The neural network image captioning model based on adversarial training

K P Korshunova[1]

[1] The Branch of National Research University "Moscow Power Engineering Institute" in Smolensk, Russia

Abstract. The paper represents the model for image captioning based on deep neural networks and adversarial training process. The model consists of a convolutional network as image encoder, a recurrent network as natural language generator and another convolutional network as an adversarial discriminator. The structure of the model, the training algorithm, some experimental results and evaluation using popular metrics are proposed.

1. Introduction

Nowadays complex artificial intelligence tasks that require processing of combination of visual and linguistic information has received increasing attention from both the computer vision and natural language processing communities. These tasks are called multimodal. They are challenging because of requiring accurate computational visual recognition, comprehensive world knowledge, and natural language generation. In addition to computer vision and natural language processing problems there are some problems related to the combination of the fields. One of the most challenging tasks is automatic Image Captioning [1], [2] known from 1990s [3], [8].

2. Image Captioning Task

Automatic Image Captioning systems generate one or more descriptive sentences in natural language given a sample image.

The task is the intersection of two data analysis fields: pattern recognition and natural language processing. In addition to visual objects, attributes and relations recognizing it requires further describing them as a natural language text [2].

The task of generating image descriptions can be understood as translation from one representation (visual features) to another (text features). In this aspect it is similar to machine translation task that is to transform data representation written in one language/modality (an input image I) into its representation in the target language/modality (a target sequence of words C) by maximizing the likelihood p(C|I) [22].

Automatic Image Captioning systems include two subsystems: “encoder” and “decoder”. An “encoder” reads the source data (raw pixels of the given image) and transforms it into a rich fixed-length vector representation, which in turn is used as the initial hidden state of a “decoder” that generates the target descriptive sentence in natural language.

The most successful Image Captioning approaches are based on deep neural networks: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). General Image Captioning approach (Figure 1): convolutional neural network (first pre-trained for an image

classification task) is used as an image “encoder”, then the last hidden layer is used as an input to the RNN decoder that generates sentences [22], [12], [5], [24], [10].

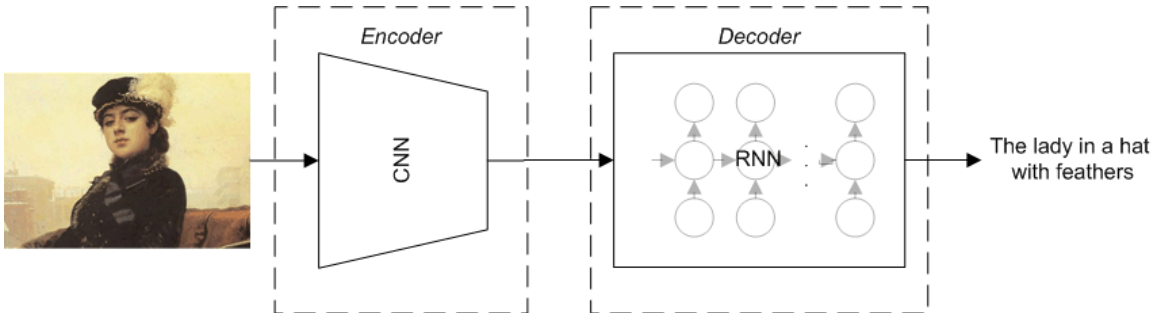


Figure 1. General Image Captioning approach.

3. The neural network image captioning model based on adversarial training

Generative Adversarial Nets (GANs [9]) that implement adversarial training have been used to produce samples of photorealistic images, to model patterns of motion in video, to reconstruct 3D models of objects from images, to improve astronomical images, etc. [23].

However in this paper we propose image captioning approach based on the Sequence Generative Adversarial Nets (Sequence GANs [14]).

GANs represent a combination of two neural network: one network (generative model G) generates candidates and the other (discriminative model D) evaluates them. Typically, the generator G learns to map from a latent space to a particular data distribution of interest, while the discriminator D discriminates between instances from the true data distribution and candidates produced by the generator. This is the implementation of adversarial training: the generative model’s training objective is to increase the error rate of the discriminative model (i.e., "fool" the discriminator network by producing novel synthesised instances that appear to have come from the true data distribution).

3.1. The structure of the model

The general structure of the proposed neural network model is represented in the Figure 2.

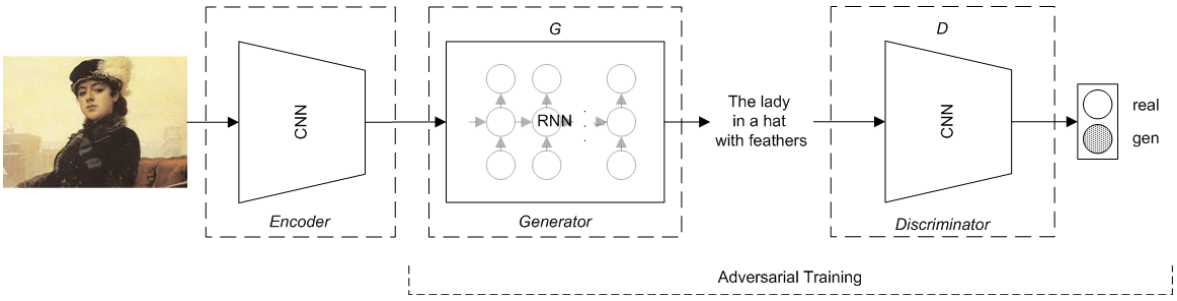


Figure 2. The general structure of the neural network image captioning model based on adversarial training.

The model consists of:

1) convolutional neural network that is used as an image “encoder”;
2) recurrent network that produces natural language descriptions;
3) another convolutional neural network that is used as the discriminator during adversarial training process.

as image encoder, recurrent network as natural language generator and convolutional network as a adversarial discriminator.

VGG16 model [19] is used for image encoding (CNN), LSTM (Long-Short Term Memory [11]) recurrent network is used for generating text descriptions (G). We choose the convolutional network

as the discriminator (D) as this kind of deep networks have recently been shown of great effectiveness in text (token sequence) classification [13].

3.2. Training algorithm

The training process of the proposed model consists of the following steps:

Step 1. Initialization and pre-training:

1.1. Pre-train CNN and G;

1.2. Generate negative samples using CNN and G;

1.3. Pre-train D;

Step 2. Training (N epochs):

2.1. Train G for g epochs;

2.2. Generate negative samples using CNN and G;

2.3. Train D for d epochs.

We use the reinforcement learning (RL) modification [20] to train the proposed model. The generative model is treated as an agent of RL. In the case of adversarial training the discriminative net D learns to distinguish whether a given data instance is real or not, and the generative net G learns to confuse D by generating high quality data.

The discriminator provides the adversariness of the training process. The CNN and the generator G work during production of the model: raw pixels of the given image are read and transformed into a rich fixed-length vector representation by the encoder CNN, then generator G generates the target descriptive sentence in natural language from this representation.

3.3. Experiments results

We have performed some experiments on the challenging public available Microsoft COCO Caption dataset [6]. It includes images from Microsoft Common Objects in COntext (COCO) [16] database. All data are divided into training set and validation set. We use 32,000 images and 160,000 corresponding text descriptions (five per image) as training set and 40,000 pairs “image-sentence” as validation set.

Several sample descriptions provided by the model after 75 training epochs are represented in the Figure 3.

In many cases descriptions made by the proposed model can describe the content of the depicted scenes (despite grammatical and semantic inaccuracies). However there are some gross mistakes.

3.4. Evaluation

Although it is sometimes not clear whether a description should be deemed successful or not given an image, prior art has proposed several evaluation metrics [17], [15], [7], [21], [4]. These metrics are based on evaluating the similarity of two sentences (candidate caption and reference caption). We use popular metrics BLEU-1, BLEU-2, BLEU-3, BLEU-4 [17], ROUGE-L [15], CIDEr [21].

We compare the proposed neural network model based on adversarial training to an CNN+RNN baseline.

The image captioning performance of the proposed (GAN) and known (CNN+RNN) models are represented in the Table 1 and Figures 4-5.

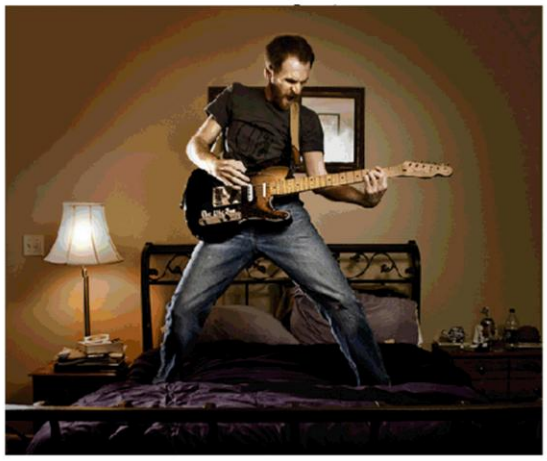
A cat standing on a big shelf behind a glass case

The bathroom with a set of sink and toilet

A giraffe walking in the grass near a

A view of the bus hanging on the road

A skateboarder is doing a trick on a skateboard



Two bears are standing on motorcycles and buildings

Figure 3. Sample image descriptions.

Table 1. The image captioning performance of the models.

Training epochs	Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	CIDEr
10	CNN+RNN	0,290	0,128	0,054	0,025	0,239	0,041
	GAN	0,309	0,139	0,060	0,027	0,253	0,039
25	CNN+RNN	0,293	0,131	0,056	0,026	0,241	0,042
	GAN	0,318	0,145	0,062	0,027	0,258	0,044
50	CNN+RNN	0,297	0,133	0,058	0,027	0,244	0,045
	GAN	0,326	0,151	0,066	0,031	0,262	0,043
75	CNN+RNN	0,297	0,133	0,058	0,027	0,244	0,044
	GAN	0,314	0,148	0,067	0,031	0,251	0,054
100	CNN+RNN	0,296	0,132	0,057	0,026	0,244	0,044
	GAN	0,324	0,155	0,074	0,037	0,272	0,054

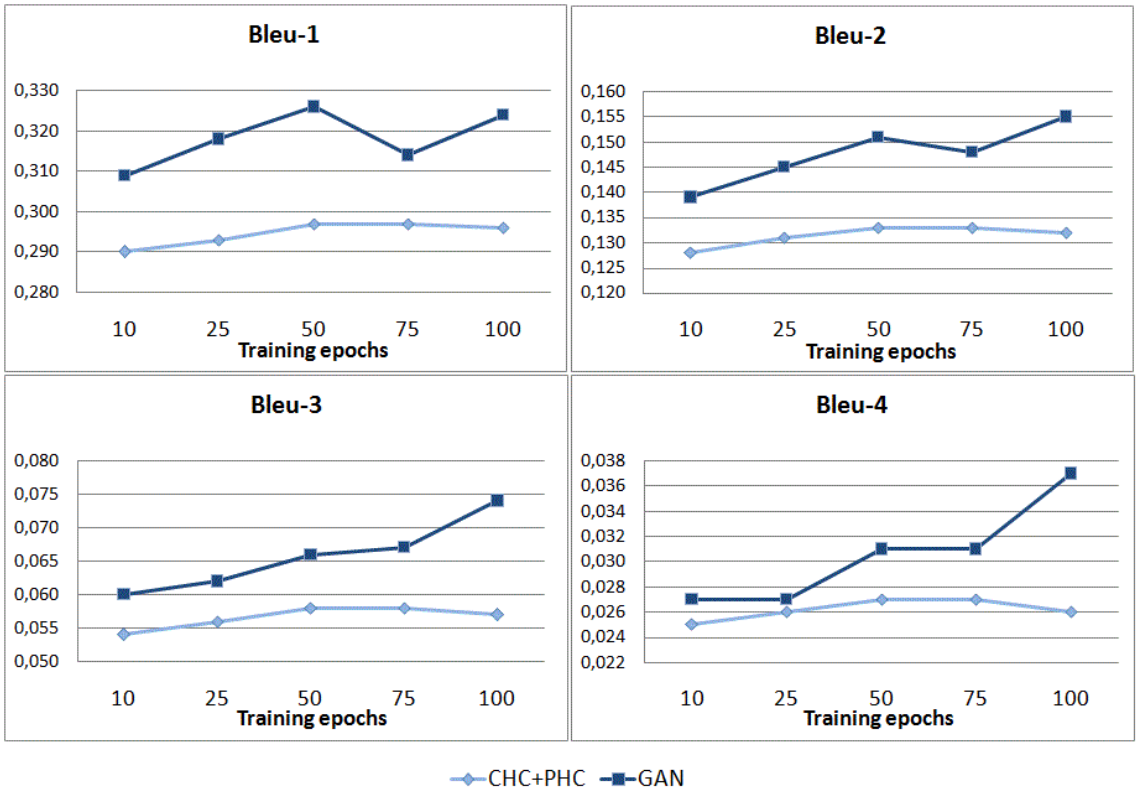


Figure 4. BLEU values w.r.t. the training epochs.

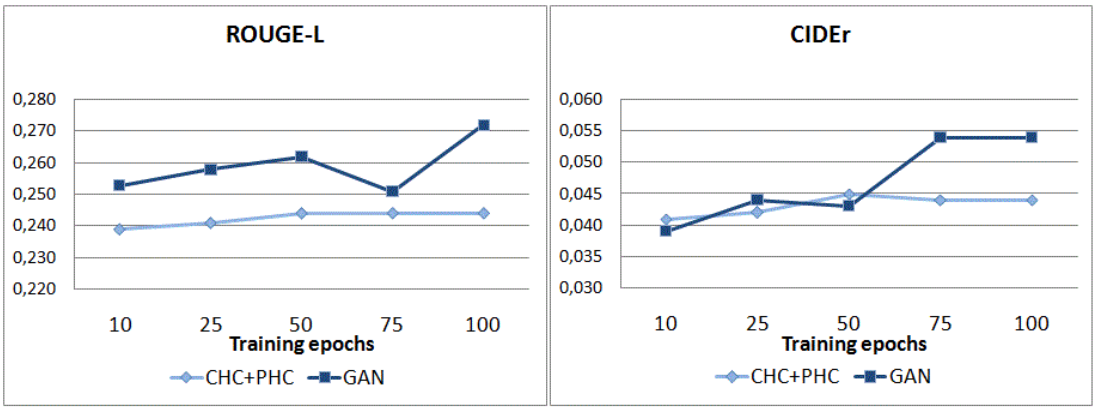


Figure 5. ROUGE-L and CIDEr values w.r.t. the training epochs.

Table 1 and Figures 4-5 show that the proposed image captioning model based on adversarial training outperforms the compared baseline (CNN+RNN) in various metrics. The best improvement is achieved for 100 training epochs. Obviously, the performance of the proposed model depends on the detailed model structure and training strategy. Choosing the attributes of the model structure (number of layers, etc.) and values of the training process parameters (number of training and pre-training epochs) is the problem for further research.

4. Conclusion

In this paper, we proposed a neural network image captioning model based on adversarial training. The model combines a convolutional neural net for image processing and Sequence Generative Adversarial Net for generating text descriptions. Some experimental work to measure the effectiveness of the model has been performed on the challenging Microsoft COCO Caption dataset. It shows that the proposed model could provide better automatic Image Captioning compared to known baseline of CNN and RNN.

5. References

[1] Borisov V. V., Korshunova K. P. Direct and Reverse Image Captioning problem definition. Postanovka priamoi i obratnoi zadachi poiska i generirovaniia tekstovykh opisanii po izobrazheniiam. *Energetika, informatika, innovatsii - 2017 (elektroenergetika, elektrotekhnika i teploenergetika, matematicheskoe modelirovanie i informatsionnye tekhnologii v proizvodstve). [Power engineering, computer science, innovations - 2017. Proceedings of the VII international scientific conference]*. Smolensk, 2017, pp 228-230 (in Russian).

[2] Korshunova K. P. Automatic Image Captioning: Tasks and Methods. *Systems of Control, Communication and Security,* 2018, no. 1, pp. 30–77. Available at: http://sccs.intelgr.com/archive/2018-01/02-Korshunova.pdf (in Russian).

[3] Abella A., Kender J. R., Starren J. Description Generation of Abnormal Densities found in Radiographs // *Proc. Symp. Computer Applications in Medical Care, Journal of the American Medical Informatics Association*. 1995. pp 542-546.

[4] Anderson P., Fernando B., Johnson M., Gould S. SPICE: Semantic propositional image caption evaluation // *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9909 LNCS. 2016. pp 382-398.

[5] Chen X., Zitnick C. L. Mind's eye: A recurrent visual representation for image caption generation // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2015. pp 2422-2431.

[6] Chen X., Fang H., Lin T. Y., Vedantam R., Gupta S., Dollár P., Zitnick C. L. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv.org, 2015. Available at: https://arxiv.org/abs/1504.00325 (accessed: 01 February 2018).

[7] Denkowski M., Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language // *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014. pp 376-380.
[8] Gerber R., Nagel N. H. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences // *Proceedings of the International Conference on Image Processing.* 1996. pp 805-808.
[9] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Bengio Y. Generative Adversarial Networks // *Proceedings of NIPS*. 2014. pp 2672–2680.
[10] Gu J., Cai J., Wang G., Chen T. Stack-Captioning: Coarse-to-Fine Learning for Image Captioning // *Association for the Advancement of Artificial Intelligence*. 2018.
[11] Hochreiter S., Urgen Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 1997, Vol. 9 (8), pp 1735-1780.
[12] Karpathy A., Fei-Fei L. Deep visual-semantic alignments for generating image descriptions // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2015.
[13] Kim, Y. Convolutional Neural Networks for Sentence Classification // *EMNLP*. 2014. pp 1746–1751.
[14] Lantao Yu, Weinan Zhang, JunWang, Y. Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient // *JAMA Internal Medicine,* 177(3). 2017. pp 326–333.
[15] Lin C. Y. Rouge: A package for automatic evaluation of summaries // *Proceedings of the Workshop on Text Summarization Branches out (WAS 2004).* 2004. Vol. 1. pp 25-26.
[16] Lin T. Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L. Microsoft COCO: Common objects in context. *European conference on computer vision*, 2014, pp740-755.
[17] Papineni K., Roukos S., Ward T., Zhu W. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp 311-318.
[18] Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF // *Journal of Documentation*. 2004. № 60 (5). pp 503-520.
[19] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // arXiv.org. 2014. – URL: https://arxiv.org/abs/1409.1556 (accessed: 01 February 2018).
[20] Sutton R. S., Barto G. Reinforcement learning: an introduction. *University College London, Computer Science Department, Reinforcement Learning Lectures,* 2017.
[21] Vedantam R., Zitnick C. L., Parikh D. CIDEr: Consensus-based image description evaluation // *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2015. pp 4566-4575.
[22] Vinyals O., Toshev A., Bengio S., Erhan D. Show and Tell: A Neural Image Caption Generator // *Conference on Computer Vision and Pattern Recognition.* 2015. pp 1-10.
[23] Generative Adversarial Network // Wikipedia, the free encyclopedia. 2018. URL: https://en.wikipedia.org/wiki/Generative_adversarial_network (accessed: 01 May 2018).
[24] Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R., Bengio Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention // *International Conference on Machine Learning.* 2015.

Acknowledgments
This work was supported by the Russian Foundation for Basic Research (Grant No. 18-07-00928)