Joint Representation for Image and Text

Dexiong Chen Xi Shen

March 24, 2017

Abstract

Image captioning is an important but challenging task in computer vision and neural language processing, in the sense that one needs to integrate vision, language understanding and learning all in a single model. In this paper, we concentrate on two state-of-the-art models and compare the advantages and drawbacks of each model. Besides, we present a novel model combining two models, which outperforms these two models regarding retrieval and generation tasks.

1 Introduction

Modeling the statistics of images and associated text descriptions in order to automatically generate descriptions for the content of an image is a challenging but important task. It has several applications including early education for children and navigation for the blind. The main difficulty that makes it differ from other computer vision problems is that one needs to integrate language understanding, vision and learning in a single model. Since the last five years, various models have been proposed with the emergence of powerful convolutional neural networks (CNNs) [15, 18, 7] for image classification.

In this paper, we will investigate the state-of-the-art models of multimodal learning for image and text, reimplement some of the models and develop a novel model in order to achieve the following tasks simultaneously. The first one is the sentence-image retrieval task that includes image search and image annotation tasks. Both of the resulting images or texts are extracted from the database. The other one, which is more challenging, is the image captioning task. The objective is to generate a sentence that describe the content of a query image.

Our contributions are as follows. First, we develop an encoder decoder model based on two previous works [10, 19] which combines two model in a natural way such that each model serves as a regulariser for another. Second, the combined model is competitive to the original two models in all of the above tasks only using a less powerful CNN [15] to extract image features and without any fine tuning techniques.

2 Related Work

The problem of describing images with sentences has largely been studied recently. A pronounced progress has been made and a large number of models have been proposed since the MSCOCO image captioning challenge [1]. Principally, these models can be

[1] More details can be found on the challenge website: `http://mscoco.org/dataset/#captions-challenge2015`

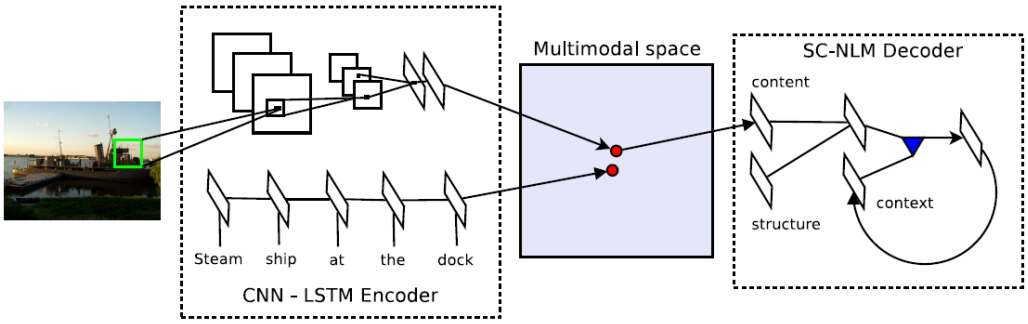

Figure 1: MNLM architecture [10]

divided into two categories. The first one treats the problem as a retrieval task, by projecting image features and text features onto a latent space. Then the most suitable annotation in the training set is transferred to a test image [6, 4, 3]. The drawback of this approach is obvious: it limits the variety of the descriptions. The other category is based on Recurrent Neural Networks (RNNs), which allows to generate description sentences by modeling the conditional probability of the next word in a sentence on all previous words [12, 8, 19]. This approach provides a more flexible outputs and an end-to-end learning framework.

In this section, we present two state-of-the-art models that are highly related to our model, where both of the methods link the problem of image captioning and the Neural Machine Translation (NMT).

Multimodal Neural Language Model (MNLM). MNLM [10] contains an encoder part to obtain the multimodal representation for image and text, and a decoder part to generate caption. The whole architecture is shown in Figure 1. The encoder aims to decrease the distance of associated image-description pairs and in the meantime increase the distance of unmatched pairs by minimizing the following loss for a batch of input images and annotations [10]

$$\text{Kiros} = \sum_x \sum_k \max\{0, \alpha - s(x,v) + s(x,v_k)\} + \sum_v \sum_k \max\{0, \alpha - s(v,x) + s(v,s_k)\}. \quad (1)$$

Our model takes advantage of the basic idea of MNLM by projecting images and sentences onto the same multimodal space, though we replace the decoder by a more recent model.

Neural Image Caption (NIC) Model. Similar to MNLM model, NIC [19] model also uses a CNN to extract image features. Instead of embedding images and sentences to a multimodal space, NIC seeks to minimize the negative log-likelihood of the correct words in a sentence by conditioning the probability of the next word on all previous words with an long Short-Term Memory (LSTM, [5]), which yields the following loss function [19]

$$\text{NIC}(I,S) = -\sum_{t=1}^{N} \log p_t(S_t). \quad (2)$$

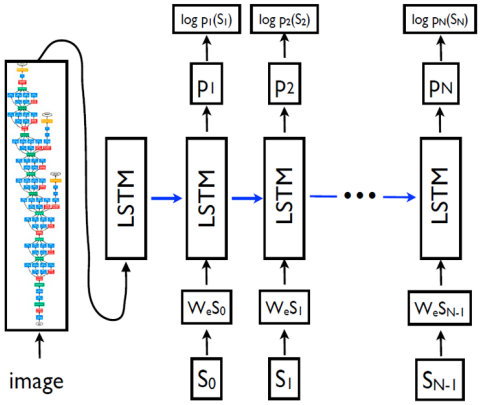

Figure 2: NIC architecture [19]

In contrast, Mao et al. [12] proposed using a similar loss called log-perplexity (log-PPL) corresponding to the average of the above term

$$\text{log-PPL} = -\frac{1}{N}\sum_{t=1}^{N}\log p_t(S_t), \tag{3}$$

which seems more reasonable for us since longer sentences should not be less desirable than shorter ones.

3 Our Model

3.1 Motivation

Following prior work [19], we have reimplemented NIC model. However, we observe that it is difficult to tune parameters to achieve the same or competitive performance as given in the original article. As explained in [20], this indicates that either the model has overfitted or the objective function (likelihood) is not aligned with human judgment. Indeed, another paper [21] also points out the a breakdown in correlation between the log-likelihood and evaluation metrics for the generation task such as BLEU [13] during the latter stage of training. This observation motivates us to find a fashion to regularize the objective function.

The primary idea occurring to us is to use image features obtained by MNLM to train NIC model. We expect that NIC model would be easier to train since images and sentences are already projected onto a latent space. After implementation of this pipeline, we notice that the model can still generate semantically correct sentences while the generated sentences are highly similar or even identical for similar images. By consequence, we design a new model that combines the above two models and expect to train two models simultaneously in order to avoid overfitting. It turns out that indeed we have gained a few BLEU scores without using any techniques to deal with overfitting but only dropout [23].

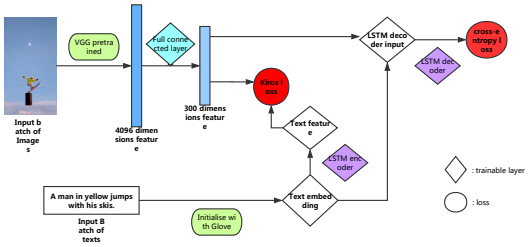

Figure 3: The architecture of our model

3.2 Architecture

By summarizing the previous explanations and related works, the pipeline of our model can be viewed as follows. We first extract image features using a CNN and sentence features using a LSTM. Then, we project image and sentence features onto a multimodal space via Kiros's loss. The resulting image features are then used to reconstruct the sentences via NIC model. The architecture is displayed in Figure 3. And the loss function for the whole model is

$$\lambda \text{Kiros} + \text{log-PPL}, \tag{4}$$

where λ is a regularization parameter that can be determined with cross-validation. In our experiments, due to the time limitation of the project, we chose $\lambda = 10$ for all experiments.

Interpretation. Inspired by MNLM, our model can be viewed as an encoder-decoder model, where we learn a multimodal space to encode images and then use a LSTM to decode images to the associated sentences. On the other hand, our model can also be considered as a regularized version of MNLM or NIC model. In fact, one loss serves as a regularization term for another, which handles overfitting problem in each of the models.

Training Details. We use the previous loss function 4 to train our model, following the same configurations described in MNLM and NIC models. The CNN weights are initialized and fixed with VGGNet [15]. The word embedding is initialized with GloVe [14]. We use ADAM [9] to train our model. Dropout for LSTM and embedding is added for large dataset, which leads to little improvement in terms of BLEU scores.

We also observe that reversing the sentences in the LSTM encoder (in MNLM) leads better generation results. This phenomenon has also been discovered in NMT such as [17].

Ranking and Generation. It turns out that our model can achieve simultaneously the three objectives given at the beginning of the paper. NIC model can also complete the ranking tasks using normalized likelihood [19, 12]. However, representing images and sentences in a multimodal space is much faster and more natural than normalized likelihood, which shows worse performance for larger datasets.

For the retrieval tasks, we the cosine similarity to compare images and sentences in the multimodal space, same as [10]. However, a weighted score can also be considered by taking into account normalized likelihood score. For the generation task,

we use a beam search approach as used in [19] which outperforms greedy or sampling approaches.

4 Experiments

Datasets We use the Flickr8k [6], Flickr30k [22] datasets in our experiments. They contains respectively 8000 and 31000 images and each image is associated with 5 sentence descriptions. The training, validation and test sets are chosen following the splits of Karpathy and Fei-Fei [8]. In future experiments, we will add experiments on the MSCOCO [11] dataset.

Evaluation Metrics. We use the standard Recall@K (R@K) and Median rank (Med r) for image-sentence retrieval tasks. Regarding the generation task, we use the BLEU score [13], which indicates the precision of word n-grams between generated and reference sentences as used in [19]. Generally, BLEU score with $n = 4$ is shown to be well correlated with human evaluations.

4.1 Ranking Results

Our test result on Flickr8k and Flickr30k are shown in Table 1 and we compare the performance to other state-of-the-art models, including MNLM and NIC models.

The model performs well on image retrieval and image annotation task. Especially it reaches almost the state-of-the-art level on Flickr8K without tuning the CNN. On the larger dataset Flickr30k, the performance remains the same level, which confirms that our model is scalable and not sensitive to the size of dataset. We also emphasize that it is reasonable to consider the combination of the cosine similarity and the log-PPL (Equation 3) to form a new score function for the ranking tasks, as Vinyals et al. [19], Mao et al. [12] used log-PPL and normalized log-PPL to rank images and texts at the test phase. However, the normalized log-PPL [12] loss is not of the same magnitude as the log-PPL loss, and the cosine similarity is also different from the Kiros loss used in the training phase. We have tested some combinations of two losses, but it turns out to improve the performance of only the image retrieval task. For the future work, we expect to find a good score function that combines these two losses in such a way that improves the performance of both tasks.

To understand the behavior of our model, we also show some qualitative results here. For the image retrieval, two examples are given in Figure 4 and Figure 5, the yellow score on the top left of the image is the cosine similarity between the image and the sentence. For the image annotation task, we show two examples in Figure 6 and Figure 7, besides, the top-5 sentences and their scores are given in the images.

4.2 Generation Results

The BLEU score [13] of the generated descriptions on Flick8k and Flickr30K are shown in Table 2. The results are competitive to the NIC model, which used a more powerful CNN. As the VGGNet [15] is originally used on classification, it might be better to finely tune the CNN in order to extract image features which would be more appropriate to generate descriptions.

Model	Image Annotation				Image Search			
	R@1	**R@5**	**R@10**	**Med r**	**R@1**	**R@5**	**R@10**	**Med r**
Flickr8K								
DeFrag	12.6	32.9	44.0	14	9.7	29.6	42.5	15
m-RNN-AlexNet	14.5	37.2	48.5	11	11.5	31.0	42.4	15
NIC	20	-	61	6	19	-	64	5
MNLM	18.0	40.9	55.0	8	12.5	37.0	51.5	10
Our reimplemented MNLM	17.8	46.9	59.4	6.5	15.2	38.8	52.3	9
Our Model	23.4	49.8	63.0	6	15.6	41.1	55.8	8
Flickr30K								
DeFrag	16.4	40.2	54.7	8	10.3	31.4	44.5	13
m-RNN-AlexNet	18.4	40.2	50.9	10	12.6	31.2	41.5	16
m-RNN-VggNet	35.4	63.8	73.7	3	22.8	50.7	63.1	5
NIC	17	-	56	7	17	-	57	7
DeepVS	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
MNLM	23.0	50.7	62.9	5	16.8	42.0	56.5	8
Our reimplemented MNLM	24.1	51.2	62.9	5	19.0	44.4	57.7	7
Our model	21.7	51.0	64.1	5	18.9	45.4	58.2	7

Table 1: Ranking results on Flickr8k and Flickr30k

Figure 4: Image retrieval result for "*A group of dogs runs beside a pond through a field.*", from left to right the scores are : 0.628, 0.616, 0.586

Figure 5: Image retrieval result for "*A boy wearing a teal shirt is riding a skateboard on a sidewalk.*", from left to right the scores are : 0.550, 0.530, 0.486

Figure 6: Image annotation result, example 1

Figure 7: Image annotation result, example 2

Model	Flickr8K				Flickr30K			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
DeepVS	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7
m-RNN-VggNet	-	-	-	-	60	41	28	19
NIC	63	41	27	-	66.3	42.3	27.7	18.3
Our reimplemented NIC	58.9	39.4	25.8	16.9	58.7	37.8	23.7	15.0
Our Model	61.2	41.6	27.5	18.1	62.5	42.2	28.2	18.7

Table 2: BLEU scores on Flickr8k and Flickr30k

Some generation examples are shown in Figure 8 9 10. They are all different from the existed descriptions which are on the left sides of the images. We also show some descriptions with errors generated by out model in Figure 11. This is due to the low frequency of some gesture or vocabulary in the training set or the main object in the image is ambiguous for the model to recognize.

Figure 8: Sentence generation example 1

Descriptions in the database :

- The dogs are in the snow in front of a fence.
- The dogs play on the snow.
- Two brown dogs playfully fight in the snow.
- Two brown dogs wrestle in the snow.
- Two dogs playing in the snow.

Sentence generated:

- Two brown dogs are playing in the snow

Figure 9: Sentence generation example 2

Descriptions in the database :

- A baby sits on a tire and holds a toy.
- A little kid, with a red toy, is sitting on a tire.
- Boy in pajamas, with red and yellow toy, seated on tire near plants.
- Child holding plastic car sits on a tire.
- The child is sitting on the street playing with a red toy.

Sentence generated:

- A little boy in a blue shirt is playing with a red toy.

Descriptions in the database :

- A boy wearing a teal shirt is riding a skate board on a sidewalk.
- A man wearing a green shirt is skateboarding on a ramp beside a white fence.
- A skateboarder in a light green shirt.
- A young male wearing a blue shirt rides a skateboard in an outdoor park at night.
- A young man skateboards at night by a fence.

Sentence generated:

- A skateboader performs a trick on a skateboard.

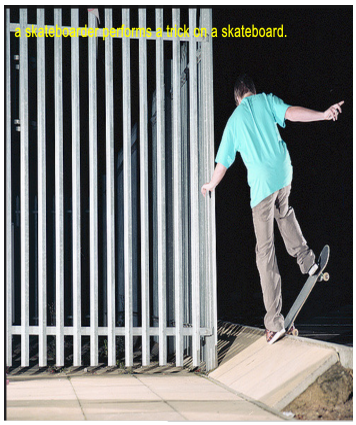

Figure 10: Sentence generation example 3

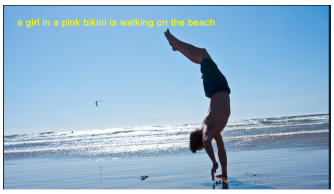

Figure 11: Other sentence generation examples. The generated sentences are : Left), *Two men stand outside of a building*; Middle), *A young boy jumping on a tramppling*; Right), *A girl in a pink Bikini is walking on the beach.*

5 Discussions and Conclusions

For the re-implemented models, we observe that, there is a breakdown in correlation between the validation set log-likelihood and BLEU score for NIC model, the same phenomenon is captured by [21], while, we do not have this problem in our model. For the re-implemented MNLM model, it performs better with more data, since more difference between the unrelated image and sentence will be learned.

The hybrid model takes about 2 hours on training Flickr8k and 10 hours on training Flickr30K on a GPU NVIDIA GeForce GT 750M with Dropout [16] on both the encoder and decoder LSTM. We find that adding Dropout on the model induces great improvement. It might be helpful to tune the hyper-parameter λ in Equation 4, by now we just take $\lambda = 10$ to make the two losses converge to a same magnitude. We take batch size equal to 100 to make sure that for each epoch, the model have enough updates [2]. Our code is available in the Github[2].

Besides this model, we also have tried Deep canonical correlation analysis (DCCA [1])

[2]https://github.com/XiSHEN0220/Deep-CCA-for-Image-Description.git

with LSTM to extract sentence features. However, it seems that DCCA works unpleasantly with the features learned by LSTM. Thus we have abandoned this idea.

To conclude, the hybrid model is an end-to-end system. By now it already performs quite well on all the three tasks, but we still have lots of trials which might possibly improve the performance, such as attaching a fine-tune CNN, try other word embedding and so on. Regarding the future work, we can eventually analyze the word embedding space learned by the model as in [8] as well as the transfer learning of the model as performed in [19].

References

[1] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML (3)*, pages 1247–1255, 2013.

[2] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[3] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer, 2010.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[6] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] A. Karpathy and L. 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.

[9] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[14] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[17] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. 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.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention.

[22] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[23] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.