ورقة بحثية مجلة مفتوحة

أنظمة توصيف الصور Image Captioning

د. سامر سليمان* م. وسيم أحمد**
*(قسم هندسة الميكاترونيكس ، جامعة المنارة
البريد الإلكتروني: samer.sulaiman@manara.edu.sy)
**(قسم هندسة الميكاترونيكس ، جامعة المنارة
(البريد الإلكتروني: wassim_eng@yahoo.ca)

الملخص

تعتبر عملية توصيف الصور على أنها تلك العملية التي يتم فيها توليد مجموعة من النصوص التي تقوم بتوصيف محتويات الصور والتي باتت تعتمد بصورة رئيسية على عمليات التعلم العميق. تستخدم هذه التقنية على مجال واسع في الوقت الراهن من أجل مساعدة ضعيفي البصر على التعرف على العناصر التي تتواجد أمامهم اثناء السير والتي تقوم شركة Nvidia بتطويرها.

تعتبر أنظمة تصنيف الصور من نمط end-to-end Sequence-to-Sequence والتي تقوم بتحويل تسلسل من الصور والتي هي عبارة عن مجموعة من البيكسلات إلى تسلسل من الكلمات, وبالتالي فإننا بحاجة لمعالجة اللغات أو العبارات والصور الخاصة بها.

من أجل القسم اللغوي, نستخدم الشبكات العصبية التكرارية RNN ونستخدم الشبكات الملتفة CNN من أجل معالجة الصور واستخراج السمات الأساسية الخاصة بها.

كلمات مفتاحية – Conventional Neural Network, Recurrent Neural Network, Caption Generation.

I. مقدمة

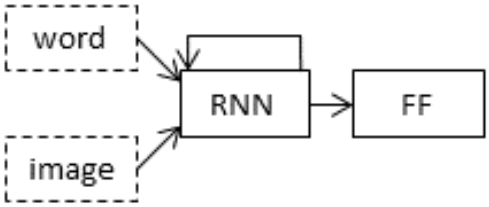
على افتراض أن الشخص يرى الصورة الموضحة في الشكل (1), فإذا طلب منا توصيف هذه الصورة, سيكون الجواب "الكلب يجلس على القماش الأزرق" أو "الكلب البني يلعب بالكرة الصفراء" , عند توصيف الصورة فإننا نحاول توليد تسلسل منطقي للصورة ومحتوياتها وهو مايتم عبر شبكات CNN في حين يتم توليد التسلسل المنطقي للعبارة باستخدام شبكات RNN.

الشكل 1. صورة دخل لنظام التوصيف

II. نماذج شبكات أنظمة توصيف الصور:

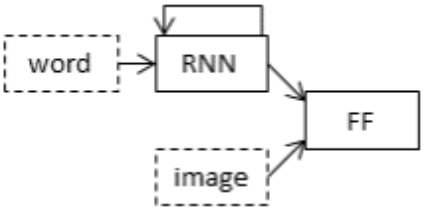
يوجد نوعين أساسيين من البنى الخاصة بالشبكات العصبية وهي:

بنية الحقن: يتم تقديم بيانات الصور مع بيانات اللغة, ومن ثم يتم عرض خليط بيانات الصور والكلمات معاً ويتم تدريب شبكة RNN على الخليط الناتج. وبالتالي, ضمن كل خطوة من عملية التدريب, تقوم شبكة RNN باستخدام خليط كلا بيانات القسمين للتنبؤ بالكلمة التالية, وتقوم بعدها الشبكة بعمليات ضبط معلومات الصور خلال عملية التدريب.



الشكل 2. بنية الحقن في نظام توصيف الصور

بنية الدمج: يتم تقديم بيانات الصور مع بيانات اللغة, ومن ثم يتم عرض خليط بيانات الصور والكلمات معاً ويتم تدريب شبكة RNN على الخليط الناتج. وبالتالي, ضمن كل خطوة من عملية التدريب, تقوم شبكة RNN باستخدام خليط كلا بيانات القسمين للتنبؤ بالكلمة التالية, وتقوم بعدها الشبكة بعمليات ضبط معلومات الصور خلال عملية التدريب.



الشكل3. بنية الدمج في نظام توصيف الصور

تمتلك تقنية الحقن ثلاث نماذج رئيسية:

الحقن الأولي :Init–Inject نستخدم وبصورة طبيعية ضمن شبكات RNN شعاع الحالة الأساسية Initial State Vector والذي يسند إلى قيمة صفرية بالنسبة للأبعاد المعطاة. من أجل هذه الحالة, يتم الحصول على شعاع سمات الصور باستخدام شبكات CNN والتي يجب ان تكون من نفس حجم شعاع الحالة الخفية لشبكة RNN ويتم تمرير شعاع الصورة كحالة أولية لشبكة RNN.

الحقن المسبق :Pre–Inject ضمن هذه الحالة, نقوم بتمرير شعاع الصورة ككلمة أولى. ومن أجل شبكة RNN, وضمن كل خطوة, يجب ان يتم إرسال الكلمات ضمن شعاع مرمز ضمن كل خطوة زمنية كتتابع. لذلك, من أجل الخطوة الزمنية الأولى نقوم بإرسال شعاع الصورة.

الحقن Par–Inject: في هذه الحالة, و ومن أجل كل خطوة, نقوم بدمج شعاع الكلمة وشعاع الصورة ضمن فضاء مضمن متساوي الأبعاد وتمريرها ليتم تدريبها ضمن RNN.

III. الخطط الوصفية المعتمدة في أنظمة توصيف الصور:

بفرض أننا نريد توصيف المشهد الموضح بالشكل (1) السابق كبشر. لذلك, نقوم في البداية بالتعرف على العناصر الموجودة في الصور, مثل ا والقماشة والكرة. ومن ثم نقوم بتوليد التوصيف والذي يتضمن الكلمات التي تصف العناصر ضمن الصورة. بين الباحث أن كل قسم متسلسل من الكلمات ضمن الوصف يتعلق بجزء محدد من الصورة, ولكن بالنسبة للآلة تكون هذه المواقع غير معروفة. ولذلك يجب القيام بعمليات mapping لتوليد توصيف النموذج المولد. يشبه هذا العمل العمل الذي قوم به البشر بتوليد قطاعات من الجمل عبر ملاحظة الأجزاء الرئيسية من الصور والحصول على العناصر ومن ثم صياغتها كعبارة.

يقوم النموذج المقترح بتوليد فضاء مضمن متعدد النماذج باستخدام كلا النموذجين لإيجاد حالات الضبط بين المقاطع الموجودة ضمن الكلمات والمواقع المتعلقة بها ضمن الصورة. يستخدم نموذج RCNN (Regional conventional neural network) لتحديد موضع منطقة العنصر ضمن الصورة, وتم تدريب شبكة CNN على مجموعة بيانات ImageNet والتي تتضمن 200 صنف من اجل التعرف على العناصر المختلفة.

من أجل نمذجة اللغة, تقترح الطريقة استخدام BRNN والتي تقدم الأداء الأفضل للعلاقات الداخلية ضمن النموذج خلال n– grams من الجمل. تم الاعتماد على نظم A300d, word2vec المضمنة من أجل الحصول على شعاع الكلمات. هذه الطريقة قامت بإنشاء معدل الارتباط بين الصورة والجملة

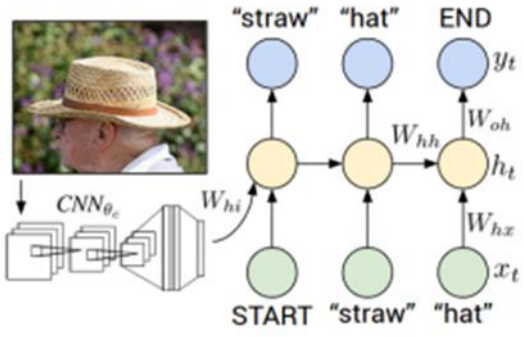
وهو عبارة كتابع للكلمات المفصولة ومعدل مناطق الصور. إذا كان هذا المعدل عالي بالتالي ستكون الجملة تدعم توصيف الصورة.

A. النموذج المطور متعدد الأبعاد لشبكة RNN:

يتعامل القسم الثاني من نموذج التوصيف الصوري مع النموذج المطور متعدد الابعاد لشبكة RNN من أجل توليد الوصف. حيث يقوم النموذج باستخلاص بيكسل الصورة I وتسلسل من شعاع كلمات الدخل X1,X2,...,Xn ومن ثم حساب الحالات المخفية لهذا التسلسل h1,h2,..,hn لإعطاء تسلسل خرج y1,y2,..,yn. ويتم تمرير أشعة سمات الصور مرة واحدة فقط كحالة خفية أولى ويتم حساب الحالة الخفية التالية من شعاع الصورة I والحالات السابقة ht−1 والدخل الحالي xt .

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$

يتم الحصول على الخرج الحالي yt من خلال استخدام طبقة softmaxعلى تابع التفعيل للحالة الخفية المعطاة. يوضح الشكل (4) التالي آلية عمل هذه الطريقة.



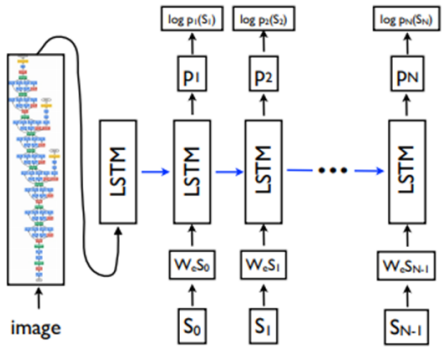
الشكل 4. النموذج المطور متعدد الأبعاد لشبكة RNN

B. بنية غوغل لتوصيف الصور:

تم وضع هذا الوصف بالورقة البحثية المعنونة Show and tell: A Neural Image Caption Generatoe عام 2015. حيث يتم استخدام نموذج LSTM بدلاً من نموذج RNN. يتم في هذه الحالة إدخال شعاع سمات الصورة I إلى تتابع LSTM في اللحظة t=−1 ولمرة واحدة ومن ثم وبدءاً من اللحظة t=0 يتم إدخال شعاع تتابع الكلمات. حيث تكون الكلمة الأخيرة المدخلة لكل خطوة هي الكلمة ذات الاحتمالية الأعلى والتي تم الحصول عليها من تابع التفعيل للحالة الخفية في المرحلة المعتبرة, تحكم المعادلات التالية هذا النموذج:

$$\begin{aligned} x_{-1} &= \text{CNN}(I) \\ x_t &= W_e S_t, \quad t \in \{0 \dots N-1\} \\ p_{t+1} &= \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\} \end{aligned}$$

من أجل الحصول على التنبؤ الأفضل للتوصيف, يستخدم النموذج Beam Search والتي تحدد K عبارة ذات أفضلية كمرشح في كل خطوة زمنية t. في كل خطوة زمنية t+1 يتم الأخذ بعين الاعتبار الكلمات K ومع K عبارة وK كلمة يكون لدينا k2 جملة محتملة ذات احتمال أعلى. حيث يتم انتخاب الكلمات ذات K الأعلى من أجل الترشيح للخطوة الزمنية K+1.



الشكل 5. بنية غوغل لتوصيف الصور

C. بنية مايكروسوفت Microsoft لتوصيف الصور:

قدمت مايكروسوفت بنيتها في الورقة البحثية المعنونة Rich Image Captioning in the Wild حيث تم الاعتماد على بنية الترميز–فك الترميز حيث يتم توصيف الصور من دون الأخذ بعين الاعتبار الكيانات مثل البيانات العامة للمشهد landmarks والمشاهير وغيرها حيث يعالج النموذج هذه البيانات بصورة منفصلة. يعتمد هذا النموذج على deep Resnet ضمن الشبكات الالتفافية ويتم الكشف فيما إذا كانت الصورة تتضمن بعضاً من المشاهير أو العلامات المشهورة. حيث يتم الاعتماد على شعاع الدخل للسمات كدخل للشبكة مع شعاع الصورة بحد ذاته. يتضمن النموذج بصورة أساسية:

1- نموذج ResNet العميق من اجل استخلاص السمات في الصورة.

2- نموذج لغوي لتوصيف جيل المشاركين ومرتبتها.

3- نظام تعرف للكيانات للعلامات الأرضية والمشاهير.

4- مصنف للتنبؤ بمعدل الثقة.

يوضح الشكل (6) التالي بنية مايكروسوفت لتوصيف الصور.

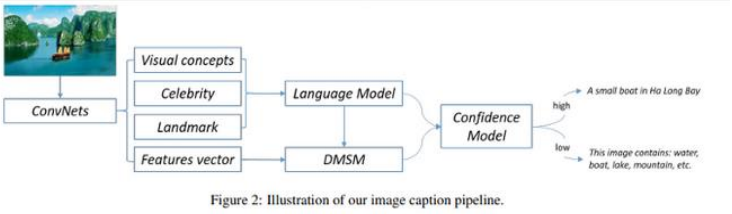
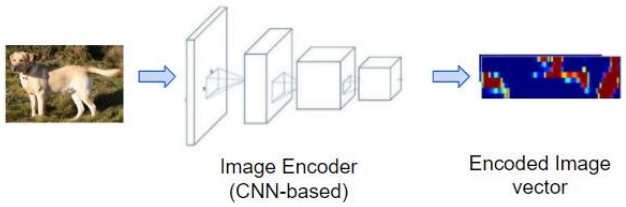


Figure 2: Illustration of our image caption pipeline.

الشكل 6. بنية مايكروسوفت لتوصيف الصور

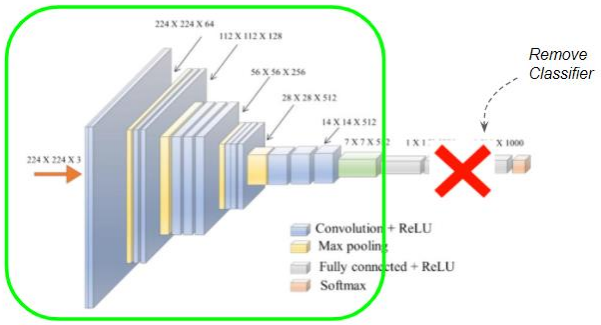
IV. بنية نظام توصيف الصور:

في هذه المرحلة يتم أخذ الصورة المصدر كدخل وتوليد عرض مرمز لها بحيث يتم التقاط جميع السمات الرئيسية ضمن الصورة.



الشكل 7. مرحلة توليد نظام ترميز الصورة

تعتمد هذه المرحلة على شبكات CNN والتي تعتمد بصورة رئيسية على التعليم المنقول transfer learning يمكن الاعتماد في هذه المرحلة على نموذج مدرب مسبقاً وإزالة قسم التصنيف الأخير حيث يمكن الاعتماد على العديد من النماذج مثل VGGNet,ResNet وغيرها من نماذج الشبكات العصبية الملتفة.



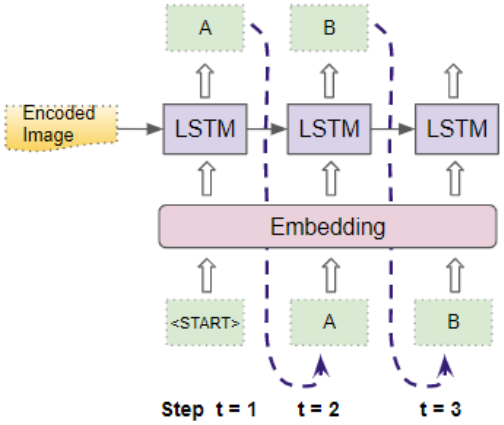
الشكل 8. نموذج VGGNet

العمود الفقري لهذه المرحلة يتألف من عدة كتل CNN والتي تقوم باستخلاص العديد من السمات من الصورة لتوليد خريطة سمات تقوم بالتقاط العناصر الأكثر أهمية في الصورة. تبدأ العملية باستخلاص العناصر الجيومترية البسيطة مثل المنحنيات وأشباه الدوائر في الطبقات الأولى حتى الوصول إلى الطبقات الأخيرة التي قد تلتقط العيون والأنف مميزةً على سبيل المثال بين الوجوه والعجلات.

a. فك ترميز التسلسل:

يتم أخذ العرض المرمز في هذه المرحلة وإخراج تسلسل من التوكنات tokens التي تقوم بوصف الصورة. وهنا سيتم استخدام

الشبكات من نمط RNN والتي تتضمن مجموعة من عناصر الذاكرة طويلة–قصيرة الأمد LSTM والتي يتم تغذيتها من طبقة مضمنة Embedding layer.



الشكل 9. فك ترميز الواصفات باستخدام LSTM

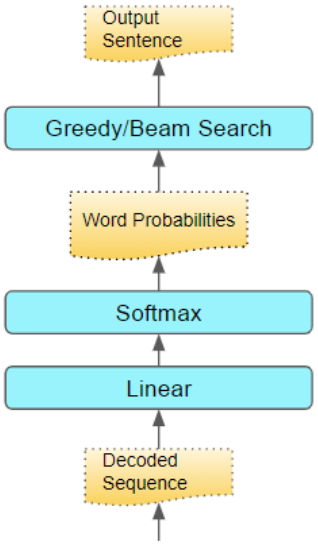
يتم في هذه المرحلة الحصول على أشعة سمات الصورة كحالة أولية وتحديد قيمة أولية بأصغر تسلسل دخل يتضمن فقط توكن البدء start token وتقوم بفك ترميز شعاع صورة الدخل وتوليد التسلسل المطلوب.

يتم توليد هذا التنبؤ باستخدام حلقة وتوليد قيمة توكن وحيدة في كل مرحلة ويعاد تمريرها إلى الدخل في الشبكة للدورة التالية, بعد عدد من التكرارات يتم توليد توكن END والذي يدل على انتهاء التسلسل.

b. مولد الجمل:

ومهمة هذه المرحلة الحصول على تسلسل قيم التوكن وإخراج وصف والذي هو جملة من الكلمات باللغة المرغوبة والتي تصف الصورة.

تتألف هذه المرحلة من طبقة خطية ملحقة بطبقة softmax حيث تقوم بتوليد احتمالية لكل كلمة ضمن الصيغة اللغوية للغة المرغوبة ولكل موقع ضمن الجملة. تعتبر هذه الاحتمالية على أنها درجة التشابه والتي تحدث فيها هذه الكلمة ضمن الجملة, نعتمد هنا على طريقة البحث الجشع لتوليد الجملة النهائية من خلال انتقاء الكلمة ذات الاحتمال الأعلى في كل موضع.

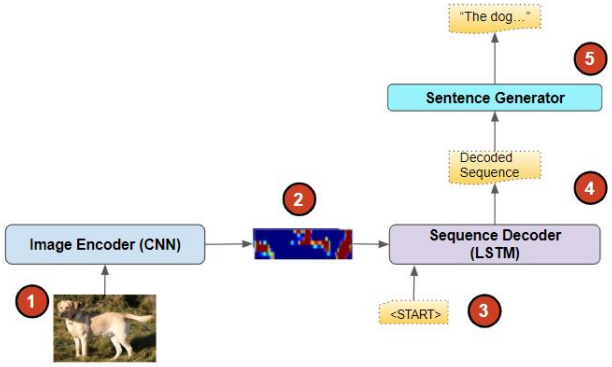


الشكل 10. بنية نظام مولد الجمل

حيث ستكون جملة الخرج هي الوصف المتنبأ به للصورة المدروسة.

c. بنية المرمز – مفكك الترميز:

يسمى النموذج الأكثر استخداماً في نظم توصيف الصور بنموذج الحقن كما ذكرنا سابقاً حيث يقوم بالربط المباشر بين ناتج ترميز الصور إلى تسلسل مفكك الترميز ومن ثم يلي ذلك مولد الصور.



الشكل 11. بناء نظام الحقن بعد دمج المراحل في نظام توصيف الصور

V. البنية متعددة المراحل Multi-Modal:

وهو النموذج الثاني الأكثر استخداماً في توصيف الصور ويسمى عادة بنموذج الدمج ويستطيع تقديم نتائج أكثر فعالية. بالإضافة إلى وصل مرمز الصور مع تسلسل فك ترميز الجمل فإن هذين النموذجين يعملان بصورة مستقلة, أي بالنسبة لهذا النموذج:

- تقوم شبكة CNN بمعالجة الصور فقط.
- تقوم شبكة LSTM بمعالجة توصيف الصور فقط.

يتم بعد ذلك الدمج بين خرجي الشبكتين ضمن طبقة متعددة النماذج (والتي قد تكون طبقة خطية أو طبقة softmax). من المزايا الإضافية لهذا النموذج السماح لنا باستخدام التعليم المنقول من أجل ترميز الصور ومفكك ترميز الجمل. يمكن لنا هنا استخدام النموذج اللغوي المدرب مسبقاً من أجل مفكك ترميز الجمل. يوجد العديد من الطرق التي يتم فيها عملية الدمج ولعل أهم هذه النماذج حساب درجة الاتصال, الجداء, إلا ان الطريقة الأسهل والتي تعطي نتائج مقبولة هي عملية الجمع.

VI. بنية العمود الفقري لكشف العناصر:

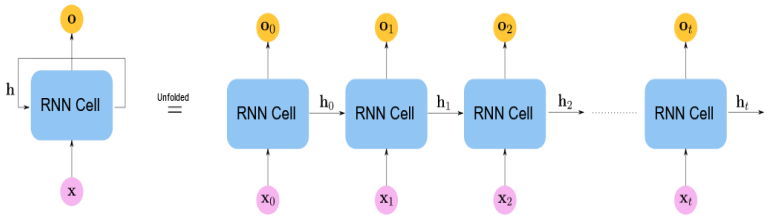
يتم استخدام هذا النموذج من اجل تحديد صنف محدد ضمن الصورة ككل, عادة في الصور المختلفة لدينا العديد من العناصر ذات الاهتمام وبالتالي وبدلاً من استخدام العمود الفقري لتصنيف الصور يمكننا استخدام العمود الفقري المدرب مسبقاً من أجل كشف العناصر واستخلاص السمات من كامل الصورة.

يقوم نموذج كشف العناصر بإحاطة العناصر في الصورة بصناديق لتحديدها بالإضافة إلى عنونة ماهية هذه العناصر وتحديد إحداثياتها ضمن الصورة وبالتالي تستطيع تقديم توصيف مرمز للصورة أكثر غنى والذي يمكن الاعتماد عليه لاحقاً لمفكك ترميز الجمل للأخذ بعين الاعتبار جميع هذه العناصر ضمن الصورة.

VII. شبكات RNN:

تعتبر شبكات RNN إحدى أهم شبكات التعلم العميق من أجل معالجة البيانات ذات النمط التتابعي, اعتبر هذا النمط من الشبكات الأفضل في معالجة البيانات التتابعية ضمن فترة من الزمن حيث يحتاج النموذج العميق ذو التغذية الأمامية لمجموعة من البارامترات لكل معامل في التتابع. بالإضافة إلى ذلك, قد لا يكون من الممكن التعميم لتحقيق التتابع على طول المتحولات المدروسة.

تقوم هذه الشبكات بتطبيق الوزن نفسه على جميع المعاملات في التتابع كما تقوم بتقليل عدد المعاملات وتسمح للنموذج بالتعميم على كامل طول المتحولات ضمن التتابع. تسمح RNN بتعميم النموذج إلى بنى المعطيات أكثر من البيانات التتابعية كما في البيانات المكانية أو الجغرافية.



الشكل 12. تركيب شبكة RNN

وتعتبر الشبكات من نمط Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) إحدى أقوى أدوات التصنيف الديناميكية المعتمدة, يرتكز نموذج تعلم الآلة على تطوير الخوارزميات التي تسمح بتطوير الأداء بشكل ذاتي ديناميكي اعتماداً على التدريب. أي كلما زادت عمليات التدريب للخوارزمية سيكون أداء هذه الخوارزمية أكثر فعالية, تتم تلك العملية من خلال خلق تابع تصنيف من بيانات التدريب المتاحة. يتم بعد ذلك قياس الأداء للمصنف المصمم عن طريق تطبيق البيانات غير الملاحظة ومطابقة الخرج.

من أجل تحويل نموذج التغذية الأمامية إلى نموذج ديناميكي فإننا بحاجة إلى إعادة تمرير الإشارات من العينات الزمنية

السابقة timesteps من جديد إلى الشبكة, وتسمى هذه الشبكات مع وصلات من نمط recurrent بشبكات Recurrent Neural Network(RNN), هذه الشبكات محدودة في البحث عن العينات الزمنية بالمجال السابق بما لا يتعدى عشر عينات زمنية فقط والسبب في ذلك كون إشارة التغذية الخلفية تتعرض للتلاشي وهو ماتم الإشارة إليه بالنمط Long Short-term Memory Recurrent Neural Network(LSTM-RNN) تستطيع الشبكات من هذا النمط بالتعلم حتى 1000 عينة زمنية timesteps وذلك وفقاً لدرجة التعقيد الذي قد تحتويه الشبكة.

تسمح هذه البنية التسلسلية للشبكات التكرارية بإضافة بعد جديد من التعقيد إلى عمليات البث الخلفي, تمتلك كل حلقة في شبكات RNN على زوجها الخاص من الدخل-خرج بينما تتشارك جميع الحلقات بنفس الأوزان وبالتالي يجب تحديد الفترة التي يجب أن تتعدل فيها الأوزان.

تمثل الخطوة الاولى في LSTM تحديد أي البيانات التي سنقوم بحذفها من حالة الخلية, يتم ذلك من خلال طبقة السيغمويد Sigmoid layer والتي تسمى بطبقة بوابة النسيان forget gate layer حيث تنظر إلى الخرج h_(t-1) والدخل x_t وتقوم بغجراج رقم قيمته بين الصفر والواحد لكل عدد ضمن حالة الخلية C_(t-1) تمثل القيمة 1 عبارة "احتفظ بهذه القيمة" في حين تمثل القيمة 0 "تخلى عن هذه القيمة."

يسمح استخدام خوارزمية Tensorflow في شبكات RNN بالقيام بالعمليات الحسابية ضمن أطوال متغيرة, في حين تؤمن شبكات RNN التقليدية إجراء التدريب على شبكة RNN بقيم أطوال ثابتة للبيانات, أي إذا كان لدينا سلسلة زمنية مؤلفة من 200 خطوة وبالتالي فإننا نقوم ببناء مخطط شبكة ثابت ب 200 خطوة, RNN يعتبر توليد النموذج الأول بطيئاً كما لا يمكننا إدخال قيم تسلسل أكبر من 200 والذي قمنا بتحديده مسبقاً.

في الشبكات الديناميكية يتم حل هذه المشكلة باستخدام حلقة while والتي تقوم بضبط حجم المخطط عند التنفيذ, يكون بناء النموذج أسرع بالإضافة للقدرة على إضافة مجالات ذات أحجام متغيرة. تختلف أيضاً شبكات RNN الديناميكية بطريقة إعداد المداخل والمخارج والتي أيضاً تكون متغيرة وفقاً لحجم batch.

VIII. التطبيق العملي:

تعتبر أنظمة توصيف الصور إحدى مشاكل الذكاء الاصطناعي والتي يجب أن يتوافق فيها الوصف النصي مع محتويات صورة ما. تحتاج هذه النظم التوصيفية جميع الطرق المعتادة بدءاً من الابصار الحاسوبي بهدف فهم محتويات الصورة ونموذجاً لغوياً من حقل معالجة اللغات الطبيعية من اجل تحويل عملية فهم الصورة إلى كلمات بتسلسل صحيح.

يوجد العديد من مجموعات البيانات التي يمكن استخدامها بفعالية في أنظمة توصيف الصور, بدءاص من Flicker8K, Flicker30k, Flicker30k تتألف هذه القاعدة من 31000 صورة تم الحصول عليها من موقع Flicker وتسمح هذه المجموعات من البيانات ببناء نظام مبسط يمكن تحميله وتشغيله بفعالية على أي معالج CPU. تتألف مجموعة البيانات من صورة نموذجية مرتبطة مع خمسة توصيفات نصية والتي تقدم توصيفاً واضحاً للكيانات الساكنة والأحداث ضمنها. تم اختيار الصور من ست مجموعات Flicker مختلفة من دون تضمنها أي أشخاص معروفين أو مواقع حيث تم اختيارها بمهارة وفعالية لتأمين مجال واسع من المشاهد والأحداث.

تتضمن هذه القاعدة بصورة مسبقة على مجموعة تدريب مؤلفة من 6000 صورة ومجموعة تحديث وتطوير مؤلفة من 1000 صورة ومجموعة اختبار مؤلفة أيضاً من 1000 صورة.

IX. خطوات التنفيذ:

a. تجهيز البيانات الصورية:

سيتم الاعتماد في هذه المرحلة على نموذج مدرب ومهيأ مسبقا من أجل توقع محتويات الصور, يوجد العديد من النماذج المهيأة والمتاحة ومن أكثرها شهرة نموذج Oxford Visual Geometry Group(VGG).

يمكن استخدام هذا النموذج بصورة مسبقة باستخدام Keras حيث وعند استخدام هذا النموذج للمرة الأولى سيتم تحميل النموذج عن الانترنت وهو بحدود 500 ميغابايت. حفاظاً على الفعالية ورفع سرعة أداء النموذج فإنه يمكن لنا القيام بعملية الحساب المسبقة لسمات الصور باستخدام النموذج المدرب مسبقاً وحفظ الناتج في ملف. ومن ثم يمكن لنا طلب هذه السمات وتغذيتها للنموذج كتوقع لصورة محددة في مجموعة البيانات. تسمح هذه الهملية بجعل عملية التدريب أكثر سرعة وتستهلك مقداراً أقل من الذاكرة.

عند استخدام نموذج VGG سنقوم بإزالة الطبقة الأخيرة من النموذج المحمل وهي المرحلة التي يتم فيها التنبؤ بتصنيف الصورة وهي مهمة غير مهمة لنا حالياً وإنما كامل الاهتمام على التوصيف الداخلي للصورة قبل القيام بعملية التصنيف وهي السمات التي قام النموذج باستخلاصها من الصورة.

يؤمن keras ايضاً أداة لضبط حجم الصورة التي تم تحميلها إلى القياس المناسب للنموذج وهو 224X224X3. إن سمات الصور سيتم تمثيلها بشعاع بطول 4096 عنصر. بعد نهاية عملية التحميل سيتم تخزين السمات المستخلصة ضمن ملف من نوع pickle للاستخدام لاحقاً حيث بحجم 127 ميغابايت.

b. إعداد البيانات النصية:

تتضمن مجموعة البيانات توصيفات مختلفة من أجل كل صورة ونص التوصيف الخاص بالصورة والذي يتطلب حداً أدنى من التنظيف. تمتلك كل صورة محدد وحيد، ويمكن استخدام هذا المحدد على اسم الصورة ورقم الملف للتوصيف.

في المرحلة التالية، سيتم التنقل عبر لائحة من توصيفات الصور وإعادة معجماً لدلائل وتوصيفات الصور حيث يشير كل دليل إلى مجموعة من واحد أو أكثر من التوصيفات النصية. في المرحلة التالية, نحتاج لعملية تنظيف النص الوصفي حيث تمت عملية tokenized لتبسيط معالجته والتعامل معه. سيتم تنظيف النص بالطرق التالية من أجل تقليل حجم العبارات المعجمية والتي سنحتاج للتعامل معها.

- تحويل كل المحارف إلى الحجم الصغير.
- إزالة جميع علامات الترقيم.
- إزالة جميع الكلمات المؤلفة من حرف وحيد أو أقل من عدد محارف محدد.
- إزالة كل الكلمات الحاوية على أرقام.

بعد إجراء عملية التنظيف يمكن لنا تحديد حجم المعجم الناتج حيث نحتاج لتلك العبارات الفعالة بأقل حجم محدد حيث تسبب المعاجم الأقل حجماً نماذجاً أصغر والتي سيتم تدريبها لاحقاً. بعد انتهاء هذه المرحلة يمكن لنا تخزين معجم دلائل الصور وتوصيفاتها إلى ملف جديد كم النمط txt حيث يتضمن دليل صورة ووصف في كل سطر.

بعد إجراء العمليات السابقة تم الحصول على 8092 توصيفاً صورياً والمعجم المنظف احتوى على 8763 كلمة. وتم تسجيل هذه النتائج في ملف txt نصي.

c. تحميل البيانات:

في البداية يجب تحميل الصورة المعدة مسبقاً والبيانات النصية بحيث يمكن لنا استخدامها لإجراء fit للنموذج, سيتم تدريب البيانات على جميع الصور والتوصيفات الخاصة بها ضمن مجموعة التدريب وخلال هذه العملية سنقوم بإدارة أداء النموذج على مجموعة التطوير واستخدام هذا الأداء لتحديد تخزين النموذج المناسب.

تم تعريف مجموعات التدريب والتحديث مسبقاً ضمن ملفات Flickr_8k.trainImages.txt Flickr_8k.devImages.txt حيث يتضمنان قوائم من أسماء ملفات الصور. ومن هذه الأسماء, يمكن لنا استخلاص دلائل الصور واستخدامها لفلترة الصور وتوصيفات كل مجموعة. بعد ذلك يمكن لنا تحميل الصور وتوصيفاتها باستخدام المجموعة المدربة مسبقاً لتدريب الدلائل المحدثة. سيقوم النموذج بتطوير عملية التوصيف لصورة محددة ويتم توليد التوصيف كلمة تلو الكلمة في كل لحظة وسيتم تزويد تسلسل الكلمات المولد سابقاً كدخل. وبالتالي فإننا سنكون بحاجة للكلمة الأولى لبدء عملية التوليد وآخر كلمة للإشارة إلى انتهاء التوصيف.

عادة يتم تعريف توكن البداية وتوكن النهاية والتي يتم تحميلها للتوصيفات كما هي حيث تعتبر هذه العملية بالغة الأهمية قبل ترميز النص النهائي وبالتالي نضمن أن التوكنات قد تم ترميزها بصورة صحيحة.

بعد ذلك, يمكننا تحميل سمات الصورة من اجل مجموعة بيانات محددة حيث يتم في البداية تحميل جميع البيانات ومن ثم إعادة مجموعات جزئية ذات اهتمام من أجل مجموعة معطاة من واصفات الصور. سيحتاج التوصيف النصي للترميز إلى أرقام قبل تقديمها إلى النموذج كدخل أو مقارنة حالات التنبؤ للنموذج. الخطوة الأولى بعملية الترميز هي عملية إعداد خريطة إسناد لكل كلمة إلى قيمة عدد صحيح محددة وهي عملية يؤمنها keras ضمن صنف Tokenizer والذي يمكنه تعلم عملية الإسناد هذه من بيانات التوصيف المحملة. كذلك سيتم تحويل معجم الوصف إلى لائحة من القيم النصية strings ومن ثم ضبطها من خلال Tokenizer. في هذه المرحلة فقط يمكننا ترميز النص.

سيتم تقسيم كل وصف إلى كلمات وسيؤمن النموذج بكلمة واحدة والصورة ومن ثم توليد الكلمة التالية. بعد ذلك يتم تزويد الكلمتين الناتجتين إلى النموذج كدخل مع الصورة لتوليد الكلمة التالية وبالتالي سيتم تدريب النموذج المدروس.

على سبيل المثال, إن تتابع الدخل little girl running in field سيقسم إلى ستة أزواج دخل–خرج من أجل تدريب النموذج.

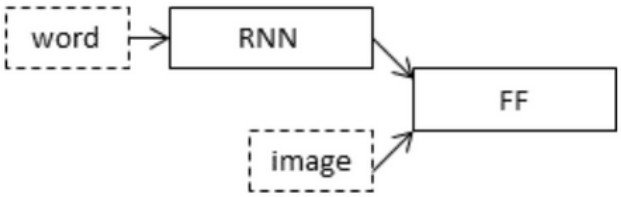
X1,	X2 (text sequence),	y (word)
photo	startseq,	little
photo	startseq, little,	girl
photo	startseq, little, girl,	running
photo	startseq, little, girl, running,	in
photo	startseq, little, girl, running, in,	field
photo	startseq, little, girl, running, in, field,	endseq

الشكل 13. توليد تتابع الدخل لعبارة little girl running in field

ستزود هذه الكلمات إلى النموذج النهائي من أجل القيام بعملية توليد الوصف النهائي, حيث ستجمع الكلمات معاً لإعطاء النموذج المطلوب, حيث يجب أن نقوم بتحديد طول التسلسل الأعظمي والمعجم المستخدم لجميع توصيفات الصور وتحويل تلك البيانات إلى أزواج دخل–خرج من البيانات من أجل تدريب النموذج .

لدينا مصفوفتي دخل للنموذج: الأولى من أجل سمات الصور والأخرى من أجل النص المرمز, كما يوجد خرج وحيد للنموذج والذي يتضمن الكلمة المرمزة التالية ضمن التتابع النصي. يتم ترميز نص الدخل كقيم صحيحة integer والتي سيتم تغذيتها إلى طبقة الكلمات المضمنة في حين سيتم تزويد سمات الصور بصورة مباشرة إلى جزء آخر من النموذج. بعد ذلك سيقوم النموذج بالتنبؤ بخرج والذي سيكون توزعاً احتمالياً عبر كل الكلمات في المعجم. وبالتالي فإن خرج هذا التوزع الاحتمالي سيأخذ القيمة 0 في جميع المواقع إلا الموقع الصحيح للكلمة تأخذ القيمة 1. سوف نحتاج حساب العدد الأكبر من الكلمات ضمن التوصيف الأطول

يتألف النموذج المقترح من الشكل التالي:



الشكل 14. نموذج الحقن المقترح

يمكن توصيف هذا النموذج بثلاثة أقسام:

- مستخلص سمات الصور photo feature extractor : وهي تتألف من 16 طبقة لنموذج VGG المدرب مسبقاً على مجموعة بيانات ImageNet حيث كما ذكرنا سابقاً تم حذف طبقة الخرج وتستخدم السمات المتنبأ بها من هذا النموذج كدخل.
- معالج التسلسل Sequence Processor: وهي طبقة الكلمات المضمنة لمعالجة النص الدخل متبوعة بطبقة شبكة Long-Short term memory Recurrent Net.

• مفكك ترميز Decoder: إن خرج كل من مستخلص السمات ومعالج التسلسل هو شعاع ذو طول ثابت يتم دمجها معاً ومعالجتها من خلال طبقة كثيفة Dense layer لاستنتاج التنبؤ النهائي.

يحتاج نموذج مستخلص سمات الصور إلى شعاع من سمات الصور بطول 4096 عنصر يتم معالجته في الطبقة الكثيفة للحصول على 256 عنصر لتوصيف الصورة. في حين يحتاج نموذج معالجة التسلسل لتتابع دخل بطول 34 كلمة يتم تغذيتها إلى طبقة مضمنة والتي تقوم باستخدام قناع لتجاهل القيم المحشوة ومن ثم إلى طبقة LSTM مع 256 وحدة ذاكرة. كما يستخدم كلا النموذجان طبقة dropout بمقدار 50 بالمئة. الهدف من ذلك تجنب الوصول إلى حالات overfitting عند تدريب مجموعة البيانات بالإضافة إلى زيادة سرعة تدريب الشبكة.

يقوم مفكك الترميز بدمج كلا خرج النموذجين عن طريق عملية جمع ومن ثم يتم إضافتها إلى طبقة كثيفة تحتوي على 256 عصبون ومن ثم إلى طبقة الخرج التي تحتوي على تابع softmax عبر كامل خرج المعاجم للكلمة التالية في التتابع.

d. ضبط النموذج Fitting The Model:

يقوم النموذج بعملية التدرب بسرعة عالية إلا أنه يتعرض لحالة overfitting لذلك سيتم إدارة مهارة النموذج المدرب على مجموعة التحديث وعند وصول النموذج إلى أعلى عملية تدريب في نهاية الدورة سيتم تخزين كامل النموذج إلى ملف ومن ثم يتم تطبيق النموذج الناتج على بيانات التدريب. يمكن القيام بذلك في Keras باستخدام ModelCheckpoint حيث يتم تدريبه للوصول إلى أدنى خطأ على بيانات الاختبار ومن ثم تخزين النموذج الذي يحتوي على ناتج التدريب وخطأ التقييم إلى ملف. سوف يتم ضبط كل مرحلة على 20 دورة ولكن وبسبب حجم البيانات الكبير كل دورة ستحتاج على الأقل إلى 30 دقيقة على هاردوير حديث.

يوضح الشكل التالي ناتج العملية:

Dataset: 6,000
Descriptions: train=6,000
Photos: train=6,000
Vocabulary Size: 7,579
Description Length: 34
Dataset: 1,000
Descriptions: test=1,000
Photos: test=1,000

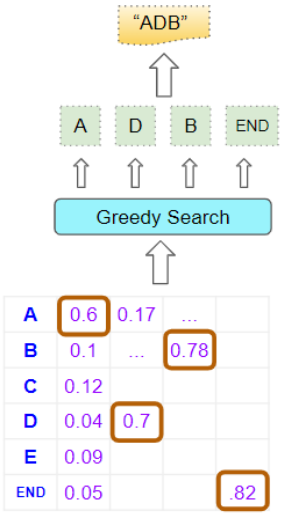
كذلك سنحصل على نواتج الاختبار والتدريب:

1 Train on 306,404 samples, validate on 50,903 samples

ستنتهي العملية ويتم تخزين النموذج في ملف من نمط h5.

e. الفارق بين Greedy Search, Beam Search:

يعتبر Greedy Search إحدى الطرق المستخدمة للحصول على الكلمة ذات الاحتمال الأعلى في كل موقع والتنبؤ بذلك، تتمتع بالسرعة وسهولة الفهم وعادة ماتعطي نتيجة صحيحة.



الشكل 15. تقنية البحث الجشع

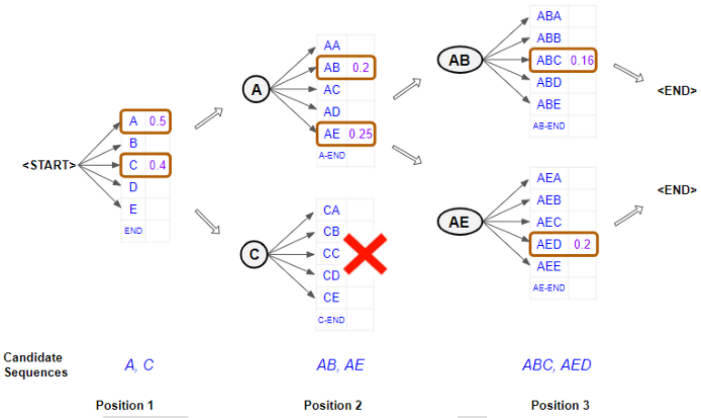
أما Beam Search وهي عملية تطوير للبحث الجشع حيث يتضمن هذا التحديث نقطتين رئيسيتين:

• في البحث الجشع نأخذ فقط الكلمة ذات الاحتمالات الأعلى في كل موقع, أما في بحث Beam فإننا نأخذ أفضل N كلمة.

• في البحث الجشع, نقوم بدراسة كل موقع بصورة مستقلة وعند تحديد الكلمة الأكثر أهمية للموقع لا نهتم بما كان قبلها أو بعدها في حين يختار Beam search التسلسل N

الأفضل للعبارات واحتمالية عمليات الدمج لكل الكلمات السابقة مع الكلمة الجديدة.

وبالتالي فأن هذا البحث يتمتع بعمومية أعلى في عمليات ضبط التسلسل.



الشكل 16. تقنية Beam Search

X. الاستنتاجات:

تمكن النظام المدروس من التعرف على الصور المقدمة بدقة بنسبة 72 بالمئة , ويعود ذلك إلى استخدام نموذج مجموعات صور ذات الحجم الأصغر, يسمح تدريب الشبكات العصبية بنماذج أكبر من مجموعات الصور إلى رفع فعالية المصنف لمستويات أعلى بالإضافة إلى إمكانية استخدام نماذج أخرى من المبنية لزيادة فعالية التعرف الصوري إلا أنها CNNشبكات تعاني من حاجتها إلى كيان مادي ذو مواصفات عالية.

المراجع:

[1]. (2021) Comparison of Architectures. [online]. https://arxiv.org/pdf/1703.09137.pdf
[2]. (2021) Andrej Karpathy's Architecture. [online]. https://cs.stanford.edu/people/karpathy/cvpr2015.pdf
[3]. (2021) Google's Architecture, [online]. https://arxiv.org/pdf/1411.4555.pdf
[4]. (2021) Microsoft's Architecture. [online]. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/06/ImageCaptionInWild-1.pdf
[5]. (2021) RNN's Applications. [online]. https://arxiv.org/pdf/1708.02043.pdf
[6]. (2021) Attention implementation in captioning. [online]. https://arxiv.org/pdf/1502.03044.pdf

منشورات المؤلف (د. سامر سليمان):

[1]. Sulaiman, S., Lehnert, R., Dai, Q. (2009). Improved QoS-Aware Awitching Mechanism for PIM-SM Protocol. 3th IEEE IMSAA.
[2]. Türk, S., Sulaiman, S., Haidine, A., Michaelis, Th., Lehnert, R. (2009). Approaches for the migration of optical backbone networks towards Carrier Ethernet: GLOBECOM 2009/EFSOI09.
[3]. Sulaiman, S., Haidine, A., Lehnert, R. (2009). Performance Evaluation of Center Search Algorithms Used for Dynamic Rendevzous Point Relocation
[4]. Sulaiman, S., Haidine, A., Lehnert, R., Türk, S. (2009). Comparative Study of Multicast Protection Algorithms Using Shared Links in 100GET Transport Network.
[5]. Dai, Q., Lehnert, R. Sulaiman, S. (2008). An Adaptive Packet Dropping Algorithm for Improved VoIP Quality at ADSL.
[6]. Sulaiman, S. (2008). Optimization of Multicast Distribution Trees. Workshop der Fachgruppe 5.2.1, TU Dortmund.
[7]. Sulaiman, S. (2006). Investigation of End-2-End delay in IP-Multicast Networks: *Workshop der TU Dresden, der Universität Twente*.
[8]. Sulaiman, S. (2003). IP Multicast Routing Protocols: MMB/Dagstuhl Seminar Performance of MobileSystem, SchlossDagstuhl, 9.-12.
[9]. Baumann, M., Marandin, D., Sulaiman, S. (2002). Combined Modelling of TCP and MultiRED in DiffServ Networks: Proc. 2nd Polish-German Teletraffic Symposium PGTS 2002; Gdansk; 23.-24.9.
[10]. Baumann, M., Marandin, D., Sulaiman, S. (2005). Combined modelling of TCP and multiRED in DiffServ networks: European Transactions on Telecommunications, Vol. 16, No. 3, pp. 217-224.
[11]. Alkheir, J., Sulaiman, S., Mualla, R. (2020). Performance Evaluation on the Effect of Different Text Representation Models on the Image Captioning Systems. *Tishreen University Journal*, Vol. 42, No. 4, print ISSN: 2097-3081.
[12]. Alkubeily, M., Sulaiman, S., Esber, GM. (2021). Designing a Virtual Platform for Modeling Nodes in

Wireless Sensor Networks at the Central Processing Unit Level. *Journal of Engineering Sciences and Information Technology*, Vol. 5, No. 5.

[13]. Alkubeily, M., Sulaiman, S., Esber, GM. (2021). Performance Evaluation of the Kernel Based Wireless Sensor Network Simulator Using an Authentication Algorithm. *Tishreen University Journal*, Vol. 43, No. 4.

[14]. Alkheir, J., Sulaiman, S., Mualla, R. (2021). Using Image Pre-classification to improve the accuracy of the image captioning systems. *Journal for the Engineering sciences*, Vol.37 No.2

[15]. Alkheir, J., Sulaiman, S., Mualla, R. (2020). Performance Evaluation on the Effect of Different Text Representation Models on the Image Captioning Systems. *Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series*, Vol. 42 No.4.

[16]. Alkubeily, M., Sulaiman, S., Esber, GM. (2020). WINDOWS FORM APPLICATION FOR VIRTUAL MINIMIZED PLATFORM KERNEL FOR WIRELESS SENSOR NETWORK SIMULATOR. *Far East Journal of Electronics and Communications*, Vol. 42 No.4.