Performance Evaluation on the Effect of Different Text Representation Models on the Image Captioning Systems

Dr. Jafar Alkheir*
Dr. Samer Sulaiman**
Rasha Mualla***

(Received 4 / 3 / 2020. Accepted 16 / 7 / 2020)

□ ABSTRACT □

This research deals with one of the most important and recent topics in the field of machine learning in general and deep learning in particular, which is image Captioning systems. In this research, an image-captioning system is built based on the ResNet50 model, which is a deep learning network based on convolution neural networks CNN, through which the features of the image representation are obtained. As for the textual representation, five different models are proposed, based mainly on the GloVe and FastText models provided by Twitter and Facebook, respectively. The effect of different vocabulary dictionaries on the performance of the proposed system is studied. A global MS-COCO dataset is used, from which a subset of 10,000 images is token, 9,000 images from them are chosen for the Training and validation group. While the testing process includes 1000 images varying from the training-set. This test-set is applied to the five designed models.

To find out the precision of the results used by the five proposed systems as well as how well they match between the original description sentences and the resulting description ones, performance measures are used such Accuracy, Average of Depth Similarity, Top-1, Top-5 and BLEU. The results show the superiority of systems based on FastText models although they take longer time than GloVe models.

Keywords: Deep Learning, Natural Language Processing, Image Representation, Text Representation, FastText Model, GloVe Model.

*Professor, Computer Science Faculty, Tishreen University, Latakia, Syria, Email: Alkheir.j@gmail.com

** Assistant Professor- Computer Engineering, Department of Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Latakia, Syria, Email: Samsulg@googlemail.com

***Postgraduate Student (PhD student) - Computer Engineering, Department of Computer and Automatic Control Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Latakia, Syria, Email: Rasha_Mualla90@hotmail.com

تقييم أثر اختلاف نموذج التمثيل النصي على أداء أنظمة وصف الصور

د. جعفر الخير*

د. سامر سليمان**

رشا معلا***

(تاريخ الإيداع 4 / 3 / 2020. قُبِل للنشر في 16/ 7 / 2020)

□ ملخّص □

يناقش البحث الحالي أحد أهم وأحدث المواضيع في مجال تعلم الآلة عموماً والتعلم العميق خصوصاً وهو أنظمة وصف الصور. تم في هذا البحث بناء نظام لوصف الصور يعتمد على النموذج ResNet50 وهو نموذج تعلم عميق مدرب مبني على أساس الشبكات العصبية الالتفافية CNN والذي يولد أشعة سمات التمثيل الصوري. أما في التمثيل النصي فقد اعتمدت خمس نماذج وصف مختلفة تعتمد بالأساس على نموذجي GloVe و FastText المقدمين من قبل تويتر وفيسبوك بالترتيب، حيث تم دراسة تأثير اختلاف معاجم المفردات على أداء نظام الوصف المقترح. استخدمت مجموعة بيانات MS-COCO العالمية حيث أخذت مجموعة جزئية منها مؤلفة من 10000 صورة، بحيث خصص 9000 صورة منها لمجموعة التدريب Training والتحقق Validation، أما لعملية الاختبار فقد تم اختيار 1000 صورة من مجموعة البيانات مختلفة عن صور التدريب والتحقق ، وتم تطبيق الاختبارات على النماذج الخمسة المصممة.

لمعرفة دقة الوصف الناتج عن النماذج المقترحة ومدى درجة تطابق جمل الوصف الفعلية مع جمل الوصف الناتجة، تم استخدام قياسات الأداء التالية Accuracy , Average of Depth Similarity ،Top-1, Top-5 BLEU. بينت النتائج العملية تفوق الأنظمة المعتمدة على نماذج FastText على الرغم من أنها تستغرق زمناً أطول من نماذج GloVe.

الكلمات المفتاحية: التعلم العميق، معالجة اللغات الطبيعية، التمثيل الصوري، التمثيل النصي، نموذج FastText، نموذج GloVe.

* أستاذ -كلية الهندسة المعلوماتية- جامعة تشرين- اللاذقية- سورية، Email: Alkheir.j@gmail.com

** مدرس - قسم هندسة الحاسبات والتحكم الآلي- كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين-اللاذقية – سورية، Email: Samsulg@googlemail.com

*** طالبة دراسات عليا (دكتوراه) - قسم هندسة الحاسبات والتحكم الآلي- كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين- اللاذقية – سورية، Email: Rasha_Mualla90@hotmail.com

مقدمة:

يعد التعلم العميق Deep Learning أحد فروع تعلم الآلة Machine Learning التي تعتمد بشكل أساسي على الشبكات العصبونية في بناء نماذجها. يضم التعلم العميق طرقاً كثيرة منها التعلم بمشرف Supervised Learning، والتعلم بدون مشرف Unsupervised Learning، ولعل الاختلاف الأساسي بين تعلم الآلة والتعلم العميق يكمن في كيفية استخلاص السمات فبينما تميل الطرق التقليدية لتعلم الآلة إلى استخدام تقنيات تعتمد على الشكل والحواف واللون والانحناءات والنسيج وغيرها، تستخدم طرق التعلم العميق تقنيات أكثر تطوراً تعتمد على استخلاص السمات بشكل هرمي بدءاً من السمات البسيطة وصولاً إلى السمات الأكثر تعقيداً للحصول على نموذج سمات قوي ولتحقيق هذه التقنيات تم بناء شبكات عصبونية ذات طبقات متعددة ومن هنا ظهرت تسمية التعلم العميق التي تعني أن الشبكة مع كل زيادة عمق أو طبقة فيها تتعلم شيئاً جديداً أكثر تفصيلاً وتعقيداً [1].

أما بما يخص نظم وصف الصور Image Captioning Systems، ومع التطور الكبير والضخم في مجموعات بيانات الصور لم تعد الطرق التقليدية في استخلاص سمات الصور ووصفها كافية حيث ظهرت بدلاً منها تقنيات أخرى جديدة لاستخلاص سمات الصور مثل الشبكات العصبونية الالتفافية Convolutional Neural Networks (CNN) [2] والتي تستطيع أخذ صورة كدخل لها ثم باستخدام عدد من عمليات الالتفاف (الترشيح) والانتخاب Pooling (لتقليل الأبعاد) تحول في الخرج إلى شعاع من السمات يعمل كمجموعة من الواصفات لتلك الصورة، حيث تم الاستفادة من زيادة عمق الشبكة وعدد طبقاتها في زيادة فعالية استخلاص السمات الأمر الذي كان غير ممكن في الشبكات التقليدية القديمة التي كانت زيادة عدد طبقاتها تضاعف زمن التدريب فيها دون جدوى تذكر.

أدى التطور الكبير والمتسارع في نظم الذكاء الصنعي وتعلم الآلة إلى تحسين وتطوير النماذج المصممة لوصف الصور Image Captioning سواء بتسميات Captions أو بجمل كاملة Full-length sentences. وعلى الرغم من تطور هذه النماذج إلا أنها لا تزال غير قادرة على تقليد العقل البشري تماماً في وصفه للمشاهد التي يراها. لذلك لا يزال هذا المجال قيد البحث والدراسة والتطوير [3].

هناك الكثير من التطبيقات المهمة التي يمكن الاستفادة من نظم وصف الصور فيها بدءاً من تطبيقات التفاعل مع البشر مروراً بتطبيقات تعليم الأطفال ووصولاً لتطبيقات المساعدة الصحية ومساعدة المكفوفين وذوي الاحتياجات الخاصة، وانتهاءً بتطبيقات استرداد الصور والبحث الصوري في محركات البحث الضخمة وغيرها من التطبيقات الكثيرة التي تشكل خدمة عالية القيمة والهدف للإنسان.

الدراسات المرجعية ذات الصلة:

نُشرَت مؤخراً العديد من الدراسات والأبحاث العلمية في مجال وصف وتسمية الصور سواء بالاعتماد على التسميات التوضيحية أو باستخدام جمل وصف كاملة، حيث تم الاعتماد في معظمها على اللغة الإنكليزية. تركزت معظم تلك الدراسات في الأعوام بين 2016 و2019. يتألف أي نظام وصف صور من دعامتين أساسيتين هما النموذج الصوري (سمات الصور) أو نموذج استخلاص السمات والنموذج النصي وتكمن عملية وصف الصور في القدرة على الربط بين النموذجين لإنتاج عملية الوصف. اعتمدت الدراسات السابقة في مرحلة استخلاص السمات على شبكات استخلاص السمات مثل AlexNet [4,5]، VGG-16 [6]، ResNet [7]، GoogleNet [5]، DenseNet [8] وغيرها، أما في النموذج النصي فقد تم استخدام نماذج نصية مثل LSTM [9]، RNN [10]،

CNN [11]، GRU [12]. وفي سبيل تطبيق تلك النماذج واختبارها تم استخدام مجموعات بيانات صور عالمية مثل MS-COCO، Flickr8k، Flickr30k [13,14].

تختلف النماذج الصورية عن بعضها البعض في البنية الداخلية المكونة لها حيث أنها جميعاً شبكات عصبونية ذات معماريات مختلفة. فمثلاً تتألف شبكة Alexnet [4,5] من 61 مليون بارامتر وحققت دقة وصف Top-1 مساوية لـ 57.1 ودقة وصف Top-5 مساوية لـ 80.2، في حين تضم شبكة VGG 138 مليون بارامتر وحققت دقات الوصف Top-5 ,Top-1 مساوية لـ 70.5و 91.2 على التوالي. أما شبكة ResNet50 فتتألف من 25.6 مليون بارامتر فقط وهو أقل بكثير من باقي الشبكات على الرغم من أن دقة الوصف في هذه الشبكة كانت الأفضل حيث حققت 75.2 وفقاً لمعيار Top-1 ودقة 93 وفقاً لمعيار Top-5 [15]. أما بالنسبة للنماذج النصية فقد تم استخدام الشبكات العصبونية التكرارية RNN Recurrent Neural Networks ومن ثم تطوير نموذج الشبكات ذات الذاكرة المتكيفة (طويلة-قصيرة) LSTM Long-Short Term Memory من شبكة RNN حيث تستطيع هذه الشبكات التعامل مع تشكيل الجمل كونها تتضمن عناصر الذاكرة، وقد استخدمت شبكات LSTM ونسخ معدلة منها بشكل كبير في معظم النماذج النصية المستخدمة في نظم وصف الصور. هناك بعض الدراسات القليلة [16] التي استخدمت شبكات CNN نفسها لتوليد النموذج النصي وعلى الرغم من تقليل تعقيد هذه الأنظمة إلا أن دقتها كانت أقل.

تكمن المشكلة الأساسية التي تعاني منها أنظمة وصف الصور في التعقيدات الحسابية المكلفة زمنياً وقد جرى استخدام العديد من الطرق لحل هذه المشكلة مثل تقليل عدد بارامترات الشبكة كما في شبكة ResNet [17] في حين تم تقليل حجم المعجم النصي (عدد المفردات) vocabulary dictionary في دراسات أخرى [18] وعادة ما يكون حجم المعجم بين 10000 إلى 40000 كلمة [3] ففي دراسة Zhang [18] تم تقليل حجم معجم المفردات بمقدار 39 مرة لكن دقة الوصف انخفضت إلى 12.6 فقط وفقاً لمعيار Bleu-1. ولحل مشكلة انخفاض دقة الوصف مع تقليل الأبعاد استخدمت دراسات أخرى طرق أخرى لتقليل الأبعاد مثل طرق التخفيض Dense Captioning كما في دراسة Johnson [19] والتي تقوم أولاً بتحديد المناطق البارزة في الصورة ثم تولد وصفاً لكل منها. تعتمد هذه الطريقة على شبكة CNN للحصول على سمات كل منطقة في الصورة. وتتضمن طبقة Dense Layer تستخدم لتخفيض أبعاد السمات التي يتم الحصول عليها من شبكة CNN (للحصول على الأجزاء المهمة من الصورة) ثم يتم استخدام شبكة LSTM من أجل بناء النموذج النصي. وفي دراسة أخرى Yang [20] تم حل مشكلة عدم ملاءمة طرق Dense Captioning للصور التي تتضمن تداخل أو تراكب للمكونات حيث تم الاعتماد على السمات البصرية والتسميات التوضيحية المخمنة لكل سمة وتسمح الطريقة بإيجاد موقع مناسب لكل مكون في الصورة ليتم دمج السمات الصورية مع السمات النصية بطريقة دمج السياق Context Fusion من أجل ربط السمات الصورية مع السمات النصية للحصول على نموذج نصي قوي. حققت الدراسة قيمة معامل أداء mAP=8.03 متفوقة على قيمته في دراسة Johnson التي كانت 5.64.

تم الاستفادة من سرعة الحسابات التي تقدمها نظم المعالجة التفرعية في دراسة أخرى وتسخيرها لتقليل حسابات أنظمة وصف الصور مع المحافظة على دقة وصف عالية كما في دراسة في Wang [21] الذي استخدم بنية تفرعية مركبة من شبكتي RNN وLSTM لوصف الصورة باستخدام التسميات التوضيحية تعتمد على تقسيم وحدات RNN وLSTM المخفية إلى عدد من الأجزاء متساوية الحجوم والتي تعمل بشكل متوازي لتوليد التسميات التوضيحية للصور. طبق الباحث نموذجه على مجموعة بيانات Flickr8k وقد تفوق على نموذج GoogleNet من حيث معيار Bleu-1

كما أنه استهلك فقط ما يقارب نصف الموارد التي تستهلكها الطرق السابقة في وصف الصور ومنها من استخدم نماذج أكثر تطوراً مثل شبكة ResNet50 و VGG وحققت دقات وصف جيدة ففي دراسة [22] تمكن الباحثون من التوصل لدقة وصف 95.9% و88.6 لشبكتي ResNet وVGG على التوالي مستفيدين من قوة الربط بين النموذجين الصوري والنصي.

تم تقييم نماذج وصف الصور المصممة بالاعتماد على عدة طرق تقييم منها BLEU، CIDEr، METEOR، Top-1، Top-5، الخ. حيث يعتبر Bleu معيار لتقييم دقة الوصف بحيث إذا كان الوصف الناتج عن عملية الاختبار قريباً جداً من الوصف المستخدم في التدريب فإن قيمة bleu ستكون عالية وإلا ستكون منخفضة، ومجال هذا المعيار هو بين 0 و100% [23]. أما بالنسبة لمعيار CIDEr فهو يقيس درجة التطابق بين الجمل الناتجة عن الوصف والجمل الحقيقية من وجهة نظر متوافقة جداً مع مفهوم البشر في وصف الصور ويسعى هذا المعيار إلى انتقاء أفضل الجمل التي تعطي الوصف للصور بشكل يتوافق مع التفكير البشري [24]. يوضح الجدول (1) مقارنة بين أهم الدراسات بين الأعوام 2016 وحتى 2019 في مجال نظم وصف الصور من حيث نموذج استخلاص السمات والنموذج النصي ومجموعة البيانات وتقييم أداء نموذج الوصف.

الجدول (1) مقارنة بين أهم الدراسات بين الأعوام 2016 وحتى 2019 في مجال نظم وصف الصور.

البحث وتاريخ النشر	نموذج استخلاص السمات	النموذج النصي	مجموعة البيانات	Bleu-1	Bleu-2	CIDEr	Top-1 Top-5
[25] 2016	VGG	LSTM	MS-COCO	71.4	50.5	63.8	-
[5] 2016	GoogleNet	LSTM	MS-COCO	50.0	31.2	61.8	-
[4] 2016	AlexNet, VGG	LSTM	MS-COCO	67.2	49.2		-
			Flickr30k	62.1	42.6		
[26] 2017	ResNet	LSTM	MS-COCO	74.2	58	108.5	-
			Flickr30k	67.7	49.4	53.1	-
[27] 2017	VGG ResNet	RNN LSTM	MS-COCO	91	83.1	102.9	-
[28] 2017	VGG	LSTM	MS-COCO	-	-	102	-
			Flickr30k	-	-	76.7	-
[29] 2017	VGG ResNet	LSTM	MS-COCO	-	-	102	-
			Flickr30k	-	-	101.1	-
[30] 2018	GoogleNet ResNet	LSTM	MS-COCO	76.4	60.4	112.5	-
[20] 2019	ResNet VGG	LSTM	MS-COCO	-	-	110.1	-
[31] 2019	ResNet	LSTM	MS-COCO	75.8	59.6	110.5	-

أجريت بعض الدراسات في جامعة تشرين [22و32]، حيث تم بناء نظام وصف صور حيث جرى استخدام ثلاثة نماذج مختلفة هي VGG16, VGG19, ResNet50 وتم استخدام شبكة RNN في عملية التمثيل النصي. تم استخدام مجموعة بيانات MS-COCO وقد بينت النتائج العملية تفوق نموذج ResNet50 بدقة وصف 95%.

ركزت معظم الدراسات الحديثة في آخر سنتين على موضوع تقليل الزمن مع الحفاظ على الدقة وكان هذا منطلق بحثنا الحالي الذي يتضمن بناء نموذج وصف للصور باللغة الإنكليزية وبالاعتماد على نماذج شبكات

استخلاص السمات حيث تم استخدام نموذج ResNet50 للوصف الصوري والذي أثبت القدرة على تقليل بارامترات الشبكة مع المحافظة على دقة وصف جيدة [1]، أما في النموذج النصي فقد تم استخدام 5 نماذج نصية مختلفة لدراسة تأثير استخدام كامل قاموس مفردات اللغة التي تم تدريب نموذج وصف الصور عليها مقابل استخدام قاموس مفردات مخفض (انتقاء الكلمات المستخدمة في تدريب نظام وصف الصور المقترح)، ثم مقارنة أداء هذه النماذج من ناحية الأداء والزمن.

أهمية البحث وأهدافه:

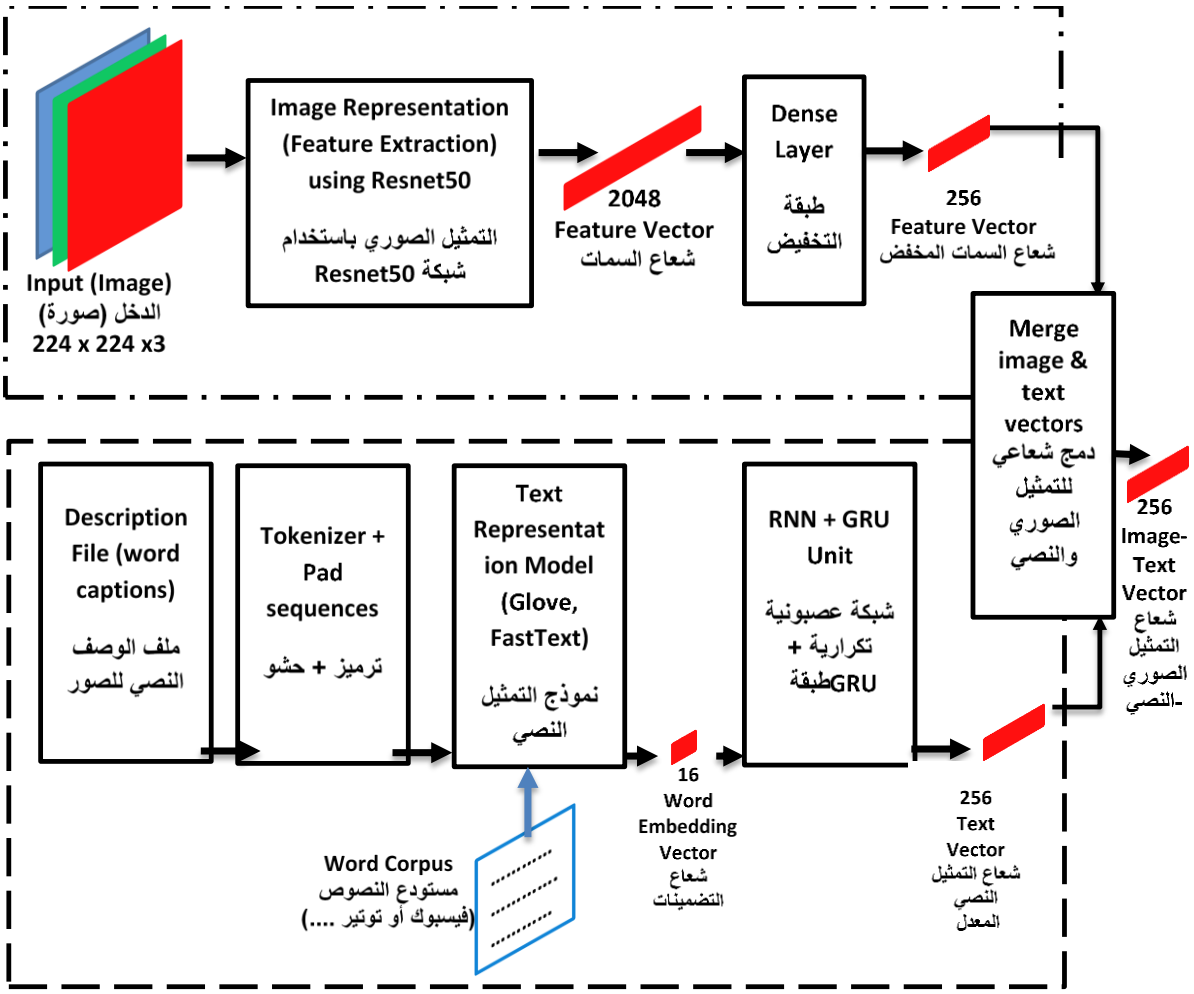
يعد مجال وصف الصور واسترجاع الصور من أكثر المجالات انتشاراً في الأعوام الأخيرة نظراً للحاجة التطبيقية الهامة له مثل التطبيقات التي تخدم فاقدي البصر وذوي الاحتياجات الخاصة حيث يتم تحويل الصور إلى نص مقروء. كذلك هناك العديد من التطبيقات التعليمية سواء للأطفال أو غير المختصين التي تتطلب معرفة محتوى الصور وتحويله إلى نصوص مقروءة. فضلاً عن تطبيقات البحث الصوري وتطبيقات الاقتراح التي تعتمد على وصف الصور من أجل تقديم نصائح محددة للزبائن وتقرب وتسهل لهم عملية البحث عن هدفهم مثل البحث عن فندق بمواصفات صورية محددة مثلاً ناهيك عن مئات التطبيقات التي تعتمد على تبادل واسترداد ملايين الصور مثل تطبيقات مواقع التواصل الاجتماعي. يهدف البحث الحالي إلى إجراء مقارنة بين خمس نماذج نصية لأنظمة وصف الصور من حيث الدقة والزمن. تعتمد هذه النماذج على النموذجين النصيين FastText و GloVe وتختلف عن بعضها البعض عن طريق تغيير عدد كلمات قاموس المفردات النصي المخصص والمستخدم من قبل كل نموذج على حدة.

طرائق البحث ومواده:

1- أدوات البحث:

حاسب بمواصفات: نظام تشغيل UbuntuX64bit، معالج intel I5G8 بسرعة 2.3 GHz وذواكر 8 GB ومساحة قرص صلب بحجم 100GB. تم استخدام لغة البرمجة Python V3.6، بيئة تطوير Anaconda Navigator، مكاتب Keras، TensorFlow، بيئة Jupiter 6.0.1. كذلك تم تحميل المرفقات الخاصة واللازمة لعمل نموذج التمثيل الصوري ResNet50 ونماذج التمثيل النصي GloVe و FastText [33,34]. أما بالنسبة لمجموعة البيانات المستخدمة فقد تم استخدام مجموعة بيانات MS-COCO العالمية حيث تم انتقاء 10000 صورة موصوفة بجملة واحدة فقط ليتم تقسيمها بشكل عشوائي إلى 9000 صورة لإنجاز عملية التدريب والتحقق (validation) و1000 صورة للاختبار.

2- البنية التصميمية لنماذج وصف الصور المستخدمة:

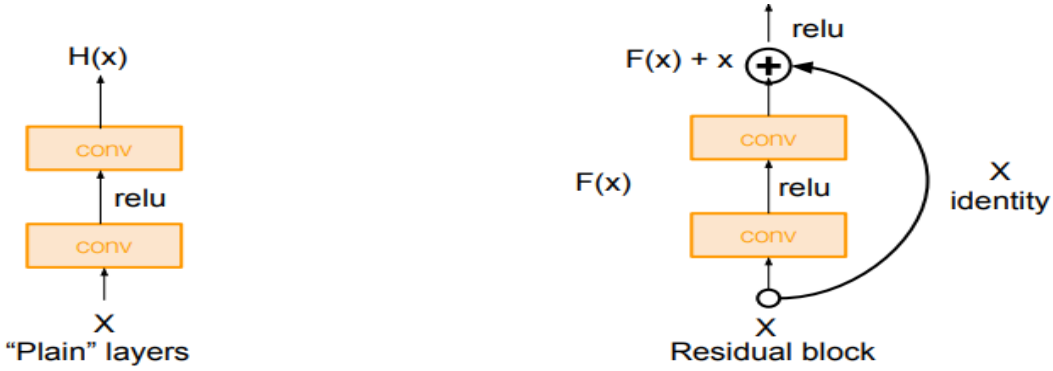


الشكل (1) نماذج التمثيل الصوري (المستطيل المتقطع والمنقط) والنصي (المستطيل المتقطع) المقترحة للبحث الحالي

2-1 مرحلة استخلاص السمات الصورية (Image Representation (Image Model:

يتم خلال هذه المرحلة الحصول على سمات صورة الدخل ذات الأبعاد 3 X 224 X 224 باستخدام شبكة (ResNet) وتم استخدام النموذج الصوري لهذه الشبكة من المرجع [35]، والذي يتألف من 50 طبقة التفافية Convolutional وظيفتها الأساسية استخلاص سمات الصور (دخل النموذج) بالاعتماد على مرشحات لازمة لإنجاز عملية الترشيح. خرج شبكة الـ ResNet هو عبارة عن شعاع من السمات بحجم 2048 عنصر. تتميز شبكات ResNet بسرعتها في عملية التدريب بسبب تقسيم بنية الشبكة إلى مجموعة بلوكات بحيث يتم تقليل أبعاد الصورة بمقدار الضعف بعد نهاية كل طبقتين من شبكة ResNet مما يسهم في تقليل عدد البارامترات اللازمة لتدريب الشبكة لذا فإن زيادة العمق في شبكات ResNet لا يزيد من مشكلة زمن التدريب [39]. يتضمن نموذج ResNet50 استخدام طبقة التفافية بمرشحات ذات حجم 1*1 قبل استخدام الطبقة ذات المرشحات 3*3 والهدف هو تقليل أبعاد أو عمق القناة Channel Depth. تم إزالة طبقة القرار Decision Layer من نموذج ResNet50 المستخدم لأننا نريد

الحصول على سمات من خرج الشبكة فقط. بسبب الحاجة إلى دمج التمثيل الصوري (سمات نموذج ResNet الـ 2048) مع التمثيل النصي. من أجل ضبط أبعاد شعاع السمات الناتج ليتوافق مع أبعاد التمثيل النصي، تم استخدام طبقة Dense Layer لتخفيض عدد السمات من 2048 إلى 256 عنصر فقط. يتضمن عنصر شبكة ResNet الواحد طبقتين التفافيتين متتاليتين واتصال تطابق Identity Connection مباشر بين دخل عنصر الشبكة وخرج الطبقة الثانية منها بحيث يتم دمجهما معاً. يوضح الشكل (1) "جزء النموذج الصوري" بنية نموذج التمثيل الصوري مع طبقة التخفيض. يوضح الشكل (2) مقارنة بين نموذج عنصر شبكة عادية وعنصر ResNet.



الشكل (2) الفرق بين عنصر شبكة عادية وعنصر شبكة ResNet

2-2 مرحلة الحصول على التمثيلات النصية (Caption Model) Text Representation:

يتم في هذه المرحلة الحصول على سمات الوصف النصي، وكدخل لهذه المرحلة يتم استخدام الملف الخاص بالتسميات النصية التوضيحية Image Description (captions) File المتاح مع مجموعة البيانات المستخدمة وهو مؤلف من 10000 جملة بمعدل جملة واحدة لكل صورة في مجموعة البيانات. يتألف الوصف النصي لكل صورة عادة من اسم الصورة متبوعاً بجملة تصفها. تتضمن عملية الحصول على التمثيل النصي المراحل التالية الموضحة في الشكل (1) "الجزء الخاص بالتمثيل النصي":

- مرحلة الترميز والحشو Tokenizer & Padding: يتم في هذه المرحلة استخدام تابعي المكتبة Keras المتاحين وهما تابع الترميز Tokenizer وتابع الحشو pad Sequences. يتم الاستفادة من تابع Tokenizer في تقسيم جمل الوصف إلى كلمات، حيث يتم تمثيل كل كلمة بدليل Index موافق لها ضمن فضاء النموذج النصي المدرب والمؤلف من أوزان الكلمات، يوضح الشكل (A-3) مثال عن ترميز كلمة "dog" حيث أن الدليل Index الخاص بها هو 37 على خرج Tokenizer. أما بالنسبة لتابع الحشو فيستخدم لجعل طول ترميز كل الكلمات ثابت وموحد ومساوٍ لعدد أوزان الكلمات المستخدم ضمن النموذج النصي المدرب، حيث يتم الحشو بأصفار عند الحاجة. خرج هذه المرحلة هو مصفوفة (عدد الأمثلة، الطول الأعظمي) تتضمن أرقام صحيحة تمثل معرفات IDs موافقة للكلمات. يوضح الشكل (B-3) أول 20 وزن لكلمة dog.

```
vocab = tokenizer.word_index
vocab['<eos>'] = 0 # add word with id 0
print('dog', vocab['dog'])
```

```
dog 37
```

(A)

```
import embedding
embedding_weights = embedding.load(vocab, 100, 'glove.twitter.27B.100d.filtered.txt')
print('dog', embedding_weights[vocab['dog']][:20])
```

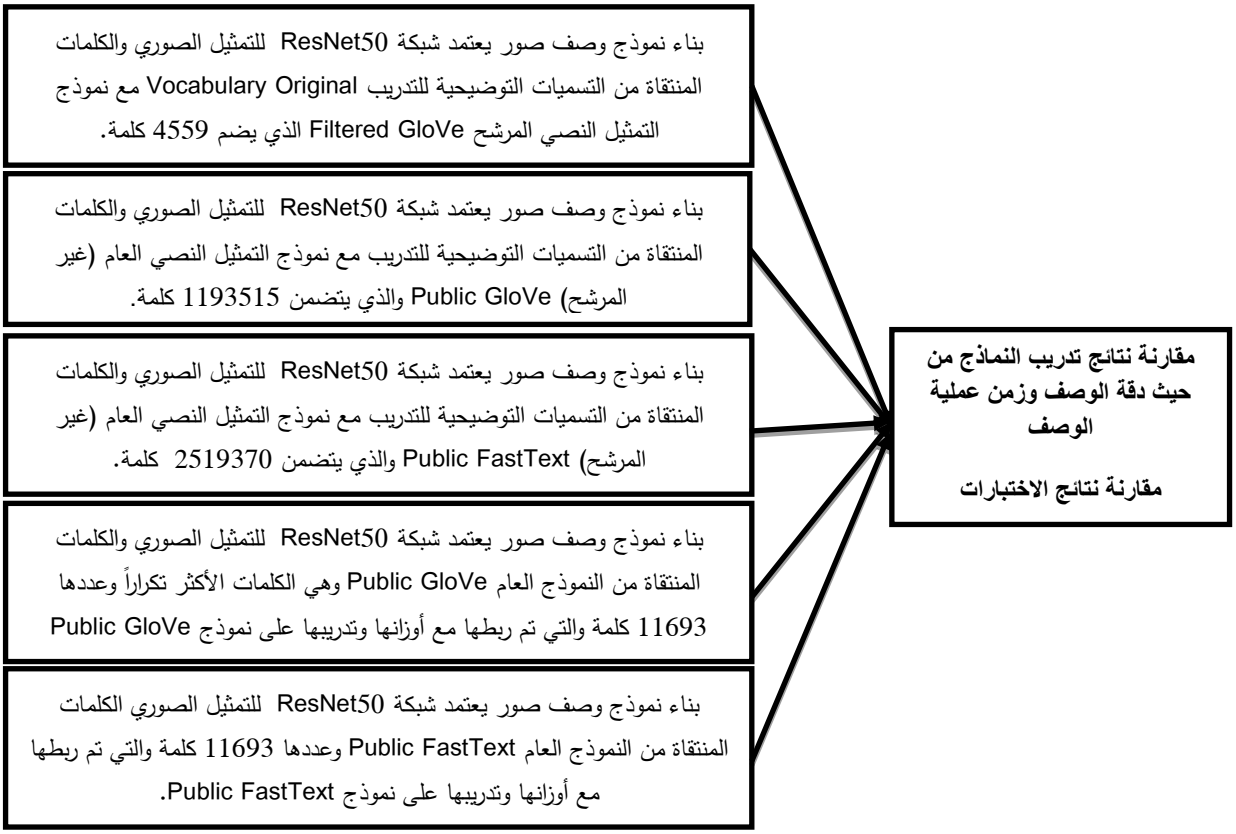
```
loading embeddings from "glove.twitter.27B.100d.filtered.txt"
```

```
dog [ 0.50779  -1.0274    0.48136  -0.09417   0.44837  -0.52291   0.51498
 -0.038927  0.35867  -0.065994 -0.82882   0.76179  -3.803    -0.010576
  0.21654   0.59712   0.37424  -0.022629 -0.010331 -0.33966 ]
```

(B)

الشكل (3) تمثيل كلمة "dog": (A) الدليل index الخاص بكلمة dog على خرج Tokenizer، (B) أول 20 وزن لكلمة dog من أصل 100 وزن.

- في المرحلة الثالثة، وهي عصب مرحلة التمثيل النصي، والذي ينتج عنه نموذج التمثيل النصي. تم في بعض الدراسات السابقة استخدام نموذج Filtered GloVe لبناء التمثيل النصي لكن الدراسة الحالية ستركز على استخدام كامل نموذج GloVe بدون ترشيح إضافة لاستخدام نموذج FastText بدون ترشيح ايضاً ودراسة تأثير استخدام كامل قاموس المفردات أو أجزاء منه ومقارنة أداء هذه النماذج كما هو موضح بالشكل (4).



الشكل (4) سيناريوهات العمل المقترح

تم في هذا البحث تحقيق خمس نماذج للوصف النصي. يتضمن النموذج الأول استخدام الكلمات المنتقاة من التسميات التوضيحية للتدريب Vocabulary Original (الكلمات التي تم تدريب نموذج التمثيل الصوري عليها) مع نموذج التمثيل النصي المرشح Filtered GloVe والذي يضم 4559 كلمة. أما النموذج الثاني فيتم فيه استخدام Vocabulary Original مع نموذج التمثيل النصي العام (غير المرشح) Public GloVe والذي يتضمن 1193515 كلمة. أما النموذج الثالث فهو يمثل استخدام Vocabulary Original مع نموذج التمثيل النصي العام (غير المرشح) Public FastText والذي يتضمن 2519370 كلمة. النموذج الرابع يتم فيه استخدام الكلمات المنتقاة من النموذج العام Public GloVe وعددها 11693 كلمة والتي تم ربطها مع أوزانها وتدريبها على نموذج Public GloVe النصي، في حين أن النموذج الأخير تضمن استخدام الكلمات المنتقاة من النموذج العام Public FastText وعددها 11693 كلمة والتي تم ربطها مع أوزانها وتدريبها على نموذج Public FastText النصي.

تتضمن هذه المرحلة طبقة مهمتها الأساسية الحصول على تمثيل للكلمات Word Representation بحيث يتم البحث ضمن النموذج النصي المدرب المستخدم لاختيار الرموز الموافقة لرموز الكلمات المستخدمة في التسميات التوضيحية الناتجة من المرحلة الثانية، وتعطي كنتيجة مجموعة من التضمينات Embedding والتي تستخدم في المرحلة اللاحقة لبناء التمثيل النصي. خرج هذه المرحلة شعاع تمثيل نصي بحجم 16 عنصر word vector 16، والعدد 16 هو الحد الأعظمي لطول الجملة (عدد كلمات جملة الوصف) Maxlen وهو رقم اختياري وتم اختيار الرقم 16 اعتماداً على أطول جملة استخدمت في وصف الصور.

- يتم في المرحلة الأخيرة من التمثيل النصي إدخال شعاع التضمينات ذو الطول 16 إلى وحدات Gated (Recurrent Units GRU) ذات حجم 256 بحيث يتم استخدام آخر طبقة مخفية ضمن GRU من أجل تمثيل كامل الجملة بدلاً من الحصول على خرج وحدة GRU والتي تمثل التنبؤ على مستوى الكلمة. خرج هذه المرحلة هو شعاع وصف نصي بحجم 256 (256 Word Vector) وذلك ليكون حجم شعاع التمثيل النصي متوافقاً مع حجم شعاع التمثيل الصوري.

2-3 النماذج النصية المستخدمة في مرحلة الحصول على التمثيلات النصية:

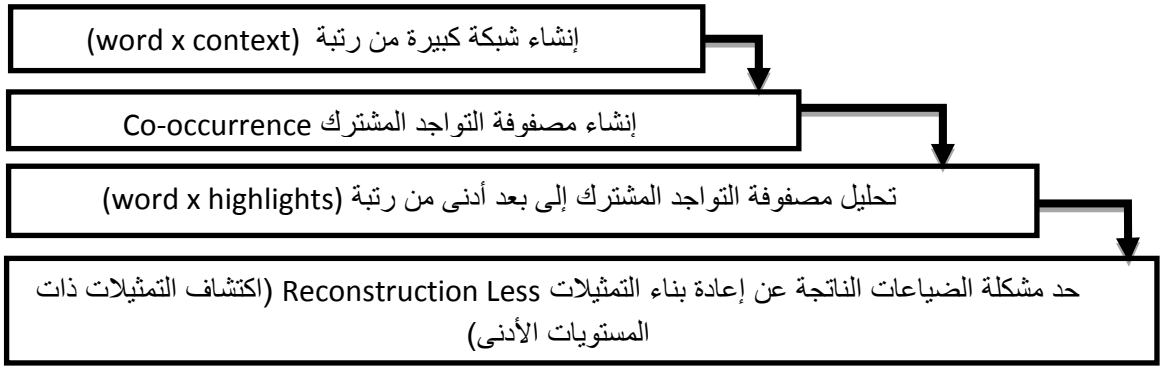
تم استخدام نموذجين نصيين مدربين هما النموذج Global Vectors For word Representation (GloVe) المتاح من قبل Twitter و FastText الخاص بموقع الـ Facebook.

نموذج Global Vectors For word Representation (GloVe):

هو نموذج تمثيل للكلمات Word Representation يعتبر بمثابة تطوير لتقنية تحويل الكلمة إلى شعاع Word2vec المعروفة في أنظمة وصف الصور. يستخدم هذا النموذج من أجل تحقيق التعلم الفعال لأشعة الكلمات وتم إنشاء النموذج من قبل الباحث Pennington وآخرون في جامعة ستانفورد [35].

يتضمن التدريب في نموذج GloVe إنشاء مصفوفة إحصائية تحدد عدد مرات تكرار الكلمات Word co-occurrence matrix ضمن نص في مستودع نصي كامل مثل تويتر أو ويكيبيديا أو غيرها من المواقع التي تتضمن نصوص خاصة بمختلف صنوف المستخدمين. إن خرج تمثيل GloVe المدرب هو نموذج مدرب يعتبر الأفضل من حيث تجميع العدد الأكبر الممكن من تضمينات الكلمات مقارنة بنماذج سابقة مثل Word2vec. حيث يعتبر نموذج Word2vec نموذجاً تخمينياً فقط أي يركز على تحسين السعة التخمينية (الكلمة الهدف –الكلمة المرجعية، أشعة الكلمات) أي تحسين الضياع Loss في تخمين الكلمات الهدف من كلمات السياق مع الأخذ بالحسبان تمثيلات أشعة الكلمات وذلك عن طريق بناء شبكة عصبونية ذات تغذية أمامية يتم تدريب الأشعة فيها على تحسين قدرة التخمين. أما في نموذج GloVe، تبنى أشعة الكلمات من خلال إنجاز عملية تقليل أبعاد Dimension Reduction على مصفوفة التواجد المشترك للكلمات ويعتبر نموذجاً معتمداً على العد Count-based أي عدد تكرار تواجد الكلمة ضمن مستودع الكلمات Corpus. ويتضمن نموذج GloVe المراحل الموضحة في الشكل (5) وهي [35]:

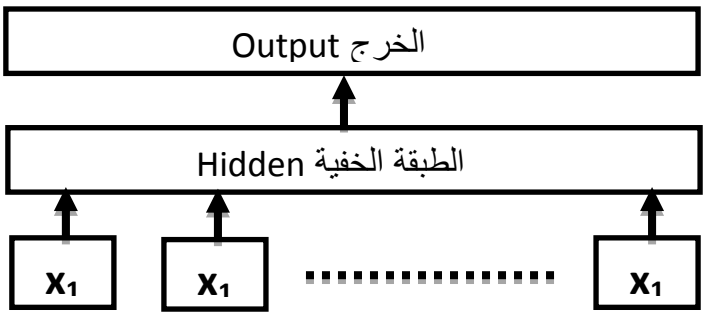
- إنشاء شبكة كبيرة من رتبة (word x context) تمثل معلومات التكرار أو التواجد المشترك للكلمات تعني كم مرة تكررت كل كلمة ضمن مستودع النصوص. ثم تحديد لكل كلمة ضمن الأسطر كم من الممكن رؤية هذه الكلمة في بعض الظروف في العمود الواحد.
- تخزن هذه المعلومات ضمن مصفوفة الترابط أو التواجد المشترك.
- بعدها يتم تحليل مصفوفة التواجد المشترك إلى بعد أدنى من رتبة (word x highlights) بحيث يعود كل سطر فيها إلى تمثيل شعاع لكلمة واحدة.
- تتضمن المرحلة الأخيرة تطبيق عملية حد من مشكلة الضياعات الناتجة عن إعادة بناء التمثيلات Reconstruction Less والتي تحاول اكتشاف التمثيلات ذات المستويات الأدنى التي يمكنها توضيح الجزء الأضخم من التغيير في المعلومات ذات المستوى الأعلى.



الشكل (5) مراحل تدريب نموذج GloVe

نموذج FastText:

يستخدم نموذج FastText لإنجاز مهام معالجة اللغات الطبيعية NLP مثل التمثيلات النصية وتصنيف الكلام والجمل (تصنيف النصوص/ المستندات / تحليل الشعور...). إنَّ تمثيل FastText موزّع تحت ترخيص BSD مما يعني أنه متاح للتعديل والاستخدام من أجل المشاريع الخاصة والتوزيع والاستخدام للأغراض التجارية [34]. يوضح الشكل (6) معمارية نموذج FastText لجملة واحدة تتضمن أشعة دخل ذات حجم n grams وهي الاشعة $x_1, x_2,$ ...، x_n، يتم تغذية أشعة الدخل هذه لدخل الشبكة بحيث تحول إلى صيغة تضمينات ويتم حساب قيمة الحالة المخفية اعتماداً على دمج وزن هذه التضمينات وفق صيغة محددة.



الشكل (6) معمارية نموذج FastText

الفرق بين نموذجي GloVe و FastText:

يعامل GloVe كل كلمة في مستودع النصوص على أنها كيان فريد بحد ذاته ويولد شعاع تمثيل لكل منها. تشبه GloVe في هذا الأمر نموذج Word2vec حيث كلاهما تعتبران الكلمة على أنها الوحدة الأصغر التي يمكن تدريبها. لكنها تختلف عنها في أن GloVe ليست تخمينية مثل Word2vec وإنما تعتمد على حساب عدد مرات تكرار الكلمة وتنشئ لذلك مصفوفة التواجد المشترك وبهذه الفكرة تتفوق على Word2vec. أما FastText فهو يعامل كل كلمة على أنها مؤلفة من مجموعة من المحارف Characters ngram لذا يكون شعاع الكلمة مؤلفاً من مجموعة من المحارف المتسلسلة للكلمة الواحدة. يسمح هذا الفرق لنموذج FastText بتوليد تضمينات نصية أفضل للكلمات حتى لو كانت هذه الكلمة نادرة الاستخدام لكن محارفها في n gram ستكون غالباً مشتركة مع محارف من كلمات أخرى. والميزة الثانية هي في التعامل مع الكلمات من خارج القاموس [36].

4-2 دمج التمثيل الصوري والنصي لإنشاء النموذج الكامل:

من أجل تحقيق إمكانية دمج التمثيل النصي مع الصوري يجب الانتباه لمطابقة كل جملة من جمل الوصف مع السمات الموافقة لها الناتجة من مرحلة استخلاص السمات أو التمثيل الصوري. دخل هذه المرحلة هو شعاع السمات الصورية Image Feature Vector الناتجة من مرحلة التمثيل الصوري وشعاع التمثيل النصي Word Vector الناتج من مرحلة التمثيل النصي وكل منهما بطول 256 عنصر. تم تطبيق طريقة الربط Concatenation المعتمدة على dot، وتتضمن استخدام بعض التوابع الوظيفية الموجودة ضمن مكتبة الـ Keras والتي تسمح بمشاركة الأوزان بين أجزاء النموذج المراد إنشاؤه وذلك لإنشاء أزواج التمثيل الصوري والنصي الصحيح والخاطئ (الإيجابي والسلبي)، والهدف من إنشاء هذه النماذج هو ربط السمات الصورية مع أشعة التمثيل النصي الصحيحة من جهة ومن جهة أخرى مع أشعة التمثيل النصي الضجيجية أو الخاطئة، بحيث يتعلم النموذج الكامل أثناء عملية التدريب على ربط السمات الصورية مع التمثيل النصي الصحيح لها وعدم الربط بين السمات الصورية والتمثيل النصي السلبي أو الضجيجي (استبعاد التوصيفات الخاطئة). ناتج عملية الدمج سيكون ناتج النموذج والذي سيستخدم لاحقاً ضمن تابع حساب الخسارة (Loss function) والذي يعطي بالعلاقة (1) [37]:

$$Loss = \sum_i \max(0, 1 - p_i + n_i) \qquad (1)$$

حيث p_i يمثل نتيجة زوج التمثيل الصوري والنصي الإيجابي وn_i نتيجة زوج التمثيل الصوري والنصي السلبي. يمكن حساب تابع الخسارة السابق اعتماداً على توابع مكتبة الـ Theano أو TensorFlow. يعتمد تابع الخسارة المقترح على وسيطين هما التسمية التوضيحية الصحيحة والتسمية التوضيحية الناتجة على خرج نموذج التدريب، وبالاعتماد على هذين الوسيطين يمكن حساب دقة النظام المقترح من خلال حساب عدد المرات التي تكون فيها قيمة نتيجة زوج التمثيل الصوري والنصي الإيجابي أعلى من قيمة نتيجة زوج التمثيل الصوري والنصي السلبي.

3- معايير تقييم أداء النماذج Evaluation Metrics:

لمعرفة دقة النماذج المصممة واختبار أدائها وقدرتها الصحيحة على إنتاج التسميات التوضيحية المناسبة للصور، لا بد من استخدام مجموعة من معايير أو قياسات الأداء التي تحدد درجة الدقة والصحة في وصف صور الاختبار. تم الاعتماد على عدة معايير من أجل تقييم الأداء، ولعل أهم عامل أداء يمكن الاستناد عليه في تقييم النتائج هو **الدقة Accuracy** وتمثل نسبة عدد المرات التي يتشابه أو يتطابق فيها الخرج الفعلي للنموذج مع الخرج المتوقع لعينات التدريب (Training Dataset) من أصل كامل العينات، وهذا ما يسمى دقة التدريب (Training Accuracy)، كما يتم حساب دقة التحقق (validation Accuracy) بعد كل تكرار لعملية التدريب ويتم فيه أيضاً حساب دقة التعرف على الخرج الصحيح من الخرج المتوقع لكن لعينات التحقق فقط (Validation Dataset) وهذا ما يسمى دقة التحقق (Validation Accuracy). المعيار الثاني هو معيار تشابه القيمة الأعلى **Top-1 Similarity** والذي يحسب مدى تشابه أفضل توصيف ناتج للصورة مع الجملة الأصلية التي تصف الصورة من ملف التسميات التوضيحية، وعلى هذا الأساس سيتم أخذ قيمة المتوسط الحسابي لجميع قيم التشابه وذلك من أجل كل عينات الاختبار المدخلة. بالإضافة للمتوسط الحسابي، تمت دراسة فاصل الثقة بنسبة 95% (Confidence Interval of average of top-1 similarity) لهذا المتوسط الحسابي والذي يوضح بشكل أفضل مدى دقة قيمة هذا المتوسط. أما المعيار الثالث فهو أفضل تشابه للقيم الخمس العليا **Top-5 Similarity** ويحسب مدى تشابه الجملة الأصلية التي تصف الصورة مع كل جملة من الجمل الخمس الأفضل توقعاً وأخذ قيمة أعلى تشابه. كذلك يتم أخذ قيمة المتوسط الحسابي

لجميع قيم التشابه العظمى وذلك من أجل كل عينات الاختبار المدخلة ودراسة فاصل الثقة بنسبة 95% للمتوسط. وبعد معرفة أفضل تشابه لمجموعة الجمل الخمس العليا مع الجملة الأصلية، يتم دراسة معيار عمق أفضل تشابه **Depth of Best Similarity** والذي يحسب ترتيب أفضل تشابه بين التوقعات الخمس العليا. يساعد هذا المعيار في فهم فيما إذا كان أفضل تشابه يتحقق بشكل أكبر عند عمق معين بدلاً من أن يكون عند الاقتراح الأول أو العمق 0. يؤخذ المتوسط الحسابي لجميع قيم العمق كما يستخدم فاصل الثقة بنسبة 95% لهذا المتوسط. المعيار الرابع يحسب أدنى وأعلى قيمة للتشابه **Max and Min Values of Similarity** أي القيمة الأعلى والأدنى التي تم الحصول عليها بين كل التشابهات في top−5 ومن أجل كل عينات التدريب أو الاختبار. والمعيار الخامس هو حساب عمق أدنى وأعلى قيمة للتشابه **Depth of Max and Min Values of Similarity** [22]. المعيار الأخير هو معيار **BLEU (Bilingual Evaluation understudy)** وهو معيار لتقييم دقة الوصف بحيث إذا كان الوصف الناتج عن عملية الاختبار قريباً جداً من الوصف المستخدم في التدريب فإن قيمة **BLEU** ستكون عالية وإلا ستكون منخفضة وبالتالي القيم المنخفضة لـ **BLEU** تعني أن الوصف غير جيد [38].

النتائج والمناقشة:

1- نموذج التدريب:

يعتمد نموذج التدريب لنظام وصف الصور المقترح على تحقيق عملية دمج التمثيل الصوري مع التمثيل النصي الصحيح، وعليه تم الاعتماد على مجموعة عشوائية من التسميات التوضيحية الصحيحة والضجيجية لتدريب هذا النموذج على اختيار التمثيل النصي الأقرب والانسب للتمثيل الصوري الناتج عن الصورة المراد وصفها. بناءً على ما سبق يمكن اعتبار دخل مرحلة التدريب مؤلف من ثلاث مدخلات هي: أشعة السمات الصورية المستخدمة (مصفوفة حجم أسطرها 2048 عنصر)، والتسمية التوضيحية الصحيحة بالإضافة إلى التسمية التوضيحية الضجيجية وتمثل كل منهما بمصفوفة حجم أسطرها مساوٍ لـ 16 عنصر (الحجم الذي تم اقتراحه لعدد الكلمات الأعظمي للجملة المستخدمة في وصف الصورة). تتضمن مجموعة بيانات التدريب مجموعة صور جزئية مأخوذة بشكل عشوائي من المجموعة المعيارية MS−COCO [13] وتتألف من 10000 صورة متضمنة 9000 صورة لإنجاز عملية التدريب والتحقق و1000 صورة للاختبار، حيث صور التحقق تستخدم لحل مشكلة الـ Overfitting. سيتم حساب تابع الخسارة بالإضافة إلى الدقة ضمن مرحلة التدريب والتي ستستخدم لاحقاً في عملية تقييم أداء النماذج المقترحة. تم تحميل النماذج النصية المدربة لكل من FastText و GloVe من المواقع التالية [33،34].

تم تكرار عملية التدريب 10 مرات وفي كل مرة سيتعلم نظام وصف الصور كيفية ربط التمثيل النصي الصحيح مع التمثيل الصوري الصحيح إضافة لتعلمه كيفية عدم ربط التمثيل النصي الضجيجي مع التمثيل الصوري الصحيح، الأمر الذي سيؤدي إلى تحسين أداء مرحلة التدريب مع كل تكرار. بنهاية مرحلة التدريب سنكون قد حصلنا على الأوزان النهائية الصحيحة والتي تقود النموذج المصمم لإعطاء الوصف الصحيح لصور الاختبار لذا يتم حفظ حالة النموذج النهائية وهي الأوزان لتستخدم في عملية الاختبار.

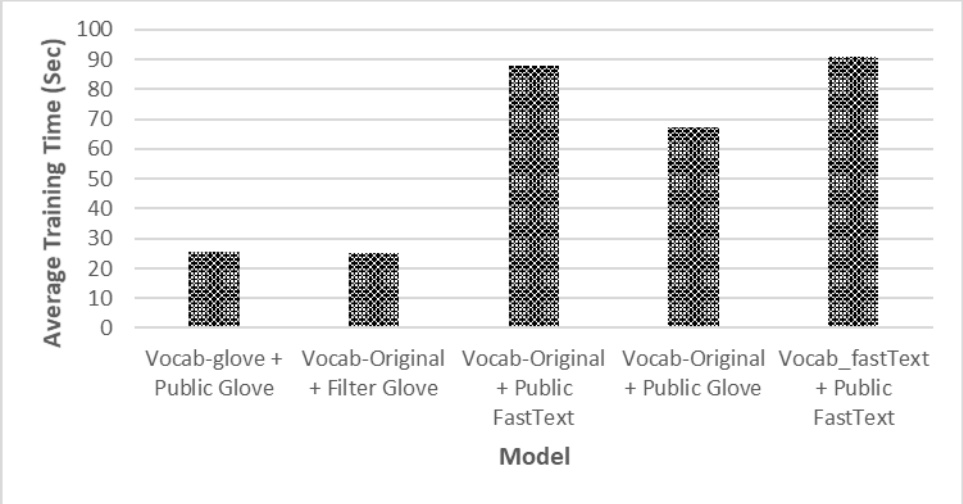
2- نتائج مرحلة التدريب:

يمكن تلخيص نتائج التدريب ضمن الجدول (2) والذي يظهر مقارنة بين النماذج الخمس المستخدمة وفق معايير تقييم الأداء المقترحة. في حين يوضح الشكل (7) مقارنة زمنية لمرحلة التدريب بين النماذج الخمس، حيث

يتضح أن نماذج FastText تستغرق وقتاً أطول في عملية التدريب مقارنة بنماذج GloVe والسبب الأساسي في ذلك يعود إلى حجم قاموس المفردات في نموذج FastText الذي يتضمن عدد مفردات أكبر بكثير من نموذج GloVe.

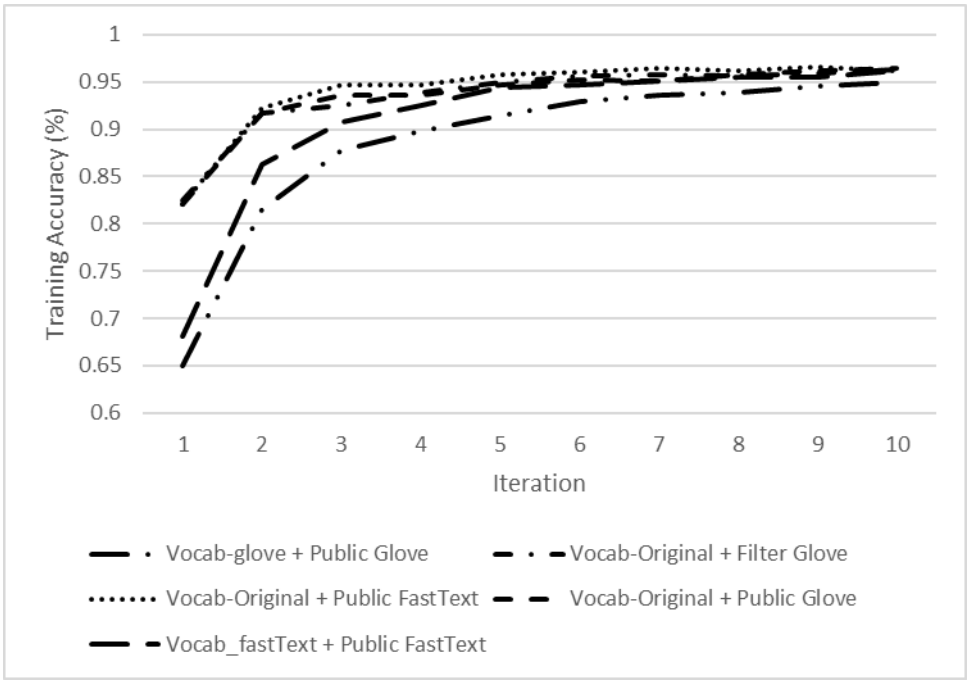
الجدول (2) مقارنة بين النماذج الخمس المستخدمة وفق معايير تقييم الأداء.

Training Time (s)	Validation Accuracy	Training Accuracy	Depth	النموذج المقترح
25.5	0.8645	0.88572	1.808	Vocab-GloVe + Public GloVe
25	0.9294	0.93496	1.894	Vocab-Original + Filter GloVe
88	0.9309	0.94088	1.843	Vocab-Original + Public FastText
67.1	0.9262	0.9345	1.763	Vocab-Original + Public GloVe
90.9	0.8879	0.90915	1.856	Vocab_FastText + Public FastText

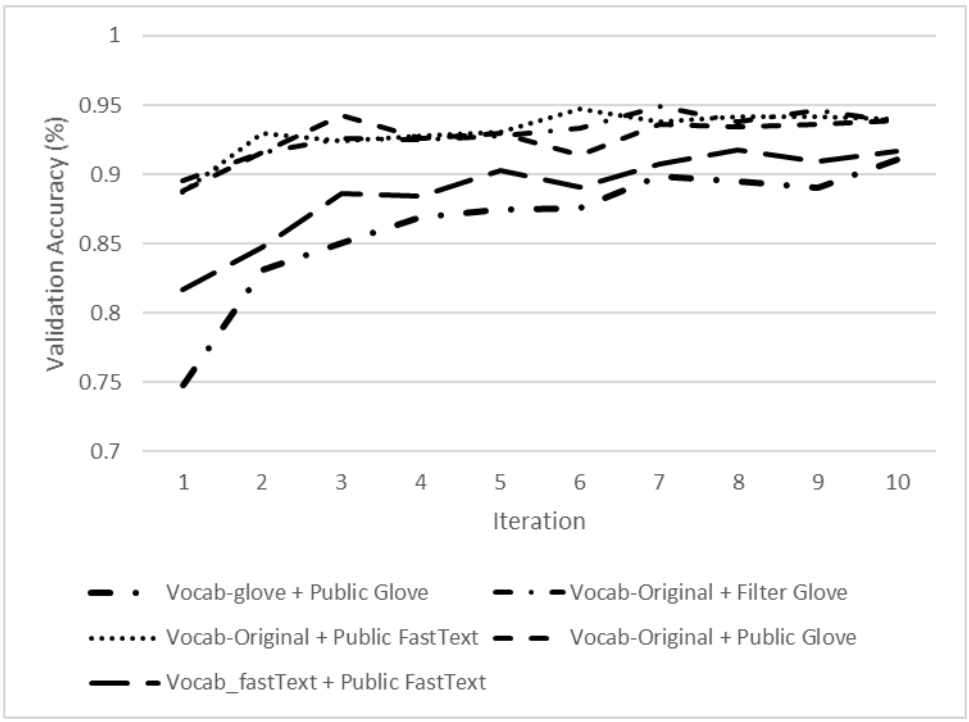


الشكل (7) أزمنة مرحلة التدريب للنماذج الخمس

من أجل مقارنة نتائج دقة التدريب Training Accuracy ودقة التحقق Validation Accuracy، تم حساب الدقة في كل تكرار من تكرارات عملية التدريب، ومن ثم تم رسم مخططات التقارب لتوضيح تلك المقارنة في الشكل (8).



(**A**)



(**B**)

الشكل (8) مخططات الانحدار لقيم دقة التدريب ودقة التحقق لكل تكرار من تكرارات الخوارزمية (A) قيم دقة التدريب، (B) قيم دقة التحقق

يتضح من الشكل (8) أنَّ دقة التدريب والتحقق لأنظمة الوصف التي اعتمدت معجم المفردات الأصلي Original Vocabulary كانت الأفضل من حيث دقة التدريب أو التحقق، مع تفوق بسيط لنظام الوصف FastText على نموذج GloVe.

3- سيناريو الاختبار Testing:

بعد انتهاء مرحلة التدريب تم اختبار النماذج المقترحة، وتم اقتراح سيناريو الاختبار التالي:

تم اختيار 1000 صورة من مجموعة البيانات المعيارية MS-COCO بشكل عشوائي ولكل صورة جملة واحدة توصِّفها. تم حساب السمات الصورية لصور الاختبار وتوليد التمثيل النصي الموافق لها عن طريق ضرب هذه السمات مع كل نموذج مدرب من النماذج الخمسة المقترحة، ليتم توليد قيم التوقع لكافة التسميات التوضيحية المستخدمة في مرحلة التدريب. يتم بعدها حفظ وترتيب التسميات التوضيحية المولدة اعتماداً على قيمة التوقع الناتج عن عملية الربط بشكل تنازلي، ليتم اختيار أفضل تسمية توضيحية كخرج للـ Top-1 واختيار أفضل خمس تسميات توضيحية كخرج للـ Top-5. يستخدم الخرجين السابقين كدخل لمرحلة التقييم باستخدام معاملات التقييم السابقة.

4- نتائج الاختبار:

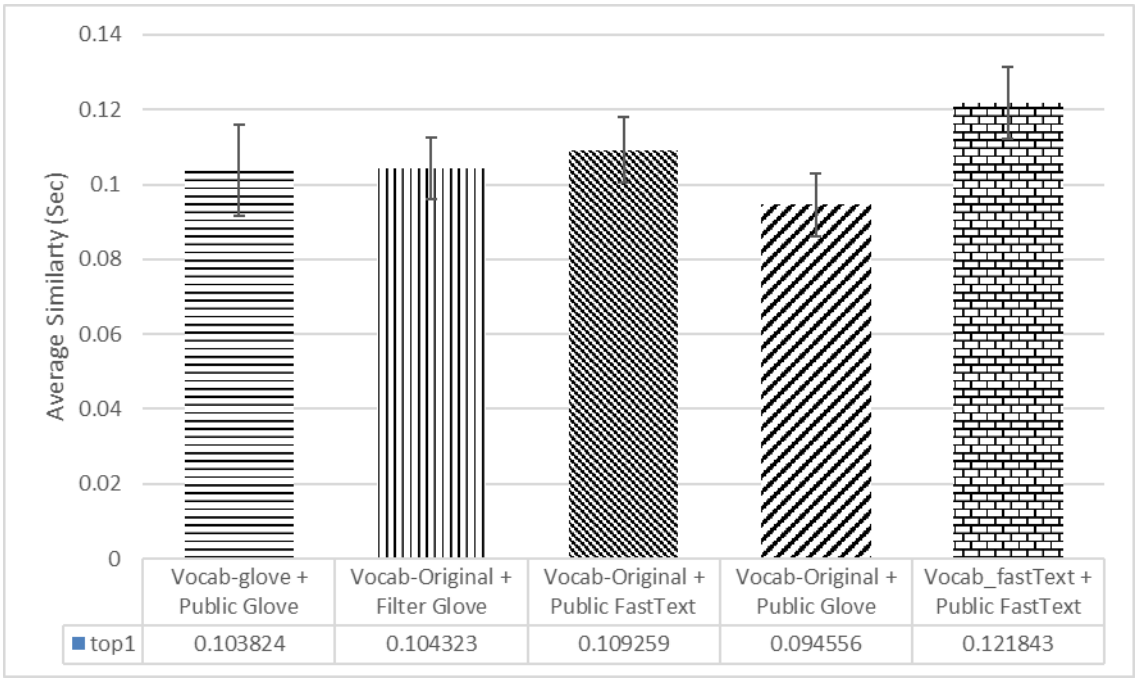
تم تطبيق الاختبارات على النماذج الخمسة المصممة باستخدام مجموعة بيانات صور عشوائية مؤلفة من 1000 صورة غير مستخدمة في مرحلة التدريب من مجموعة البيانات المعياريةMS-COCO . تم تطبيق معيار BLEU من أجل تقييم أداء النماذج الخمسة والنتائج موضحة ضمن الجدول (3). حيث تبين نتائج المعيار BLEU تفوق نموذج FastText سواء باستخدام كامل معجم المفردات (Vocab_origin + Public FastText) أو باستخدام معجم المفردات الخاص بالنموذج FastText ذاته (Vocab_FastText + Public FastText).

يتضمن الجدول (3) نتائج معيار الأداء BLEU للنماذج الخمسة، وكما هو واضح تفوق نموذج (Vocab_FastText+ Public FastText).

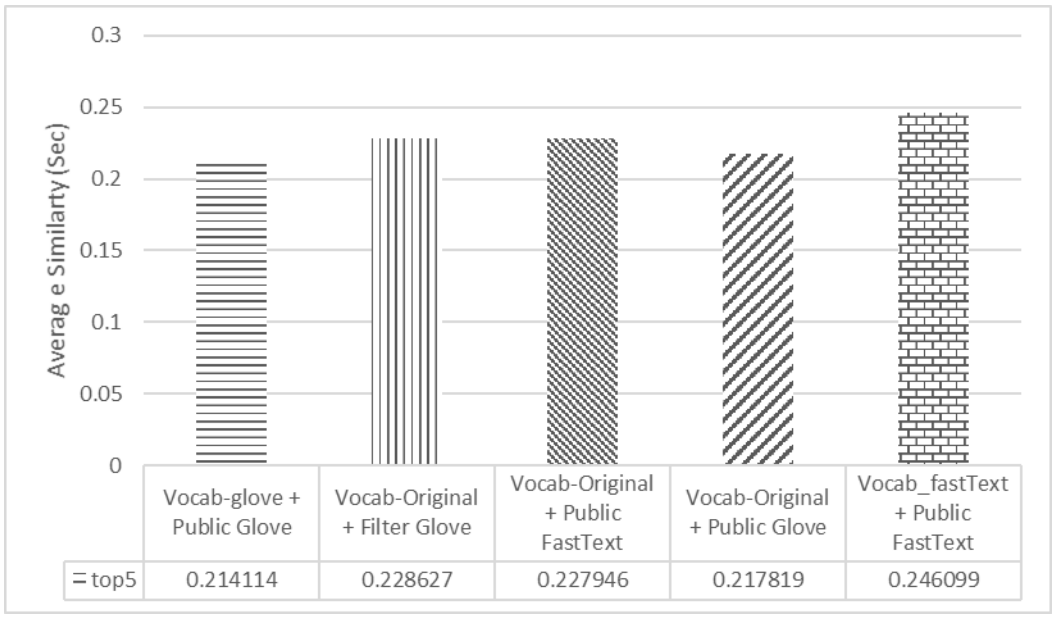
جدول (3) نتائج معيار الأداء BLEU للنماذج الخمسة

النموذج / القياس	Vocab-fastText + Public FastText	Vocab-Original + Public GloVe	Vocab-Original + Public FastText	Vocab-Original + Filter GloVe	Vocab- GloVe + Public GloVe
BLEU	48.1	34.9	40.9	35.1	35.3

يوضح الشكل (9) نتائج قياسات Top-1 وTop-5 لمرحلة الاختبار للنماذج الخمسة المقترحة، وكما هو واضح أثبتت قياسات الأداء أنَّ نموذج Vocab_origin + Public FastText تفوق على باقي النماذج، كما أن نماذج FastText عموماً تفوقت في الأداء على نماذج GloVe والسبب الأساسي يعود لطريقة توليد التضمينات المتاحة من قبل FastText والتي تتفوق على مقابلتها في GloVe من حيث العمل على مستوى أحرف الكلمة وبالتالي إمكانية التعرف على الكلمات من خارج قاموس المفردات. لمزيد من الدقة في القياسات تم حساب القيم المتوسطة والعليا والدنيا لدرجة عمق الحصول على أفضل تشابه عند معاينة Top-5. تمثل خطوط الخطأ Error Bars الموضحة في النتائج فاصل الثقة بنسبة 95% (Confidence Interval) للقيمة المتوسطة المحسوبة.



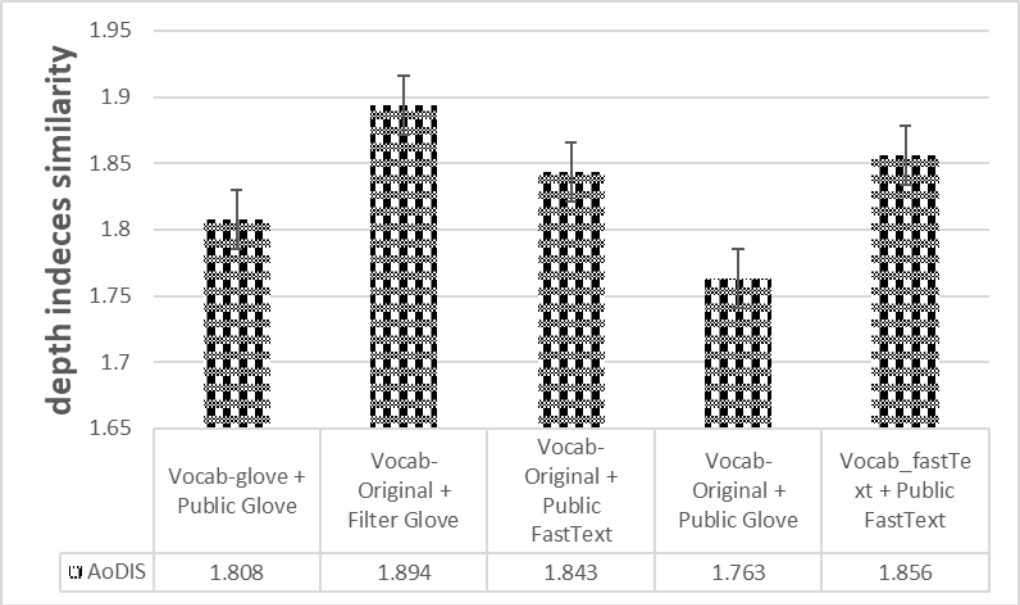
(A)



(B)

الشكل (9) نتائج قياسات الأداء لمعايير Top-1, Top-5 للنماذج الخمسة المقترحة

يوضح الشكل (10) أن القيم المتوسطة لدرجة العمق متقاربة وهي بحدود 2 لكل النماذج، لكن هناك بعض الاختلافات في القيم العليا والدنيا كما هو موضح في الجدول (4)، وبحساب فاصل الثقة للقيمة المتوسطة تبين صغر قيمته مما يعني أنه يمكن اعتماد القيمة المتوسطة لدرجة العمق لجميع النماذج مما يعني أن أفضل توصيف يمكن الحصول عليه يقع بين 1 و2 ويمكن اعتبار ثاني أفضل توصيف ناتج بين التوصيفات هو الخرج النهائي للأنظمة الخمسة وتجاهل باقي التوصيفات.



الشكل (10) درجة عمق التوصيف الأكثر تشابهاً في الـ Top-5

جدول (4) نتائج معيار الأداء BLEU للنماذج الخمسة

Vocab- GloVe + Public GloVe	Vocab-Original + Filter GloVe	Vocab-Original + Public FastText	Vocab-Original + Public GloVe	Vocab-fastText + Public FastText	النموذج / القياس
1	3	3	4	1	Max
0	0	0	0	0	Min

الاستنتاجات والتوصيات:

تم في هذا البحث بناء خمس أنظمة مختلفة لوصف الصور من أجل دراسة تأثير اختلاف نماذج التمثيل النصي على أداء أنظمة وصف الصور، حيث:

- بينت النتائج أن الأنظمة المبنية بالاعتماد على نماذج FastText تتفوق على مقابلاتها في نماذج GloVe والسبب يعود إلى اختلاف قاموس المفردات النصي المستخدم لتمثيل وتوصيف الصور إضافة لبنية نماذج FastText المدربة مسبقاً على مستوى أحرف كلمات الوصف مما يؤدي إلى توليد أوزان ثابتة للكلمة (300) بمجالات دقة أكبر، في حين أنه متغير ويأخذ 4 قيم مختلفة في نماذج GloVe تبعاً لعدد الأوزان المستخدمة وهي 25d, 50d, 100d, 200d والتي تم استخدام النموذج ذات البعد 100 في هذا البحث.
- بينت النتائج أن نماذج FastText تستغرق وقتاً أطول من نماذج GloVe وذلك بسبب زيادة مفردات الوصف التي ستزيد بالضرورة من زمن عملية التدريب.
- توضح النتائج وقياسات الأداء أنه عند اختيار أفضل ثاني جملة للوصف من بين أفضل خمس جمل لوصف الصور في الأنظمة الخمسة هو الخيار الأفضل. في الدراسات المستقبلية، يمكن دراسة تأثير اختلاف اللغات على النماذج المقترحة مثل استخدام اللغة العربية كلغة وصف.

- كما يمكن تغيير طرق دمج أزواج التمثيل النصي مع الصوري بالإضافة إلى تحسين طرق اختيار التمثيلات النصية الصحيحة والضجيجية والتوصل لطريقة مثلى محسنة عن الطريقة العشوائية المستخدمة في هذا البحث.

References:

[1] Alom, Md Zahangir, et al. "Improved Inception-Residual Convolutional Neural Network for Object Recognition." arXiv preprint arXiv:1712.09888 (2017).
[2] Joel P, Interactive Content-Based Image Retrieval with Deep Neural Networks, Springer International Publishing AG, pp.77-88, 2017.
[3] Staniute R, Šešok D, A Systematic Literature Review on Image Captioning, applied science journal, pp.1-20, May 2019.
[4] Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, the Netherlands, 15–19 October 2016; pp. 988–997.
[5] Mathews, A.P.; Xie, L.; He, X. Senticap: Generating image descriptions with sentiments. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
[6] Fan, C.; Crandall,D.J.DeepDiary: Automatic captiongeneration for lifelogging image streams. In Proceedings of the European Conference on Computer Vision, Amsterdam, the Netherlands, 11–14 October 2016; pp. 459–473.
[7] Vedantam, R.; Bengio, S.; Murphy, K.; Parikh, D.; Chechik, G. Context-aware captions from context-agnostic supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 251–260.
[8] Wang, Q.; Chan, A.B. CNN+CNN: Convolutional decoders for image captioning. arXiv 2018, arXiv:1805.09019.
[9] Huang, Q.; Smolensky, P.; He, X.; Deng, L.; Wu, D. Tensor product generation networks for deep NLP modeling. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, LA, USA, 1–6 June 2018; pp. 1263–1273
[10] Kinghorn, P.; Zhang, L.; Shao, L. A region-based image caption generator with refined descriptions. Neurocomputing 2018, 272, 416–424.
[11] Kilickaya, M.; Akkus, B.K.; Cakici, R.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. Data-driven image captioning via salient region discovery. IET Comput. Vis. 2017, 11, 398–406.
[12] Shen, K.; Kar, A.; Fidler, S. Lifelong learning for image captioning by asking natural language questions. arXiv 2018, arXiv:1812.00235.
[13] https://cocodatset.org , last access at 1/9/2019.
[14] https://www.flickr.com/photos/tags/dataset/, last access at 1/9/2019.
[15] Yan, S.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning using adversarial networks and reinforcement learning. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 248–253.
[16] Aneja, J.; Deshpande, A.; Schwing, A. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5561–5570.
[17] KOUSTUBH. ResNet, AlexNet, VGGNet, Inception: Understanding Various Architectures of Convolutional Networks. Available online: https://cv-tricks.com/cnn/understand-ResNet-alexnet-vgg-inception/ (accessed on 24 May 2019).

[18] Mishra, A.; Liwicki, M. Using deep object features for image descriptions. arXiv 2019, arXiv:1902.09969.
[19] Johnson J, Karpathy A, Fei-Fei L. Densecap: Fully convolutional localization networks for dense captioning. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016 (pp. 4565-4574).
[20] Yang, L., Tang, K.D., Yang, J. and Li, L.J., 2017, July. Dense Captioning with Joint Inference and Visual Context. In CVPR (pp. 1978-1987).
[21] Wang M, Song L, Yang X, Luo C. A parallel-fusion RNN-LSTM architecture for image caption generation. In Image Processing (ICIP), 2016 IEEE International Conference on 2016 Sep 25 (pp. 4448-4452). IEEE.
[22] Mualla R, Alkheir J, Image Description and Retrieval Based on Deep Learning and Natural Language Processing, Tishreen University Journal, Syria, 2019.
[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.
[24] Vedantam R, C. Zitnick L, Parikh D, CIDEr: Consensus-based Image Description Evaluation, arXiv:1411.5726v2 [cs.CV], 3 Jun 2015.
[25] Sugano, Y.; Bulling, A. Seeing with humans: Gaze-assisted neural image captioning. arXiv 2016, arXiv:1608.05203.
[26] 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, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
[27] Dai, B.; Lin, D. Contrastive learning for image captioning. In Advances in Neural Information Processing Systems; MIT Press: London, UK, 2017; pp. 898–907.
[28] Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional GAN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2970–2979.
[29] Xu, K.; Wang, H.; Tang, P. Image captioning with deep LSTM based on sequential residual. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 361–366.
[30] Jiang, W.; Ma, L.; Jiang, Y.G.; Liu, W. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 499–515.
[31] Zhang, M.; Yang, Y.; Zhang, H.; Ji, Y.; Shen, H.T.; Chua, T.-S. More is better: Precise and detailed image captioning using online positive recall and missing conceptsmining. IEEE Trans. Image Process. 2019, 28, 32–44.
[32] Mualla R, Alkheir J, "Development of an Arabic Image Description System", International Journal of Computer Science Trends and Technology (IJCST) – 6(3), 2018, pp:205-213.
[33] Pennington J., Socher R., Manning, C. D.," GloVe: Global Vectors for Word Representation", Available: https://nlp.stanford.edu/projects/glove/ , last access at 15/12/2019.
[34] https://FastText.cc/doc/en/crawl-vectors/html, , last access at 20/12/2019.
[35] Pennington J, Socher R, Manning C, "GloVe: Global Vectors for Word Representation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp:1532–1543.
[36] https://towardsdatascience.com , last access at 20/11/2019.

[37] FAVRE B, "Deep learning for NLP Joint text and image representations", 2017.
[38] Papineni K, Roukos S, Ward T and Zhu WJ, BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.
[39] Gracelyn Sh.,"Implementing a ResNet model from scratch", Available: https://towardsdatascience.com/implementing-a-resnet-model-from-scratch-971be7193718, last access at 1/1/2020.