Image Captioning using an LSTM network

Fangzhou Ai
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
faai@eng.ucsd.edu

Yue Qiao
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
yuq021@eng.ucsd.edu

Zunming Zhang
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
zuz008@eng.ucsd.edu

Abstract

In this report, we ...

1 Introduction

Automatically generating a natural language description of an image, a problem known as image captioning, has recently received a lot of attention in Computer Vision. The problem is interesting not only because it has important practical applications, such as helping visually impaired people see, but also because it is regarded as a grand challenge for image understanding which is a core problem in Computer Vision. Generating a meaningful natural language description of an image requires a level of image understanding that goes well beyond image classification and object detection. The problem is also interesting in that it connects Computer Vision with Natural Language Processing which are two major fields in Artificial Intelligence.

In this assignment, we will explore the power of Recurrent Neural Networks to deal with data that has temporal structure. Specifically, we will generate captions for images. In order to achieve this, we will need an encoderdecoder architecture for this assignment. Simply put, the encoder will take the image as input and encode it into a vector of feature values. This will then be passed through a linear layer for providing the input to the LSTM. It will be trained to predict the next word at each step. we used a pre-trained convolutional network as the encoder and an LSTM model as the decoder and tried to fine tune the encoder and train the decoder by backpropagating error into it. The error will come from caption generation. The training uses images and several captions for each image generated by humans. We then run the network in generative mode to generate captions on images it has never seen before. We also tried to replace the LSTM in the encoder part with Vanilla RNN and GRU and compared the performances of them.

Preprint. Under review.

2 Related Work

Recently, several approaches have been proposed for image captioning. We can roughly classify those methods into three categories. The first category is template based approaches that generate caption templates based on detecting objects and discovering attributes within image. For example, the work [11] was proposed to parse a whole sentence into several phrases, and learn the relationships between phrases and objects within an image. In [8], conditional random field (CRF) was used to correspond objects, attributes and prepositions of image content and predict the best label. Other similar methods were presented in [15, 10, 9]. These methods are typically hard-designed and rely on fixed template, which mostly lead to poor performance in generating variable-length sentences. The second category is retrieval based approach, this sort of methods treat image captioning as retrieval task. By leveraging distance metric to retrieve similar captioned images, then modify and combine retrieved captions to generate caption [10]. But these approaches generally need additional procedures such as modification and generalization process to fit image query.

Inspired by the success use of CNN [7, 19] and Recurrent Neural Network [1, 13, 14]. The third category is emerged as neural network based methods [4, 3, 6, 17, 18]. Our work also belongs to this category. Among those work, Kiro et al.[5] can be as pioneer work to use neural network for image captioning with multimodal neural language model. In their follow up work [6], Kiro et al. introduced an encoder-decoder pipeline where sentence was encoded by LSTM and decoded with structure-content neural language model (SC-NLM). Socher et al.[16] presented a DT-RNN (Dependency Tree-Recursive Neural Network) to embed sentence into a vector space in order to retrieve images. Later on, Mao et al.[12] proposed m-RNN which replaces feed-forward neural language model in [6]. Similar architectures were introduced in NIC [17] and LRCN [2], both approaches use LSTM to learn text context.

Our work is based on the LSTM method to do the image captioning. In our approach, we used a pre-trained concolutional network as the encoder and an LSTM model as the decoder to do the image captioning. We also tried to replaced the LSTM with the Vanilla RNN and GRU to compared the model performances.

3 Methods

3.1 Dataset

For this image captioning task, we used the dataset from the well-known Common Objects in Context (COCO) repssitory. COCO is a large-scale object detection, segmentation, and captioning dataset. In this report, we used a subset (around 1/5) of the COCO 2015 Image Captioning Task. The training set contains around 82k images with roughly 410k captions while the test set has around 3k images with almost 15k captions. The original images in the dataset are of different sizes and aspect ratios, which we are resizing to 256x256 before the training.

3.2 Model

3.2.1 Baseline Model

Our captioning system us implemented based on a Long Short-Time Memory (LSTM) network (baseline model). For the encoder part, we use a forzen pretrained convolutional network, namly ResNet50, as the encoder. We removed the last layer of pre-trained abd added a trainable linear layer with outputs a feaure vector of a fixed suze for each image. This seted the initial state of the LSTM network based on the image. For baseline model, we resized the image to 256x256 and hidden size of 512.

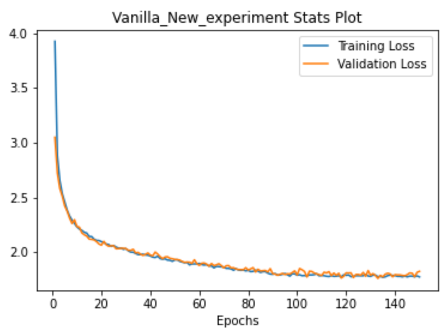
3.2.2 Vanilla RNN and GRU

For the model comparison, we also tried Vanilla RNN and GRU. For these two models, we simply replaced the LSTM module in the encoder with either a Vanilla or a GRU, while the others remained to be same. Then, trained and compared the performance of these three different models.

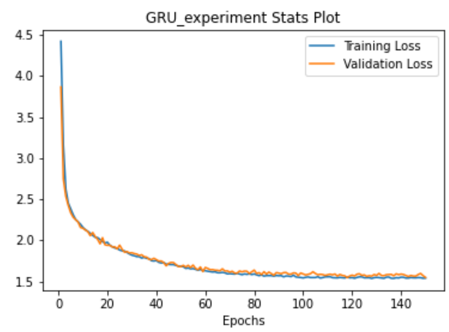
4 Results

4.1 Learning Curve

The learning curve for the Vanilla RNN and GRU could be found in Figure 1.



(a) Vanilla RNN Loss



(b) GRU Loss

Figure 1: Training and Validation loss for Vanilla RNN and GRU

4.2 Cross Entropy Loss

The test set Cross Entropy Loss of different could be found in Table 1

Table 1: Cross Entropy Loss for different models

Model	Testset Cross Entropy Loss
Baseline	TBD
Vanilla RNN	1.819
GRU	1.586
Finee tuning model	TBD

4.3 BLEU Score

For this report, we set the weight for BLEU1 to be [1, 0, 0, 0] and BLEU4 to be [0.25, 0.25, 0.25, 0.25]. The BLEU scores for different models could be found in Table 2

Table 2: BLEU scores for different models

Model	BLEU1	BLEU4
Baseline	TBD	TBD
Vanilla RNN	54.59	8.93
GRU	65.16	17.17
Finee tuning model	TBD	TBD

4.4 Images Visualization of the Best Performance Model

5 Discussion

5.1 Baseline with different Temperatures

5.2 Baseline LSTM vs Vanilla RNN vs GRU

5.2.1 Learning Curve

5.2.2 Cross Entropy Loss

5.2.3 BLEU Score

5.3 Baseline model vs fine tuning model

5.3.1 Learning Curve

5.3.2 Cross Entropy Loss

5.3.3 BLEU Score

6 Individual contributions to the project

Fangzhou Ai

Yue Qiao

Zunming Zhang

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[4] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International conference on machine learning*, pages 595–603. PMLR, 2014.

[6] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[8] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 2891–2903, 2013.

[9] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, 2012.

[10] Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362, 2014.

[11] Siming Li, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, 2011.

[12] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[13] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[14] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5528–5531. IEEE, 2011.

[15] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, 2012.

[16] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[17] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[19] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.