مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (41) العدد (2) 2019
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (41) No. (2) 2019

دراسة أداء أنظمة وصف الصور المعتمدة على نماذج مختلفة للتعلم العميق

د. جعفر الخير *

رشا معلا **

(تاريخ الإيداع 13 / 12 / 2018. قُبِل للنشر في 22/ 4 / 2019)

□ ملخّص □

تم في البحث الحالي إنجاز عملية وصف الصورة Image Description عبر وضع تسميات توضيحية لمكونات الصورة Image Captioning بثلاث نماذج مختلفة والمقارنة بينها. من أجل بناء نماذج الوصف، تم استخدام مكتبة Keras وهي مكتبة تستخدم كإطار عمل لتعلم الآلة Machine Learning Framework والتي تضم أهم المكاتب اللازمة لإنشاء وتدريب شبكات التعلم العميق. تم استخدام ثلاثة نماذج مطبقة على مكتبة Keras وذلك لاستخدامها في استخلاص سمات الصور وهي نموذج شبكة ResNet50 ونموذج شبكة VGG16 بالإضافة إلى نموذج شبكة VGG19. تتميز هذه النماذج باعتمادها على بنية الشبكة العصبونية الالتفافية Convolutional Neural Network (CNN) وأكثر دقة في استخلاص سمات الصورة من النماذج السابقة. أما لعملية التدريب وربط الوصف مع السمات فقد تم استخدام شبكة عصبونية تكرارية Recurrent Neural Network (RNN).

تم استخدام مجموعة بيانات MSCOCO العالمية حيث تم أخذ مجموعة جزئية منها مؤلفة من 10000 صورة، حيث تم أخذ 9000 صورة منها لمجموعة التدريب Training و1000 صورة لمجموعة التحقق Validation. أما لعملية الاختبار فقد تم استخدام صور من الحياة الطبيعية من خارج مجموعتي التدريب والتحقق.

تمت مقارنة النماذج الثلاثة باستخدام معايير تقييم مختلفة وهي Top−1 وTop−5والعمق والدقة والتي تحدد مدى قرب الوصف الناتج من الوصف الفعلي للصورة. من النتائج تبين أن النموذج ResNet50 يتفوق على النموذجين VGG16 وVGG19 من ناحية دقة الوصف ومدى التشابه مع الوصف الصحيح للصور المدروسة. كما تم ملاحظة أن النماذج الثلاثة تعطي وصفاً أدق وأكثر تشابهاً للصورة عند حساب القيمة المتوسطة لأفضل ثالث توصيف من خرج النظام.

الكلمات المفتاحية: التعلم العميق، نظام وصف الصور، التمثيل الصوري، التمثيل النصي، نماذج شبكات التعلم العميق.

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

** ماجستير في هندسة الاتصالات المعلوماتية قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية – جامعة تشرين–اللاذقية – سوريا، Email: rasha_mualla90@hotmail.com

مجلة جامعة تشرين للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (41) العدد (2) 2019
Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (41) No. (2) 2019

Performance Evaluation of Image Description Systems Based on Different Deep Learning Models

Dr. Jafar Alkheir [*]
Rasha Mualla **

(Received 13 / 12 / 2018. Accepted 22 / 4 / 2019)

□ ABSTRACT □

In the current research, an image description process is done by generating captions of the image's components by three different models and comparing them. In order to build the description models, the Keras library was used as Machine Learning Framework that includes the most important libraries needed to establish and train deep learning networks. Three models were applied to the Keras library to extract features from images, the ResNet50, the VGG16 and the VGG19 network models. These models were characterized by their reliance on the Convolutional Neural Network (CNN) and more accurately in extracting image attributes from previous models. For the training process and matching the description with the features, a Recurrent Neural Network (RNN) was used.

The MSCOCO dataset was used with a subset of 10,000 images taken, 9,000 of which were taken for Training and 1000 for Validation. For the testing process, images of normal life were taken from outside the training and verification groups.

The three models were compared using several measurements which are Top-1, Top-5, depth and accuracy. These metrics define how closely the resulting description of the actual description of the image. The results show that the ResNet50 model outperforms both models VGG16 and VGG19 in terms of the accuracy and the convergence with the correct description of the studied images. Furthermore, the three models give a better accurate and convergence when they use the average value of the best third description of the system output.

Keywords: Deep Learning, Image Description System, Image Representation, Text Representation, Deep Learning Network Models.

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

****Master of Tele-informatics , Department of computer and automatic control Engineering, Faculty of Mechanical and electrical Engineering, Tishreen University, Latakia, Syria, Email : rasha_mualla90@hotmail.com**

مقدمة:

يعتبر التعلم العميق Deep Learning (DL) واحد من أحدث وأهم مجالات تعلم الآلة Machine Learning (ML) وأكثرها تطوراً واستخداماً من قبل الشركات العالمية ضمن مجالات متعددة من التطبيقات. يوجد العديد من تطبيقات التعلم العميق خصوصاً فيما يخص مجال الصورة مثل التعرف على الصور Image Recognition ووصف الصور Image Description ووضع تسميات توضيحية لمكونات الصورة Image Captioning إلخ... .

إن الفكرة الأساسية التي أدت لظهور مفهوم التعلم العميق DL هي قصور تقنيات الذكاء الصنعي التقليدية في تحقيق الأهداف المطلوبة منها خصوصاً مع تزايد وتضخم البيانات التي تعالجها؛ فمثلاً، مع زيادة عدد الطبقات في الشبكات العصبية التقليدية فإن العديد من المشاكل تبدأ بالظهور منها زمن التدريب الكبير وتباطؤ عملية التدريب خصوصاً في حال استخدام خوارزميات تعتمد على التدرج gradient في التعلم مثل الانتشار الخلفي للخطأ backpropagation وعندها فإن قيمة التدرج ستقل مع زيادة عمق الشبكة وعدد طبقاتها مما يجعل عملية التعلم أبطأ مع زيادة العمق كونها تعتمد بشكل أساسي على قيمة التدرج. وهناك مشكلة أخرى تظهر بسبب زيادة عدد عصبونات الطبقة الواحدة وهي مشكلة الـ overfitting وفيها تتسرع عملية التدريب بشكل كبير وتنتهي بوقت قصير بسبب القفزات الكبيرة في قيم التدرج gradient والتي ستسرع الوصول للهدف مما يجعل أداء الشبكة مناسباً لعينات التدريب لكنه سيكون سيئاً جداً مع عينات الاختبار المختلفة عن عينات التدريب [1،2،3،4].

إضافة لكل ما سبق فإن المشكلة الحقيقية في الشبكات العصبية التقليدية هي أن عملية التدريب تستغرق وقتاً طويلاً مع زيادة عدد العصبونات وعدد الطبقات والسبب يعود لزيادة الأوزان والإنحيازات التي يجب استخدامها مع توسيع تلك المعمارية، وإضافةً لكل ذلك فإن الشبكات العصبية التقليدية تكون كاملة الاتصال أي أن كل عصبون في أي طبقة سيتصل مع كل العصبونات في الطبقة المجاورة وهذا يخلق فضاء كبير من الأوزان التي يجب معالجتها ضمن مرحلة التدريب.

تم حل المشاكل السابقة باستخدام شبكات التعلم العميق وذلك من خلال عدة أمور منها جعل عملية الاتصال بين العصبونات غير كاملة أي ليس بالضرورة للعصبون أن يتصل مع كل عصبونات الطبقة المجاورة إضافة إلى تقليل البيانات مع زيادة عمق الشبكة من خلال استخدام خوارزميات الانتخاب pooling، والاحتفاظ بالبيانات المهمة منها فقط. أيضاً تم حل بعض المسائل التي تتطلب ربط الخرج بالدخل والتي كان من غير الممكن حلها بالشبكات التقليدية حيث يوجد أنواع من شبكات التعلم العميق تستطيع ربط الخرج بالدخل أو ما يسمى بعنصر الذاكرة. إضافة لكل ذلك تم حل مشاكل Overfitting و Gradient Vanishing من خلال عدة طرق تستخدمها شبكات التعلم العميق مثل شبكات Recurrent Neural Network (RNN) وشبكات Long-Short Term Memory (LSTM) التي تحتفظ بعناصر ذاكرة تمكنها من حفظ خرجها في فترات زمنية محددة واستخدامها لاحقاً [5،6،7،8]. إذاً هناك الكثير من التطبيقات التي كانت صعبة الإنجاز أو ربما مستحيلة بدون شبكات التعلم العميق.

الدراسات المرجعية ذات الصلة:

عبر العقد الأخير من الزمن، تطور استخدام الشبكات العصبونية والتعلم العميق بشكل كبير في مجال معالجة اللغات الطبيعية Natural Language Processing (NLP) حيث بين Collobert وآخرون أن إطار التعلم العميق البسيط قد تفوق على كل الطرق التقليدية المستخدمة لمعالجة اللغات الطبيعية [9]. كما وضح أن العديد من التطبيقات مثل التعرف على المكونات وعنونة الصور المعتمدة على التسميات المعنونة وغيرها، تستخدم التعلم العميق كحل فعال

لها وخصوصاً الشبكات العصبونية الالتفافية Convolutional Neural Network (CNN) والشبكات العصبونية التكرارية RNN.

قدم الباحث Pennington طريقة مشهورة لتضمين كلمات الوصف في اللغات الطبيعية [10] وتحويلها لنموذج وصف قابل للمعالجة من قبل الآلة. تسمى الطريقة المتبعة من قبله طريقة العد "Count Method"، حيث يتم عد عدد مرات ظهور الكلمة وتشكيل مصفوفة التكرار co-occurrence للكلمة ثم يتم تطبيع المصفوفة الناتجة لمجال موحد لكل الكلمات وتحويل المصفوفة إلى شعاع ببعد أصغري باستخدام تابع الخسارة Loss Function.

قدم الباحث Yang نموذج متعدد الشبكات لوصف الصور [11] يتم فيه استخلاص المعلومات الأساسية لمكونات الصورة ومواقعها في المشهد. تم استخدام شبكة RNN المعتمدة على وحدات LSTM لتوليد وصف للصور. تم تطبيق الاختبارات على مجموعة جزئية من مجموعة بيانات MSCOCO وتم التوصل لدرجة 0.28 على مقياس Bleu-4، وهو إحدى أهم المقاييس المستخدمة في تقييم أداء خوارزميات التعلم العميق. لم تقدم الدراسة أي تحسين ملحوظ في الأداء عن الدراسات السابقة في مجال وصف الصور، كما أنها لم تعطِ إمكانية لاستعادة الصور من خلال تقديم وصف للنموذج المصمم.

طور Huang [12] طريقة جديدة لتصميم نموذج لمعالجة اللغات الطبيعية باستخدام الشبكات العصبونية والتعلم العميق، حيث تم الاعتماد على شبكة Tensor Product Generation Network. كان الهدف توليد تسميات توضيحية لوصف الصور. تفوق النموذج المصمم على أداء شبكة LSTM ذات التعلم العميق ذو الذاكرة طويلة – قصيرة الأمد. تم توليد تسلسل من الفئات القواعدية (أسماء، ضمائر، صفات) والحصول على كلمات وصف الصور وفقاً لفئاتها القواعدية. حققت الدراسة درجة أداء 0.305 وفق مقياس Blue-4 متفوقة على أداء LSTM والذي كان 0.292. من عيوب الدراسة أنها أخذت بعين الاعتبار فقط توليد وصف للصور دون إمكانية الحصول على الصور من خلال وصفها.

استخدم الباحث Shah [13] شبكات التعلم العميق LSTM ونموذج Show & Tell لوصف الصور وتم تدريب النموذج على صور من مجموعة بيانات MSCOCO وتم تطبيق الاختبارات العملية على جهاز حاسب Intel Xeon E3 processor المؤلف من 12 نواة و32 جيغابايت من الذاكرة وتم استخدام مكتبة TesnsorFlow لتدريب الشبكة المقترحة . تم اختبار النموذج على صور من مجموعة بيانات MSCOCO وأعطى النموذج وصف بكلمات للصور وليس بجمل كاملة.

في البحث [14]، تم الاستعانة بالشبكات الالتفافية CNN وشبكات الذاكرة طويلة-قصيرة الأمد LSTM من أجل بناء نظام وصف للصورة باللغتين العربية والانكليزية ومقارنة تأثير اختلاف اللغات على أداء النظام المقترح. تم في هذا البحث بناء ملفات الوصف الخاصة باللغة العربية كما تم استخدام مجموعة بيانات صور Flickr8k. بينت النتائج العملية أن نظام الوصف باللغة الإنكليزية أعطى دقة وصف أعلى كما أن النتائج أوضحت أن بناء نظام كامل للوصف بالاعتماد على مجموعة بيانات وصف عربية قد أعطى دقة أعلى من ترجمة خرج نظام الوصف باللغة الإنكليزية.

مؤخراً، قام الباحث He [15] وعدد من الباحثين الآخرين بتقديم نموذج جديد لشبكة تعلم عميق تعتمد مبدأ Residual Learning وسميت شبكات ResNet لتسهيل عملية تدريب الشبكات العميقة جداً Very Deep Networks والتي تقدم أداء عالي لكنها تتطلب وقت تدريب طويل. تم في ذلك البحث إعادة بناء الطبقات بحيث تتعلم باستخدام التوابع

من النوع Residual مع الأخذ بالحسبان قيمة الدخل بدلاً من استخدام توابع غير مرجعية. تمكنت تلك الشبكات المصممة من رفع الأداء مع زيادة العمق وكانت سهلة التعلم وسهلة التعديل والتطوير. تم اختبار النموذج المقترح باستخدام عدد طبقات مختلفة [34، 50، 110، 152] وذلك من أجل اختيار عدد الطبقات الأمثل للنموذج. بيَّن الباحثون أنه على الرغم من أن الشبكة أكثر عمقاً من نماذج طبقات VGG [16] إلا أنها أكثر سهولة في التدريب منها وأقل تعقيداً في الحسابات.

في هذا البحث سيتم بناء ومقارنة ثلاثة أنظمة لوصف الصور بالإضافة إلى الحصول على الصورة من خلال وصفها معتمدةً على ثلاثة نماذج مختلفة هي: ResNet50, VGG16, VGG19، كما وسيتم استخدام معايير تقييم جديدة لتقييم أداء الأنظمة المقترحة والتي سيتم شرحها لاحقاً.

أهمية البحث وأهدافه:

يعتبر البحث الأول من نوعه في جامعة تشرين من حيث تطرقه إلى تطوير نظم وصف الصور باللغة الإنكليزية وبدقة وصف عالية، حيث يمكن الاستفادة منها في التطبيقات التي تحتاج عملية وصف آلية بلغة طبيعية للصور مثل التطبيقات الموجهة لفاقدي البصر التي تجعلهم قادرين على معرفة ماهية المشهد المحيط بهم، والتطبيقات الهادفة لتعليم الأطفال بحيث يتم تحويل الصور إلى محتوى نصي مقروء ومسموع لتعليمهم كيفية التحدث، عدا عن التطبيقات الأمنية التي تتضمن عمليات التحكم وكاميرات المراقبة وإصدار التقارير المباشرة اعتماداً على ماهية محتويات الصور التي تلتقطها تلك الكاميرات، وانتهاءً بتطبيقات التواصل الاجتماعي التي تتضمن تبادل واسترداد الملايين من الصور.

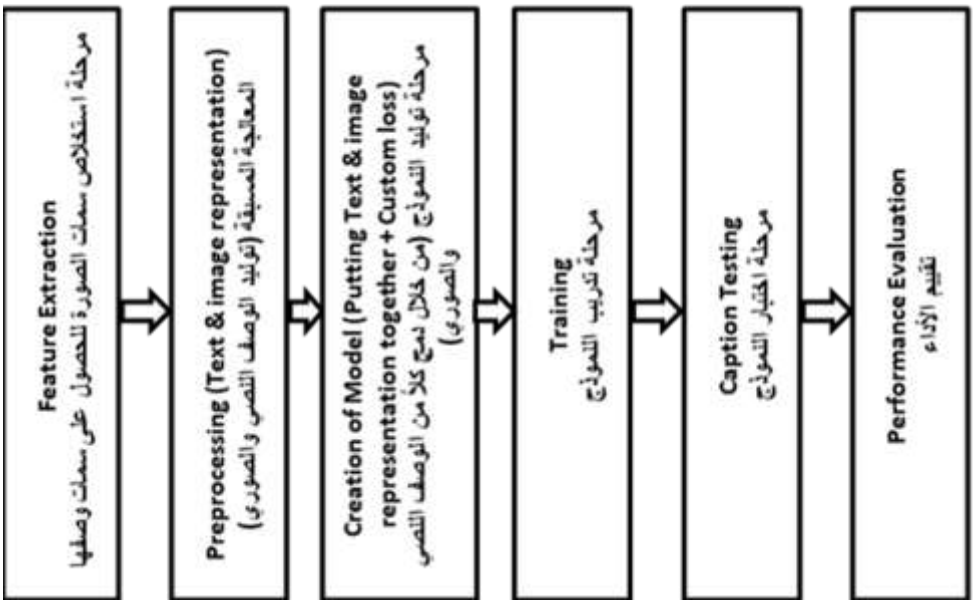
يهدف البحث لمقارنة كلاً من نتائج النماذج الثلاث وإبراز محاسنها ومساوئها من خلال تقديم تسميات توضيحية لوصف المكونات الموجودة في الصورة باللغة الإنجليزية.

طرائق البحث ومواده:

يعتمد هذا البحث على المنهج التحليلي والتجريبي في عملية المقارنة بين النماذج الثلاثة المقترحة ومناقشة النتائج. تم الاعتماد على مجموعة من الأدوات البرمجية والمادية. حيث تم استخدام جهاز حاسب بالمواصفات التالية: معالج Intel Xeon 2.83 GHZ، ذواكر 3GB. تم الاعتماد على نظام التشغيل Ubuntu 18.04 ولغة الـ Python 3.6 ومنصة الـ JUPYTER. كذلك تم استخدام مكتبتي Keras وTensorFlow من أجل بناء نماذج شبكات الـ DL. تم استخدام خوارزميات التعلم العميق مثل CNN, RNN ونماذج شبكات التعلم العميق ResNet50, VGG16, VGG19. بالنسبة لمجموعة البيانات، تم استخدام 10000 صورة منتقاة بشكل عشوائي من مجموعة البيانات المعيارية MSCOCO بالإضافة إلى ملفات الوصف الخاصة بها Captions والتي تتألف من خمس جمل لكل صورة، حيث تم تعديل هيكلية هذا الملف لتصبح جملة واحدة لكل صورة، تمثل أفضل وصف للصورة لتخفيف العبء على مرحلة التدريب [17].

النظام المقترح:

يتألف النظام المقترح من ست مراحل أساسية موضحة في الشكل (1). حيث تم تقييم أداء ثلاثة نماذج باستخدام مجموعة إضافية من معايير التقييم المختلفة عن المستخدمة في الدراسات المرجعية والتي أظهرت نتائج المقارنة بشكل أفضل.



الشكل (1) نظام وصف الصور المقترح

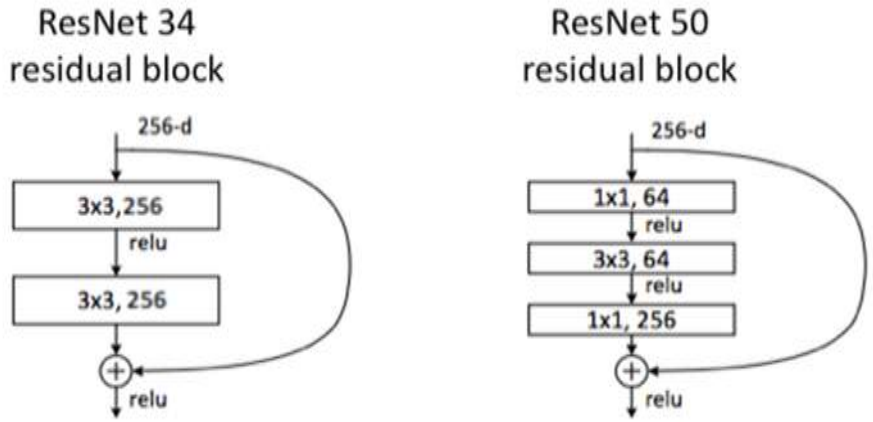
مرحلة استخلاص السمات:

تم استخدام عدة نماذج لإنجاز عملية الوصف الصوري وهي نماذج ResNet50, VGG16, VGG19 وجميعها مطبقة ضمن مكتبة الـ Keras. السبب الأساسي في اختيار هذه النماذج هو العمق الكبير المستخدم لبناء الشبكة والذي يجعل منها أكثر كفاءة في استخلاص سمات الصور. حيث بينت العديد من الدراسات أن زيادة العمق تزيد من فرصة الحصول على سمات أدق وأكثر وضوحاً وتتضمن معلومات أفضل عن مكونات الصورة [15،16].

عندما نقوم بداية بتحديد الفرق بين النموذجين VGG وResNet، قد نجد أن عدد البارامترات المكونة لكل منهما هو العامل الأكثر تأثيراً، حيث أن VGG لديها ما يقارب 138 مليون بارامتر في حين يملك نموذج الـ ResNet ما يقارب 25.6 مليون بارامتر فقط مع العلم أنّ VGG مكون من 16 أو 19 طبقة، بينما الـ ResNet يتكون من 34 أو 50 أو 110 أو 152 طبقة (تم في هذا البحث استخدام عدد الطبقات 50 اعتماداً على الدراسة [15] والموضح في الشكل(2)). يؤثر هذا الفرق عملياً بشكل أساسي على المساحة المطلوبة لتخزين الشبكة.

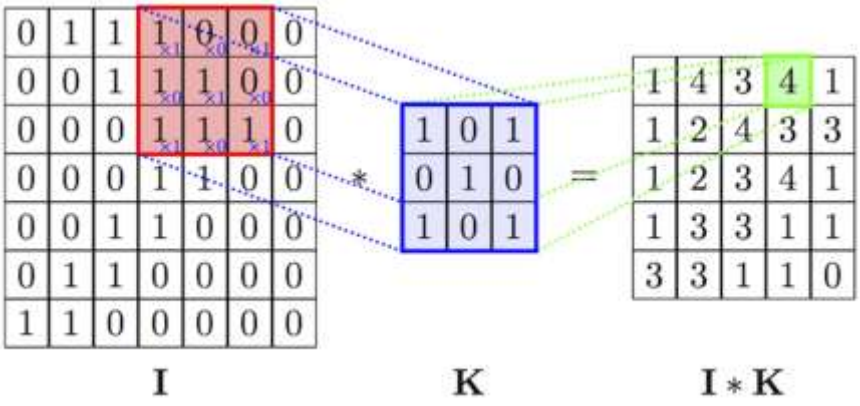
حسب ما هو معروف قبل نشوء نموذج شبكات Residual او المعروفة بـ ResNet فإن زيادة العمق يجب أن تزيد من دقة الشبكة، طالما أنه تم حل مشكلة الOverfitting. إن الإشارة المطلوبة لتغيير الأوزان تنشأ من نهاية الشبكة بمقارنة القيمة الحقيقية وقيمة الخرج المتنبئ بها. مما يؤدي إلى ظهور مشكلة أن هذه الإشارة تصبح صغيرة جدًا في الطبقات الأقدم (الأولى) مع زيادة العمق. وبمعنى أخر تصبح الطبقات الأقدم عديمة النفع مع زيادة عدد الطبقات، وهذا ما يسمى تلاشي الانحدار (Vanishing Gradient Problem). المشكلة الثانية في تدريب الشبكات الأعمق هي إجراء التحسين على كمية بارامترات ضخمة وبالتالي إضافة طبقات جديدة سيؤدي إلى خطأ تدريبي أعلى. وهذا ما يسمى مشكلة الانحطاط (Degradation Problem). تسمح الشبكات من النموذج Residual بتدريب مثل هذه

الشبكات العميقة من خلال بناء شبكة مكونة من وحدات تسمىResidual Models والتي تختلف عن بعضها البعض باختلاف عدد الطبقات المستخدمة من حيث استخدام أنواع مختلفة من مرشحات التحويل كما هو موضح في الشكل(2)



الشكل (2) هيكلية شبكات الـ ResNet34 و ResNet50

بعد معرفة أهم المشاكل التي عالجتها شبكات الـ ResNet والتي عانت منها شبكات الـ VGG، يجب معرفة السبب خلف السرعة الكبيرة في التدريب التي تحققها شبكات الـ ResNet، باعتبار أن سرعة الشبكات العصبية الالتفافية تعتمد بشكل كبير على حجم الدخل. لنفترض صورة رمادية (قناة لونية واحدة) بأبعاد 100×100. عند تطبيق مرشح تحويل 3×3 مع حشوة 1 و0 عليها، فإن هذه العملية تتطلب حوالي 163K FLOPs هنا يمكن ملاحظة أنه من أجل صورة بأبعاد 100×100 سيتم الحصول على صورة جديدة بأبعاد 98×98. من أجل حساب قيمة كل بيكسل من الصورة الناتجة ذات الأبعاد 98×98، يجب القيام بـ 9 عمليات ضرب و8 عمليات جمع، أي ما مجموعه 17 عملية لكل قيمة. وعليه، إذا تم حساب كل ما سيتم الحصول عليه من عمليات (98*98 *17) يصبح مساوياً FLOPs 268,163. الآن لو تم تطبيق نفس المرشح على صورة أكبر ولتكن 200x200. فإن الصورة ستحتوي على مساحة أكبر بـ 4 مرات، لذلك سيتم الحصول على أكثر من 4 أضعاف قيمة الـ FLOPs السابقة بشكل تقريبي. يوضح الشكل(3) عملية تطبيق المرشح على بكسلات الصورة.



الشكل (3) مبدأ عمل المرشحات المستخدمة في شبكات الـ DL

عادة ما يتم تقسيم بنية الشبكة العصبونية إلى مجموعة من البلوكات، وفي نهاية كل بلوك يتم تقليص بعدي الارتفاع والعرض بعامل تقسيم يساوي إلى 2 بعد كل طبقتين، بينما يقوم أول بلوك والمكون من طبقة واحدة فقط ضمن الـ

ResNet50بتقليل ارتفاع الصورة وعرضها بمقدار 4 وذلك بسبب وجود عملية تقسيم إضافية على 2 ضمن الطبقة المكونة للبلوك.

من هيكلية VGG، يمكن رؤية أن الطبقتين الأولى والثانية تطبقان عملية الالتفاف على الصورة الكاملة 224×224 في بداية الشبكة وهي عملية مكلفة للغاية. وعليه فإن الطبقة الأولى تحتاج إلى 170M FLOPs لإنجاز عملية الالتفاف، ولكنها تنتج خرجاً بحجم 64X224X224 من الصورة المدخلة بحجم 3X224X224. بما أن الطبقة في ResNet50تطبق مرشح التحويل نفسه، فيجب أن يكون عدد الـ FLOPs قريبًا من (64/3)*170M. لكن في الواقع، هي تقريبًا 3,7 مليار فقط. وبالتالي فإن طبقة واحدة في VGG لديها ما يقارب عدد الـ FLOPs المطلوب ضمن شبكة الـ .ResNet50. لتجنب هذه المشكلة الحسابية في حساب كل هذه البارامترات فإنّ ResNet50 تعالج هذه المشكلة في الطبقة الأولى. حيث يتم تقليل عدد الصفوف والأعمدة بواسطة عامل يساوي إلى 2 ويستخدم فقط 240M FLOPs، ليتم بعدها تطبيق عملية التجميع max pooling والتي ستؤدي إلى تخفيض آخر بعامل 2 أيضاً. بعد ذلك، تتباطىء عملية الترشيح ضمن المرشحات الالتفافية في المراحل اللاحقة مقارنة مع عملية الترشيح في المراحل الأولى. تستخدم ResNet50 فكرة شبكات أرق، ولكن أعمق (Thinner but Deeper Networks) عن طريق استخدام أنوية Kernels بشكل متناوب بين عمليات الالتفاف وتوابع التنشيط غير الخطية، الأمر الذي يظهرها وكأنها تستخدم أنوية Kernels أقل في ResNet50 مقارنة مع VGG.

تم استخدام الكتل الإلتفافية في ResNet50 لتقليل عدد العمليات بشكل أكبر، وذلك أثناء استخدام شبكات ذات عدد أكبر من المرشحات في الطبقات الالتفافية. تعتمد ResNet50 على فكرة استخدام الطبقة الالتفافية 1x1 لتقليل عمق القناة (Channel depth) قبل تطبيق الطبقة الالتفافية 3x3، والتي كانت مستخدمة في النماذج السابقة، وإعادة استخدامها مرة أخرى بعد طبقة 3x3 الالتفافية.

مرحلة التحضير (ما قبل التدريب):

يعتمد النظام المقترح على التعليم المتري (metric learning)، حيث سيتم استخدام العملية Dot بين التمثيل الصوري والنصي (image and text representation) من أجل مطابقة التسمية التوضيحية مع الصورة عن طريق اختيار القيمة الأعلى لعملية الـ Dot والتي تمثل مطابقة التسمية التوضيحية مع الصورة بينما يمثل انخفاض هذه القيمة اختلاف التسمية التوضيحية مع الصورة. وعليه فإن النظام المقترح سيقوم بحساب ناتج عملية الـ dot على زوج التمثيل الصوري والنصي للحصول على قيمتين الأولى تعبر عن المطابقة بين زوج التمثيل الصوري والنصي (الإيجابية) والثانية تعبر عن عدم المطابقة بين زوج التمثيل الصوري والنصي (السلبية). يتم استخدام هذه القيم في مرحلة التدريب باستخدام الحد الأقصى لفقدان الهامش (maximum margin loss)، الأمر الذي سيجعل قيمة زوج التمثيل الصوري والنصي المطابق أعلى من واحد زائد قيمة زوج التمثيل الصوري والنصي الغير مطابق.

إن الطريقة المثلى لعملية إيجاد زوجي التمثيل الصوري والنصي المطابق والغير مطابق هي تجريب جميع الأزواج، إلا أن هذه العملية مكلفة جداً لذلك سيتم استخدام مجموعة جزئية عشوائية من أزواج التمثيل الصوري والنصي من أجل إيجاد الزوج المطابق والغير مطابق كما هو مقترح في الدراسة المرجعية [17].

التمثيل الصوري (Image representation):

من أجل الحصول على التمثيل الصوري المناسب للنظام المقترح، تم استخدام النماذج الثلاثة المقترحة (,ResNet50 VGG16, VGG19) كل نموذج على حدى بعد إزالة طبقة القرار (Decision Layer) من هذه النماذج وذلك من

أجل الحصول على سمات الصور التي ستستخدم في مرحلة التدريب. فمثلا سيتم الحصول على شعاع سمات مؤلف من 2048 سمة في حال استخدام النموذج ResNet50، بينما سنحصل على شعاع سمات مؤلف من 25088 سمة عند استخدام نموذجي الـ VGG.

كما ذكرنا سابقاً، تحتاج عملية التدريب إلى مقارنة أزواج التمثيل الصوري والنصي، وعليه يجب أن يكون حجم كلا التمثيلين متساوٍ. من أجل ذلك تم إضافة طبقة جديدة إلى النماذج المقترحة وهي طبقة الـ Dense layer والتي تحول حجم شعاع سمات الصور (التمثيل الصوري) من 2048 في حال النموذج ResNet50 و25088 في حال نموذجي الـ VGG إلى 256 وهو نفس الحجم الذي سنحصل عليه للتمثيل النصي والذي سيتم شرحه في الفقرة التالية.

التمثيل النصي (Text representation):

من أجل الحصول على التمثيل النصي سيتم استخدام الملف الخاص بالتسميات التوضيحية الخاصة بمجموعة الصور المدروسة في هذا البحث (10000 صورة) والمؤلف من اسم الصورة متبوعاً بجملة توضيحية واحدة لتلك الصورة. هنا يجب الانتباه إلى توافق كل سطر من الملف الخاص بالتسميات التوضيحية مع السطر المقابل له ضمن التمثيل الصوري.

من أجل عملية المقارنة مع التمثيل الصوري باستخدام العملية dot يجب تحويل الجمل المستخلصة من الملف السابق إلى سلسلة من الأعداد الصحيحة متساوية بالحجم مع سطر التمثيل الصوري. يتم تحقيق ذلك باستخدام التابع Tokenizer و pad_sequences الموجودين ضمن مكتبة الـ Keras. بعد ذلك سيتم عرض هذه السلسلة بفضاء من 100 بعد (Dimintional space=100) للتضمينات (embeddings) والتي سيتم إدخالها إلى وحدات متكررة البوابات (Gated Recurrent Units GRU). هنا سيتم استخدام آخر حالة مخفية ضمن الـ GRU من أجل تمثيل النص بأكمله عوضاً عن الحصول على خرج الـ GRU والذي يمثل التنبؤ على مستوى الكلمة. وفي النهاية يمكن ترشيح هذه التضمينات (embeddings) لتتضمن فقط الكلمات المستخدمة في التسميات التوضيحية.

مرحلة إنشاء النموذج

كمرحلة أولى سيتم دمج التمثيل الصوري والنصي عن طريق استخدام بعض التوابع الوظيفية الموجودة ضمن مكتبة الـ Keras والتي تسمح بمشاركة الأوزان بين أجزاء النموذج المراد إنشاؤه وذلك لإنشاء أزواج التمثيل الصوري والنصي الإيجابي والسلبي. يحتوي النموذج المقترح على ثلاث مدخلات وهي: سمات الصورة (والتي تم استخراجها باستخدام النماذج الثلاثة المقترحة) وتسمية توضيحية صحيحة بالإضافة إلى تسمية توضيحية غير متطابقة (تسمية توضيحية ضجيجية) والتي يتم اختيارها بشكل عشوائي من تسميات الصور التدريب الأخرى. الخطوة التالية ستتضمن اعداد الطبقات التي سيتم استخدامها في أمكان متعددة من هذا النموذج. هذه الطبقات هي: طبقة التضمينات GloVe وطبقة الـ GRU المستخدمة للحصول على التمثيل النصي بالإضافة إلى طبقة الـ Dense والتي ستستخدم للحصول على التمثيل الصوري. بعد ذلك سيتم انشاء مسارات التنفيذ (Piplines) لكل مدخل على حدى عن طريق استدعاء الطبقات المناسبة لكل مدخل مع اتاحة مشاركة الأوزان لمسار التنفيذ لكل من التسمية التوضيحية الصحيحة والضجيجية.

سيتم استخدام تابع الدمج الخاص بمكتبة الـ Kears من أجل تحقيق عملية الدمج بين التمثيل الصوري والنصي باستخدام العملية dot والتي سيتم تطبيقها على سلسلة الأزواج الناتجة للصور المستخدمة في عملية التدريب مع التسميات التوضيحية المقابلة لها. ناتج عملية الدمج سيكون ناتج النموذج والذي سيستخدم لاحقاً ضمن تابع حساب الخسارة (Loss function).

بناءً لما سبق يمكن إنشاء عدة نماذج اعتماداً على المدخلات والمخرجات كما يلي: نموذج التدريب بخرج بصيغة سلسلة من نتائج الأزواج الإيجابية والسلبية، النموذج الثاني الخاص بتوليد مسار التنفيذ للتمثيل الصوري بالإضافة إلى النموذج الخاص بتوليد مسار التنفيذ للتمثيل النصي. سيتم استخدام نموذجي توليد مسارات التنفيذ في مرحلة التنبؤ في حالة استخدام النظام لصور جديدة مطلوب توصيفها.

تابع الخسارة وتابع الدقة

سيتم استخدام تابع الخسارة المقترح في الدراسة المرجعية [17] والذي يعطى بالعلاقة التالية:

$$Loss = \sum_{i} \max(0, 1 - p_i + n_i) \quad (1)$$

حيث pi يمثل نتيجة زوج التمثيل الصوري والنصي الإيجابي وni 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 عن خرج نموذج التدريب.

بالاعتماد على وسيطي تابع الخسارة يمكننا أيضاً حساب دقة النظام المقترح عن طريق حساب عدد المرات التي تكون فيها قيمة نتيجة زوج التمثيل الصوري والنصي الإيجابي أعلى من قيمة نتيجة زوج التمثيل الصوري والنصي السلبي.

سيناريو التدريب:

بعد الانتهاء من بناء النظام المقترح، سيتم تدريب كل نموذج من النماذج المستخدمة (ResNet50, VGG16, VGG19) باستخدام مجموعة صور مأخوذة بشكل عشوائي من المجموعة المعيارية MSCOCO مؤلفة من 10000 صورة، تم تخصيص 9000 صورة للتدريب و1000 صورة للتحقق (validation)، والتي تستخدم لحل مشكلة الـ Overfitting. تم استخدام ثلاث مدخلات هي: سمات الصور والتي تُمثل بمصفوفة حجم اسطرها يعتمد على النموذج المستخدم (2048 لـ ResNet50 و25088 لـ VGG)، والتسمية التوضيحية الصحيحة بالإضافة إلى التسمية التوضيحية الضجيجية والتي تُمثل بمصفوفة حجم اسطرها مساوٍ لـ 16، وهو الحجم الذي تم اقتراحه ليناسب طول الجملة الأعظمي المستخدمة في وصف الصور. سيتم حساب تابع الخسارة بالإضافة إلى الدقة ضمن مرحلة التدريب والتي ستستخدم لاحقاً في عملية تقييم أداء النظام عند استخدام النماذج المقترحة. سيتم تكرار عملية التدريب 10 مرات، في كل تكرار تتم عملية انتقاء عشوائية جديدة لمصفوفة التسمية التوضيحية الضجيجية. بناءً على ذلك سيتعلم النموذج مع كل عملية تكرار كيف يربط التسميات التوضيحية الصحيحة لمكونات الصورة مع السمات المستخلصة، بالإضافة إلى كونه سيتعلم عدم الربط بين التسميات التوضيحية الخاطئة (الضجيجية) مع هذه السمات، الأمر الذي سيؤدي إلى تحسين أداء مرحلة التدريب.

بعد انتهاء مرحلة التدريب سيتم حفظ الأوزان النهائية المستخدمة لإنشاء التسميات التوضيحية بالإضافة إلى حساب وحفظ التمثيل الصوري والنصي لكل الصور، والتي سيتم استخدامهما في عملية الاختبار لاحقاً.

معايير التقييم Evaluation Metrics:

تعد عملية التقييم مرحلة أساسية من مراحل تصميم وبناء نماذج التعلم العميق. يساعد التقييم على فهم دقة توقع النموذج المستخدم للخرج ليكون مطابقاً للخرج الصحيح إضافة لمهمته الأساسية في معرفة إذا كان بالإمكان أن نعتمد النموذج ليستخدم في قطاع العمل المناسب أو يجب إضافة تحسينات أو تعديلات أو حتى الانتقال بشكل جذري إلى نموذج آخر يحقق الغاية المنشودة. عند تقييم أي نموذج يجب اختيار معايير تقييم تناسب النموذج المستخدم وتعطي

نتائج واضحة يمكن فهمها بطريقة قد تمكننا من معرفة المواضع التي يمكن أن نعمل عليها لتحسين النموذج. للقيام بعملية حساب مدى تشابه كل تسمية توضيحية متوقعة من قبل النظام المقترح مع التسمية التوضيحية الصحيحة، يستخدم مفهوم مدى اختلاف شعاعين هندسيين عن بعضهما عن طريق حساب الزاوية بينهما ثم حساب تجيب (cosine) هذه الزاوية. الزاوية التي قياسها 0° يكون cosine(0°)=1(وبالتالي يكون الشعاعان الناتجان عن التسمية التوضيحية المتوقعة والصحيحة منطبقان والتوصيفان متشابهان تماماً. الزاوية التي قياسها 90° يكون cosine (90°)=0(وبالتالي التوصيفان مختلفان تماماً. كلما اقتربت قيمة التجيب من الواحد كلما كانت التسمية التوضيحية المتوقعة أقرب للصحيحة. ليتم حساب تجيب زاوية التشابه هذه يجب تحويل كل تسمية توضيحية إلى شعاع رياضي يعبر عن هذا التوصيف وذلك باستخدام التابعين TfidfVectorizer و fit_transform من مكتبة sklearn. أخيراً تتم عملية حساب تجيب الزاوية لأشعة التسميات التوضيحية المتوقعة مع شعاع التسمية التوضيحية الصحيحة باستخدام التابع cosine_similarity وتخزن في مصفوفة ليتم حساب معايير التقييم التي نريده فيما بعد.

تم الاعتماد على عدة معايير تقييم منها ثلاثة معايير جديدة (4، 5، 6) لإظهار نتائج المقارنة بشكل أفضل. وفيما يلي سيتم وصف كل معيار على حدى:

1. الدقة **Accuracy**: أثناء التدريب يتم حساب دقة التعرف على الخرج الصحيح من الخرج المتوقع لعينات التدريب (Training Dataset) وهذا ما يسمى دقة التدريب (Training Accuracy)، كما يتم حساب دقة التحقق بعد كل تكرار لكامل 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). تتم الاستفادة من هذين المعيارين في الحفاظ على مسار التدريب وعدم ظهور مشكلة الـ Overfiting ومعرفة مدى دقة النموذج ليكون التوقع صحيحاً.

2. تشابه القيمة الأعلى **Top-1 Similarity**: ستتم دراسة مدى تشابه الجملة المتوقعة والتي تمثل أفضل توصيف 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 ستتم 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.

3. أفضل تشابه للقيم الخمس العليا **Top-5 Similarity**: هنا ستتم دراسة مدى تشابه الجملة الأصلية التي تصف الصورة مع كل جملة من الجمل الخمس الأفضل توقعاً والتي تمثل أعلى خمس توصيفات للصورة وأخذ قيمة أعلى تشابه. بناءً على ما سبق سيتم أخذ قيمة المتوسط الحسابي لجميع قيم التشابه العظمى وذلك من أجل كل عينات الاختبار المدخلة. هنا أيضاً ستتم دراسة فاصل الثقة بنسبة 95% لهذا المتوسط الحسابي.

4. عمق أفضل تشابه **Depth of Best Similarity**: بعد معرفة أفضل تشابه لمجموعة الجمل الخمس العليا مع الجملة الأصلية ستتم دراسة على أي عمق كان هذا التشابه، أو بمعنى أخر ما هو 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 من أجل كل عينات الاختبار، وذلك لتقدير مركز العمق الأمثل. ويستخدم فاصل الثقة بنسبة 95% لهذا المتوسط لدراسة دقته وتوزع قيم العمق حوله.

5. أدنى وأعلى قيمة للتشابه **Max and Min Values of Similarity**: يوضح معيار أعلى قيمة للتشابه، القيمة الأفضل التي تم الحصول عليها بين كل التشابهات في top-5 ومن أجل كل عينات الاختبار. في حين يوضح معيار أدنى قيمة للتشابه، القيمة الأدنى التي تم الحصول عليها بين كل التشابهات في top-5 ومن أجل كل عينات الاختبار.

6. عمق أدنى وأعلى قيمة للتشابه **Depth of Max and Min Values of Similarity**: يدل على العمق الذي يظهر فيه كل من القيمتين العليا والدنيا لقيمة التشابه، والتي تحدثنا عنها سابقاً.

سيناريو الاختبار **Testing**:

ليتم اختبار النماذج المقترحة، قمنا بعدة خطوات متتالية بدءاً من اختيار بيانات الاختبار إلى اختيار توابع الاختبار وانتهاءً بتطبيق هذه التوابع على هذه البيانات ولكل نموذج على حدى. بشكل عام كان مسار التنفيذ للاختبار (Testing Pipeline) من الشكل التالي:

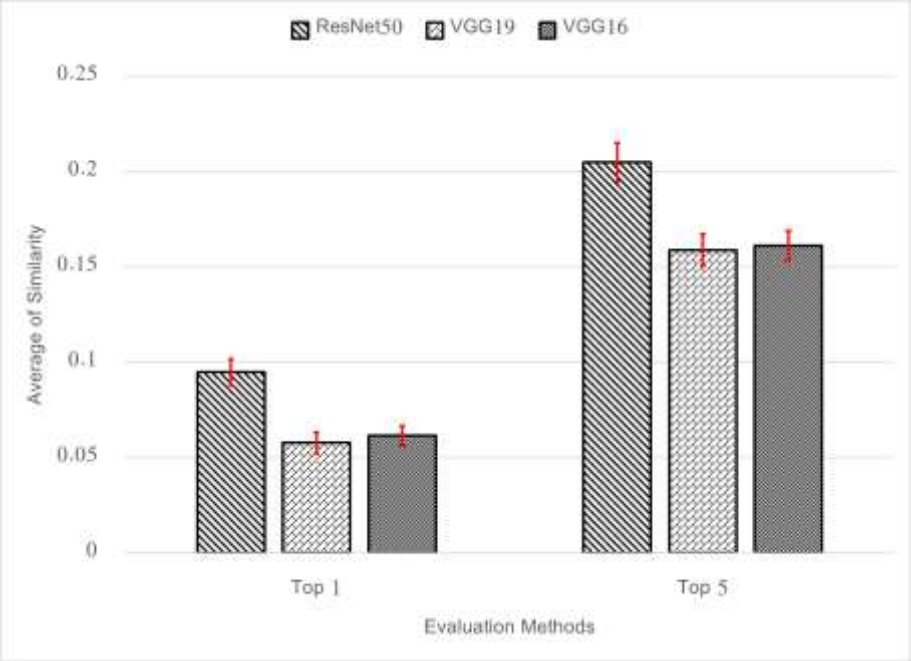
تم اختيار 1000 صورة من مجموعة البيانات المعيارية MSCOCO بشكل عشوائي ليتم الاختبار عليها، وكل صورة لها توصيف واحد.

تابع توليد التسميات التوضيحية (Generate Captions): يعمل هذا التابع على توليد أو توقع توصيفات للصورة الممررة الجديدة. يجب تحميل كل من نموذج التمثيل الصوري والنصي الناتج عن مرحلة التدريب (Load image_model and captions_model) والذي سيتم استخدامه لاستنتاج قيم التمثيل الصوري والنصي للصور الجديدة المدخلة (predicte image_representations and cations_representations). وبناءً على ذلك، لتوليد توصيف للصور الجديدة المدخلة، سيتم استخراج السمات الممثلة لها عن طريق تابع extract_features وتوليد التمثيل الصوري لهذه السمات عن طريق تابع الـ predict الخاص بالنموذج image_model. بعدها يتم دمج التمثيل الصوري والنصي باستخدام العملية (Dot Product). سيتم حفظ وترتيب التسميات التوضيحية المولدة اعتماداً على قيمة التوقع الناتج عن العملية Dot بشكل تنازلي، ليتم اختيار أفضل تسمية توضيحية كخرج لـ Top-1 واختيار أفضل خمس تسميات توضيحية كخرج لـ Top-5. يستخدم الخرجين السابقين كدخل لمرحلة التقييم.

مناقشة النتائج:

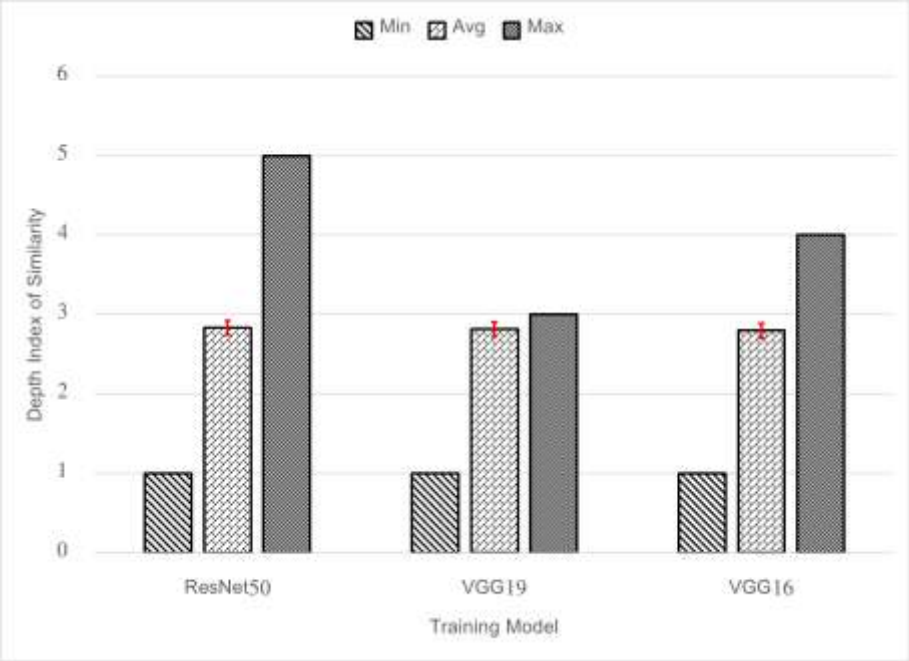
من أجل تقييم ومقارنة أداء النماذج الثلاثة المقترحة في هذا البحث، تم الاعتماد على أربعة معايير تقييم كما هو موضح سابقاً. تم الحصول على دقة النظام لكل نموذج أثناء مرحلة التدريب، حيث تفوق نموذج الـ ResNet50 على النموذجين الخاصين بـ VGG وحصل على دقة 95.9% بينما حصل كل من VGG19 وVGG16 على 89.1% و88.6%على التوالي. هنا لا بد من التنويه إلى زيادة زمن التدريب بشكل لاخطي لنماذج الـ VGG مع زيادة عدد عينات التدريب. لذلك ولتلافي هذه الزيادة اللاخطية تم تقسيم مجموعة التدريب عند استخدام النموذجين VGG16 وVGG19 إلى مجموعتين تتألف كل منهما من 5000 عينة ومن ثم دمج مصفوفتي التدريب الناتجة بمصفوفة واحدة. يبين الشكل(4) القيمة المتوسطة لنسبة التشابه بين الوصف المقترح من النظام باستخدام النماذج الثلاثة في الحالتين Top-1 وTop-5. بينت النتائج تفوق النموذج ResNet50 سواء في حال حساب التشابه لأفضل وصف مقترح للصورة Top-1 ولأفضل خمس توصيفات للصورة Top-5. بينما يظهر تقارب أداء النموذجين VGG16 وVGG19 مع أفضلية بسيطة للنموذج VGG16. من أجل بيان مدى موثوقية القيمة المتوسطة الناتجة، تم حساب

فاصل الثقة لكل نتيجة والذي يظهر مدى تقارب القيم الناتجة ضمن مجال القيمة المتوسطة. يظهر الشكل(4) قيمة صغيرة لفاصل الثقة مما يتيح لنا قبول القيمة المتوسطة كقيمة موثوقة يمكن الاعتماد عليها.



الشكل(4) القيمة المتوسطة لتشابه الوصف المقترح مع الوصف الصحيح للصور

يبين الشكل(5) القيمة المتوسطة والدنيا والعليا لدرجة عمق الحصول على أفضل تشابه عند معاينة أفضل خمس توصيفات للصور Top-5. من النتائج يتضح أن القيمة المتوسطة لعمق التوصيف الأكثر تطابقاً بحدود الثلاثة في جميع النماذج، إلا أن الاختلاف يظهر في القيمة العليا. لإظهار تأثير هذا الاختلاف تم حساب فاصل الثقة للقيمة المتوسطة. تبين النتائج صغر قيمة فاصل الثقة الأمر الذي يؤدي إلى امكانية اعتماد القيمة المتوسطة كنتيجة موثوقة واعتبار الاختلاف في القيمة العليا عبارة عن حالات شاذة يمكن تجاهلها. بناءً على ما سبق يمكن اعتبار أن أفضل توصيف للصور يمكن الحصول عليه ضمن مجال العمق بين 2 و4 وبالتالي يمكن تعديل خرج النظام ليصبح ثالث أفضل توصيف مقترح.



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

يمكن تلخيص النتائج السابقة ضمن الجدول (1) والذي يظهر مقارنة بين النماذج الثلاثة المستخدمة ResNet50 VGG16, VGG19، وفق معايير تقييم الأداء المقترحة.

الجدول (1) مقارنة بين النماذج الثلاثة المستخدمة ResNet50 VGG16, VGG19، وفق معايير تقييم الأداء.

النموذج المستخدم	Top-1	Top-5	Depth	Accuracy
ResNet50	0.094768	0.204743	1.835	%95.9
VGG16	0.061396	0.161092	1.799	%88.6
VGG19	0.057725	0.158807	1.814	%89.1

مما سبق يمكن القول إن نموذج ResNet50 كان الأفضل في عملية التوصيف من ناحية كل معايير الأداء ويعود ذلك لعدة أسباب الأول أن عمق الشبكة فيه أكبر من نموذجي VGG16،VGG19 والسبب الثاني استخدامه لخطوط التغذية الأمامية لتغيير الهدف وجعله مرتبطاً بقيمة الدخل.

الاستنتاجات والتوصيات:

تم في هذا البحث بناء ومقارنة ثلاثة نماذج لتوصيف الصور باستخدام الشبكات ResNet50, VGG16, VGG19. بينت نتائج الاختبار تفوق النموذج ResNet50 على النموذجين VGG16 وVGG19 من ناحية دقة الوصف ومدى تشابهه مع الوصف الصحيح للصور المدروسة. بالإضافة إلى ذلك تم ملاحظة أن النماذج الثلاثة يمكن أن تعطي وصف أدق وأكثر تشابهاً للصورة عند اختيار أفضل ثالث توصيف مقترح كقيمة متوسطة من خرج النظام. وعليه تم اقتراح تعديل خرج النظام المدروس ليصبح ثالث أفضل توصيف.

اعتمد النظام المدروس على الاختيار العشوائي لزوجي التمثيل الصوري والنصي المطابق والغير مطابق والتي تستخدم في عملية التحقق ضمن مرحلة التدريب، والذي لا يمثل الطريقة المثلى في عملية إيجاد زوجي التمثيل. وعليه يمكن

تحسين هذه الطريقة للوصول إلى طريقة مثلى لاختيار زوجي التمثيل والذي يتم العمل عليه كخطوة مستقبلية. كما ويمكن دراسة تأثير اختلاف اللغات على النظام المقترح عن طريق استخدام اللغة العربية كلغة وصف.

المراجع:

[1] Barbu A., Bridge A., Burchill Z., Coroian D., Dickinson S.,Fidler S., Michaux A., Mussman S., Narayanaswamy S.,Salvi D.,Video in sentences out. arXiv preprint arXiv:1204.2742, 2012

[2] Barnard K., P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. JMLR, 2003.

[3] Brostow, G., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-de_nition ground truth database. PRL 30(2) (2009),pp: 88-97

[4] Bell, S., Upchurch, P., Snavely, N., Bala, K.: OpenSurfaces: A richly annotated catalog of surface appearance. SIGGRAPH 32(4) (2013).

[5] Mao J., W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014.

[6] Mikolov T., M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, 2010.

[7] Mikolov T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[8] Russakovsky, O., Deng, J., Huang, Z., Berg, A., Fei-Fei, L.: Detecting avocados to zucchinis: what have we done, and where are we going? In: ICCV, (2013).

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in EMNLP, vol. 14, 2014, pp. 1532–1543.

[11] Yang A, Ahang Y, Rehman S, Huang Y, "Image Captioning with Object Detection and Localization", Electronic Engineering, Tsinghua University, Beijing, China, 2017.

[12] Huang Q , Smolensky P , He X , Deng L , Wu D, "Tensor Product Generation Networks for Deep NLP Modeling", Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, 2018, pp:1263-1273.

[13] Shah P, " Image Captioning using Deep Neural Architectures", Chhotubhai Gopalbhai Patel Institute of Technology, 2018, pp:1-7.

[14] 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.

[15] He K, Zhang X, Ren Sh, Sun J, " Deep Residual Learning for Image Recognition", arXiv:1512.03385v1 [cs.CV] 10 Dec 2015.

[16] Simonyan K, Zisserman A, " VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION", Conference Paper at ICLR, 2015.

[17] Favre B, "Deep learning for NLP Joint text and image representations", 2017.