Homework1 - Introduce a New NN with Memory

學號：104064510　　姓名：李冠毅
學號：(交大)0556083　　姓名：李季紘

1. Paper Title：

Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge [arXiv]

2. Brief Introduction of paper：

此篇論文主要目標為 Image Captioning，其中設計的架構於 2015 年 MSCOCO[1] Image Captioning Challenge 與微軟團隊並列第一名，而本篇便對於整體架構做詳盡的介紹。

過往 Captioning 手法時常利用 Object Detection 技術先提取關鍵字並排列，或者利用 CNN 中的 Feature 與文字一同進行 Multimodal training，不過此篇論文所使用的架構卻格外簡易，首先針對圖片進行 Image Classification，隨後就將圖片的代表文字做為 LSTM-based Sentence Generator 的第一個 Input，而後續所有文字都是由一連串的 LSTM 模組產生整個句子，於 Training 階段時會在降低 Negative log likelihood 的同時找出 LSTM[2]各個 gate 的 Weight 值，Inference 階段除了架構本身產生文句的方法以外，論文也有提即可使用一項「BeamSearch」技術，在指定 k 值後，會在文字組合產生的過程中不斷找出 k 個最適合的文字組合以利於後續文句的產生，文中提及這可提升一些 Performance。

由於 Image Captioning 的衡量基準尚未有一個統一標準，因此論文中也有運用多種評分運算法（主要分析 BLEU[3]分數）與其他過往應用相同衡量基準的論文進行比較，大部分都有很明顯的提升幅度，然而在加入人類評分做為 Ground Truth 比較後，尚還有相當大的差距，說明雖然分數高，不過仍未達人類可接受標準。

實驗部分除了針對 MSCOCO dataset 以外，也有對 Flickr8k[4]、Flickr30k[5]等 dataset 做測試，Dataset 之間也有做交叉比對做 Transfer Learning，實驗數據內主要都是基於 MSCOCO 進行 Training，相較於其他的 Dataset 有著更多的 Image 以及更良好的 Description，而且應用在其他 Dataset 上面也有著不錯的 BLEU score，這裡作者特別提及當前 Image Captioning 的資料量其實不多，而從目前資料量最多的 MSCOCO dataset 可以得到良好的 Transfer Learning 結果來看，他們相信之後隨著 Dataset 數量的增加，此項 end-to-end 網路的 Performance 也會越來越好。

最後這篇論文有再針對當初 2015 年時參加 MSCOCO Image Captioning Challenge 時提升 BLEU score 所使用的方法做出一連串說明，包誇運用更好的 Image Model、對 Image Model 進行 Fine-tuning、調整 BeamSearch Size、Scheduled Sampling[6]、Ensembling[7]…等等，各項都有助於提升 BLEU score。

3. Properties Discussion

這個 End-to-End NN 架構中具又相當好替換架構的特性，因為前面是用 CNN 做 classafication，所以之後如果有比較好的 CNN 架構，就可以直接進行替換，不需要額外針對複雜的 multimodel 做 training。

在 LSTM 架構的部分，使用了 GOOGLE 所提的 BeamSearch[8]，他的特性是藉由分析多種 LSTM 可能的組合，可避免某一 LSTM cell 錯誤的決策影響後續 LSTM cell 產生結果，但有可能因此造成運算量增加；作者原先預期 Beam size 越大產生的結果應會越好，然而經實驗後發現 Beam size 其實適量就好，以這篇 PAPER 為例，使用的大小為 3，此處判定過大的 size 可能反而造成 overfitting 的問題，因此對於 testing data 會有不好的結果。

LSTM 適合用在翻譯以及 image captioning，除此之外，我們認為也有與 Object Detection 網路結合的可能性，透過 NN memory 的特性，例如在使用 Faster R-CNN[9]偵測 Video Sequences 的時候，也許能幫助減少 frame 與 frame 之間的 bounding box 大小與位置變化差異，不過在未經實驗證實的狀況下，判斷也有可能只會造成運算量的增加而未有太多的實質效益。

4. References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” arXiv:1405.0312, 2014.

[2] S. Hochreiter and J. Schmidhuber, “Long short-term memory,”Neural Computation, vol. 9, no. 8, 1997.

[3] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in ACL, 2002.

[4] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using amazon’s mechanical turk,” in NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 2010, pp. 139–147.

[5] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” in ACL, 2014.

[6] S. Bengio, O. Vinyals, N. Jaitly, and N.Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in Advances in Neural Information Processing Systems, NIPS, 2015.

[7] L. Breiman, “Bagging predictors,” Machine Learning, vol. 24, pp. 123–140, 1996.

[8] Google, "Beam Search," (9.30) Available: https://www.youtube.com/watch?v=UXW6Cs82UKo

[9] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448. doi: 10.1109/ICCV.2015.169