وطرق multimodal space. من المعروف أن مجموعات توصيف الصور تحتوي التوصيفات المقابلة للصور كنصوص. في الطرق المعتمدة على visual space، تمرر سمات الصورة والتوصيفات المقابلة لها مستقلتين إلى وحدة فك الترميز اللغوي language decoder. لكن في حالة الفضاء متعدد الوسائط multimodal space، يُتعلم فضاء multimodal space من الصور والنصوص المقابلة لها، ثم يمرر هذا التمثيل متعدد الوسائط multimodal إلى وحدة فك الترميز اللغوي language decoder.

1.1.1.2 الفضاء الصوري Visual Space

توجد الكثير من الأعمال حول الفضاء الصوري visual space [5-13]، وتتضمن المنهجيات الآتية: Supervised Learning, Other Deep Learning, Dense Captioning, Captions for the Whole Scene, Encoder-Decoder Achitectures, Compositional Architectures, Attention-Based methods, Semantic Concept-Based mathods, Novel Object-based methods, Stylized Caption methods.

2.1.1.2 الفضاء متعدد الوسائط Multimodal Space

البنيـة التقليديـة لمنهجيـات multimodal space تتضـمن جـزءاً للترمـيز encoder ومكونـاً لفهم الصـورة وجـزءاً للتمثيل في فضاء multimodal و decoder. مكون فهم الصورة يستخدم Convolutional Neural Network لاستخراج سمات الصورة. المرمز encoder يستخرج سمات الكلمات ويتعلم تمثيلاً كثيفاً للتضمين embedding لكل كلمة. بعد هذا يجري إدخال السياق إلى الطبقات المتكررة recurrent. تسقط طريقة multimodal سمات الصورة وسمات النص في فضاء مشترك.

من الأعمـال الحديثـة في هذا المجـال هو عمـل Chen وآخـرين [14]، حيث اقترحوا نموذجاً معتمداً على التمثيل متعدد الوسائط multimodal. تستطيع منهجيتهم توليد توصيفات جديدة من الصور واسترجاع السمات البصرية من النصوص. تستطيع منهجيتهم أيضاً إنشاء إسقاط بين الصورة وخرج نظام التوصيف، لكن لا تستطيع قلب الكلمات إلى تمثيل لسمات الصورة. تتضمن المنهجية طبقة recurrent إضافية مخفية بصرية مع RNN لتوليد الصورة من نص.

2.1.2 التعلم بالإشـراف Supervised Learning وطـرق التعلم العميـق الأخـرى Other Deep Learning

في التعلم بالإشراف supervised learning تأتي بيانات التدريب مع الخرج الصحيح الذي يدعى label. بعكس التعلم بغـير إشـراف unsupervised learning، حيث يُتعامـل مـع بيانـات بـدون label. التعلم بـالتعزيز reinforcement learning هو منهجية تهدف إلى جعل الآلة تكتشف الصنف label بمكافأة إيجابية أو سلبية reward. هنـاك عـدد من منهجيـات توصـيف الصـور الـتي تسـتخدم Generative Adversarial Networks (GANs) الـتي هي من نمـط تعلم unsupervised مـع طريقـة reinforcement learning. تصـنف هـذه المنهجيات مع "الطرق الأخرى من التعلم العميق".

1.2.1.2 منهجيات التعلم بالإشراف Supervised Learning

تتضـمن هـذه المنهجيـات: Encoder–Decoder، Compositional Architecture, Attention–based methods, Semantic concept–based, Stylized captioning, Novel object–based, Dense image captioning.

2.2.1.2 طرق التعلم العميق الأخرى Other Deep Learning

ركز الباحثون مؤخراً على المنهجيات المعتمدة على reinforcement learning و unsupervised learning لتوصيف الصور، لأنه يصعب أحياناً توصيف الصور بدقة لحجوم كبيرة.

تواجه طرق reinforcement learning عدداً من القيود، مثل صعوبة إيجاد تابع قيمة والمعلومات التي تستطيع تحديد الفعل الواجب القيام به انطلاقاً من حالة ما.

منهجيات policy gradient هي نوع من التعلم بالتعزيز reinforcement learning يمكنه اختيار سياسة محددة لسلوك action ما باستخدام طرق optimization و gradient descent، حيث يمكن للسياسة تضمين معرفة في المجـال الخـاص بالسـلوك action ليضـمن تقـارب النمـوذج. إذن تحتـاج منهجيـات policy gradient إلى parameters أقل من منهجيات تابع القيمة.

تستخدم منهجيـات التعلم العميـق الحاليـة أنواعـاً من المرمـزات encoders لاسـتخراج سـمات الصـورة. تـدخل هـذه السـمات إلى مفككـات ترمـيز decoders معتمـدة على الشـبكات العصـبونية لتوليـد التوصـيفات. يحـدث هـذا وفقـاً للخطوات الآتية:

1. تعمل CNN مع RNN لتوليد التوصيفات.
2. تعمل CNN مع RNN أخرى على تقييم التوصيفات وإرسال التقييم إلى الشبكة الأولى بهدف توليد توصيفات بجودة أعلى في المرات القادمة.

قدم Ren وآخرون [15] طريقة جديدة لتوصيف الصور تعتمد على reinforcement learning. تشمل بنية هذه الطريقة شبكتين تتنبآن معاً بالكلمة الآتية الأفضل عند كل خطوة زمنية. تعمل "policy network" كدليل محلي وتساعد على توقع الكلمة الآتية بناءً على الحالة الحالية، وتعمل "value network" كموجه عام وتقيّم المكافأة مع مراعاة جميع الامتدادات الممكنة للوضع الحالي. هذه الآلية قادرة على ضبط الشبكات عند توقع الكلمات الصحيحة، لذلك يمكن أن تولد توصيفات جيدة مشابهة لتوصيفات البشر. تستخدم الطريقة نموذج actor–critic reinforcement learning model لتدريب الشبكة بأكملها، ويُستخدم التضمين الدلالي المرئي لحساب قيمة المكافأة الفعلية في التنبؤ بالكلمة الصحيحة. كما أنه يساعد في قياس التشابه بين الصور والجمل مما يمكنها من تقييم صحة التوصيفات المولدة.

اقترح Rennie وآخرون [16] طريقة أخرى لتوصيف الصور على أساس reinforcement learning، حيث تُستخدم خوارزمية استدلال في وقت الاختبار لعمل normalisation للمكافأة بدلاً من تقدير إشارة المكافأة و normalisation في وقت التدريب. يظهر أن فك الترميز decoding في وقت الاختبار فعال جداً لتوليد توصيفات للصور بجودة عالية.

3.1.2 التوصيف الكثيف Dense Captioning والتوصيف للمشهد بأكمله Captions for the Whole Scene

1.3.1.2 التوصيف الكثيف Dense Captioning

خلافاً لأساليب توصيف الصور الأخرى التي تستخدم مناطق مختلفة من الصورة للحصول على معلومات من الأغراض المختلفة لتوليد توصيف واحد للصورة بأكملها، طريقة توصيف الصور الكثيف تحدد جميع المناطق البارزة للصورة ومن ثم تولد وصفاً لكل منطقة من تلك المناطق.

يتضمن الأسلوب النموذجي لهذه الفئة الخطوات الآتية:

(1) إنشاء مقترحات المناطق المختلفة للصورة المحددة.
(2) استخدام شبكة CNN للحصول على سمات الصورة على أساس المنطقة.
(3) استخدام مخرجات الخطوة 2 في نموذج توليد اللغة لإنشاء توصيف لكل منطقة.

بدأت هذه الطريقة بعمل Johnson وآخرين [17]، حيث اقترحوا بنية شبكة fully convolutional localization network تتكون من شبكة CNN وطبقة تحديد مكاني localisation كثيفة ونموذج لغة LSTM.

2.3.1.2 التوصيف للمشهد بأكمله Captions for the Whole Scene

تتبع كل هذه المنهجيات لتصنيف "الطرق الأخرى المعتمدة على التعلم العميق"، وتتضمن: Encoder–Decoder Architecture, Compositional Architecture, Attention–based, Semantic Concept–based, Stylized Captioning, Novel Object–based Image Captioning, Other Deep Learning Network–based.

4.1.2 بنية الترميز وفك الترميز Encoder–Decoder Architecture والبنية التركيبية Compositional Architecture

1.4.1.2 بنية الترميز وفك الترميز Encoder–Decoder Architecture

تعمل طرق توصيف الصور هذه على شبكات عصبونية بطريقة end–to–end. هذه الأساليب تشبه إلى حد بعيد الترجمة الآلية العصبونية المعتمدة على إطار عمل وحدتي ترميز encoder وفك ترميز decoder. في شبكة كهذه، تُستخرج سمات الصور العامة من خرج الطبقات المخفية لشبكة CNN ثم تدخل في LSTM لإنشاء سلسلة من الكلمات.

يتضمن الأسلوب النموذجي لهذه الفئة الخطوات الآتية:

(1) استخدام CNN للحصول على نوع المشهد واكتشاف الأغراض وعلاقاتها.
(2) استخدام خرج الخطوة 1 باستخدام نموذج لغوي لتحويلها إلى كلمات وعبارات مدمجة تنتج توصيفاً للصورة.

2.4.1.2 توصيف الصور المعتمد على البنية التركيبية Compositional Architecture

تتكون الأساليب القائمة على البنية التركيبية من عدة كتل بناء وظيفية مستقلة: أولاً، تستخدم CNN لاستخراج المفاهيم الدلالية من الصورة، ثم يستخدم النموذج اللغوي لإنشاء مجموعة من التوصيفات المرشحة. عند إنشاء التوصيف النهائي، يعاد ترتيب هذه التوصيفات باستخدام نموذج تشابه عميق multimodal.

يتضمن الأسلوب النموذجي لهذه الفئة الخطوات الآتية:
(1) الحصول على سمات الصورة باستخدام شبكة CNN.
(2) الحصول على المفاهيم المرئية (مثل خصائص الأغراض) من السمات المرئية.
(3) إنشاء توصيفات متعددة بنموذج لغوي باستخدام معلومات الخطوتين 1 و 2.
(4) إعادة ترتيب التوصيفات التي أنشئت باستخدام نموذج تشابه عميق multimodal لاختيار توصيفات عالية الجودة للصورة.

اقترح Ma وآخرون [18] طريقة توصيف صور تركيبية معتمدة على الشبكات. تستخدم هذه الطريقة كلمات هيكلية لتوليد توصيفات ذات مغزى. كما تستخدم أيضاً طريقة متعددة المهام مشابهة لطريقة التعلم متعدد الحالات، وطريقة تحسين متعددة الطبقات لتوليد كلمات هيكلية، ثم تستخدم طريقة الترجمة الآلية القائمة على وحدة فك الترميز LSTM decoder لترجمة الكلمات الهيكلية إلى توصيفات للصور.

اقترح Wang وآخرون [19] بنية RNN–LSTM متوازية اندماجية لتوليد توصيفات للصور. تقسم بنية هذه الطريقة الوحدات المخفية في RNN و LSTM إلى عدد من الأجزاء ذات الحجم نفسه، وتعمل الأجزاء بالتوازي بنسب متقابلة لإنشاء توصيفات للصور.

5.1.2 تصنيفات أخرى

1.5.1.2 توصيف الصور المعتمد على الانتباه Attention-based Image Captioning

استخدمت الأساليب القائمة على وحدة فك الترميز ووحدة الترميز العصبونية Neural encoder-decoder بصورة أساسية في الترجمة الآلية. باتباع هذا التوجه استُخدمت هذه المنهجية أيضاً في مسألة توصيف الصور ووجدت فعالةً جداً. عند توصيف صورة، تستخدم CNN كمرمز encoder لاستخراج السمات المرئية من صورة الدخل وتستخدم RNN كوحدة فك ترميز decoder لتحويل هذا التمثيل كلمة بكلمة إلى وصف بلغة طبيعية للصورة. ومع ذلك، فإن هذه الأساليب غير قادرة على تحليل الصورة بمرور الوقت في أثناء إنشاء توصيف للصورة. لا تأخذ هذه الأساليب في الاعتبار الجوانب المكانية للصورة ذات الصلة بأجزاء توصيفاتها، وإنما تنشئ توصيفات تنظر إلى المشهد ككل. أصبحت الآليات القائمة على الانتباه شائعة بطريقة متزايدة في التعلم العميق لأنها تستطيع معالجة هذه القيود، حيث يمكنها التركيز ديناميكياً على الأجزاء المختلفة من صورة الدخل في أثناء إنتاج سلسلة الخرج. يتمثل الاختلاف الأساسي بين الأساليب القائمة على الانتباه attention والطرق الأخرى في أنها تستطيع أن تركز على الأجزاء البارزة من الصورة وتوليد الكلمات المقابلة في الوقت نفسه.

الطريقة النموذجية لهذه الفئة تعتمد الخطوات الآتية:

(1) يُحصل على معلومات الصورة بناءً على المشهد بأكمله باستخدام شبكة CNN.

(2) يولد جزء التوليد اللغوي كلمات أو عبارات بناءً على مخرجات الخطوة 1.

(3) يركز مكون الانتباه على المناطق البارزة للصورة المعينة عند كل خطوة زمنية لنموذج توليد اللغة بناءً على الكلمات أو العبارات التي أنشئت حتى ذلك الوقت.

(4) تُحدّث التوصيفات ديناميكياً حتى نهاية عمل نموذج توليد اللغة.

كان Xu وآخرون [20] أول من أدخل طريقة توصيف الصور القائمة على الانتباه، وتصف هذه الطريقة المحتويات البارزة في الصورة تِلْقَائِيّاً. يطبق في هذه الطريقة أسلوبان مختلفان: الانتباه القاسي العشوائي stochastic hard attention والانتباه الحتمي الناعم deterministic soft attention لتوليد مصفوفة الانتباه. تستخدم معظم الأساليب المستندة إلى CNN الطبقة العليا من ConvNet لاستخراج معلومات الأغراض البارزة من الصورة. من

عيوب هذه الأساليب أنها قد تفقد معلومات معينة مفيدة لإنشاء توصيفات مفصلة. من أجل الحفاظ على المعلومات، تستخدم طريقة الانتباه السمات من الطبقة التلافيفية convolutional الأخيرة بدلاً من طبقة كاملة التوصيل fully connected layer.

اقترح Jin وآخرون [21] طريقة أخرى قائمة على الانتباه لتوصيف الصور. هذه الطريقة قادرة على استخراج المعنى المجرد بناءً على العلاقة الدلالية بين المعلومات المرئية والمعلومات النصية، ويمكنها أيضاً الحصول على معلومات دلالية ذات مستوى أعلى من خلال اقتراح سياق خاص بالمشهد. يتمثل الاختلاف الأساسي بين هذه الطريقة والطرق الأخرى القائمة على الانتباه في أنها تقدم مناطق مرئية متعددة للصورة بقياسات متعددة ويمكن لهذه التقنية استخراج المعلومات المرئية المناسبة للأغراض في الصورة. لاستخراج السياق الخاص بالمشهد، يستخدم أولاً Latent Dirichlet Alocation (LDA) لإنشاء قاموس من جميع التوصيفات في مجموعة البيانات. يستخدم عصبون متعدد الطبقات للتنبؤ بشعاع الموضوع لكل صورة ثم تستخدم وحدة LSTM موجهة بالمشهد تحوي طبقتين متتاليتين لإنشاء وصف للسياق العام للصورة.

اقترح Wu وآخرون [22]. طريقة الانتباه القائمة على المراجعة لتوصيف الصور. نموذج المراجعة هذا يستطيع تنفيذ خطوات مراجعة متعددة مع الانتباه إلى حالات CNN المخفية. ناتج CNN هو عدد من أشعة الحقائق التي تستطيع التعبير عن الحقائق الإجمالية للصورة. تعطى هذه الأشعة كمدخلات في آلية الانتباه الخاصة ب LSTM. على سبيل المثال، يمكن لوحدة المراجعة مراجعة ما يأتي أولاً: ما هي الأغراض الموجودة في الصورة؟ ثم تستطيع مراجعة المواضع النسبية للأغراض، ويمكن لمراجعة أخرى استخراج معلومات السياق العام للصورة. تمرر هذه المعلومات إلى وحدة فك الترميز لإنشاء توصيفات للصور.

اقترح Pedersoli وآخرون [23]. آلية انتباه تعتمد على المناطق لتوصيف الصور. تسقط الأساليب السابقة القائمة على الانتباه مناطق الصورة فقط إلى حالة نموذج لغة RNN، أما هذا الأسلوب فيربط مناطق الصورة بكلمات التوصيف المعطاة لحالة RNN ويمكنه التنبؤ بكلمة التوصيف الآتية ومنطقة الصورة المقابلة في كل خطوة زمنية من RNN. كما أنه قادر على التنبؤ بالكلمة التالية بالإضافة إلى مناطق الصورة المقابلة في كل خطوة زمنية من RNN لتوليد توصيفات الصور. من أجل تحديد مساحات الانتباه تستخدم طرق توصيف الصور السابقة القائمة على الانتباه إما موضع شبكة تفعيل Activation CNN أو مقترحات الأغراض. في المقابل، تستخدم هذه الطريقة محولاً convolutional transformer قابلاً للتدريب من البداية إلى النهاية إلى جانب طرق تنشيط CNN وطرق

اقتراح الأغراض. تساعد مجموعة من هذه التقنيات هذه الطريقة في حساب مجالات الانتباه التكيفية للصورة. في التجارب تُظهر هذه الطريقة أن آلية الانتباه الجديدة هذه مع شبكة المحولات المكانية يمكن أن تنتج توصيفات للصور عالية الجودة.

اقترح Lu وآخرون [24] طريقة أخرى قائمة على الانتباه لتوصيف الصور. تعتمد هذه الطريقة على نموذج الانتباه التكيفي مع راصد بصري. تركز طرق التوصيف الحالية للصور القائمة على الانتباه على الصورة في كل خطوة زمنية من RNN لكن هناك بعض الكلمات أو العبارات (على سبيل المثال: a, of). التي لا تحتاج إلى الانتباه إلى الإشارات المرئية. بالإضافة لذلك، يمكن أن تؤثر هذه الإشارات المرئية غير الضرورية على عملية إنشاء التوصيفات وتؤدي إلى تدهور الأداء العام. لذلك يمكن أن تحدد طريقتهم المقترحة متى ستركز على منطقة الصورة ومتى ستركز فقط على نموذج توليد اللغة. بمجرد أن يقرر النظر إلى الصورة، يجب عليه اختيار الموقع المكاني للصورة. تتمثل المساهمة الأولى لهذه الطريقة في تقديم طريقة انتباه مكاني جديدة يمكنها حساب السمات المكانية من الصورة. ثم في أسلوب الانتباه التكيفي، قاموا بإدخال امتداد LSTM جديد. بصورة عامة، تعمل LSTM كوحدة فك ترميز يمكنها إنتاج حالة مخفية في كل خطوة زمنية، لكن هذا الامتداد قادر على إنتاج راصد بصري إضافي يوفر خياراً احتياطياً لوحدة فك الترميز. كما أن لديها بوابة تستطيع التحكم في مقدار المعلومات التي ستحصل عليها وحدة فك الترميز من الصورة.

اقترح Liu وآخرون [25] طريقة لتوصيف الصور يمكنها تقييم وتصحيح خريطة الانتباه في كل خطوة زمنية. من المنطقي إنشاء توصيل متوافق بين مناطق الصورة والكلمات التي أنشئت ومن أجل تحقيق هذا قدمت هذه الطريقة مقياس تقييم كمّيّاً لحساب خرائط الانتباه. وهي تستخدم مجموعة بيانات Flickr30k entities ومجموعة بيانات MSCOCO لقياس كل من خريطة الانتباه بالحقيقة الأساسية والتسميات الدلالية لمناطق الصورة. من أجل تعلم وظيفة الانتباه بطريقة أفضل، اقترح نموذج الانتباه الخاضع للإشراف حيث يُستخدم نوعان من نماذج الانتباه الخاضع للإشراف هنا: الإشراف القوي مع معلومات المحاذاة strong supervision with alignment annotation والإشراف الضعيف مع وضع العلامات الدلالية weak supervision with semantic labeling. في الإشراف القوي مع معلومات المحاذاة، يمكن تعيين كلمة حقيقة أساسية مباشرةً في المنطقة. ومع ذلك، فإن محاذاة الحقيقة الأساسية ليست ممكنة دائماً لأن جمع البيانات والتعليق عليها غالباً ما يكون مكلفاً للغاية. يجرى إشراف ضعيف باستخدام المربعات المحيطة أو أقنعة التجزئة في مجموعة بيانات MSCOCO. تُظهر

الطريقة في التجارب أن نموذج الانتباه الخاضع للإشراف له أداء أفضل في تحديد الانتباه بالإضافة إلى توصيف الصورة.

اقترح Chen وآخرون [26] طريقة أخرى قائمة على الانتباه لتوصيف الصور. تأخذ هذه الطريقة في الاعتبار كلاً من الانتباه المكاني والقناة لحساب خريطة الانتباه. لا تراعي طرق التوصيف الحالية للصور المعتمدة على الانتباه سوى المعلومات المكانية لإنشاء خريطة الانتباه. من العيوب الشائعة لطرق الانتباه المكاني هذه أنها تحسب التجميع الموزون فقط على خريطة السمات الانتباهية لذلك تفقد هذه الأساليب المعلومات المكانية تدريجياً. علاوة على ذلك، هي تستخدم المعلومات المكانية فقط من آخر طبقة convolutional في شبكة CNN. مناطق الاستقبال لهذه الطبقة كبيرة جِدّاً بحيث تجعل الفجوة محدودة بين المناطق، لذلك لا تحصل على اهتمام مكاني كبير للصورة. ومع ذلك في هذه الطريقة تستخرج سمات CNN ليس فقط من المواقع المكانية ولكن أيضاً من قنوات مختلفة وطبقات متعددة لذلك تحصل على اهتمام مكاني كبير. بالإضافة إلى هذا، في هذه الطريقة يعمل كل مرشح لطبقة convolutional ككاشف دلالي بينما تستخدم طرق أخرى مصادر خارجية للحصول على معلومات دلالية.

قدم Tavakoli وآخرون [27] طريقة قائمة على الانتباه لتوصيف الصور، معتمدة على الأماكن البارزة في الصورة من القاعدة إلى القمة يمكن أن تستفيد من المقارنات مع طرق توصيف الصور الأخرى القائمة على الانتباه. ووجدوا أن البشر يصفون الأشياء الأكثر أهمية أولاً بدلاً من الأشياء الأقل أهمية. كما يوضح عملهم أيضاً أن الطريقة تعمل أفضل على البيانات من خارج مجموعة التدريب.

طبقت طريقة Anderson وآخرين [28] في كل من المقاربتين من الأعلى إلى الأسفل ومن الأسفل إلى الأعلى، وتستخدم Faster R-CNN في آلية الانتباه التصاعدي لمقترحات المناطق لتحديد المناطق البارزة في الصورة. لذلك يمكن أن تنتبه هذه الطريقة لمناطق على مستوى الغرض بالإضافة إلى مناطق الصور البارزة الأخرى.

قدم Park وآخرون [29] نوعاً مختلفاً من طرق توصيف الصور المعتمد على الانتباه. يمكن أن تولد هذه الطريقة توصيفات للصور تتناول المواضيع الشخصية في الصورة وهي تهتم بصورة أساسية بمهمتين: التنبؤ بالوسم hashtag وتوليد المنشورات. تستخدم هذه الطريقة Context Sequence Memory Network (CSMN) للحصول على معلومات السياق من الصورة. يحوي وصف صورة من المنظور الشخصي الكثير من التطبيقات في شبكات التواصل الاجتماعي لكن وصفها ليس بالأمر السهل لأنه يتطلب موضوعاً وعاطفةً وسياقاً للصورة. لذلك،

تأخذ الطريقة بالاعتبار المعرفة السابقة حول مفردات المستخدم أو أنماط الكتابة من المستندات السابقة لإنشاء توصيفات الصور. من أجل العمل مع هذا النوع الجديد من توصيفات الصور فإن طريقة CSMN لها ثلاث مساهمات: أولاً، يمكن أن تعمل ذاكرة هذه الشبكة كمستودع وتحتفظ بأنواع متعددة من معلومات السياق. ثانياً، صممت الذاكرة بطريقة تمكنها من تخزين جميع الكلمات التي أنشئت مسبقاً تسلسلياً. بسبب ذلك لا تعاني هذه الطريقة من مشكلة vanishing gradient. ثالثاً، يمكن لشبكة CNN المقترحة أن ترتبط بوصلات ذاكرة متعددة تساعد في فهم المفاهيم السياقية.

استخدم Sugano وآخرون [30] طريقة تستعمل معلومات النظرة البشرية مع آلية الانتباه للشبكات العصبونية العميقة في إنشاء توصيفات للصور. تدمج هذه الطريقة معلومات النظرة البشرية في نموذج LSTM القائم على الانتباه، واستخدموا مجموعة بيانات SALICON لإجراء التجارب وحققت نتائج جيدة.

2.5.1.2 توصيف الصور المعتمد على المفهوم الدلالي Semantic Concept-Based Image Captioning

تنتبه الأساليب المعتمدة على المفهوم الدلالي انتقائياً لمجموعة من مقترحات المفاهيم الدلالية المستخرجة من الصورة. بعد ذلك تدمج هذه المفاهيم في حالات مخفية ومخرجات RNN.

تتبع الطرق في هذه الفئة الخطوات العامة الآتية:

(1) يستخدم برنامج الترميز encoder المستند إلى CNN لترميز سمات الصورة والمفاهيم الدلالية.

(2) تدخل سمات الصورة في مدخلات نموذج توليد اللغة.

(3) تضاف المفاهيم الدلالية إلى الحالات المخفية المختلفة لنموذج اللغة.

(4) ينتج جزء تكوين اللغة توصيفات ذات مفاهيم دلالية.

وسع Karpathy وآخرون [31] طريقتهم في عمل آخر، حيث يمكن لطريقتهم الجديدة إنشاء توصيفات بلغة طبيعية للصور وكذلك لمناطقها الجزئية. تستخدم هذه الطريقة مزيجاً جديداً من شبكات CNN على مناطق الصورة، وشبكات عصبونية متكررة ثنائية الاتجاه Bidirectional Recurrent Neural Networks على الجمل،

وتضميناً multimodal مشتركاً يربط بين الطريقتين. كما يوضح أيضاً بنية شبكة عصبونية متعددة الوسائط متكررة multimodal recurrent neural network تستخدم المحاذاة الناتجة لتدريب النموذج على إنشاء توصيفات جديدة لمناطق الصورة. في هذه الطريقة تُستخدم علاقات شجرة التبعية Dependency Tree Relations (DTR) للتدريب على ربط مقاطع الجملة بمناطق الصورة التي تحوي إطار سياق ثابت. تعتبر الأجزاء المتجاورة من الجمل متحاذية في فضاء التضمين وتكون أكثر وضوحاً وقابلية للتفسير وغير ثابتة الطول.

طريقة Yao وآخرين [32] لها بنىً مختلفة لدمج السمات مع تمثيلات الصور. الأساس أن يقدم نوعان من التمثيلات البنيوية هنا. في المجموعة الأولى، تدخل فقط السمات إلى LSTM أو تمثيلات الصور إلى LSTM أولاً ثم السمات والعكس صحيح. في المجموعة الثانية، يمكن التحكم بالخطوة الزمنية لوحدة LSTM. يقرر إذا كان سيدخل تمثيل الصورة والسمات مرة واحدة أو عند كل خطوة زمنية. جربت هذه الأنواع من البنى على مجموعة بيانات MSCOCO ومقاييس التقييم الشائعة.

3.5.1.2 توصيف الصور القادر على التعرف على الأغراض الجديدة Novel Object-based Image Captioning

مع أن أساليب توصيف الصور القائمة على التعلم العميق قد حققت نتائج واعدة، إلا أنها تعتمد إلى حد كبير على مجموعات بيانات الصور المقرونة بالتوصيفات. يمكن لهذا النوع من الأساليب فقط إنشاء توصيف للأغراض التي هي داخل سياق التوصيفات، لذلك تتطلب هذه الأساليب مجموعة كبيرة من أزواج التدريب بين الصور والجمل. يمكن أن تولد أساليب توصيف الصور المستندة إلى الأغراض الجديدة توصيفات لأغراض جديدة غير موجودة في مجموعات بيانات توصيفات الصور المقترنة.

تتبع الطرق في هذه الفئة الخطوات الآتية:

(1) تدريب المصنف المعجمي المنفصل ونموذج اللغة على بيانات صور غير مقترنة وبيانات نصية غير مقترنة.

(2) تدريب نموذج توليد التوصيفات العميق على بيانات التوصيفات المقترنة للصورة.

(3) الجمع بين كلا النموذجين معاً للتدريب بطريقة مشتركة مما يؤدي إلى إنشاء توصيفات للأغراض الجديدة.

اقترح Yao وآخرون [33]. آلية نسخ لتوليد توصيفات للأغراض الجديدة. تستخدم هذه الطريقة مجموعة بيانات منفصلة للتعرف على الأغراض لتطوير المصنفات للأغراض الجديدة. تدمج هذه الطريقة الكلمات المناسبة في التوصيفات الناتجة باستخدام وحدة فك ترميز RNN مع آلية النسخ. تضيف بنية الطريقة هذه شبكة جديدة للتعرف على الأغراض غير المعروفة مسبقاً من الصور غير المقترنة ودمجها مع وحدة LSTM لإنشاء التوصيفات.

4.5.1.2 التوصيف ذو النص المنمط Stylized Captioning

تُنشئ أنظمة توصيف الصور الحالية توصيفات تستند فقط إلى محتوى الصورة ويمكن أن نقول أنها واقية، لكنها لا تنظر إلى أسلوب كتابتها بطريقة منفصلة. لكن يمكن أن تكون التوصيفات ذات الأسلوب أكثر تعبيراً وجاذبية من الوصف المجرد للصورة.

تتبع طرق هذه الفئة الخطوات العامة الآتية:

(1) يستخدم مرمز الصور المستند إلى CNN للحصول على معلومات الصورة.

(2) تجهز مجموعة نصية منفصلة لاستخراج مفاهيم التوصيف المنمط المختلفة (مثلاً: رومانسي، فكاهي) من بيانات التدريب.

(3) ينشئ جزء التوليد اللغوي توصيفات مبسطة وجذابة باستخدام معلومات الخطوة 1 والخطوة 2.

اقترح Gan وآخرون نظاماً جديداً لتوصيف الصور يسمى StyleNet. يمكن أن تولد طريقتهم توصيفات جذابة تستخدم أنماطاً مختلفة. تتكون بنية هذه الطريقة من CNN و LSTM معدلة تستطيع فصل الحقيقة عن الأسلوب في كتابة التوصيفات. يستخدم تدريب متعدد المهام من السلاسل لتحديد عوامل النمط ثم إضافة هذه العوامل في وقت التشغيل لإنشاء توصيفات جذابة. من المثير للاهتمام أنه يستخدم مجموعة لغة خارجية أحادية اللغة للتدريب بدلاً من الصور المقترنة بالتوصيفات ومع ذلك فإنه يستخدم مجموعة بيانات جديدة لتوصيف الصور تسمى FlickrStyle10k ويمكنها إنشاء توصيفات بأنماط مختلفة.

في محادثاتنا اليومية، والاتصالات، والعلاقات الشخصية، واتخاذ القرار نستخدم العديد من التعبيرات النمطية غير الواقعية مثل العواطف والفخر والعار. ادعى Mathews وآخرون أن التوصيفات التلقائية للصور تفتقد هذا الجانب

غير الواقعي، لذلك اقترحوا طريقة تسمى SentiCap. يمكن أن تولد هذه الطريقة توصيفات للصور تحوي مشاعراً إيجابية أو سلبية. يقدم هذا نموذجاً جديداً يستخدم switching RNN يجمع بين جزئي CNN + RNNs يعملان بالتوازي. في كل خطوة زمنية يولد نموذج التبديل هذا احتمال التبديل بين وحدتي RNN تقوم إحداهما بإنشاء توصيفات مستخدمةً كلمات واقعية والأخرى تستخدم كلمات بمشاعر، ثم تؤخذ مدخلات من الحالات المخفية لكل من وحدات RNN لتوليد التوصيفات. يمكن أن تؤدي هذه الطريقة إلى إنشاء توصيفات للصور بنجاح في ضوء المشاعر المناسبة.

2.2 دراسات لمنهجيات استخراج السمات في توصيف الصور الآلي

1.2.2 التوصيف الآلي للصور بالاستعانة بمعلومات أغراض الصورة

اقترح Yin & Ordonez [3] نموذجاً لترجمة السلاسل إلى سلاسل تقوم فيه شبكة LSTM بترميز سلسلة من الأغراض ومواضعها كسلسلة دخل، ويقوم نموذج LSTM بفك ترميز هذا التمثيل لإنشاء توصيفات للصور. يستخدم نموذجهم نموذج YOLO لكشف الأغراض واستخراج معلومات الأغراض من الصور (فئات الأغراض ومواقعها) وزيادة دقة التوصيفات. اقترحوا أيضاً في عملهم نموذجاً يستخدم نموذج تصنيف الصور VGG مدرباً مسبقاً على ImageNet لاستخراج السمات المرئية. يأخذ المرمز encoder في كل خطوة زمنية كدخل زوجاً من البيانات، فئة الغرض (مرمزة بطريقة one–hot vector)، وشعاع الموقع الذي يتضمن إحداثيتي X و Y، وعرض وارتفاع المربع المحيط بالغرض معالجة بأسلوب normalisation على النطاق [0,1] بالنسبة لأبعاد الصورة المدخلة. درّب النموذج باستخدام خوارزمية backpropagation لكن بدون نشر مقدار الخطأ إلى نموذج كشف الأغراض. تظهر دراستهم أن نموذجهم زاد في الدقة عند دمجه مع وحدتي CNN و YOLO.

طور Vo–Ho وآخرون [2]. نظام توصيف صور يستخرج سمات الأغراض من YOLO9000 و Faster R–CNN. يعالج كل نوع من السمات بوحدة انتباه attention لإنتاج سمات محلية تمثل الجزء الذي يركز عليه النموذج حَالِيّاً. يدمج نوعا السمات ويدخلان في وحدة LSTM لتوليد احتمالات للكلمات الموجودة في قائمة المفردات المحددة عند كل خطوة زمنية، وتُستخدم استراتيجية البحث الشعاعي beam search لمعالجة النتائج من

أجل اختيار أفضل توصيف. لاستخراج السمات من الصور استخدموا شبكة ResNet CNN. من أجل كل صورة دخل قاموا أولاً باستخراج قائمة من سمات الأغراض باستخدام YOLO9000، مع الاحتفاظ فقط بالسمات العشرين ذات نسبة الثقة الأعلى، ثم قسموا كل سمة إلى كلمات وأزالوا الكلمات المتكررة بحيث تحتوي القائمة كلمات فريدة فقط. تمثل كل كلمة i بما في ذلك الرمز المميز <NULL>. بشعاع حسب تمثيل one-hot vector طوله عدد المفردات. بعد ذلك ضمّنوا كل كلمة في فضاء متعدد الأبعاد باستخدام طريقة word embedding، واستخدموا وحدات LSTM للتوليد اللغوي.

اقترح Lanzendörfer وآخرون [34] نموذجاً للإجابة على الأسئلة من الصورة Visual Question Answering (VQA) استناداً إلى نظام iBOWIMG. يستخرج النموذج السمات من شبكة Inception V3 بالإضافة إلى سمات الأغراض المستخرجة من نموذج YOLO لكشف الأغراض، ويستخدم نموذجهم أسلوب الانتباه attention. ترمز مخرجات YOLO كأشعة من قياس (80 × 1). لإعطاء المزيد من السمات التي تحمل معلومات إلى نموذج iBOWIMG، ويحوي كل عمود عدد الأغراض المكتشفة من النوع الذي يوافقه. تولد ثلاثة أشعة ترميز للأغراض لكل من مجالات ثقة الكشف البالغة 25%. و 50%. و 75%، ثم تضم بسمات الصورة وسمات الإجابة الآلية على الأسئلة.

قدم Herdade وآخرون [35]. نموذج Encoder-Decoder يدمج بطريقة صريحة معلومات العلاقات المكانية بين الأغراض المكتشفة من خلال الانتباه الهندسي geometric attention. استخدموا أولاً نموذجاً لكشف الأغراض لاستخراج السمات المرئية والهندسية من جميع الأغراض المكتشفة في الصورة، ثم استخدموا نموذج تحويل لعلاقات الأغراض Object Relation Transformer لإنشاء نصوص توصيفات الصور. استخدموا نموذج Faster R-CNN مع ResNet-101 كنماذج CNN الأساسية لكشف الأغراض واستخراج السمات. تنشئ شبكة Region Proposal Network (RPN) صناديق تحديد الأغراض المكتشفة باستخدام خرائط سمات من ResNet-101 كمدخلات. يجري تجاهل الصناديق المحيطة المتداخلة باستخدام تقاطع Intersection-over-Union (IoU) عند عتبة 0.7 باستخدام non-maximum suppression. تجاهلوا أيضاً جميع المربعات المحيطة التي لها ثقة تنبؤ أقل من عتبة 0.2 ثم طبقوا mean-pooling على الصورة لإنشاء شعاع سمات من 2048 بعداً لكل مربع تحديد لغرض. بعد ذلك تدخل أشعة السمات هذه في نموذج Transformer.

درس Wang وآخرون [36]. توصيف الصور بطريقة end–to–end باستخدام تمثيلات قابلة للتفسير بدرجة عالية حصلوا عليها من نماذج كشف الأغراض في الصور. لقد أجروا دراسة مفصلة لفعالية عدد من السمات المستندة إلى كشف الأغراض بهدف توصيف الصور، ووجدوا أن معلومات عدد مرات التكرار للغرض وحجمه وموقعه مفيدة وتكمل دقة التوصيفات الناتجة. اكتشفوا أيضًا أن لفئات معينة من الأغراض تأثيراً أكبر من غيرها على توصيف الصور الآلي.

اقترح Sharif وآخرون [1]. الاستفادة من العلاقات اللغوية بين الأغراض الموجودة في الصور لزيادة جودة توصيف الصور، حيث عملوا على الاستفادة من "تضمين الكلمات" لاستخراج دلالات الكلمات واستخدام الترابط الدلالي للأغراض. يرمز النموذج المقترح التقارب المكاني والدلالي لأزواج الكائنات في تضمين العلاقات التي تستخدم معلومات العلاقة اللغوية. كما أنها تستخدم NASNet لاستخراج المعلومات الدلالية العامة للصورة. بهذه الطريقة يمكن تعلم العلاقات الدلالية الحقيقية غير الظاهرة في المحتوى المرئي للصورة، مما يسمح لوحدة فك الترميز decoder بالتركيز على أهم علاقات الأغراض والسمات المرئية ويعطي توصيفات ذات مغزى أكثر.

قام فاريش Variš وآخرون [37]. بالبحث في فضاء التضمين المشترك بين النماذج النصية والصورية. تناول بحثهم فضاءات تمثيل جديدة للنصوص والصور، وخلصت نتيجتهم إلى أن السمات الصورية المتنبأ بها والكلمات المعبرة عن صنف الغرض يمكن أن تكون في فضاء التمثيل ذاته.

2.2.2 مقارنة نماذج CNN لاستخدامها في استخراج السمات

أجرى Holliday & Dudek [38] تقييماً واسع النطاق لشبكات CNN كوحدات استخراج سمات لمقارنة السمات المرئية في بيئة تحوي تغييرات كبيرة في المظهر والمنظور والقياس المرئي. يغطي تقييمهم 82 طبقة مختلفة من اثنتي عشرة بنية مختلفة لشبكات CNN تنتمي إلى أربع عائلات: AlexNets و VGG Nets و ResNets و DenseNets، مقيّمين فائدتها في مهام متطابقة واختلافات كبيرة في المنظور والمظهر، ووجدوا في نتائجهم اختلافات كبيرة في كل من المتانة robustness وحجم السمات feature size بين البنى المختلفة. وفقاً لبحثهم كانت أفضل السمات الإجمالية هي مخرجات الكتلة transition block الثالثة لبنى DenseNet، وخاصة DenseNet121 و DenseNet161 اللتين تتخذان توازنات مختلفة قليلاً من حيث الدقة وحجم السمات.

أجرى Valey وآخرون [39]. مقارنة بين أحدث النماذج المدربة مسبقاً على تصنيف الصور الدقيق باستخدام مجموعة بيانات Stanford Cars-196. من المثير للاهتمام أن أفضل شبكتين من حيث الدقة في بحثهم (DenseNet161 و DenseNet121). هما شبكتا CNN مدربتان مسبقاً موصىً بهما في المرجع [38]. كوحدات استخراج سمات.

أجرى Irvin وآخرون [40]. تجربة لاختبار أداء شبكات ResNet152 و DenseNet121 و Inception V4 و SEResNeXt101 على مجموعة بيانات CheXpert، وكان أداء شبكة DenseNet121 هو الأفضل فيها.

أوضح Rajpurkar وآخرون [41]. أن نماذج DenseNet استخدمت في تسعة من أفضل عشرة نماذج في منافسة CheXpert كجزء من النماذج الكلية، مع أن أداء DenseNet لم يكن هو الأفضل على ImageNet.

أجرى Ke وآخرون [42] دراسة حول أداء استخراج السمات لستة عشر نموذج CNN معروفة على مجموعة بيانات لصور صدرية بالأشعة السينية CheXpert. لم يجدوا علاقة بين الأداء على ImageNet والأداء على مجموعة بيانات الصور الطبية، ومع ذلك اكتشفوا أن اختيار بنية CNN يؤثر على الأداء أكثر من اختيار عائلة النموذج بالنسبة للمسائل الطبية. كما لاحظوا أن التدريب المسبق على ImageNet يعطي زيادة للدقة في جميع البنى، والزيادة أقل في النماذج الأكبر.

يحتوي عمل Sharif وآخرين [1]. مقارنة بين ست شبكات CNN كمستخرجات للسمات العامة من أجل نموذجهم لتوصيف الصور. قاموا باختبار Inception V3 و Densenet201 و InceptionResNet V2 و Resnet152 V2 و Xception و NASNet واستخدموا مجموعة بيانات Flickr30k للاختبار. أعطت NASNet أفضل النتائج على أغلب المعايير، وهذا يتوافق مع تجاربنا الموضحة في الفصل الرابع.

3.2 مقارنة المنهجية المستخدمة في هذا البحث مع منهجيات الأبحاث السابقة

يختلف هذا البحث عن الأبحاث السابقة بأنه يستخدم جميع سمات الأغراض من أجل جميع الأغراض في الصورة إلى جانب سمات CNN، بينما يستخدم بعض الأبحاث السابقة بعض سمات الأغراض وبعضها يستخدم بعض سمات بعض الأغراض، ولا تدمجها جميع الأبحاث السابقة مع سمات CNN. يختلف أيضاً هذا البحث عن الأبحاث السابقة بأنه يتضمن أشمل دراسة على علمنا حتى الآن حول أثر بنية CNN على جودة السمات المستخرجة منها بهدف توصيف الصور آلياً. يبين الجدول 1.2 خلاصة عن أهم الأبحاث المشابهة لهذا البحث.

Method	Reference	Year
CNN+Attention+LSTM	Xu et al [20]	2015
CNN+Object Detection+LSTM Encoder+ Attention+LSTM Decoder	Yin & Ordonez [3]	2017
CNN+Object Detection+ Attention+LSTM Decoder	Vo–Ho et al [2]	2019
Object Detection+Object Relation Transformer+Positional Encoding	Herdade et al [35]	2019
Study of CNN features across CNN families	Holliday & Dudek [38]	2020
CNN+Object Detection+Linguistic Features+Attention	Sharif et al [1]	2020
CNN+Object Detection+Attention+Potisional Encoding+GRU Decoder	Ours	2021

الجدول 1.2. مقارنة مع منهجيات الأبحاث السابقة.

4.2 مجموعات البيانات

1.4.2 مجموعة بيانات MSCOCO

مجموعة بيانات Microsoft COCO مجموعة بيانات كبيرة جِدّاً للتعرف على الصور وتقسيمها وتوصيفها. هناك ميزات مختلفة لمجموعة بيانات MSCOCO مثل تقسيم الأغراض object segmentation، والتعرف ضمن السياق recognition in context. فيها أغراض متعددة لكل فئة، في أكثر من 300,000 صورة، وأكثر من مليوني نموذج، و 80 فئة للأغراض، و 5 توصيفات لكل صورة. تستخدم العديد من منهجيات توصيف الصور مجموعة البيانات هذه في التجارب.

2.4.2 مجموعة بيانات Flickr30K

Flickr30K هي مجموعة بيانات لتوصيف الصورة التلقائي وفهم اللغة الأساسي. تحوي 30,000 صورة مجموعة من موقع Flickr مع 158 ألف توصيف أنشأها معلقون بشر، ولا توفر أي تقسيم ثابت للصور للتدريب والاختبار والتحقق ويمكن للباحثين اختيار النسب الخاصة بهم. تحوي مجموعة البيانات هذه أيضاً كاشفات للأغراض الشائعة، ومصنفاً لونياً، وتحيزاً نحو اختيار أغراض أكبر.

3.4.2 مجموعة بيانات Flickr8K

Flickr8k مجموعة بيانات شائعة فيها 8,000 صورة مجموعة من موقع Flickr. تتكون بيانات التدريب من 6,000 صورة و 1,000 صورة لكل من قسمي الاختبار testing والتحقق validation. تحوي كل صورة في مجموعة البيانات هذه 5 توصيفات مرجعية كتبها بشر.

4.4.2 مجموعة بيانات Visual Genome

مجموعة بيانات Visual Genome هي مجموعة بيانات لتوصيف الصور. لا يتطلب توصيف الصور التعرف على أغراض الصورة فحسب، بل يتطلب أيضاً التفكير في تفاعلاتها وخصائصها. على عكس مجموعات البيانات الثلاث الأولى حيث يقدم شرح للمشهد بأكمله، تحوي مجموعة بيانات الجينوم المرئي توصيفات منفصلة لمناطق متعددة في الصورة. تتكون من سبعة أجزاء أساسية: أوصاف المناطق، الأغراض، الخصائص، العلاقات، رسوم بيانية للمناطق، رسوم بيانية للمشهد، وأزواج إجابات وأسئلة. تحوي أكثر من 108,000 صورة. كل صورة فيها وسطياً 35 عنصراً و 26 سمة و 21 علاقة زوجية بين الأغراض.

5.4.2 مجموعة بيانات Instagram

أنشأ كل من Tran وآخرين و Park وآخرين مجموعتي بيانات باستخدام صور من Instagram وهي خدمة شبكة اجتماعية لمشاركة الصور. مجموعة بيانات Tran وآخرين تحوي حوالي 10 آلاف صورة معظمها من المشاهير. أما Park وآخرون استخدموا مجموعة البيانات الخاصة بهم للتنبؤ بالوسم hashtag وتوليد المنشورات في شبكات التواصل الاجتماعي. تحوي مجموعة البيانات هذه 1.1 مليون منشور حول عدد كبير من الموضوعات وقوائم hashtag طويلة من 6300 مستخدم.

6.4.2 مجموعة بيانات IAPR TC-12

تحوي مجموعة بيانات IAPR TC-12 فيها 20,000 صورة. جمعت الصور من مصادر مختلفة مثل الرياضة وصور الأشخاص والحيوانات والمناظر الطبيعية والعديد من المواقع الأخرى حول العالم. تتضمن صور مجموعة البيانات هذه توصيفات صور بلغات متعددة فيها صور أغراض كثيرة أيضاً.

7.4.2 مجموعة بيانات Stock3M

تحوي مجموعة بيانات Stock3M فيها 3,217,654 صورة محملة من المستخدمين وهي أكبر ب 26 مرة من مجموعة بيانات MSCOCO وتتضمن محتوىً متنوعاً.

8.4.2 مجموعة بيانات MIT–Adobe FiveK

تتكون مجموعة بيانات MIT–Adobe FiveK من 5,000 صورة. تحوي الصور مجموعة متنوعة من المشاهد والمواضيع وظروف الإضاءة وهي تتعلق بصورة أساسية بالأشخاص والطبيعة والأشياء من صنع الإنسان.

9.4.2 مجموعة البيانات FlickrStyle10k

تتكون مجموعة بيانات FlickrStyle10k من 10,000 صورة مجموعة من موقع Flickr مع توصيفات حسب أساليب كتابة معينة. تتكون بيانات التدريب من 7,000 صورة وبيانات التحقق validation من 2,000 صورة وبيانات التجريب testing من 1,000 صورة. تحوي كل صورة توصيفات رومانسية وفكاهية وواقعية.

10.4.2 مجموعة بيانات Google Conceptual Captions

قدم Sharma وآخرون [43]. مجموعة بيانات Conceptual Captions التي تحتوي عدداً أكبر من الصور من مجموعة بيانات MSCOCO وتمثل مجموعة متنوعة من الصور وأنماط توصيفها. حققوا ذلك عن طريق استخراج توصيفات الصور وتصفيتها من مليارات صفحات الويب.

5.2 معايير التقييم

BLEU (Bilingual Evaluation Understudy). BLEU هو مقياس يستخدم لقياس جودة النص المولد آلياً، حيث تقارن مقاطع النص الفردية بمجموعة من النصوص المرجعية وتحسب الدرجات لكل منها. عند تقدير الجودة الإجمالية للنص الناتج يحسب متوسط الدرجات المحسوبة، ومع ذلك لا ينظر في الصواب النحوي هنا. يختلف أداء مقياس BLEU اعتماداً على عدد الترجمات المرجعية وحجم النص الذي أنشئ. يحظى معيار BLEU بشعبية لأنه رائد في التقييم التلقائي للنصوص المترجمة آلياً وله توافق مقبول مع تحكيم الجودة البشري لكنه مع ذلك محدود. فمثلاً تكون درجات BLEU جيدة فقط إذا كان النص المولد قصيراً، وهناك بعض الحالات التي لا تعني فيها الزيادة في درجة BLEU أن جودة النص المولد جيدة.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE هو مجموعة من المقاييس المستخدمة لقياس جودة ملخصات النصوص، تقارن تسلسل الكلمات وأزواج الكلمات و n-grams مع مجموعة من الملخصات المرجعية التي أنشأها البشر. توجد أنواع مختلفة من ROUGE مثل ROUGE-1 و 2 و ROUGE-W و ROUGE-SU4 لأداء مهام مختلفة، لكن لمعيار ROUGE مشاكل في تقييم ملخصات النصوص متعددة المستندات.

METEOR (Metric for Evaluation of Translation with Explicit ORdering). METEOR هو مقياس آخر يستخدم لتقييم النص المترجم آلياً، وتقارن فيه مقاطع الكلمات القياسية بالنصوص المرجعية. بالإضافة إلى ذلك، تراعى أيضاً المطابقة مع جذوع الكلمات في الجملة ومرادفات الكلمات. يستطيع METEOR تحقيق توافق أفضل على مستوى الجملة أو الجزء من الجملة.

CIDEr (Consensus-based Image Description Evaluation). CIDEr هو مقياس الإجماع التلقائي لتقييم توصيفات الصور. تحوي معظم مجموعات البيانات الحالية خمسة توصيفات فقط لكل صورة وتعمل مقاييس التقييم السابقة مع هذا العدد الصغير من الجمل ولا تكفي لقياس الإجماع بين التوصيفات التي أنشئت مع التحكيم البشري. يحقق معيار CIDEr الإجماع البشري باستخدام Term Frequency-Inverse Document Frequency (TF-IDF).

SPICE (Semantic Propositional Image Caption Evaluation). SPICE هو مقياس تقييم جديد لتوصيفات صور يعتمد على المفهوم الدلالي، يعتمد على التمثيل الدلالي القائم على graph يسمى scene-graph. يستطيع التمثيل البياني هذا استخراج معلومات الأغراض والسمات المختلفة وعلاقاتها من توصيفات الصورة.

يقيّم ROUGE-L مدى كفاءة وفصاحة التوصيفات المولدة، بينما يركز CIDEr على القواعد النحوية والبراعة اللغوية. أما SPICE يحلل دلالات التوصيفات التي أنشئت. ليس لمعايير BLEU و METEOR و ROUGE توافق جيد مع التقييمات البشرية للجودة، أما SPICE و CIDEr فلهما توافق أكبر ولكن يصعب رفع قيمتهما.

في الفصل الآتي نقترح عدداً من النماذج المعتمدة على التعلم العميق بهدف حل المسألة.

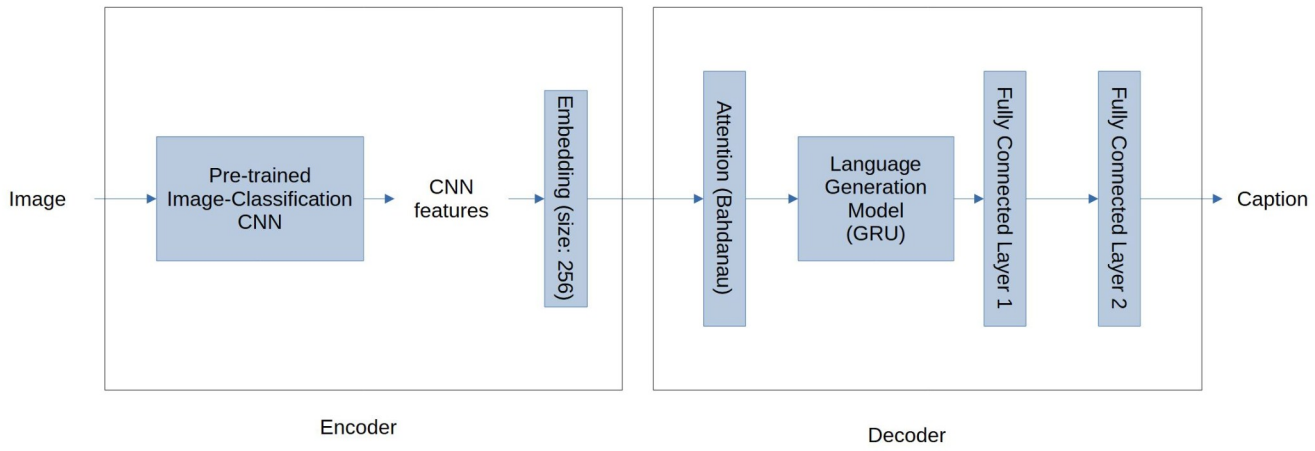
الفصل الثالث: النماذج المقترحة في هذه الأطروحة

اقترحنا في هذا البحث نموذجاً على مراحل لتوصيف الصور آلياً، وسنستعرض في هذا الفصل بالتفصيل هذه النسخ من النموذج والفروق بينها. في البداية نقترح نموذجاً ابتدائياً بسيطاً نسميه (Baseline model). ثم نعدله بناءً على منهجيات عدة لاختيار النموذج النهائي.

1.3 النموذج الأولي (Baseline Model)

يستخدم هذا النموذج بنية Encoder–Decoder لتوصيف الصور آلياً. في مرحلة ترميز الصورة، يستخدم نموذج Xception لاستخراج سمات الصورة مدرباً مسبقاً على مجموعة بيانات ImageNet، حيث تستخرج السمات من الطبقة التلافيفية الأخيرة فيه. نستخدم هنا نموذج Xception المدرب في مكتبة Keras ونزيل منه الطبقة كاملة التوصيل (التي تختص بالتصنيف) من أجل استخراج السمات. مصفوفة السمات الأساسية المستخرجة من Xception هي من أبعاد (10×10×2048)، نجري عليها عملية تسطيح حتى تصبح من أبعاد (100×2048). نستخدم حجم دفعة Batch Size من 64 صورة.

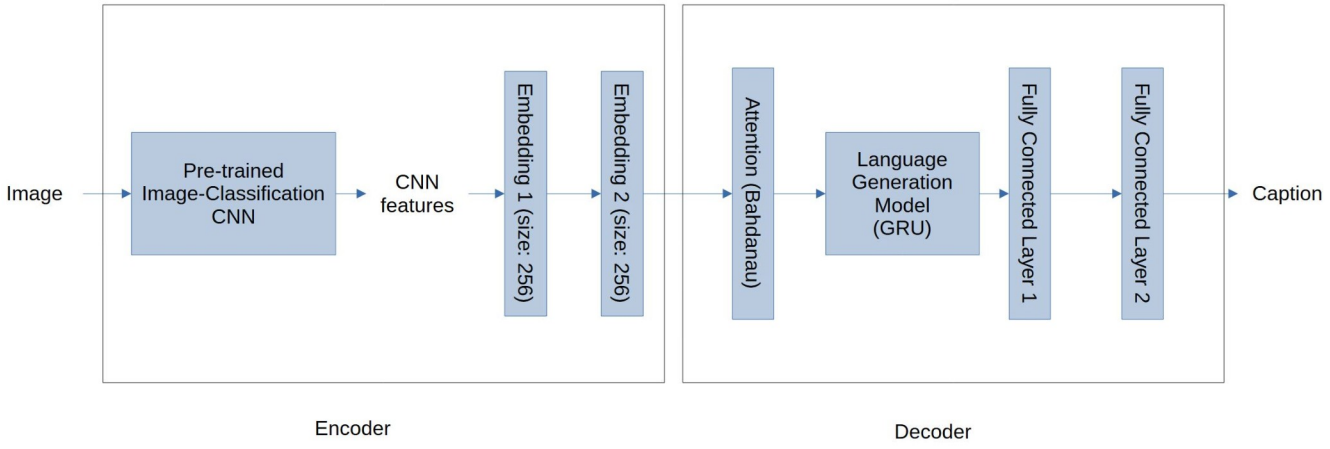
تدخل سمات شبكة CNN إلى طبقة تضمين مرتبطة كلياً embedding fully connected layer لإسقاط السمات على فضاء أصغر من أجل إدخالها إلى وحدة توليد النصوص، بعرض طبقة يساوي 256. (الشكل 1.3) يدخل خرج طبقة التضمين إلى وحدة GRU، التي نستخدمها هنا لأدائها الجيد بالمقارنة مع بقية نماذج التوليد اللغوي حسب (Gu et al., 2017) [44]. ولسرعتها وقلة استهلاكها للذاكرة مما يجعلها مناسبة لبيئة معطيات كبيرة. يتكوّن مفكك الترميز decoder للنموذج من وحدة GRU مع طبقتين كاملتي الارتباط من بعدها. في نهاية المرمز توجد طبقتان كاملتا الارتباط، الأولى بعرض 512، والثانية بعرض يساوي عدد المفردات. يتميز هذا النموذج ببساطته واستخدامه لتقنية الانتباه attention في توليد النصوص وبسرعة تدريبه وتنبؤه بالنصوص. يستخدم هذا النموذج طريقة تدريب الانتشار العكسي backpropagation ونموذج Bahdanau soft attention في مرحلة فك الترميز decoding.



الشكل 1.3. رسم توضيحي للنسخة 0 من النموذج (Baseline).

2.3 التعديل 1 (Double Word Embedding)

يعمل هذا النموذج بطريقة مشابهة للنموذج الأولي (Baseline)، لكن قمنا هنا بإضافة طبقة تضمين إضافية في مرحلة إسقاط خرج شبكة CNN على فضاء أصغر، سعياً وراء الحصول على نتائج أفضل. تأتي الفكرة من أن الأعمال السابقة تسقط سمات CNN على فضاء أصغر تمهيداً لفك ترميزها في مفكك الترميز Decoder، فإن كان الإسقاط على فضاء تضمين مفيداً في ترميز السمات فربما إسقاط خرج مرحلة التضمين على فضاء تضمين آخر يزيد القوة التعبيرية للسمات الناتجة، ولم يسبق على علمنا في الأبحاث السابقة من حاول بمثل هذا الأسلوب. بينت التجارب الموضحة في الفصل الرابع قلة فعالية هذا التعديل حيث أدى إلى نتائج أقل من النموذج الأولي، ولهذا لم نعتمده. يبين الشكل 2.3 هذا النموذج.



الشكل 2.3. رسم توضيحي للنسخة 1 (Double Word Embedding).

3.3 التعديل 2 (YOLO Bounding Boxes)

تتضمن هذه النسخة فكرة جديدة لاستخراج سمات الصورة من أجل توليد نصوص معبرة عنها. استخدم بعض الباحثين في المراجع السابقة، نماذج تصنيف الصور من أجل استخراج السمات، واستخدم البعض الآخر نماذج كشف الأغراض object detection لاستخراج السمات. في هذا النموذج، نستخدم سمات من كلا النوعين وندمجها من أجل الحصول على سمات بجودة أعلى تستطيع التعبير عن الصورة بطريقة أفضل (الشكل 3.3).

يمكننا أن نقول أنه في حالة البشر، إذا رأى الإنسان صورة لشخصين بعيدين فقد يصفها بأنها صورة لمكان عام، ولكن إن رآهما قريبين من بعضهما سيعرف أنهما يتحدثان مثلاً. يحاول هذا التعديل على النموذج الاستفادة من معلومات أنواع الأغراض ومواقعها ومساحتها في الصورة من أجل الوصول إلى فهم أقرب لفهم البشر للصور.

كما في النسختين السابقتين، تستخدم هذه النسخة من نموذجنا نموذج Xception بدون الطبقة الأخيرة كاملة التوصيل لاستخراج سمات الصورة وهو مدرب مسبقاً على ImageNet ونستخدم حجم دفعة Batch Size من 64 صورة. كما يستخدم YOLOv4 كنموذج لاكتشاف الأغراض object detection، لأنه يتميز بسرعته وقدرته على

كشف عدد كبير من أصناف الأغراض، مما يجعله مناسباً لتطبيقات المعطيات الكبيرة. نستعمل هنا نموذجاً مدرباً على MSCOCO من مكتبة YOLOv4.

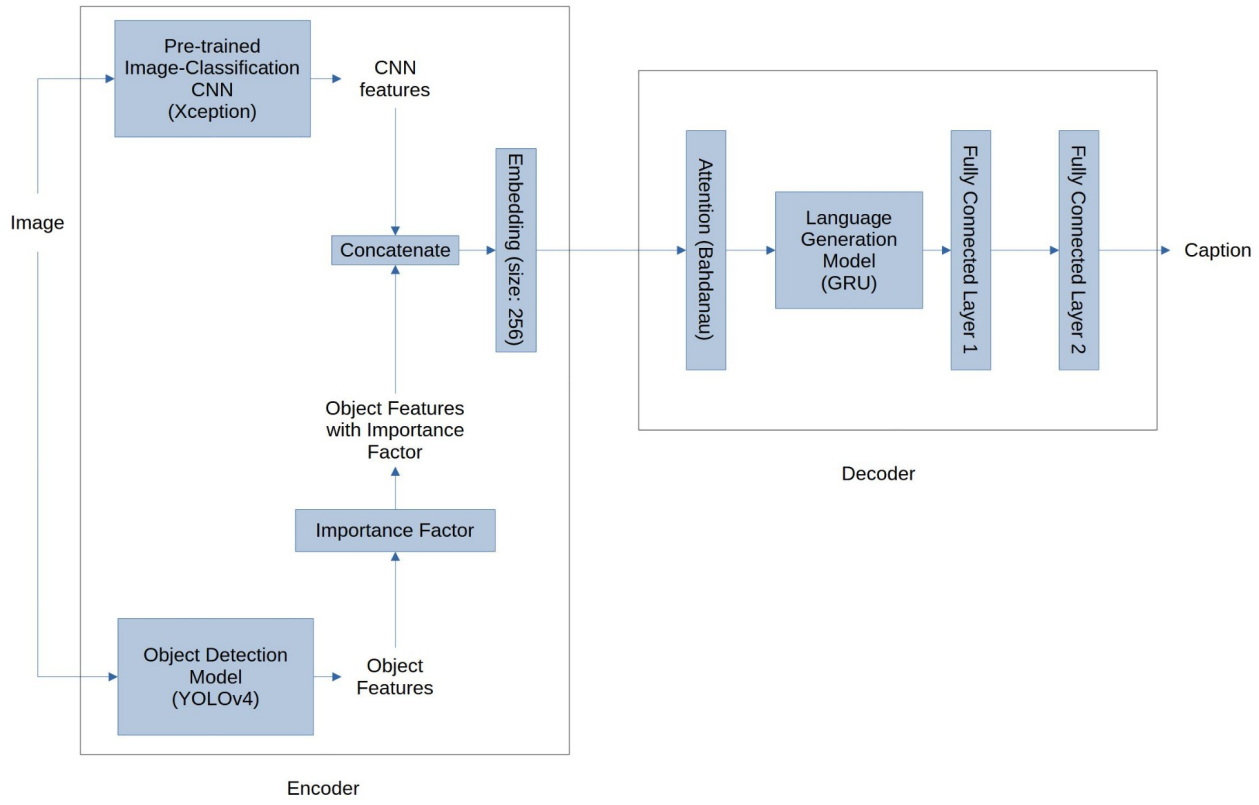
في الأبحاث السابقة التي استفادت من سمات الأغراض، استخدمت بعض النماذج مصفوفة من طول 80 عنصراً تظهر فقط عدد الأغراض من كل صنف (80 صنفاً للنماذج المدربة على MSCOCO)، فيما استخدمت نماذج أخرى عدداً محدداً مسبقاً من الأغراض على أن تكون درجات الثقة confidence rate فيها هي الأعلى. للتأكد من استخدام جميع المعلومات التي يمكن لنموذج YOLO أن يقدمها، نستفيد من جميع معلومات الأغراض على ألا يتجاوز عدد الأغراض 292 غرضاً، وهو حد أكثر من كافٍ عادةً لكشف جميع الأغراض المهمة في الصورة.

يستخرج YOLO من أجل كل صورة مصفوفة من الأغراض، بحيث يتضمن الغرض الواحد موقع X على الصورة، وموقع Y، وعرض الغرض وارتفاعه، ودرجة الثقة من نوع الغرض بين 0 و 1، ورقم نوع الغرض، وعامل أهمية importance factor جديد نقترحه هنا.
يهدف عامل الأهمية المقترح هذا إلى تمييز الأغراض الكبيرة (أي ذات المساحة الأكبر) في الواجهة على الأغراض الصغيرة، التي غالباً ما ستكون أقل أهمية، وتمييز الأغراض التي درجة الثقة في صنفها عالية على الأغراض ذات الثقة الأقل. يضاف عامل الأهمية هذا إلى كل غرض، وترتب الأغراض وفقه باستخدام خوارزمية Quick Sort بتعقيد $O(n \times \log(n))$. يحسب العامل على هذا النحو:

$$\text{Importance Factor} = \text{Confidence Rate} \times \text{Object Width} \times \text{Object Height}$$

خرج مرحلة استخراج الأغراض object feature extraction هو مصفوفة أحادية البعد، نملؤها بالحشو padding حتى تصبح بطول 2048 لتكون مناسبة لإضافتها إلى سمات Xception.

في مرحلة الضمّ concatenation نضمّ سمات object features مع سمات CNN. سمات CNN بأبعاد (100×2048)، وسمات YOLO بأبعاد (1×2048).. تصبح أبعاد مصفوفة السمات النهائية بعد ضمّ سمات YOLO كسطر أخير بأبعاد (101×2048).



الشكل 3.3. رسم توضيحي للنسخة 2 من النموذج (YOLO Bounding Boxes).

تقدم هذه النسخة من النموذج أفضل النتائج من بين النماذج المقترحة في هذا البحث. وفيما يأتي جدول يظهر فعالية استخراج السمات بهذه الطريقة:

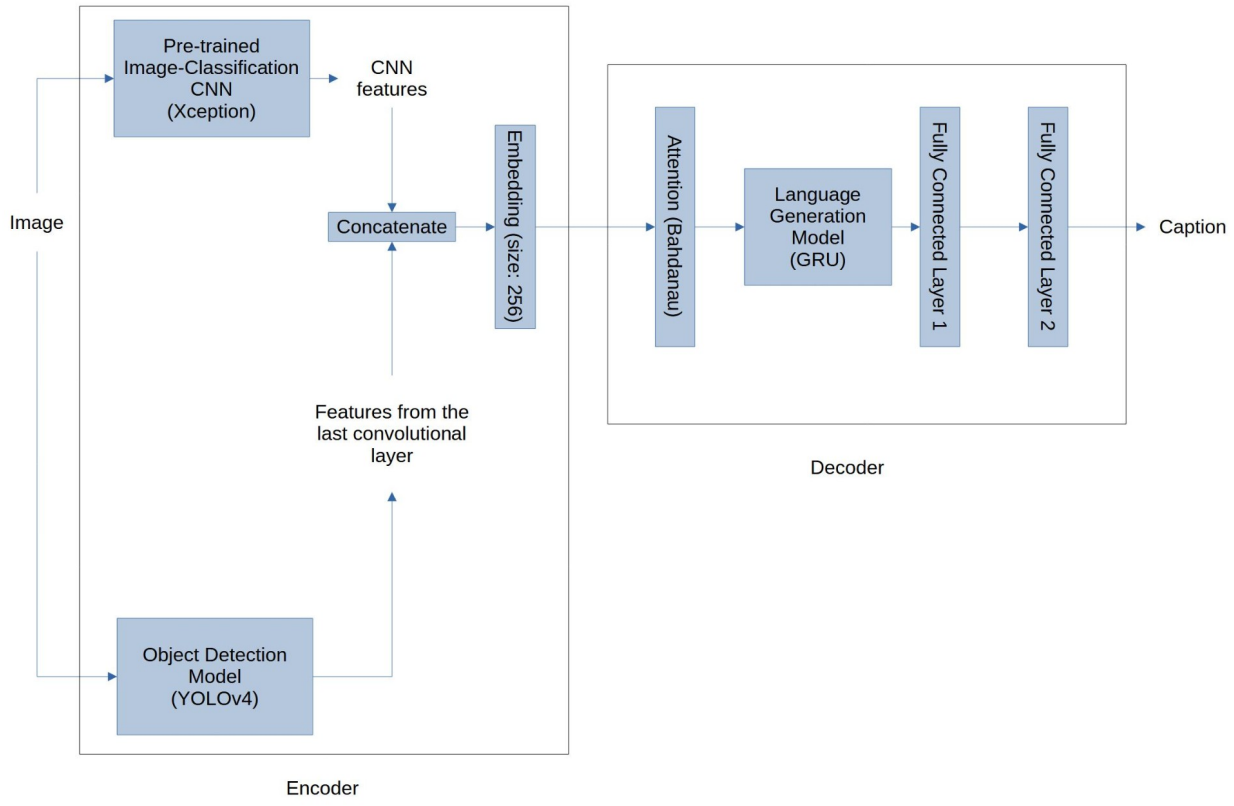
Model (Development/ Test)	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		CIDEr		ROUGE-L		SPICE	
Model with YOLO Bounding Boxes	0.544	0.474	0.361	0.288	0.234	0.170	0.150	0.099	0.190	0.167	0.522	0.361	0.403	0.361	0.132	0.112
Model without Object Features	0.485	0.399	0.319	0.220	0.208	0.117	0.135	0.062	0.176	0.123	0.337	0.148	0.372	0.293	0.118	0.074
Score Difference	0.059	0.075	0.042	0.068	0.026	0.053	0.015	0.037	0.014	0.044	0.185	0.213	0.031	0.068	0.014	0.038

الجدول 1.3. فعالية استخدام استخراج السمات حسب النسخة 2 من النموذج (YOLO Bounding Boxes).

4.3 التعديل 3 (YOLO Raw Features v1)

يشبه هذا التعديل على النموذج التعديل 2 (YOLO Bounding Boxes). في البنية، لكنه لا يستخدم عامل الأهمية المقترح في النسخة 3 من النموذج، فبدلاً من استخراج سمات الغرض من YOLO نستخرج هنا سمات من الطبقة التلافيفية الأخيرة في YOLO وهذه الخطوة هي الأولى من نوعها على نموذج استخراج سمات الأغراض على علمنا. تكمن الفكرة في رفع قيمة معايير التقييم باستخراج مصفوفة سمات من الطبقة التلافيفية الأخيرة في YOLO وأبعادها (16×16×255) نضغطها إلى (256×255). في مرحلة ضمّ السمات concatenation نحول مصفوفتي السمات إلى مصفوفتين ذات بعد واحد ثم نضمهما لتصبحا مصفوفة واحدة من بعد (1×270,080) (الشكل 4.3).

حسب نتائج التجارب المبينة في الفصل الرابع، سبب هذا التعديل هبوطاً في نتائج التقييم على مجموعتي التحقق والاختبار، ولهذا لم نعتمده.

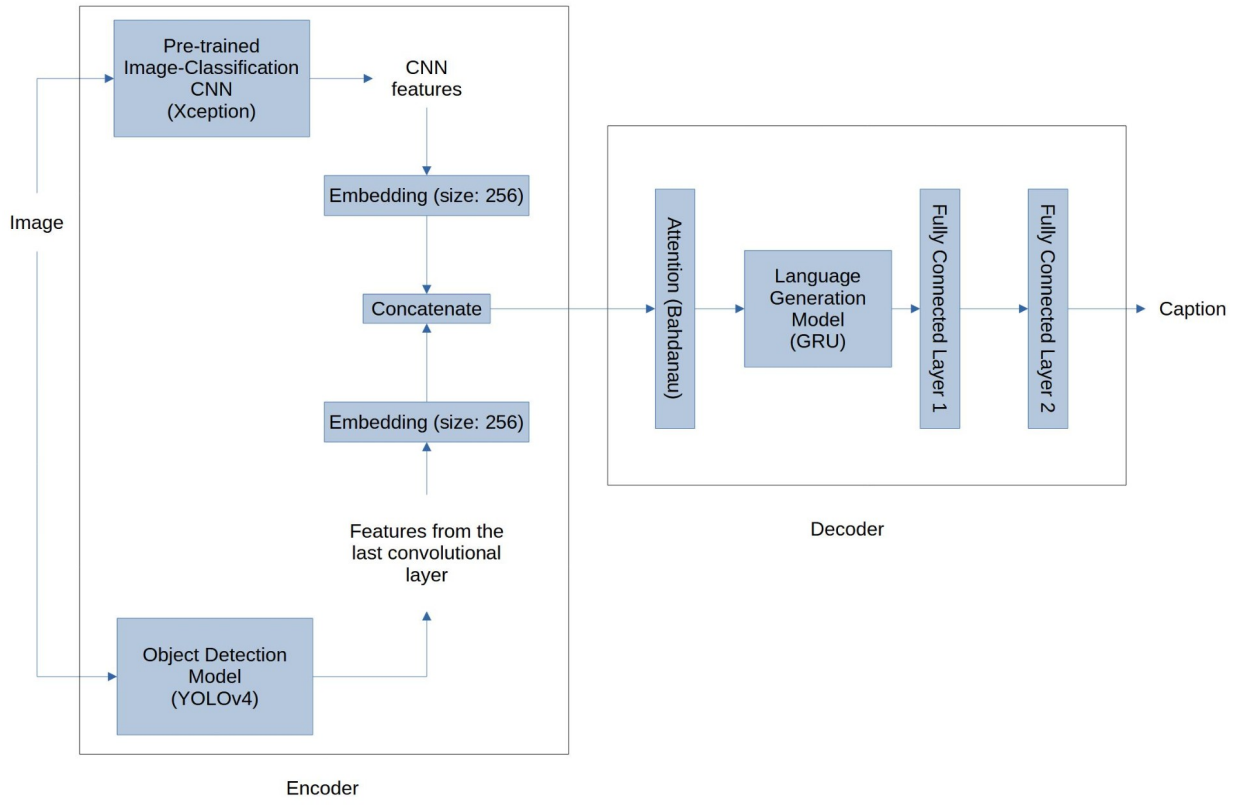


الشكل 4.3. رسم توضيحي للنسخة 3 من النموذج (YOLO Raw Features v1).

5.3 التعديل 4 (YOLO Raw Features v2)

يشبه هذا التعديل النسخة 3 (YOLO Raw Features v2). من حيث مبدأ استخراج السمات من YOLOv4، لكنه يمرر سمات YOLO في طبقة تضمين ويمرر سمات Xception في طبقة تضمين منفصلة، ثم يجري الضمّ concatenation لخرج هاتين الطبقتين لتنتج مصفوفة السمات النهائية (الشكل 5.3).. بينت هذه الطريقة فعالية جيدة على قسم التدريب training split في MSCOCO تقارن بفعالية التعديل 2، لكن كانت نتائج التجربة أقل بكثير على قسم الاختبار testing split. هذا دل على قلة قدرة النموذج الذي يعمل بهذا الأسلوب على التعميم وبسبب هذا لم نعتمد هذا التعديل.

من بين هذه النسخ حسب التجارب، تظهر فعالية النسخة 2 (YOLO Bounding Boxes). الأكبر في استخراج سمات قادرة على التعبير عن الصورة بطريقة جيدة. نرجع هذا إلى أن استخدام المعلومة التي ينتجها مكون YOLO الذي يستنتج أماكن الأغراض في الصور فعال أكثر من ترك بنية decoder تتعلمها من الصفر. حيث تظهر فعالية هذه الطريقة مع قلة حجم مصفوفة سمات YOLO بالمقارنة مع حجم مصفوفة سمات Xception.



الشكل 5.3. رسم توضيحي للنسخة 4 (YOLO Raw Features v2).

6.3 التعديل 2.1 (YOLO Bounding Boxes)

يتبع التعديل 2.1 بنية مشابهة جداً لبنية النسخة 2. الزيادة الموجودة في النسخة 2.1 هي محاولة الاستفادة من اسم صنف الغرض المكتشف باستخدام YOLO إذا كان موجوداً في فهرس كلمات التدريب. مثلاً، إذا كانت هناك كرة مكتشفة في الصورة سيكون اسم صنفها "ball"، فبدلاً من استخدام رقم الصنف الذي أعاده نموذج YOLO نستخدم رقم كلمة "ball" في فهرس مجموعة التدريب. يبدو هذا تحسيناً منطقياً على النسخة 2.1 لأنه يستخدم كلمات الفهرس الذي أنشأناه نفسها بدلاً من استخدام رقم صنف غريب عن مجموعة تدريب توصيف الصور (الشكل 6.3).

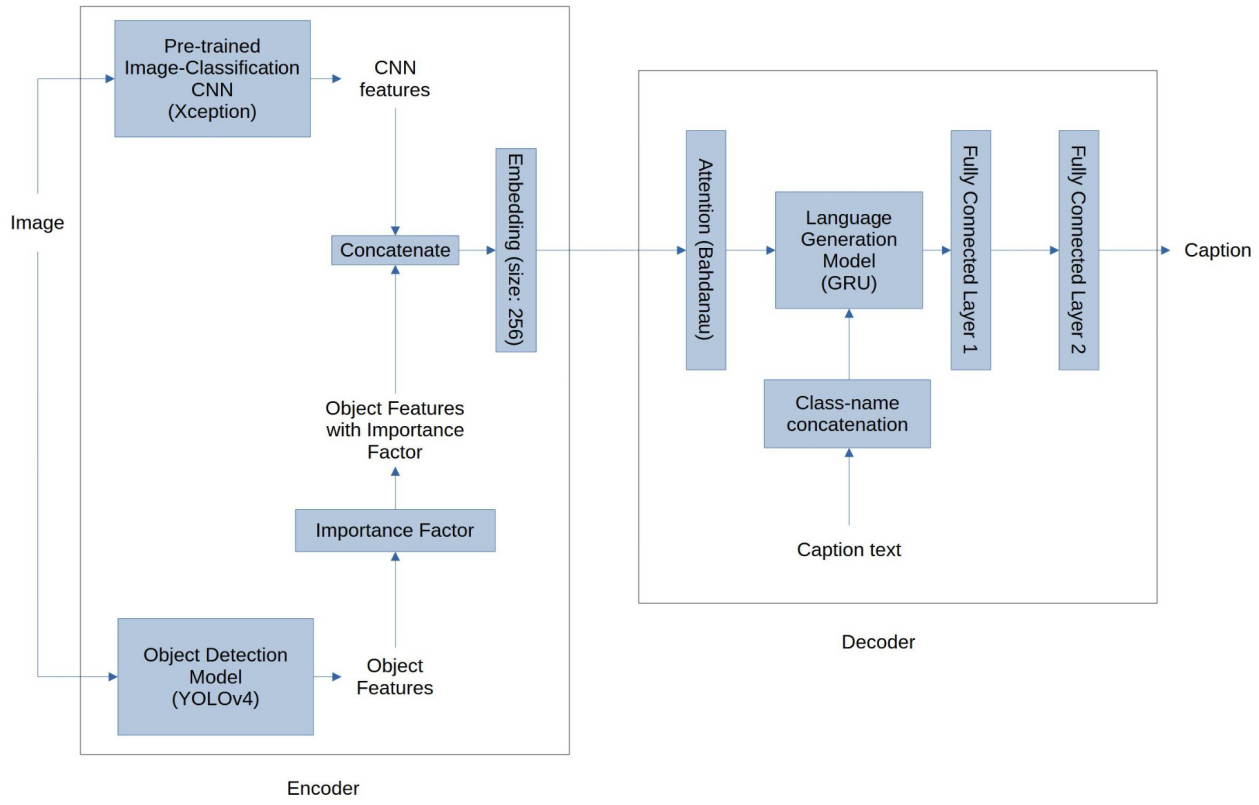
كانت هناك مشكلة تبديل أسماء أصناف الأغراض التي تتكون من أكثر من كلمة، مثل "traffic light" و "stop sign". كانت هناك مجموعة من الحلول الممكنة وهي:

- استخدام الكلمة الأولى من اسم الصنف.
- استخدام الكلمة الأخيرة من اسم الصنف.
- دمج كلمات اسم الصنف، وهو الحل الذي طبقناه. تطلب هذا أيضاً إجراء التعديل نفسه على مجموعة التدريب عند قراءتها، فحولنا جميع الكلمات التي تتكون من أكثر من كلمة إلى كلمة واحدة مضمومة. مثال:

"traffic light"==>"trafficlight"

"stop sign"==>"stopsign"

سبب هذا التعديل على أسلوب استخدام سمات الأغراض أثراً سلبياً كما تبين النتائج في الفصل الرابع، ولهذا لم نعتمد هذا التعديل في النموذج النهائي. وعلى هذا يكون النموذج النهائي الذي نقترحه في هذه الأطروحة هو النسخة 2 من النموذج (YOLO Bounding Boxes).



الشكل 6.3. تدريب النسخة 2.1 (YOLO Bounding Boxes).

الفصل الرابع: الاختبارات والنتائج

نورد في هذا الفصل نتائج الاختبارات على النماذج المقترحة في الفصل السابق بهدف مقارنتها مع بعضها ومقارنة العمل في هذا البحث مع الأبحاث السابقة. نستخدم معايير التقييم BLEU1، BLEU-2، BLEU-3، BLEU-4، METEOR، ROUGE-L، CIDEr، SPICE.

1.4 المرحلة الأولى من التجارب

الهدف الأساسي من هذه المرحلة هو تقدير الزمن اللازم للتنفيذ، واختبار سلسلة الإجراءات pipeline التي تستخدم النموذج الأولي (Baseline model). ونماذج استخلاص السمات Feature Extraction Models. جرى اختبار مجموعة من نماذج استخلاص السمات هي Xception و VGG19 و ResNet50 و Inception V3 و DenseNet121.

أجرينا التجارب على مخدم المعهد المجهز ب 32 GB من الذاكرة ووحدة CPU من نوع Intel Corei9-9900K ووحدة GPU من نوع NVIDIA GeForce RTX 2080 واستخدمت في هذا البحث مكتبة CUDA لاستخدام وحدة GPU في المعالجة على التفرع.

جرى استخدام مجموعة التدريب training split المتاحة من مجموعة المعطيات MSCOCO، المؤلفة من 82,783 صورة، حيث جرى تقسيمها إلى 80%. كمجموعة تدريب حالية، وجرى التقييم على 20%. المتبقية من المعطيات. جرى اعتماد 15000 كلمة كحجم للقاموس Dictionary Size.
يبين الجدول 1.4 نتائج التجارب المجراة التي اختبرت مجموعة نماذج استخلاص السمات، كما يبين الزمن الذي استغرقه تنفيذ كل منها.

Model (Train)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Time
DenseNet121	0.493	0.31	0.195	0.118	0.179	0.408	0.373	0.122	08:18:38
Inception V3	0.504	0.322	0.200	0.122	0.182	0.460	0.382	0.125	09:26:40
ResNet50	0.525	0.343	0.218	0.136	0.188	0.481	0.395	0.129	10:11:12
VGG19	0.485	0.308	0.189	0.115	0.170	0.375	0.364	0.113	08:05:58
Xception	0.528	0.343	0.219	0.138	0.186	0.493	0.392	0.130	12:20:05

الجدول 1.4. المرحلة الأولى من التجارب.

يمكننا أن نلاحظ من الجدول 1.4 أن استخدام النموذجين Xception و ResNet50 أدى إلى النتائج الأفضل في المرحلة الأولى من التجارب، بينما أدت بقية النماذج إلى نتائج أقل مع اختلافات بينها وأدى استخدام نموذج VGG19 إلى النتائج الأقل. يمكن تفسير التغير في نتائج التقييم في جميع المعايير بين التجارب أن اختيار نموذج CNN المستخدم في استخراج السمات feature extraction له تأثير كبير على جودة التوصيفات المنتجة وهو الذي قاد إلى فكرة تجربة أخرى لتحديد تأثير feature extraction model على معايير تقييم النصوص الناتجة.

2.4 المرحلة الثانية من التجارب

أدى ذلك الاختلاف الكبير في نتائج التقييم في المرحلة الأولى من التجارب بين نماذج استخراج السمات إلى التفكير في تأثير نموذج استخراج السمات feature extraction model على جودة التوصيف الآلي للصور. وعلى هذا أجرينا تجارباً مع تثبيت عدد الكلمات على 15,000 وتغيير المتغيرات الآتية:

1. نموذج استخراج السمات feature extraction، اخترنا النماذج: DenseNet121، DenseNet169، DenseNet201، Inception V3، InceptionResnet V2، NasNetLarge، ResNet50، ResNet101، ResNet152، VGG16، VGG19، Xception

سبب اختيار هذه النماذج هي أنها الأعلى في الدقة top 1 accuracy و top 5 accuracy من بين النماذج المدربة مسبقاً في مكتبة Keras، كما أن تضمين نماذج Keras هو الأسهل والأكثر توافقية مع مكتبة

TensorFlow المستخدمة في هذا المشروع، وضمنت هذه التجربة أيضاً أفضل النماذج المختبرة من عدد من عائلات النماذج CNN families التي اقترحها الباحثان Holliday و Dudek عام 2020 [38]، حيث قارن الباحثان نماذج CNN المدربة مسبقاً على ImageNet بهدف استخراج السمات بدون مجال معين كهدف، لكن المرجع اعتمد على الدقة accuracy والمتانة robustness لتقييم النماذج.

2. مجموعة التدريب Dataset. اخترنا أشهر ثلاثة Benchmark Datasets في مجال التوصيف الآلي للصور، وهي MSCOCO، Flickr30k، Flickr8k.

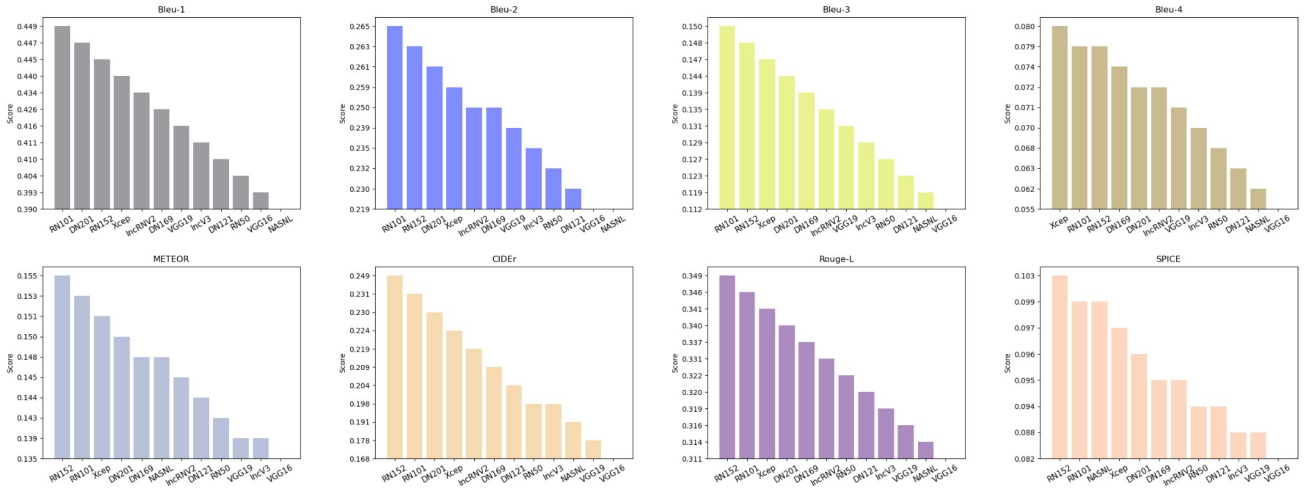
3. التقسيم spliting. في هذه التجارب خلطنا قسمي التدريب والتحقق من مجموعة البيانات وأعدنا تقسيمها عشوائياً بنسبة 20%. لمجموعة التحقق و 80%. لمجموعة التدريب. أما عند استخدام مجموعة الاختبار فاستخدمنا مجموعة التدريب ومجموعة التحقق معاً للتدريب ومجموعة الاختبار للتجريب.

في MSCOCO مجموعة الاختبار منشورة بدون labels. لهذا استخدمنا 20%. من مجموعة التدريب كمجموعة تحقق، أما المنشور من MSCOCO كمجموعة تحقق فاستخدمناه كمجموعة اختبار.

عدد التجارب في هذه المرحلة 72 تجربة، وفيما يأتي ثلاثة جداول بهذه النتائج (الجداول 2.4 و 3.4 و 4.4)، مع تغميق النتائج الثلاث الأعلى في كل عمود، مع رسوم توضيحية (الأشكال 1.4 و 2.4 و 3.4). توضح النتائج مرتبة.

Model (Train/Test)	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		CIDEr		ROUGE-L		SPICE	
VGG16	0.607	0.393	0.459	0.219	0.336	0.112	0.239	0.055	0.238	0.135	0.671	0.168	0.480	0.311	0.177	0.082
VGG19	0.576	0.416	0.434	0.239	0.316	0.131	0.225	0.071	0.244	0.139	0.627	0.178	0.477	0.316	0.186	0.088
ResNet50	0.638	0.404	0.503	0.232	0.387	0.127	0.293	0.068	0.264	0.143	0.797	0.198	0.519	0.322	0.203	0.094
ResNet101	0.654	**0.449**	0.516	**0.265**	0.399	**0.150**	0.304	**0.079**	0.259	**0.153**	0.796	**0.231**	0.519	**0.346**	0.200	**0.099**
ResNet152	0.653	**0.445**	0.517	**0.263**	0.399	**0.148**	0.301	**0.079**	0.264	**0.155**	0.815	**0.249**	0.521	**0.349**	0.204	**0.103**
Inception V3	0.604	0.411	0.457	0.235	0.336	0.129	0.243	0.070	0.243	0.139	0.683	0.198	0.484	0.319	0.183	0.088
Xception	0.631	0.440	0.492	0.259	0.373	**0.147**	0.278	**0.080**	0.252	**0.151**	0.742	0.224	0.501	**0.341**	0.196	0.097
InceptionResNet V2	0.541	0.434	0.397	0.250	0.285	0.135	0.202	0.072	0.231	0.145	0.575	0.219	0.451	0.331	0.176	0.095
DenseNet121	0.609	0.410	0.466	0.230	0.349	0.123	0.257	0.063	0.244	0.144	0.714	0.204	0.489	0.320	0.184	0.094
DenseNet169	0.593	0.426	0.453	0.250	0.337	0.139	0.245	0.074	0.250	0.148	0.709	0.209	0.491	0.337	0.190	0.095
DenseNet201	0.634	**0.447**	0.492	**0.261**	0.373	0.144	0.278	0.072	0.249	0.150	0.759	**0.230**	0.506	0.340	0.190	0.096
NASNetLarge	0.698	0.390	0.578	0.219	0.472	0.119	0.379	0.062	0.290	0.148	0.984	0.191	0.568	0.314	0.222	**0.099**

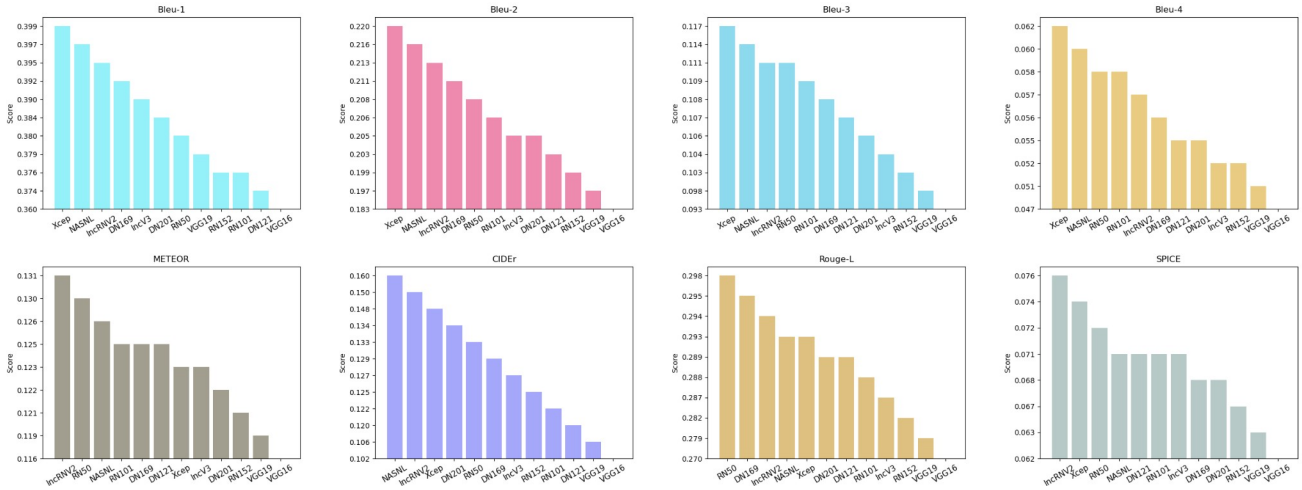
الجدول 2.4. نتائج التجارب على Flickr8k dataset.



الشكل 1.4. مخطط يوضح النتائج على Flickr8k مرتبة.

Model (Train/Test)	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		CIDEr		ROUGE-L		SPICE	
VGG16	0.457	0.360	0.287	0.183	0.179	0.093	0.110	0.047	0.155	0.116	0.279	0.102	0.342	0.270	0.098	0.062
VGG19	0.456	0.379	0.284	0.197	0.175	0.098	0.107	0.051	0.155	0.119	0.272	0.106	0.338	0.279	0.098	0.063
ResNet50	0.513	0.380	0.343	0.208	0.228	**0.111**	0.150	**0.058**	0.174	**0.130**	0.366	0.133	0.379	**0.298**	0.117	**0.072**
ResNet101	0.494	0.376	0.329	0.206	0.216	0.109	0.141	**0.058**	0.180	0.125	0.353	0.122	0.380	0.288	0.121	0.071
ResNet152	0.497	0.376	0.332	0.199	0.220	0.103	0.144	0.052	0.183	0.121	0.358	0.125	0.385	0.282	0.123	0.067
Inception V3	0.479	0.390	0.308	0.205	0.196	0.104	0.124	0.052	0.167	0.123	0.307	0.127	0.364	0.287	0.111	0.071
Xception	0.485	**0.399**	0.319	**0.220**	0.208	**0.117**	0.135	**0.062**	0.176	0.123	0.337	**0.148**	0.372	0.293	0.118	**0.074**
InceptionResNet V2	0.474	**0.395**	0.297	**0.213**	0.185	**0.111**	0.114	0.057	0.158	**0.131**	0.287	**0.150**	0.348	**0.294**	0.104	**0.076**
DenseNet121	0.445	0.374	0.281	0.203	0.176	0.107	0.109	0.055	0.165	0.125	0.269	0.120	0.349	0.289	0.107	0.071
DenseNet169	0.469	0.392	0.300	0.211	0.188	0.108	0.117	0.056	0.165	0.125	0.302	0.129	0.354	**0.295**	0.109	0.068
DenseNet201	0.477	0.384	0.305	0.205	0.192	0.106	0.120	0.055	0.164	0.122	0.311	0.134	0.355	0.289	0.107	0.068
NASNetLarge	0.515	**0.397**	0.349	**0.216**	0.237	**0.114**	0.160	**0.060**	0.187	**0.126**	0.402	**0.160**	0.398	0.293	0.129	0.071

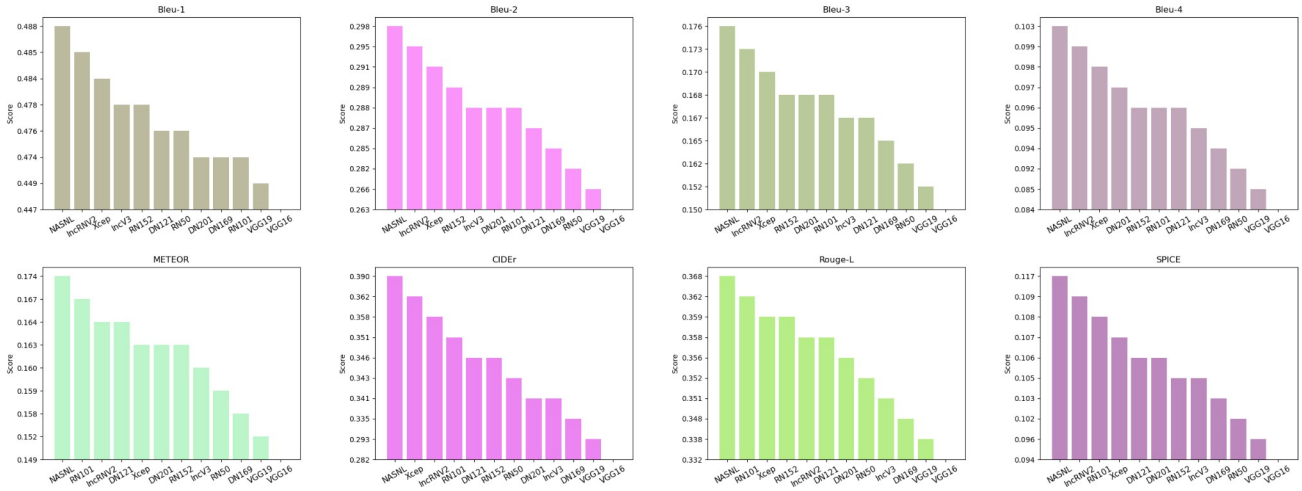
الجدول 3.4. نتائج التجارب على Flickr30k dataset.



الشكل 2.4. مخطط يوضح النتائج على Flickr30k مرتبة.

Model (Train/Test)	BLEU−1		BLEU−2		BLEU−3		BLEU−4		METEOR		CIDEr		ROUGE−L		SPICE	
VGG16	0.499	0.447	0.318	0.263	0.196	0.150	0.119	0.084	0.169	0.149	0.391	0.282	0.367	0.332	0.112	0.094
VGG19	0.502	0.449	0.321	0.266	0.199	0.152	0.121	0.085	0.174	0.152	0.409	0.293	0.374	0.338	0.117	0.096
ResNet50	0.534	0.476	0.351	0.282	0.224	0.162	0.141	0.092	0.194	0.159	0.508	0.343	0.404	0.352	0.134	**0.102**
ResNet101	0.525	0.474	0.344	0.288	0.219	0.168	0.137	0.096	0.190	0.167	0.488	0.351	0.396	**0.362**	0.132	**0.108**
ResNet152	0.537	0.478	0.353	0.289	0.225	0.168	0.142	0.096	0.190	0.163	0.506	0.346	0.399	**0.359**	0.133	0.105
Inception V3	0.518	0.478	0.334	0.288	0.210	0.167	0.130	0.095	0.186	0.160	0.473	0.341	0.392	0.351	0.127	0.105
Xception	0.530	**0.484**	0.349	**0.291**	0.223	**0.170**	0.141	**0.098**	0.189	0.163	0.510	**0.362**	0.398	0.359	0.133	0.107
InceptionResNet V2	0.519	**0.485**	0.338	**0.295**	0.211	**0.173**	0.129	**0.099**	0.181	**0.164**	0.449	**0.358**	0.388	0.358	0.124	**0.109**
DenseNet121	0.518	0.476	0.333	0.287	0.207	0.167	0.127	0.096	0.181	**0.164**	0.450	0.346	0.384	0.358	0.124	0.106
DenseNet169	0.507	0.474	0.327	0.285	0.206	0.165	0.127	0.094	0.186	0.158	0.449	0.335	0.387	0.348	0.128	0.103
DenseNet201	0.520	0.474	0.336	0.288	0.212	0.168	0.132	0.097	0.178	0.163	0.453	0.341	0.384	0.356	0.122	0.106
NASNetLarge	0.572	**0.488**	0.394	**0.298**	0.265	**0.176**	0.177	**0.103**	0.205	**0.174**	0.605	**0.390**	0.428	**0.368**	0.147	**0.117**

الجدول 4.4. نتائج التجارب على MSCOCO dataset.



الشكل 3.4. مخطط يوضح النتائج على MSCOCO مرتبة.

من هذه التجارب نستطيع تقديم جدول من نماذج CNN المقترحة لتحسين نتيجة التقييم في كل واحد من المقاييس (الجدول 5.4)، حيث نقترح نموذجاً ما لمعيار معين إذا حافظ هذا النموذج على تصنيف ضمن أفضل ستة نماذج باختلاف مجموعة البيانات بعد ترتيب النتائج.

Metric	Recommended Models
BLEU-1	Xception, InceptionResNet V2
BLEU-2	ResNet101, Xception, InceptionResNet V2
BLEU-3	ResNet101, Xception, InceptionResNet V2
BLEU-4	ResNet101, InceptionResNet V2
METEOR	ResNet101, NASNetLarge
CIDEr	Xception, InceptionResNet V2
Rouge-L	Xception, InceptionResNet V2
SPICE	ResNet101, NASNetLarge, Xception

الجدول 5.4. خلاصة نماذج CNN المرشحة لتحسين كل من معايير التقييم.

3.4 المرحلة الثالثة من التجارب

الهدف من هذه المرحلة اختبار النماذج المقترحة بهدف الوصول إلى النموذج النهائي.
يعرض الجدول 6.4 نتائج التجارب على النسخة 1 (Double Word Embedding).

Model (Train)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Dictionary size	Dataset
Model 1 (Double Word Embedding) + ResNet152	0.593	0.435	0.307	0.210	0.224	0.598	0.462	0.170	10,000	Flickr8k
Model 1 (Double Word Embedding) + ResNet152	0.562	0.409	0.287	0.196	0.228	0.576	0.456	0.174	15,000	Flickr8k

الجدول 6.4. التجارب على النسخة 1 من النموذج (Double Word Embedding).

توضح نتائج التجربتين السابقتين انخفاض معايير التقييم عند استخدام طبقتين من التضمين embedding، يمكن تفسير هذا بالتعلم المفرط overfitting بسبب عدد الطبقات الزائد. تقود هذه النتيجة إلى الثبات على طبقة واحدة من التضمين embedding كما في المراجع السابقة.

يعرض الجدول 7.4 نتائج التجارب على النسخة 2 من النموذج (YOLO Bounding Boxes).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Dictiona-ry Size	Dataset	Time
Model 2 (YOLO Bounding Boxes) + Xception	0.544	0.361	0.234	0.150	0.190	0.522	0.403	0.132	15,000	MSCOCO – training	09:59:24
Model 2 (YOLO Bounding Boxes) + Xception	0.474	0.288	0.170	0.099	0.167	0.361	0.361	0.112	15,000	MSCOCO – validation	16:54:39

الجدول 7.4. نتائج التجارب على النسخة 2 من النموذج (YOLO Bounding Boxes).

توضح التجربتان السابقتان ارتفاعاً كبيراً في دقة النتائج عند استخدام معلومات object detection على شكل tags مرتبة حسب عامل الأهمية المقدم في النموذج المقترح. يمكن تفسير هذا بأهمية المعلومة الإضافية عن أنواع الأغراض وأماكنها في الصورة في دقة التوصيف، حيث يحاكي الأمر ما يفكر به البشر.

في ما يأتي يعرض الجدول 8.4 نتائج التجارب على النسخة 3 (YOLO Raw Features v1).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Diction-ary Size	Dataset	Time
Model 3 (YOLO Raw Features v1) + Xception	0.242	0.052	0.007	0.001	0.057	0.010	0.154	0.009	15,000	MSCOCO dataset – training	11:24:19
Model 3 (YOLO Raw Features v1) + Xception	0.250	0.056	0.007	0.001	0.061	0.011	0.162	0.010	15,000	MSCOCO dataset – validation	19:53:25

الجدول 8.4. نتائج التجارب على النسخة 3 من النموذج (YOLO Raw Features v1).

قادت هاتان التجربتان إلى معرفة مدى سوء تجميع السمات غير المعالجة من YOLOv4 مع سمات CNN على شكل شعاع، حيث تشير النتائج إلى انخفاض كبير في دقة التوصيف.

تجربة للنسخة 4 (YOLO Raw Features v2)، باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على قسم التدريب MSCOCO training split. جرى التدريب على 80%. من القسم والتقييم على 20%. ولم نستخدم مجموعة التحقق، بزمن تنفيذ 18:55:24 (حوالي 19 ساعة). كانت النتائج كما يأتي:

Bleu-1: 0.547

Bleu-2: 0.360

Bleu-3: 0.232

Bleu-4: 0.148

METEOR: 0.197

CIDEr: 0.544

ROUGE-L: 0.411

SPICE: 0.139

تجربة للنسخة 4 (YOLO Raw Features v2). باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على قسم التدريب MSCOCO training split. جرى التدريب على كل مجموعة التدريب والتقييم على كل مجموعة التحقق، بزمن تنفيذ 31:35:36 (أكثر من 31 ساعة). كانت النتائج كما يأتي:

Bleu-1: 0.247

Bleu-2: 0.131

Bleu-3: 0.059

Bleu-4: 0.028

METEOR: 0.116

CIDEr: 0.057

ROUGE-L: 0.245

SPICE: 0.049

في التجربتين السابقتين، أظهرت الأولى نتيجة قريبة من نتيجة تجربة النسخة 2 (YOLO Bounding Boxes) على قسم التدريب MSCOCO training split. لكن في مرحلة الاختبار أظهرت الثانية نتيجة أقل بكثير، يمكن تفسير هذا بقلة قدرة النموذج على التعميم في حال استخدام السمات غير المعالجة من YOLOv4، ولو أن طريقة دمج سمات YOLOv4 و Xception في هذا النموذج أفضل من طريقة الشعاع في النسخة 3 (YOLO Raw Features v1).

تجربة للنسخة 2 (YOLO Bounding Boxes). باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على قسم التدريب MSCOCO training split وبدون عامل الأهمية لمعرفة مدى تأثيره. جرى التدريب على كل مجموعة التدريب والتقييم على كل مجموعة التحقق، بزمن تنفيذ 17:44:57 (أكثر من 17 ساعة). كانت النتائج كما يأتي:

Bleu-1: 0.489

Bleu-2: 0.294

Bleu-3: 0.173

Bleu-4: 0.100

METEOR: 0.162

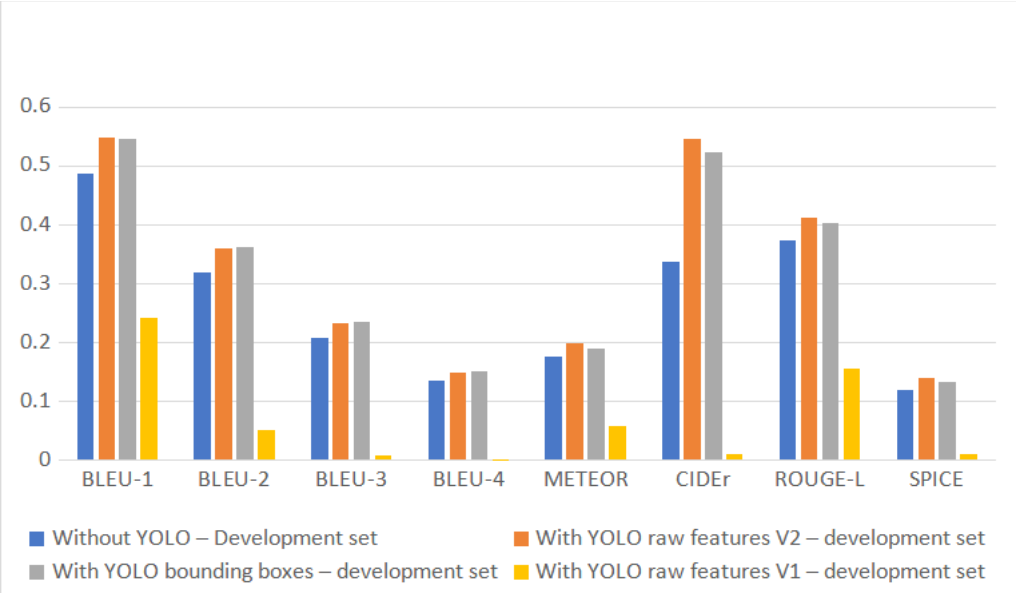
CIDEr: 0.376

ROUGE−L: 0.359

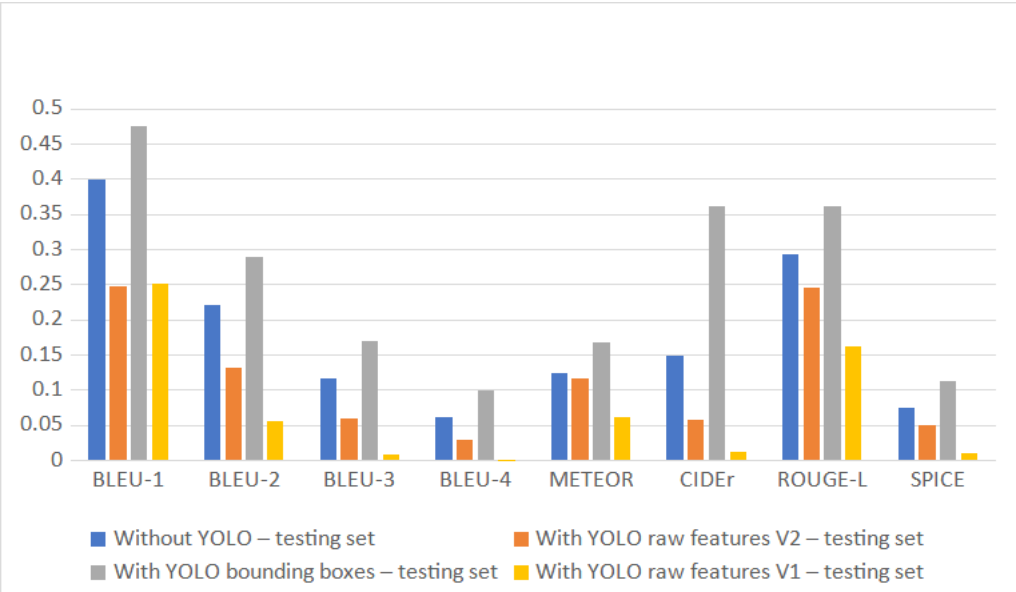
SPICE: 0.107

تبين هذه التجربة أن هذا العامل يرفع تقييمات METEOR و ROUGE−L و SPICE ويخفض الباقي.

نوضح في الشكلين الآتيين (الشكلان 4.4 و 5.4) نتائج هذه الاختبارات بطريقة clustered bar chart.



الشكل 4.4. مخطط يوضح النتائج على مجموعة التحقق MSCOCO – development set للنماذج المقترحة.



الشكل 5.4. مخطط يوضح النتائج على مجموعة الاختبار MSCOCO – testing set للنماذج المقترحة.

4.4 المرحلة الرابعة من التجارب

الهدف من هذه المرحلة مقارنة النتائج مع نتائج الأبحاث السابقة حسب التقسيمات المستخدمة فيها.

تجربة للنسخة 2 (YOLO Bounding Boxes). باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على MSCOCO dataset بدون عامل الأهمية. جرى التدريب على مجموعتي التدريب والتحقق معاً والتقييم على مجموعة الاختبار، بزمن تنفيذ 14:11:26 (أكثر من 14 ساعة). استخدمنا تقسيم Karpathy [31]. كانت النتائج كما يأتي:

Bleu-1: 0.486

Bleu-2: 0.293

Bleu-3: 0.173

Bleu-4: 0.099

METEOR: 0.164

CIDEr: 0.390

ROUGE-L: 0.358

SPICE: 0.108

تجربة للنسخة 2 (YOLO Bounding Boxes). باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على MSCOCO dataset مع عامل الأهمية. جرى التدريب على مجموعتي التدريب والتحقق معاً والتقييم على مجموعة الاختبار، بزمن تنفيذ 14:01:18 (أكثر من 14 ساعة). استخدمنا تقسيم Karpathy [31]. كانت النتائج كما يأتي:

Bleu-1: 0.492

Bleu-2: 0.296

Bleu-3: 0.174

Bleu-4: 0.101

METEOR: 0.163

CIDEr: 0.390

ROUGE-L: 0.358

SPICE: 0.108

نلاحظ من التجربتين السابقتين بقاء تقييم CIDEr على حاله من دون نقصان، وهو ما يختلف عن بحث Herdade وآخرين [35]، حيث قالوا أن طرق ترتيب الأغراض المكتشفة positional encoding methods الآلية (أي التي من ابتكار البشر) تؤدي إلى نقصان في تقييم CIDEr. جربوا في بحثهم عدداً من الطرق المصطنعة وهي الترتيب حسب حجم الصندوق المحيط بالغرض، وترتيب الأغراض من اليسار إلى اليمين، والترتيب من فوق إلى تحت، وقارنوها مع ترك الأغراض بدون ترتيب وكلها أدت إلى نقصان في تقييم CIDEr.

تجربة للنسخة 2 (YOLO Bounding Boxes). باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على Flickr30k dataset مع عامل الأهمية. جرى التدريب على مجموعتي التدريب والتحقق معاً والتقييم على مجموعة الاختبار، بزمن تنفيذ 04:53:36 (حوالي 5 ساعات). كانت النتائج كما يأتي:

Bleu-1: 0.398

Bleu-2: 0.221

Bleu-3: 0.116

Bleu-4: 0.061

METEOR: 0.129

CIDEr: 0.150

ROUGE-L: 0.298

SPICE: 0.074

يتضمن الجدول 10.4 نتائج التجربة السابقة ونتائج التجربة مع المرحلة الثانية من التجارب التي أجريت على النموذج الأولي (Baseline)ـ باستخدام مجموعة بيانات Flickr30k ونموذج Xcpetion مدرب مسبقاً على مجموعة بيانات ImageNet. يمكننا رؤية تأثير أسلوبنا في استخراج السمات على جودة التوصيف باستخدام مجموعة بيانات Flickr30k حيث ساهم في رفع تقييم ROUGE-L و CIDEr و METEOR وبقيت قيمة معيار SPICE على حالها، وهي المعايير التي لها أكبر توافقية مع تحكيم الجودة البشري وتدل على زيادة السلامة اللغوية والدلالية للتوصيفات الناتجة. أما معايير BLEU فقد تغيرت كلها بمقدار ضئيل، حيث زادت قيمة معيار BLEU-2 وقلت قيم BLEU-1 و BLEU-3 و BLEU-4. نلاحظ أن التغيير كان أكثر وضوحاً في حال استخدام MSCOCO.

تجربة للنسخة 0 من النموذج (Baseline)ـ باستخدام عدد كلمات Dictionary Size من 15,000 كلمة، باستخدام Xception ك feature extraction model على MSCOCO dataset. جرى التدريب على مجموعتي التدريب والتحقق معاً والتقييم على مجموعة الاختبار، بزمن تنفيذ 17:44:31 (أكثر من 17 ساعة). استخدمنا تقسيم Karpathy [31]. كانت النتائج كما يأتي:

Bleu-1: 0.463

Bleu-2: 0.273

Bleu-3: 0.156

Bleu-4: 0.087

METEOR: 0.157

CIDEr: 0.339

ROUGE-L: 0.345

SPICE: 0.102

أنتجت هذه التجارب تقييمات مقاربة لمثيلاتها في المرحلة الثالثة من التجارب. حسب تقسيم [31] Karpathy، أدى عامل الأهمية إلى زيادة في تقييمات BLEU من دون نقصان في المعايير الأخرى، ما عدا METEOR بنقص صغير جداً مقداره 0.001. نلخص المقارنة مع بحثين سابقين في الجدولين الآتيين (الجدولان 9.4 و 10.4):

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
Yin and Ordonez [3] baseline model	NA	NA	NA	0.21	0.215	0.759	0.464	NA
Yin and Ordonez [3] results with object features	NA	NA	NA	0.253	0.238	0.922	0.507	NA
Yin and Ordonez [3] increase	NA	NA	NA	0.043	0.023	0.163	0.043	NA
Our increase	0.029	0.023	0.018	0.014	0.006	0.051	0.013	0.006

الجدول 9.4. مقارنة مع بحث [3] Yin and Ordonez على karpathy split.

نفسر الفرق بين نتائجنا ونتائج البحث الآخر في الجدول 9.4 بأن بحثهم أعطى وزناً أكبر لسمات استخراج الأغراض واستخدم طرقاً أكثر تعقيداً في ترميز السمات. حيث يقوم نموذجهم باستخراج سمات الأغراض ثم تمريرها في نموذج ترميز LSTM، واستخراج سمات CNN ثم تمريرها في نموذج ترميز LSTM آخر، ويجمع الشعاعين الناتجين خطياً ليعطي شعاع السمات النهائي.

يبين الجدول 10.4 مقارنة مع بحث Sharif وآخرين. نفسر الفرق بين بحثنا وبحثهم بفائدة استخدامهم للسمات الإضافية التي استخرجوها من نصوص التوصيفات وأدخلوها لنظام التوصيف.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
Sharif et al. [1] Baseline model	0.4368	NA	NA	NA	0.1297	0.2517	0.2997	0.0700
Sharif et al. [1] Suggested model	0.4462	NA	NA	NA	0.1350	0.2835	0.3116	0.0741
Our baseline model	0.3990	0.2200	0.1170	0.0620	0.1230	0.1480	0.2930	0.0740
Our model (with the importance factor)	0.3980	0.2210	0.1160	0.0610	0.1290	0.1500	0.2980	0.0740
Sharif et al. increase	**0.0094**	NA	NA	NA	0.0053	**0.0318**	**0.0119**	**0.0041**
Our increase	−0.0010	**0.0010**	**−0.0010**	**−0.0010**	**0.0060**	0.0020	0.0050	0

الجدول 10.4. مقارنة مع بحث [1] .Sharif et al.

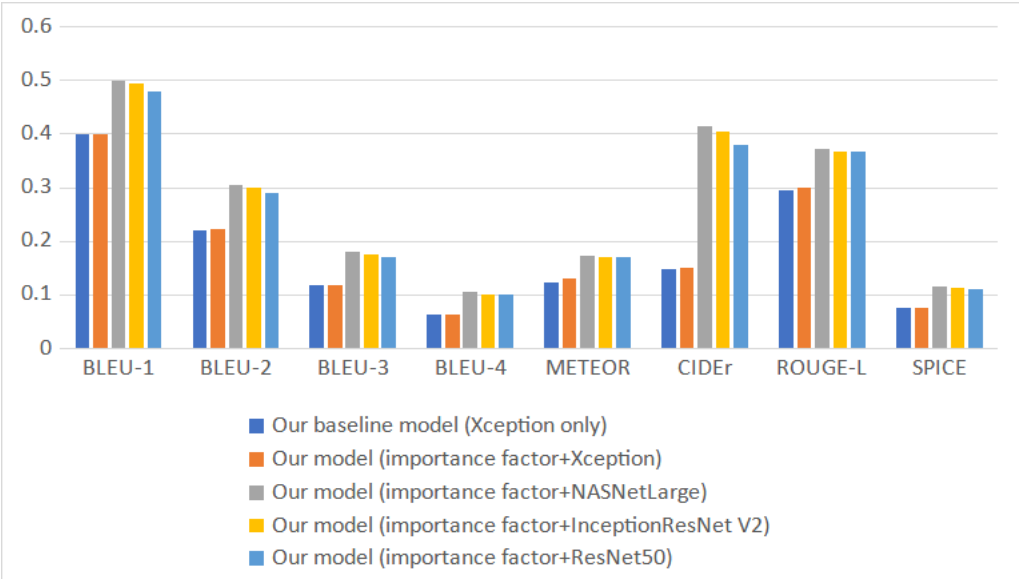
5.4 المرحلة الخامسة من التجارب

هدف هذه المرحلة تجريب النسخة 2 من النموذج (YOLO Bounding Boxes)- مع NASNetLarge و InceptionResNet V2 و ResNet50 لأن هذه النماذج كانت من بين الأفضل على عدد من معايير التقييم كما هو مبين في الجدول 4.4.. جميع تجارب هذه المرحلة استخدم فيها عامل الأهمية Importance Factor المقترح في هذا البحث، وحجم قاموس Dictionary Size من 15,000 كلمة. أجريت هذه التجارب على مجموعة بيانات MSCOCO حسب تقسيم [31] Karpathy، حيث دُرّب النموذج على مجموعتي التدريب والتحقق معاً وقُيّم على مجموعة الاختبار. يظهر الجدول 11.4 مقارنة بين نتائج هذه التجارب.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Time
Our baseline model (Xception only)	0.399	0.220	0.117	0.062	0.123	0.148	0.293	0.074	NA
Our model (Importance Factor+Xception)	0.398	0.221	0.116	0.061	0.129	0.150	0.298	0.074	14:01:18
Our model (Importance Factor+NASNetLarge)	**0.498**	**0.303**	**0.178**	**0.104**	**0.172**	**0.413**	**0.371**	**0.115**	**18:10:12**
Our model (Importance Factor+InceptionResNet V2)	0.493	0.298	0.174	0.100	0.169	0.403	0.367	0.113	12:29:27
Our model (Importance Factor+ResNet50)	0.479	0.289	0.170	0.099	0.169	0.378	0.367	0.110	11:33:15

الجدول 11.4. مقارنة بين عدة نماذج CNN في النسخة 2 من النموذج (YOLO Bounding Boxes).

بينت هذه المجموعة من التجارب قوة NASNetLarge بالنسبة لبقية نماذج CNN المجربة، ونعزو هذا إلى قدرة منهجية NASNet على تعديل بنيتها بما يتناسب مع مجموعة البيانات، لكن لم تؤدّ هذه التجارب إلى زيادة الفرق في الدقة بين حالة استخدام سمات الأغراض (التي تتضمن رقم صنف الغرض وإحداثيتيه وبعديه وقيمة عامل الأهمية) وحالة عدم استخدامها. يبين الشكل 6.4 هذه النتائج.



الشكل 6.4. مخطط يوضح مقارنة بين عدة نماذج CNN في النسخة 2 من النموذج (YOLO Bounding Boxes).

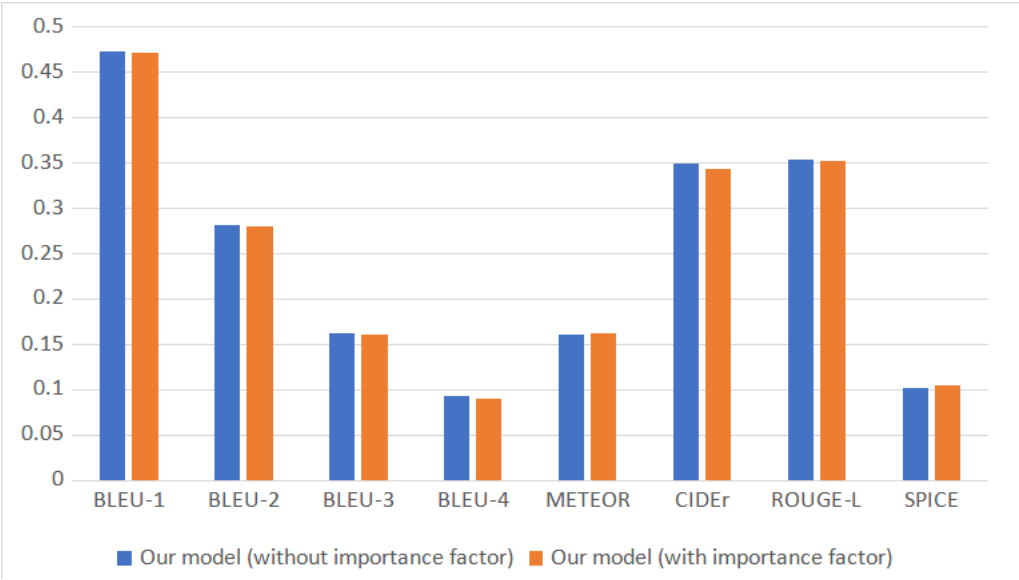
6.4 المرحلة السادسة من التجارب

الهدف من هذه المرحلة تجريب أداء النسخة 2.1 (YOLO Bounding Boxes).. أجريت تجربتا هذه المرحلة بحجم قاموس من 15,000 كلمة واستخدام Xception لاستخراج السمات إلى جانب سمات الأغراض. استخدمنا مجموعة بيانات MSCOCO حسب تقسيم [31] Karpathy. جرى التدريب على مجموعتي التدريب والتحقق معاً والتجريب على مجموعة الاختبار. نورد نتائج هذه المرحلة في الجدول 12.4.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE	Time
Model 2.1 (No importance factor)	0.472	0.281	0.161	0.092	0.160	0.348	0.353	0.101	13:12:25
Model 2.1 (With importance factor)	0.470	0.280	0.160	0.090	0.161	0.342	0.352	0.104	13:14:17

الجدول 12.4. نتائج التجريب على النسخة 2.1 (YOLO Bounding Boxes v2).

بينت هاتان التجربتان قلة فعالية النسخة 2.1، يمكن إرجاع هذا إلى نقص معلومات السياق لأن الكلمة المعبرة عن صنف الغرض يمكن أن ترد في سياقات مختلفة عديدة، وهذا هو النموذج المقترح الوحيد في هذا البحث الذي يستفيد من الكلمة المعبرة عن صنف الغرض بدلاً من رقمه. يبين الشكل 7.4 مقارنة بين نتائج هاتين التجربتين.



الشكل 7.4. مخطط يوضح مقارنة بين تجربتي المرحلة السادسة.

النسخة 0: النموذج الأولي Baseline

النسخة 1: هو النموذج الأولي مع طبقتين embedding بدلاً من واحدة

النسخة 2: مع YOLO Bounding Boxes v1

النسخة 2.1: مع YOLO Bounding Boxes v2

النسخة 3: مع YOLO Raw Features v1

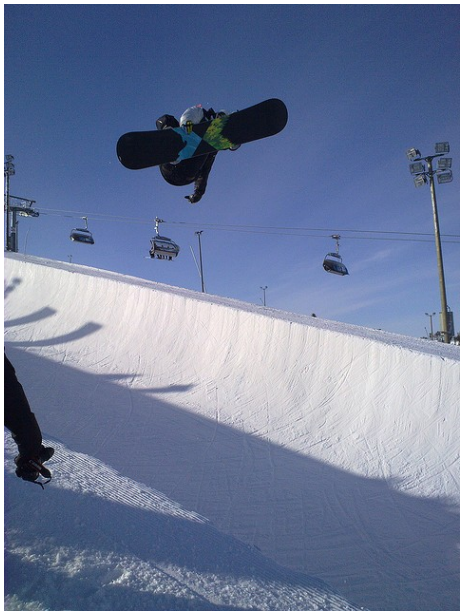
النسخة 4: مع YOLO Raw Features v2

نعرض في الجدول 13.4 مقارنة نهائية بين الجداول المقترحة، حسب تقسيم [31] Karpathy الذي يخصص 5,000 صورة للاختبار و 5,000 صورة للتحقق وباقي الصور المنمطة للتدريب. نلاحظ تفوق النسخة 2 من النموذج (YOLO Bounding Boxes) على بقية النماذج، وخصوصاً في معيار CIDEr.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
Without YOLO – testing set	0.399	0.22	0.117	0.062	0.123	0.148	0.293	0.07
With YOLO Raw Features v1 – testing set	0.25	0.056	0.007	0.001	0.061	0.011	0.162	0.01
With YOLO Raw Features v2 – testing set	0.247	0.131	0.059	0.028	0.116	0.057	0.245	0.05
With YOLO Bounding Boxes – testing set	0.474	0.288	0.17	0.099	0.167	0.361	0.361	0.11

الجدول 13.4. نتائج الاختبار على مجموعة اختبار MSCOCO.

نورد هنا مقارنة بين نتائج النسخة 0 (Baseline). ونتائج النسخة 2 (YOLO Bounding Boxes).. نلاحظ أن سمات الأغراض تسهم في زيادة جودة التوصيفات من حيث التعرف على الأغراض في الصورة والسلامة اللغوية (الشكل 8.4)، نورد في الشكل أيضاً التوصيفات الخمسة للصورة من مجموعة بيانات MS COCO من أجل المقارنة معها.



Baseline model: this is up in snow pants jumping on a big snowy mountain at night.

With object features: a skier performing a jump against some snow.

Reference caption 1: A snowboarder is in the air against the blue sky.

Reference caption 2: A snow boarder coming airborne while riding his snow board down a snowy slope.

Reference caption 3: A ski boarder is airborne on a snow covered slope.

Reference caption 4: A snowboarder catches air going down a hill.

Reference caption 5: The snowboarder is mid-jump, flying through the air.

(أ)

Baseline model: two people on skis sitting on a snowy surface.

With object features: a person standing next to snowboards attached.

Reference caption 1: A person is standing in the snow near a tree with skis and snowboards.

Reference caption 2: Snowboards resting upon a tree, with man hiding inside it like fort

Reference caption 3: A person holds a snowboard in front of a tree with snowboards leaning on it.

Reference caption 4: a person holding some skis walking through the snow

Reference caption 5: a a couple of snowboards are up against a tree

(ب)

Baseline model: a cow is standing in a open field as it grazes.

With object features: cows eat alone grazing on grasses in a hill.

Reference caption 1: A photograph of two cows grazing on a green pasture.

Reference caption 2: two cows with spots are grazing in some grass

Reference caption 3: Two black and white cows grazing in a grassy pasture.

Reference caption 4: Two black and white cows are in a grass field.

Reference caption 5: Two jersey cows grazing in a green meadow.

(ج)

Baseline model: man walking next to an old fashioned planes.

With object features: a small black and white picture of a prop plane sitting on the runway.

Reference caption 1: A black and white photo of a propeller plane.

Reference caption 2: A fighter jet on a runway with a cover over the cockpit.

Reference caption 3: A plane is covered while sitting at an airport

Reference caption 4: There is an old plane sitting on the runway.

Reference caption 5: A light aircraft parked and covered on the tarmac

(د)

Baseline model: a brown bears perch in front of their mom and another animal.

With object features: a brown bear is standing behind a group of brown bears.

Reference caption 1: Three brown bears looking out a cage at the ground below.

Reference caption 2: Three bears stand together near a fence.

Reference caption 3: A family of bears in front of a fence at a humane facility.

Reference caption 4: Three brown bears are looking outside their enclosure.

Reference caption 5: brown bears standing around looking at a metal fence

(هـ)

Baseline model: two women make homemade my diners can be judged on a table.

With object features: a group of people sitting at a blue table of food.

Reference caption 1: A group of people sitting and standing around a pot of food.

Reference caption 2: A group of people sampling food at an outdoor event.

Reference caption 3: people at an informal gathering eating under a tent

Reference caption 4: A group of people have dinner together inside of a tent.

Reference caption 5: A large group of people eating under a tent.

(و)

الشكل 8.4. مقارنة بين النتائج قبل إضافة سمات الأغراض وبعدها على بعض الصور.

الفصل الخامس: الخاتمة والآفاق المستقبلية

قمنا في هذه الأطروحة بدراسة نظم توصيف الصور المعتمدة على التعلم العميق، وقدمنا دراسة مرجعية لأحدث الأعمال التي تتناول نظم توصيف الصور المعتمدة على سمات CNN والسمات المستخرجة من نماذج كشف الأغراض في الصورة. وجدنا أن الأعمال قليلة في مجال استخدام سمات كشف الأغراض مع السمات المستخرجة من الطبقات التلافيفية الأخيرة للشبكات العميقة CNN.

أجرينا بدايةً تجارب لتوضيح أثر اختيار بنية CNN على جودة السمات المستخرجة بهدف التوصيف الآلي للصور، وبيّنا الترابط الكبير بين بنية نموذج CNN وطبيعة الصور في التجربة وأوضحنا تأثير هذا على معايير التقييم المستخدمة في مجال التوصيف الآلي للصور. من ثم اقترحنا قائمة من النماذج المدربة على ImageNet لتحسين التقييم لكل واحد من المقاييس، حيث تتكون القائمة من النماذج التي حافظت على أداء جيد مع تغيير مجموعات البيانات.

بعد ذلك اقترحنا نموذجاً للتوصيف الآلي للصور تعتمد آلية الانتباه attention وبنية Encoder–Decoder تتكون من جزأين أساسيين: جزء فهم الصورة وترميزها Encoder وجزء التوليد اللغوي Decoder. طورنا في جزء ترميز الصورة أسلوباً جديداً لاستخدام سمات كشف الأغراض في الصورة وسمات CNN. وبعد إجراء التجارب، اقترحنا النموذج النهائي الذي يعتمد على استخدام جميع سمات كشف الأغراض مع سمات CNN. تتميز النماذج المقترحة في هذا البحث ببساطتها واستقلالية الأجزاء المكونة لها عن بعضها، مما يمكن النظم المستقبلية من الاستفادة من أسلوب الترميز الذي قدمناه. أخيراً اقترحنا أسلوباً جديداً لترتيب الأغراض المكتشفة في الصورة Positional Encoding، وبيّنا أثره على معايير التقييم.

من أبرز الصعوبات التي واجهتنا في هذا البحث إيجاد نماذج مدربة مناسبة لاستخراج السمات وكشف الأغراض، والكلفة الحسابية العالية للتجارب بسبب تعقيد العمليات المطلوبة والحجم الكبير للمعطيات الواجب معالجتها.

نريد تحسين نموذجنا كمتابعة في المستقبل من أجل رفع دقة التقييمات، وذلك باتباع الطرق الآتية:

- دراسة لتطوير مكون التوليد اللغوي، بهدف توليد نصوص بجودة أعلى.
- استخدام أساليب ترميز وفك ترميز أكثر تطوراً، مثل أسلوب [45] Meshed-memory.
- إيجاد العلاقة السببية بين جودة السمات المستخرجة من بنية محددة لنموذج CNN وطبيعة الصور المعالجة، وذلك بهدف إيجاد بنية CNN أفضل لطبيعة مسألة توصيف الصور.
- استخدام نماذج كشف المسميات المدربة مسبقاً في استخراج معلومات إضافية من نصوص توصيفات الصور بهدف إغناء السمات المدخلة إلى مكون التوليد اللغوي وبالتالي زيادة دقة التوصيف.

الملاحق

الملحق أ: التقنيات المستخدمة

يقدم هذا الجزء البرمجيات المستخدمة في نظام توصيف الصور بالتعلم العميق.

أ.1 مكتبة TensorFlow 2

TensorFlow[1] منصة برمجية شاملة مفتوحة المصدر للتعلم الآلي، تتضمن بيئة شاملة ومرنة من الأدوات والمكتبات ولها مجتمع تطوير كبير يدعمها. تتيح للباحثين استخدام التكنولوجيا الحديثة في تعلم الآلة وتمكن المطورين من بناء ونشر التطبيقات التي تستخدم التعلم الآلي بسهولة. يمكن استخدامها لمجموعة كبيرة من المهام لكن لها تركيز خاص على تدريب واستخدام الشبكات العصبونية العميقة، وهي أيضاً مكتبة رياضيات ترميزية Symbolic Mathematics تعتمد على تدفق البيانات والبرمجة القابلة للتفاضل. تستخدم في البحث والتطوير البرمجي في Google.

طور TensorFlow فريق Google Brain للاستخدام الداخلي في Google، ثم أصدرت حسب ترخيص Apache 2.0 في عام 2015.

أ.2 مكتبة Keras

Keras[2] هي مكتبة برمجية مفتوحة المصدر توفر واجهة Python للشبكات العصبونية. تعمل Keras كواجهة لمكتبة TensorFlow.

1 https://www.tensorflow.org

2 https://keras.io

دعمت Keras العديد من المكتبات حتى الإصدار 2.3، بما في ذلك TensorFlow و Microsoft Cognitive Toolkit و Theano و PlaidML. أما اعتباراً من الإصدار 2.4 فتدعم مكتبة TensorFlow فقط. صممت لتمكين التجريب السريع للشبكات العصبونية العميقة، وتركز على سهولة الاستخدام وقابلية التوسيع.

أ.3 MSCOCO Evaluation Toolkit

هي مكتبة ملحقة بمجموعة بيانات MSCOCO[3] مصممة لتقييم جودة التوصيفات المنتجة حسب معايير التقييم الآتية: SPICE، ROUGE-L، CIDEr، METEOR، BLEU-4، BLEU-3، BLEU-2، BLEU-1. تتوفر هذه المكتبة فقط للغة Python 2. قبل تشغيل كود التقييم، يجب تحضير النتائج في ملف حسب الصيغة التي تتوقعها المكتبة. تنتج من تشغيل كود التقييم بنيتان من البيانات تلخصان جودة توصيف الصور مكتوبتين بصيغة JSON في ملف.

أ.4 مكتبة Numpy

هي مكتبة للغة Python تتضمن دعماً للمصفوفات والمصفوفات الكبيرة متعددة الأبعاد مع مجموعة كبيرة من التوابع الرياضية عالية المستوى للعمل على هذه المصفوفات[4].

3 https://github.com/tylin/coco-caption

4 https://numpy.org

أ.5 مكتبة Matplotlib

Matplotlib[5] هي مكتبة لرسم المخططات للغة Python ومكتبتها الرياضية NumPy. توفر واجهة برمجة تطبيقات غرضية التوجه لإنشاء الرسوم البيانية في التطبيقات باستخدام مكتبات GUI عامة مثل Tkinter و wxPython و Qt و GTK+.

أ.6 مكتبة yolov4

هي مكتبة Python تحوي تطبيقاً لنموذج YOLOv4 مدرباً على مجموعة بيانات MSCOCO ومكتوباً باستخدام TensorFlow 2. تدعم المكتبة نسختي Python 2 و Python 3[6].

أ.7 Hardware

أجريت التجارب على مخدم المعهد المجهز ب 32 GB من الذاكرة ووحدة CPU من نوع Intel Corei9-9900K ووحدة GPU من نوع NVIDIA GeForce RTX 2080 واستخدمت في هذا البحث مكتبة CUDA لاستخدام GPU في المعالجة على التفرع.

5 https://matplotlib.org

6 https://pypi.org/project/yolov4

المراجع

[1] Sharif N, Jalwana MA, Bennamoun M, Liu W, Shah SA. Leveraging Linguistically-aware Object Relations and NASNet for Image Captioning. In 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ) 2020 Nov 25 (pp. 1-6). IEEE.

[2] Vo-Ho VK, Luong QA, Nguyen DT, Tran MK, Tran MT. A Smart System for Text-Lifelog Generation from Wearable Cameras in Smart Environment Using Concept-Augmented Image Captioning with Modified Beam Search Strategy. Applied Sciences. 2019 Jan;9(9):1886.

[3] Yin X, Ordonez V. Obj2text: Generating visually descriptive language from object layouts. arXiv preprint arXiv:1707.07102. 2017 Jul 22.

[4] Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR). 2019 Feb 4;51(6):1-36.

[5] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25:1097-105.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[8] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1-9).

[9] Girshick R. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision 2015 (pp. 1440-1448).

[10] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 580-587).

[11] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015;201.

[12] Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In NIPs 1999 Nov 29 (Vol. 99, pp. 1057-1063).

[13] Ranzato MA, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732. 2015 Nov 20.

[14] Chen X, Lawrence Zitnick C. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 2422-2431).

[15] Luo Y, Huang Z, Zhang Z, Wang Z, Li J, Yang Y. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In Proceedings of the 27th ACM International Conference on Multimedia 2019 Oct 15 (pp. 2341-2350).

[16] Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 7008-7024).

[17] Johnson J, Karpathy A, Fei-Fei L. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4565-4574).

[18] Ma S, Han Y. Describing images by feeding LSTM with structural words. In 2016 IEEE International Conference on Multimedia and Expo (ICME) 2016 Jul 11 (pp. 1-6). IEEE.

[19] Wang M, Song L, Yang X, Luo C. A parallel-fusion RNN-LSTM architecture for image caption generation. In 2016 IEEE International Conference on Image Processing (ICIP) 2016 Sep 25 (pp. 4448-4452). IEEE.

[20] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning 2015 Jun 1 (pp. 2048-2057). PMLR.

[21] Jin J, Fu K, Cui R, Sha F, Zhang C. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272. 2015 Jun 20.

[22] Wu ZY, Cohen RS. Encode, review, and decode: Reviewer module for caption generation. arXiv preprint arXiv:1605.07912. 2016 May;3.

[23] Pedersoli M, Lucas T, Schmid C, Verbeek J. Areas of attention for image captioning. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 1242-1250).

[24] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 375-383).

[25] Liu C, Mao J, Sha F, Yuille A. Attention correctness in neural image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence 2017 Feb 12 (Vol. 31, No. 1).

[26] Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 5659-5667).

[27] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
2487–2496.

[28] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 6077-6086).

[29] Chunseong Park C, Kim B, Kim G. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 895-903).

[30] Sugano Y, Bulling A. Seeing with humans: Gaze-assisted neural image captioning. arXiv preprint arXiv:1608.05203. 2016 Aug 18.

[31] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3128-3137).

[32] Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 4894-4902).

[33] Yao T, Pan Y, Li Y, Mei T. Incorporating copying mechanism in image captioning for learning novel objects. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 6580-6588).

[34] Lanzendörfer L, Marcon S, der Maur LA, Pendulum T. YOLO-ing the Visual Question Answering Baseline.

[35] Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: Transforming objects into words. arXiv preprint arXiv:1906.05963. 2019 Jun 14.

[36] Wang J, Madhyastha P, Specia L. Object counts! bringing explicit detections back into image captioning. arXiv preprint arXiv:1805.00314. 2018 Apr 23.

[37] Variš D, Sudoh K, Nakamura S. Image Captioning with Visual Object Representations Grounded in the Textual Modality. arXiv preprint arXiv:2010.09413. 2020 Oct 19.

[38] Holliday A, Dudek G. Pre-trained CNNs as Visual Feature Extractors: A Broad Evaluation. In 2020 17th Conference on Computer and Robot Vision (CRV) 2020 May 13 (pp. 78-84). IEEE.

[39] Valev K, Schumann A, Sommer L, Beyerer J. A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification. In Pattern Recognition and Tracking XXIX 2018 Apr 27 (Vol. 10649, p. 1064902). International Society for Optics and Photonics.

[40] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 590-597).

[41] Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, Ng AY, Lungren MP. CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. arXiv preprint arXiv:2002.11379. 2020 Feb 26.

[42] Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. arXiv preprint arXiv:2101.06871. 2021 Jan 18.

[43] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2018 Jul (pp. 2556-2565).

[44] Gu J, Wang G, Cai J, Chen T. An empirical study of language cnn for image captioning. In Proceedings of the IEEE International Conference on Computer Vision 2017 (pp. 1222-1231).

[45] Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 10578-10587).

Abstract

Image captioning is one of the trending problems in modern Artificial Intelligence (AI). It is concerned with generating an output text describing an input image, where the output can be one or more sentences. Image captioning is important for many reasons. For example, it can be used for automatic image indexing, which is important for many applications.

The problem of image captioning has been solved recently by deep learning techniques, especially Encoder-Decoder methods. In this thesis, we present an Encoder-Decoder attention-based architecture that makes use of convolutional features extracted from the Xception model pre-trained on ImageNet, and object features extracted from the YOLOv4 model, pre-trained on MSCOCO. We also introduce a new positional encoding scheme for object features, "the importance factor", and show its effect on evaluation scores. We test our model on the MSCOCO dataset and compare it to similar works.

We also present a thorough experimental study about feature extraction using Convolutional Neural Networks (CNNs) for the task of image captioning in the context of deep learning. We perform a set of 72 experiments using 3 datasets on 12 image classification CNNs pre-trained on the ImageNet dataset. We study the effect of changing the CNN feature extractor on image captioning quality, and find a strong relationship between the model structure and the image captioning dataset. To benefit from these results, we recommend a set of pre-trained CNNs for each of the image captioning evaluation metrics we want to optimise.

The evaluation scores are calculated using the eight standard metrics in the image captioning field. Our work contributes to image captioning by introducing better representation schemes for images.

Keywords: Image Captioning; Object Features; Convolutional Neural Network; Deep Learning; Feature Extraction

Syrian Arab Republic
Higher Institute for Applied Science and Technology
Department of Information Systems

Image Captioning Using Deep Learning Techniques

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master in Big Data Systems

By

Eng. Muhammad Abdelhadie Al-Malla

Supervisors

Dr. Assef Jafar **Dr. Nada Ghneim**

December 2021