Image captioning using adversarial training and convolutional fuzzy neural networks

Kseniya Korshunova[a]

[a] *The Branch of National Research University "Moscow Power Engineering Institute" in Smolensk, 1 Energeticheskiy proezd, Smolensk, 2014013, Russian Federation*

Abstract

The paper represents the model for image captioning based on convolutional fuzzy neural networks and adversarial training process. The structure of the models, the training algorithms are proposed.

Keywords 1

Automatic Image Captioning, deep neural networks, fuzzy neural networks, adversarial training

1. Introduction

Image Captioning problem [1, 2] consists of two stages: foto or video analysis and forming the natural language descriptions according to the previous analysis. The problem is really complicated because of necessity to use a complicated combination of different data types (visual and linguistic information) processing methods [2].

The problem of Automatic Image Captioning has great perspectives in different practice fields: intelligent analysis of the big data, technological processes control, computer-human interaction, automatization in various subject areas. First attempts to solve image captioning problem were in 1900s [3, 4].

Nowadays due to increasing of the general word information and complication of practice problems generative image captioning methods are most popular. Most of them are based on deep neural networks. There is traditional image captioning architecture: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNN reads the source data (raw pixels of the given image) and transforms it into a rich fixed-length vector representation. This vector is used as the initial hidden state of a RNN that generates the target descriptive sentence in natural language [5, 6, 7]. The training process can be general for the both parts or separate for the CNN and RNN. As a rule typical back propagation algorithms are used for training [8, 9].

This paper represents the model for image captioning based on adversarial training process using hybrid convolutional fuzzy neural networks.

2. Adversarial training for automatic image captioning

Generative Adversarial Nets (GANs [10]) 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. Adversarial learning is a variant of training a probabilistic model using variational approximations that try to construct the best approximation to a complex posterior distribution, choosing it from a rich set of distributions

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia

EMAIL: ksenya-kor@mail.ru

generated by a neural network. The output will be a network that can generate new examples from a good approximation to the posterior distribution [11].

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 synthesized instances that appear to have come from the true data distribution).

Let Z – the space of hidden (latent) factors, on which the prior distribution is given $p_z(z)$, X– output data space. Then $G=G(z, \theta_g):Z\rightarrow X$ and $D=D(x, \theta_d):X\rightarrow[0,1]$. The discriminator maps objects from data space to a chunk [0,1], that is interpreted as the likelihood that the example was actually "genuine" from p_{data}, but not generated from p_{gen}. The goal of the discriminator is to maximize the target function value in (1)

$$E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_{gen}(x)}[\log(1 - D(x))], \quad (1)$$

where $p_{gen}(x)$ – distribution generated by the generator, $p_{gen}(x) = G_{z\sim p_z(z)}$. On the other hand the generator's goal is to "fool" the discriminator by minimizing the target function value in (2)

$$E_{x\sim p_{gen}(x)}[\log(1 - D(x))] = E_{z\sim p_z(z)}[\log(1 - D(G(z)))], \quad (2)$$

So we can see that the discriminator and the generator has a competition by solving optimization problem in (3)

$$\min_G \max_D V(D, G), \quad (3)$$

where $V(D, G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_z(z)}[\log(1 - D(G(z)))]$.

In this paper we propose image captioning approach based on the Sequence Generative Adversarial Nets (Sequence GANs [12]).

The model consists of following components.

1. Convolutional neural network that is used as an image "encoder".
2. Recurrent network that produces natural language descriptions. This is the generator G during adversarial training.
3. Another convolutional neural network that is used as the discriminator during adversarial training process.

VGG16 model [13] is used for image encoding (CNN), LSTM (Long-Short Term Memory [14]) recurrent network is used for generating text descriptions (G). We choose the convolutional fuzzy neural network as the discriminator (D) for text (token sequence) classification.

The discriminator provides the adversariness of the training process. Only the CNN and the generator G work during production of the model.

There are some difficulties during adversarial training in the case of discrete data (text tokens) [15]. That's why we can't using typical gradient descent backpropagation training algorithm [16]. To solve these problems we use the reinforcement learning (RL) modification [17] 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. We also use Monte-Carlo tree search [18] to compute average "reward" for the agent (the generator).

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

Step 1. Initialization and pre-training:

1.1. Pre-training G;

1.2. Generating negative samples using CNN and G;

1.3. Pre-training D;

Step 2. Training (N epochs):

2.1. Training G for g epochs;

2.2. Generating negative samples using CNN and G;

2.3. Training D for d epochs.

Note, that the CNN training is a separate process that is done before the adversarial training starts. The training process is presented in the Figure 1.

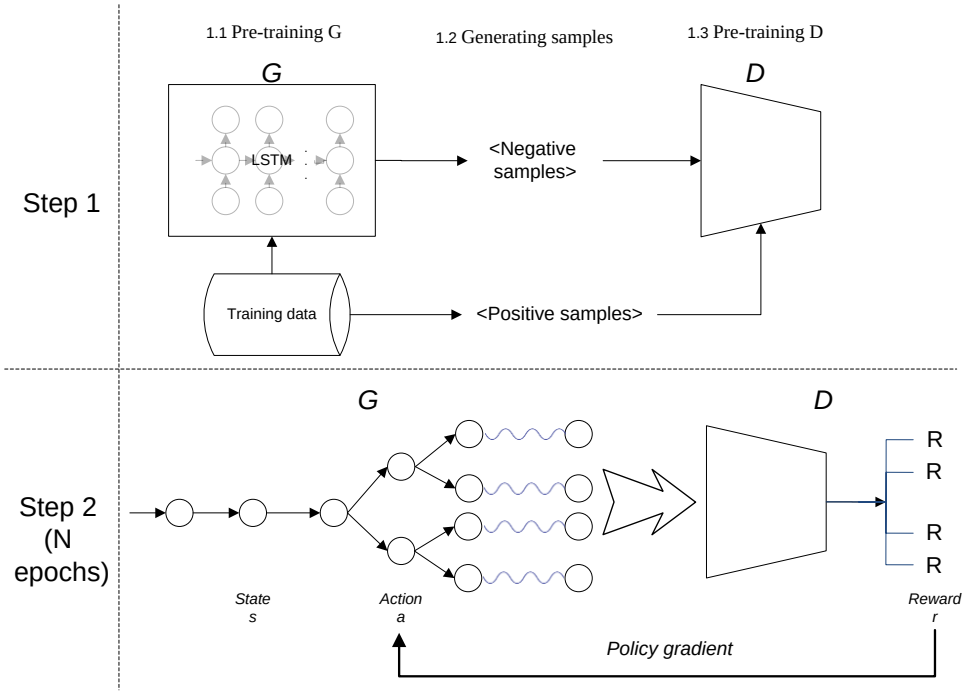

Figure 1: The adversarial training process of the proposed generative model

3. Convolutional fuzzy neural network for binary text classification

In [12] the type of CNN is used as the discriminative model. We propose the convolutional fuzzy neural network [19] for binary text classification during adversarial training.

Nowadays Convolutional Neural Networks (CNN) are one of the most powerful approaches to solve image and other data types classification problems. However, it is still difficult to detect the boundaries between classes in the classification of complex objects or complex real-world scenes. These classification objects are often characterized by uncertainty and inaccuracy in its representation and have a complex structure with non-isolated, overlapping classes. To enable a system to deal with cognitive uncertainties in a manner more like humans, one may incorporate the concept of fuzzy logic into the neural networks. For practical purposes, a fuzzy neural network is often more effective than just a fuzzy network or an ordinary (classical) neural network, as it allows indeterminate and inaccurate information processing [20].

The proposed Convolutional Fuzzy Neural Network (CFNN) model's architecture is built up of four types of layers: convolutional layer, pooling layer, Self-Organization (or Fuzzy) Layer, and fully-connected layer. To form a full Convolutional Fuzzy Neural Network architecture we stack three parts:

- a convolutional network (convolutional and pooling Layers);
- The Self-Organization Layer (The Fuzzy Layer);
- a classifier (some fully-connected layers).

In contrast to regular Convolutional Neural Network, the CFNN includes The Self-Organization Layer (The Fuzzy Layer, [21]) that is a kind of preprocessor. It is situated between the convolutional network and the classifier (a kind of postprocessor).

The convolutional network (part 1) takes an input images and form some abstract high-level properties of it by series of convolutional and pooling layers interchange.

The Fuzzy Layer performs a preliminary input data distribution into a predetermined number of clusters. Note that these clusters are not equivalent to output classes and the number of clusters and target classes can differ. The outputs of the Fuzzy Layer (part 2) neurons represents the values of the membership functions for the fuzzy clusters of input data. These membership grades indicate the

degree to which data points belong to each cluster. These values goes to the input of a classifier (part 3). Its output is the full CFNN output (the class scores).

The tuning of the proposed CFNN consists of 3 independent stages.

1. Training the convolutional network (a regular CNN corresponding to the determinate CFNN) to form some abstract properties of the input image. Backpropagation algorithm is used on this stage.
2. Tuning of the fuzzy layer parameters that is called self-organization. The Fuzzy Layer is self-organizing. It is trained in an unsupervised way using a competitive learning scheme. Self-organization of the "fuzzy layer" means choosing the positions of the clusters centers (choosing the parameters of the membership functions in the formula above). Various fuzzy clustering algorithms can be applied (C-means algorithm, Gustafson-Kessel algorithm).
3. The classifier training. The parameters of the convolutional and fuzzy layers are stable. Only fully-connected layers weights are tuning. The classifier is trained by a standard backpropagation algorithm.

These stages are presented in the Figure 2.

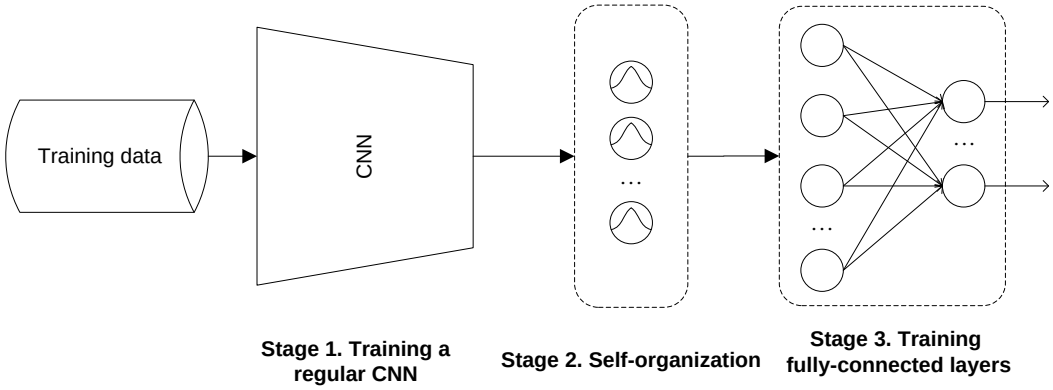

Figure 2: The tuning of the proposed CFNN

4. Automatic Image Captioning Method

The general scheme of the proposed Automatic Image Captioning Method based on adversarial training and hybrid convolutional fuzzy neural network is presented in the Figure 3.

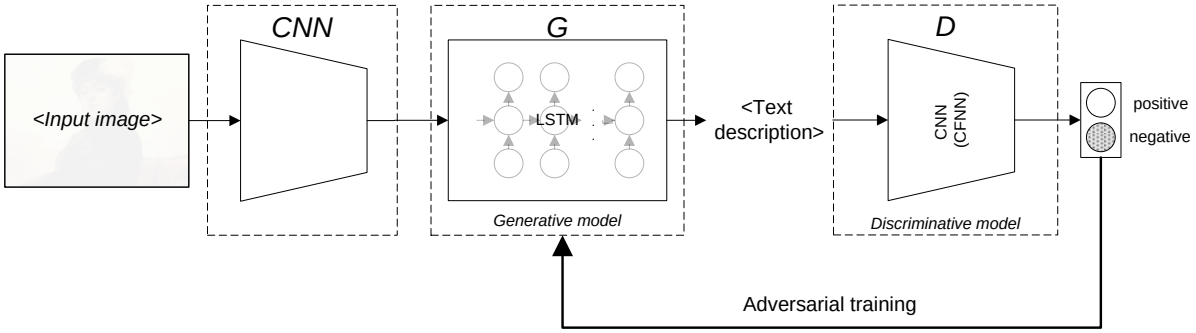

Figure 3: The general scheme of the proposed Automatic Image Captioning Method

Earlier some experimental work to measure the effectiveness of the adversarial training for image captioning has been performed [22]. It shows that the proposed method could provide better text descriptions compared to known baseline of CNN and RNN.

On the other hand some experimental work to measure the effectiveness of CFNN has been performed and show that the CFNN could provide better accuracy in image classification in less training time [19].

Now simulation modeling of the proposed method are being carried out to check the quality of solving image captioning problem using adversarial training and hybrid fuzzy neural networks.

5. Conclusion

The paper represents the automatic image captioning method based on convolutional fuzzy neural networks and adversarial training process.

6. Acknowledgements

The work was supported by grant RFBR No. 18-07-00928_a “Methods and technologies of intellectual support for research of complex hydro-mechanical processes in conditions of uncertainty on the convoluted neuro-fuzzy networks”.

7. References

[1] V. V. Borisov, K. P. Korshunova, Direct and Reverse Image Captioning problem definition. Postanovka priamoi i obratnoi zadachi poiska i generirovaniia tekstovykh opisanii po izobrazheniiam, in: Energetika, informatika, innovatsii, Power engineering, computer science, innovations - 2017. Proceedings of the VII international scientific conference, Smolensk, 1 (2017) 228-230.

[2] K. P. Korshunova, Automatic Image Captioning: Tasks and Methods, Systems of Control, Communication and Security (SCCS) 1 (2018) 30-77. URL: http://sccs.intelgr.com/archive/2018-01/02-Korshunova.pdf.

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

[4] R. Gerber, N. H. Nagel, Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences, in: Proceedings of the International Conference on Image Processing, 1996, 805-808.

[5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: A Neural Image Caption Generator, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, 1-10.

[6] A. Karpathy, L. Fei-Fei Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

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

[8] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning Internal Representations by Error Propagation, Parallel Distributed Processing, vol. 1, Cambridge, MA, MIT Press, 1986, 318-362.

[9] R. J. Williams, D. Zipser, Experimental Analysis of the Real-Time Recurrent Learning Algorithm, Connection Science, 1 (1) (1989) 87-111.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, Generative Adversarial Networks, in: Proceedings of NIPS, 2014, 2672-2680.

[11] S. I. Nikolenko, A. A. Kadurin, E. O. Arhangelskaya, Deep learning. Diving into the world of neural networks. Glubokoye obucheniye. Pogruzheniye v mir neyronnykh setey, Piter, Saint Petersburg, 2018.

[12] Y. Lantao, Z. Weinan, Y. Y. JunWang, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient, JAMA Internal Medicine, 177(3) (2017) 326-333.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv.org, 2014. URL: https://arxiv.org/abs/1409.1556.

[14] S. Hochreiter , J. Urgen Schmidhuber, Long Short-Term Memory, Neural Computation, 9 (8) (1997) 1735-1780.

[15] F. Huszar, How (not) to train your generative model: Scheduled sampling, likelihood, adversary?, arXiv.org, 2015. URL: https://arxiv.org/abs/1511.05101.

[16] I. Goodfellow, Generative adversarial networks for text, in: Conference on Neural Information Processing Systems, 2016. URL: https://arxiv.org/abs/1701.00160.
[17] R. S. Sutton, G. Barto, Reinforcement learning: an introduction, University College London, Computer Science Department, Reinforcement Learning Lectures, 2017.
[18] C. B. Browne, A Survey of Monte Carlo Tree Search Methods, in: IEEE Transactions on Computational Intelligence and AI in Games, volume 4, 1 (2012) 1-43.
[19] K. P. Korshunova, A Convolutional Fuzzy Neural Network for Image Classification in: Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, 2018, 1-4.
[20] R. Fuler, Neural fuzzy systems, Abo: Publishing House Abo Akademi University, 1995.
[21] S. Mitra, S. K. Pal, Fuzzy self organization, inferencing and rule generation, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 26(5) (1996) 608-620.
[22] K. P. Korshunova, The Neural Network Image Captioning Model Based on Adversarial Training, in: Proceedings of the II International Scientific and Practical Conference ”Fuzzy Technologies in the Industry – FTI 2018”, Ulyanovsk, Russia, October 23-25, 2018, 438-444.