Can adversarial training learn image captioning ?

Jean-Benoit Delbrouck **Bastien Vanderplaetse** **Stéphane Dupont**
ISIA Lab, Polytechnic Mons, Belgium
{jean-benoit.delbrouck, bastien.vanderplaetse, stephane.dupont}@umons.ac.be

Abstract

Recently, generative adversarial networks (GAN) have gathered a lot of interest. Their efficiency in generating unseen samples of high quality, especially images, has improved over the years. In the field of Natural Language Generation (NLG), the use of the adversarial setting to generate meaningful sentences has shown to be difficult for two reasons: the lack of existing architectures to produce realistic sentences and the lack of evaluation tools. In this paper, we propose an adversarial architecture related to the conditional GAN (cGAN) that generates sentences according to a given image (also called image captioning). This attempt is the first that uses no pre-training or reinforcement methods. We also explain why our experiment settings can be safely evaluated and interpreted for further works.

1 Introduction

Generative adversarial networks (GAN, [1]) have attracted a lot of attention over the last years especially in the field of image generation. GAN have shown great success to generate high fidelity, diverse images with models learned directly from data. Recently, new architectures have been investigated to create class-conditioned GAN [2] so that the model is able to generate a new image sample from a given ImageNet category. These networks are more broadly know as conditional-GAN or cGAN [3] where the generation is conditioned by a label.

In the field of Natural Language Generation (NLG), on the other hand, a lot of efforts have been made to generate structured sequences. In the current state-of-the-art, Recurrent neural networks (RNN; [4]) are trained to produce a sequence of words by maximizing the likelihood of each token in the sequence given the current (recurrent) state and the previous token. Scheduled sampling [5] and reinforcement learning [6] have also been investigated to train such networks. Unfortunately, training discrete probabilistic models with GAN has shown to be a very difficult task. Previous investigations require complicated training techniques such as gradient policy methods and pre-training and often struggles to generate realistic sentences. Moreover, it is not always clear how NLG should be evaluated in an adversarial settings [7].

In this paper, we propose a cGAN-like architecture that generates a sentence according to a label, the label being an image to describe. This work is related to image captioning task that proposes strict evaluation methods for any given captioning data-set. We also investigate if GAN can learn image captioning in a straightforward manner, this includes a fully differentiable end-to-end architecture and no pre-training. The generated sentences are then evaluated against to the ground truth captioning given by the task. The widely-used COCO caption data-set [8] contains 5 human-annotated ground-truth descriptions per image, this justifies our will to use a generative adversarial setting whose goal is to generate realistic and diverse samples.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 Related work

A few works can be related to ours. First, [9] proposed a Sequence Generative Adversarial Nets trained with policy gradient methods [10] and used synthetic data experiments to evaluate the training. Other works also investigated adversarial text generation with reinforcement learning and pretraining [11, 12]. Finally, the closest work related to ours is the one of [13] who proposes an adversarial setting pre-training and without reinforcement. Our model differs in the way that we use a conditional label as image to generate a sentence or image caption.

3 Adversarial image captioning

In this section, we briefly describe the model architecture used in our experiments.

As any adversarial generative setting, our model is composed of a generator G and a discriminator D. The generator G is an RNN that uses a visual attention mechanism [14] over an image I to generate a distribution of probabilities p_t over the vocabulary at each time-step t. During training, G is fed a caption as the embedded ground-truth words and D is fed with either the set of probability distributions from G or the embedded ground truth words of a real caption. D has to say if the input received is either real or fake according to the image. D is also a the same RNN as G but with different training weights. The RNN can be expressed as follows:

$$\begin{aligned} \text{if } G: \boldsymbol{h}_t &= f_{\text{gru}_1}(\boldsymbol{x}_t, \boldsymbol{h}'_{t-1}) \\ \text{if } D: \boldsymbol{h}_t &= f_{\text{gru}_1}(\boldsymbol{x}_t \text{ or } \boldsymbol{p}_t, \boldsymbol{h}_{t-1}) \end{aligned} \tag{1}$$

$$\boldsymbol{v}_t = f_{\text{att}}(\boldsymbol{h}_t, I) \tag{2}$$

$$\boldsymbol{h}'_t = f_{\text{gru}_2}(\boldsymbol{h}_t, \boldsymbol{v}_t) \tag{3}$$

$$\begin{aligned} \text{if } G: \boldsymbol{p}_t &= \text{softmax}(\boldsymbol{W}^{\text{proj}} \boldsymbol{h}'_t) \\ \text{if } D: [0,1] &= \boldsymbol{W}^{\text{ans}} \boldsymbol{h}'_t \end{aligned} \tag{4}$$

where x_t is the embedded ground-truth symbols of word t and f_{att} the attention model over image I. G and D are both trained simultaneously with the following min-max objective:

$$\begin{aligned} \min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log(D(x))] \\ + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[\log(1 - D(\tilde{x}))] \end{aligned} \tag{5}$$

where x is an example from the true data and $\tilde{x} = G(z)$ a sample from the Generator. Variable z is supposed to be Gaussian noise.

4 Tips and tricks

It is important to mind two tricks to make adversarial captioning work:

Gradient penality for embeddings As show in equation 1, the discriminator receives half of the time a probability distribution over the vocabulary from G. This is fully differentiable compared to $\arg\max p_t$. A potential concern regarding our strategy to train our discriminator to distinguish between sequence of 1-hot vectors from the true data distribution and a sequence of probabilities from the generator is that the discriminator can easily exploit the sparsity in the 1-hot vectors. However, a gradient penalty can be added to the discriminator loss to provides good gradients even under an optimal discriminator. The gradient penalty [15] is defined as $\lambda \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$ with $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$ and where ϵ is a random number sampled from the uniform distribution $U[0,1]$

Dropout as noise For the evaluation of a model to be consistent, we can't introduce noise as input of our Generator. To palliate this constraint, we provide noise only in the form of dropout to make our Generator less deterministic. Because we don't want to sample from a latent space (our model don't fall into the category of generative model), using only dropout is a good work-around in our case. Moreover, dropout has already shown success in previous generative adversarial work [16].

5 Experimentation

We use the MS-COCO data-set [8]consisting of 414.113 image-description pairs. For our experiments, we only pick a subset of 50.000 training images, 1000 images are use for validation.

Each ground-truth symbol $\boldsymbol{x}_t \in \mathbb{R}^{300}$ is a word-embedding from Glove [17]. All GRU used are of size 256, so is $\boldsymbol{h}_t$. Image I is extracted at the output of the *pool-5* layer from ResNet-50 [18]. The attention mechanism f_{att} consists of a simple element-wise product between vh_t and I :

$$\boldsymbol{v}_t = \boldsymbol{h}_t \odot \boldsymbol{W}_I$$

where $\boldsymbol{W}_I \in \mathbb{R}^{2048\times256}$ and $\boldsymbol{v}_t \in \mathbb{R}^{256}$. Finally, the size of the following matrices are: $\boldsymbol{W}^{\text{proj}} \in \mathbb{R}^{256\times|\mathbb{V}|}$ where $|\mathbb{V}|$ is the vocabulary size and $\boldsymbol{W}^{\text{ans}} \in \mathbb{R}^{256\times1}$.

As hyper-parameters, we set the batch size to 512, the gradient penalty $\lambda = 9$ and a dropout of p=0.5 is applied at the output of f_{gru_1} in the Generator. We stop training of the BLEU score on the validation set doesn't improve for 5 epochs.

6 Results

(a) Ground truth : *a group of people who are sitting on bikes*
Generated caption : *a group of people riding on the side of a car*
$\text{BLEU}{-}4 = 0.683$

(b) Ground truth : *a kitchen with a stove a sink and a counter*
Generated caption : *a kitchen with a sink stove a sink and other*
$\text{BLEU}{-}4 = 0.719$

(c) Ground truth : *a group of people standing around a kitchen*
Generated caption : *a group of people standing in a kitchen*
$\text{BLEU}{-}4 = 0.946$

Figure 1: Success case of our adversarial captioning model. The model is able to recognize groups of people, some locations and objects. We also notice the correct use of verbs.

The best configuration as described in section 5 gives a $\text{BLEU}{-}4$ score [19] of 7.30. Figure 1 shows some of the best generated captions given images. We observed that the model is able to recognize groups of people as well as some locations (such as a kitchen) and objects (such as a sink). The model also learned to use the correct verb for a given caption. For example, in Figure 1 the model is capable of making differentiate *riding* with *standing*.

(a) Ground truth : *a nude man sitting on his bed while using his phone*
Generated caption : *a <unk> <unk> <unk> next to an table*

(b) Ground truth : *two people stand using laptops in a dark room with big stars on the wall*
Generated caption : *a kitchen with a sink and tiled sink*

(c) Ground truth : *an elephant using its trunk to blow the dirt off its face*
Generated caption : *a man of people sits in a kitchen*

Figure 2: Worst generated captions ($\text{BLEU}{-}4 = 0$)

Nevertheless, we can identify two failure cases. First, the model often output sentences filled with the <unk> token. It is possible that the model hasn't been trained for long enough and on too few data. The Generator receives only a single adversarial feed back for all the words generated. It is possible some words may not have received enough gradient in order to be successfully used. In general, the pool of words used is not very large: the words used in Figure 1 are related to the ones used in Figure 2. Secondly, the model sometimes outputs well formed sentences (Figure 2 b) and c)) but unrelated to the image. Here, it is possible that the conditional information has not been taken into account.

7 Conclusion

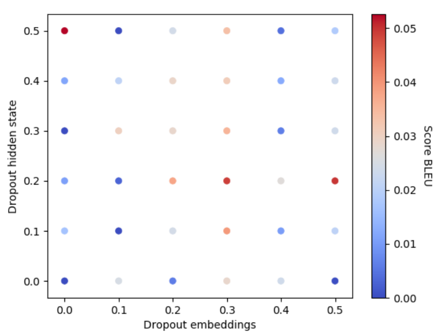

Figure 3: Result of the dropout on embedding and hidden state

In this paper, we made a first attempt on adversarial captioning without pre-training and reinforcement techniques. The task is challenging, especially since the generator G and discriminator D work with different sparsity. Nevertheless, only the WGAN with gradient penalty was able to give acceptable results. Other techniques such as the relativistic GAN [20] or WGAN-divergence [21] didn't work in our case. We also notice that the model was very sensitive to dropout. However, Figure 3 confirms our intuition that no dropout is not benefical for the generator (the bottom-left of the heat-map resulted in a BLEU score of 0).

There are a few improvements that can be made for future research. First, the attention model could be more sophisticated so that the visual signal is stronger. The size of the overall model could also be increased. Finally, the model should be trained on the full COCO training set. It is possible that enforcing an early-stop of 5 epochs for training could be an issue since the model could take time to converge.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2018.

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[4] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[5] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[6] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017.

[7] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[9] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[11] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[13] Ofir Press, Amir Bar, Ben Bogin, Jonathan Berant, and Lior Wolf. Language generation with recurrent generative adversarial networks without pre-training. *arXiv preprint arXiv:1706.01399*, 2017.

[14] 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 Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[15] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[20] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[21] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for gans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 653–668, 2018.