Semantic Alignment for Hierarchical Image Captioning

Anonymous WACV submission

Paper ID ****

Abstract

Inspired by recent progress of hierarchical reinforcement learning and adversarial text generation, we introduce a hierarchical adversarial attention based model to generate natural language description of images. The model automatically learns to align the attention over images and subgoal vectors in the process of caption generation. We describe how we can train, use and understand the model by showing its performance on Flickr8k. We also visualize the subgoal vectors and attention over images during generation procedures.

1. Introduction

Image captioning is a classical form of scene understanding problem, which is considered as one of the primary goals of computer vision. The generation of captions from images has various practical applications, such as navigation for the blind and aiding the visually impaired. More significantly, it is a core problem connecting computer vision and natural language processing.

Image captioning models take an image as input and output a sequence text that describes its content. Basically, this problem consists of two sub-tasks. One is capturing, to recognize the objects and their mutual relationships in pictures. The other is expressing, to construct a language model capable of expressing the captured information of pictures in a natural language.

For the first problem, there are lots of previous work with impressive performance, using different paradigms such as item recognition, attention mechanism, etc. Show, Attend and Tell [1] (AttendCap), in particular, is one of the representative work which bases the image feature extraction mainly on the attention mechanism. According to its experiment results, the learned attention distribution over images is impressively consistent to that of human-beings' intuitions, though not perfectly.

However, for the second problem, not much remarkable work has been proposed. Most image captioning models adopt the paradigm mentioned in [2]. They train a RNN language model via Maximum Likelihood Estimation[3] that learns to combine the inputs from various object fragments detected in the original images to form a caption. However, as is proved by Huszar[4], this method suffers from so-called exposure bias. A simple mistake in the prefix will make the language model out of hand. Scheduled sampling [5] is then proposed to address this problem, however soon proved to be fundamentally inconsistent according to Huszar[4].

It's necessary to notice that, recently, adversarial methods such as Generative Adversarial Nets[6] (GANs) have had a great impact on generative tasks. One of its variants, Sequence Generative Adversarial Nets(SeqGAN) [7], extends GANs to the language modeling tasks by adversarial reinforcement learning via Policy Gradient [8]. And then it is improved as LeakGAN [9], with the generative model replaced by a hierarchical reinforcement learning architecture called Feudal Network [10]. In this case, Feudal Network serves as the mechanism that explicitly convert the feature extracted by LeakGAN's discriminative model into a guidance that directs the rest part of the generative model to generate text with high-quality. Motivated by the drawbacks of current image captioning models and the progress of generative framework, we design a new paradigm for the image captioning problem.

In this paper, we propose a new image captioning model called HACap *i.e.* Hierarchical Adversarial-attentional Captioning to exploit both advantages of AttendCap and LeakGAN. The hierarchical method seems very natural for image captioning, since intuitively the process of cognition, understanding and expression of images is hierarchical itself. To avoid the model's being limited by human designs, instead of setting subtasks for this problem manually, we design a hierarchical architecture to automatically illustrate the hierarchy of task.

The contributions of this paper are the following:

- We propose a hierarchical image captioning model based on some ideas derived from LeakGAN and AttendCap.
- We show that this hierarchical adversarial method

presents alignment between attention over images and subgoal vectors from captioning process, which is cognitive consistency to human priors without supervision in some cases.

2. Related Works

In this section, we describe relevant background on two parts: image caption generation and text generation. The former shapes our main framework of model and the latter gives us the improved orientation of the current models. The combination of development in these two research fields shows our advantages on addressing the prominent interdisciplinary research problem.

Image captioning is regarded as a kind of translation from images to natural language and from this point, sequence to sequence training with encoder-decoder framework[11] seems natural to be applied to address this task. In the captions generation process given the image and previous word, a feed forward neural network with a multimodal log-bilinear[12] model is implemented to predict the next word, which is then replaced by recurrent neural network in[13] and [14] and further is followed by LSTM[15]. In all above methods, features are extracted by a CNN to represent the input images while another scheme[16] utilize the result of object detection from R-CNN and output of a bidirectional RNN to learn a joint embedding space.

Then attention mechanisms have been introduced to encoder-decoder frameworks by Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (AttendCap) [1], which has been practically proven effective. For the capturing task, it adopts a multilayer ConvNet as the image-end feature extractor *i.e*. the encoder of the seq2seq framework. For the expressing part, it utilize a peephole LSTM [17] as the decoder of the seq2seq framework. Before being sent into the LSTM, the features are attended due to the language context. There are two versions of its attention mechanism. One is the stochastic "hard" attention, that is to say, for each possible location of the attention, it is either fully attended or not attended at all. This is not differentiable, thus need to be trained via variational approximations or REINFORCE. The other is the "soft" attention, which allows partial attention described by a factor $\alpha_{t,i}$. It is differentiable and thus can be trained via backpropogation. Its attention mechanism has already taken significant steps in the direction of conveying information selectively. But, the feedback about images is restricted to some extent for being sent back directly from the decoder. To make better use of the decoder's feedback, we implement another hierarchical framework motivated by a series of research in text generation.

In text generation task, Long Text Generation via Adversarial Training with Leaked Information (LeakGAN) [9] is a state-of-the-art framework. As is mentioned before, the conventional methods for text generation part in image captioning tend to train a RNN language model via Maximum Likelihood Estimation. For each instance of the dataset, which contains an image and its corresponding caption, the neural network recurrently reinforces the probability of the current given token with the prefix subsequence before the occurrence of the token, which is treated as a prior. This is also a classical approach for most all2seq tasks. However, this approach can not avoid exposure bias, which is due to the discrepancy between RNN language models training and inference stage: the models whole process of sequentially generating the next token is based on previously generated tokens during inference, however itself is trained to generate tokens given ground-truth prefixes. Following improving methods such as Scheduled sampling are proposed to solve the problem but soon proved to be fundamentally inconsistent. Motivated by a variety of generative adversarial models used in continuous generative tasks, especially those of computer vision tasks such as style transferring [18], super resolution [19], etc, researchers explore the potential of generative adversarial networks for discrete tasks. Though GANs cannot be directly applied to language modeling, since the distribution of language is usually considered as discrete and not differentiable, there are some works leading the trend. Sequence Generative Adversarial Net (SeqGAN) [7], as a typical one, extends GANs to the language modeling tasks by adversarial reinforcement learning via Policy Gradient [8]. Particularly, for each time step an estimation of Q-value is produced through Monte-Carlo search [7] on the current processed prefix. An interesting fact is that the Maximum Likelihood Estimation can be regarded as policy-based reinforcement learning with only episode-replaying where each recorded action is rewarded by 1.0. Therefore SeqGAN is a natural extension of the MLE method with adversarial settings.

SeqGAN is then improved as LeakGAN, with hierarchical adversarial language model, which allows its discriminative module leak its own high-level extracted features to the generative module to further help the guidance. The generator part incorporates such informative, non-scalar signals into all generation steps through an additional MANAGER model, which takes the extracted features of current generated words and outputs a latent vector to guide the WORKER module for next-word prediction. With extensive experiments on synthetic data and various real-world tasks with Turing test, the authors of LeakGAN show that LeakGAN is highly effective in long text generation and also improves performance in short text generation scenarios (including caption-like scenarios). Another important and interesting feature of LeakGAN is that it manages to implicitly learn sentence structures through only the interactions between MANAGER and WORKER without any supervision.

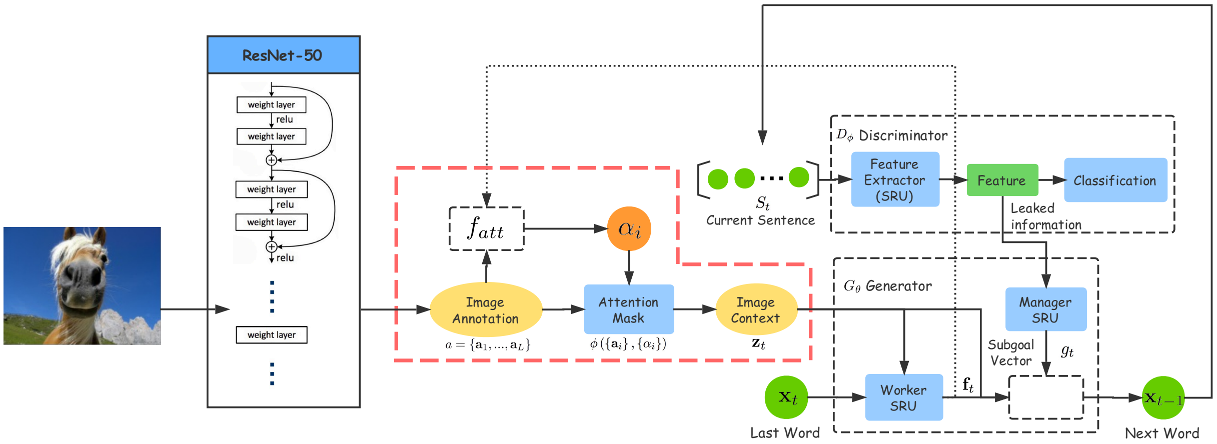

Figure 1. The architecture of our model, which take an image as input and output a sequence of words.

3. Methodology

3.1. Overview

The ultimate goal of HACap model is to output captions with an image as input. We first present an encoder that extracts features from input images. Then these features, or annotations, after being processed by an attention mechanism, are conveyed to a decoder to generate corresponding outputs. The core insight of the model lies in the correspondences between the attention over images and subgoal vectors provided by decoder.

3.2. Basic Formulation

Following the schedules of AttendCap, we formalize the image captioning problem as a special sequence-to-sequence problem. As is shown in Figure 4, for each time step, the model takes a single raw image and generates a caption $\mathbf{y}$ encoded as a 1-of-K encoded words *i.e.* C one-hot vectors.

$$y = \{\mathbf{y}_1, ..., \mathbf{y}_C\}, \mathbf{y}_i \in \mathbb{R}^K$$

where K is the size of the vocabulary and C is the length of the caption.

3.3. Encoder: Convolutional Features

A multilayer convolutional neural network is applied on the image in order to extract a set of features which are referred as annotation vectors. The extractor produces L D-dimensional vectors, each of which represent a local region of the image.

$$a = \{\mathbf{a}_1, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D$$

3.4. Decoder: Hierarchical GAN

We design a hierarchical GAN as our language model inspired by LeakGAN's structure. In our framework, there also are two modules called MANAGER and WORKER in the generative part, as well as a discriminative net which is allowed to leak high-level extracted features to MANAGER. For hierarchical generation, the MANAGER make use of guidance from the discriminator and produce subgoal vectors, which are a kind of instructive information in sentences.

Both the generator G_θ and the discriminator D_ϕ are Simple Recurrent Unit (SRU)[20] with a forget gate $\mathbf{f}_t$, a reset gater$_t$ and highway connections[21]. Given an input, which is a concat of language context $\mathbf{x}_t$ and image context $\mathbf{z}_t$ at time t, combined with the architecture of SRU, the formulas are as follows:

$$\begin{aligned}
\tilde{\mathbf{x}}_t &= \mathbf{W}(\mathbf{x}_t \oplus \mathbf{z}_t) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_f(\mathbf{x}_t \oplus \mathbf{z}_t) + \mathbf{b}_f) \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r(\mathbf{x}_t \oplus \mathbf{z}_t) + \mathbf{b}_r) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + (1 - \mathbf{f}_t) \odot \tilde{\mathbf{x}}_t \\
\mathbf{h}_t &= \mathbf{r}_t \odot g(\mathbf{c}_t) + (1 - \mathbf{r}_t) \odot \mathbf{W}_h(\mathbf{x}_t \oplus \mathbf{z}_t)
\end{aligned}$$

where $\mathbf{x}_t, \mathbf{z}_t, \mathbf{f}_t, , \mathbf{r}_t, \mathbf{c}_t, \mathbf{h}_t$ are the language context, the image context, forget gate, reset gate, memory cell and hidden state of the SRU, respectively. $g(\cdot)$ is an activation function used to produce the output state $\mathbf{h}_t$.

3.5. Attention: Instructive Information

In input terms of the decoder, the image context vector $\mathbf{z}_t$ is a dynamic representation of the relevant part of the image input at time t. We design the a mechanism Φ, the soft attention model, which was first introduced by Bahdanau et al.[22], that derives $\mathbf{z}_t$ from the annotation vectors a:

$$\mathbf{z}_t = \Phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

where Φ is a function that returns a single vector given the set of annotation vectors and their corresponding weights. It can be considered as a kind of mask over the image, representing for the relative emphasis we put on each image location, or the probability that which location is the right place to focus to generate the next word.

More specifically, in each time step, the attention model f_{att} will generate a positive weight α for each location i. We use a multilayer perceptron conditioned on the hidden features $\mathbf{f}_t$ provided by the WORKER SRU in G_θ following the formula:

$$\begin{aligned} \mathbf{e}_{t,i} =& f_{att}(\mathbf{a}_i, \mathbf{f}_t) \\ =& \mathbf{W}_e(\mathrm{ReLU}(L_{img}\mathbf{a}_i + L_{lang}\mathbf{f}_t + \mathbf{b}_a)) \\ \alpha_{t,i,n} =& \frac{\exp(\mathbf{e}_{t,i,n})}{\sum_{k=1}^{n}\exp(\mathbf{e}_{t,i,k})}. \end{aligned}$$

where $\mathbf{W}_e$, L_{img}, L_{lang} are all fully connected layers, $\mathbf{b}_a$ is bias term.

The soft attention model help speeding the training process and convergence. It also can assist to avoid facing a too sparse problem with horrible variance since the language model is already based on reinforcement learning. Besides, we use a form of doubly stochastic regularization. By construction, $\sum_t \alpha_{ti} = 1.0$ as they are the output of a softmax. As is proposed in AttendCap, this method can be interpreted as encouraging the model to pay equal attention to every part of the image over the course of generation. They observed that this penalty was important quantitatively to improving overall BLEU score and that qualitatively this leads to more rich and descriptive captions. In this approach, at each time step t, the attention model predicts a gating scalar β from previous hidden state h_{t-1} , such that,

$$\Phi\left(\{a_i\}, \{\alpha_i\}\right) = \beta \sum_{i}^{L} \alpha_i a_i$$

where $\beta_t = \sigma(f_\beta(\mathbf{f}_t))$. Note that in this case, attention weights put more emphasis on the objects in the images by including the scalar β.

4. Training Procedure

4.1. Features Extraction with Decoder

We use the ResNet[23], as the decoder to extract the annotations a_i. ResNet inserts shortcut connections, which are those skipping one or more layers, over plain networks as is shown in Figure 2 to address vanishing/exploding gradients and overfitting. The function $\mathcal{F}(\mathbf{x})$ represents the residual mapping to be learned. The operation $\mathcal{F} + \mathbf{x}$ is performed by an element-wise addition.

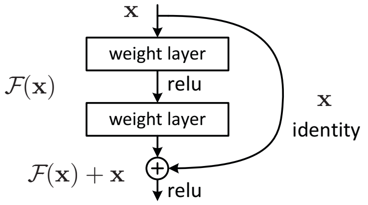

Figure 2. Residual learning: a building block.[23]

In our experiment we use the pre-trained model in tensorflow.Keras, which is acclaimed to be trained without fine tuning on ImageNet. Specifically, we use the $14 \times 14 \times 1024$ feature map as the spatial feature outputs so the decoder operates on the flattened 196×1024 encoding.

4.2. Context Construction with Attention

To convert the extracted annotations to context with information from generated language sequence, our attention mechanism utilize the features from the WORKER generator as instructions. In the course of computing attention function parameters, the WORKER generator serves as a guide from the part of expression to assist the process of capturing and there is no gradient sent back to the language model. In every time step, our attention module receives G's feature representation, e.g., the feature map of the WORKER SRU, and uses it to form soft attention.

During practical training process, we implement Doubly Stochastic Regularization method mentioned in AttendCap. After being processed by attention, the annotations with 1024 dimensions in 196 positions become context with 1024 dimensions.

4.3. Word Generation with Decoder

For generation part, at each time step, the MANAGER receives the leaked feature vector $\mathbf{f}_t$ from the discriminator D_ϕ, whose extractor is implemented by a three layers SRU. The leaked information is further combined with current hidden state of the MANAGER with 256 dimensions to produce the subgoal vector g_t , with 6 dimensions, under the guidance of which the WORKER module takes the concat of current word $\mathbf{x}_t$ and image context $\mathbf{z}_t$ as input and take a final action at current state.

We implement an end-to-end manner using a policy gradient algorithm. Both the MANAGER and WORKER modules are trained to minimize the joint loss function. In the course of training, MANAGER tends to predict advantageous directions in the discriminative feature space and WORKER is intrinsically rewarded to follow such directions.

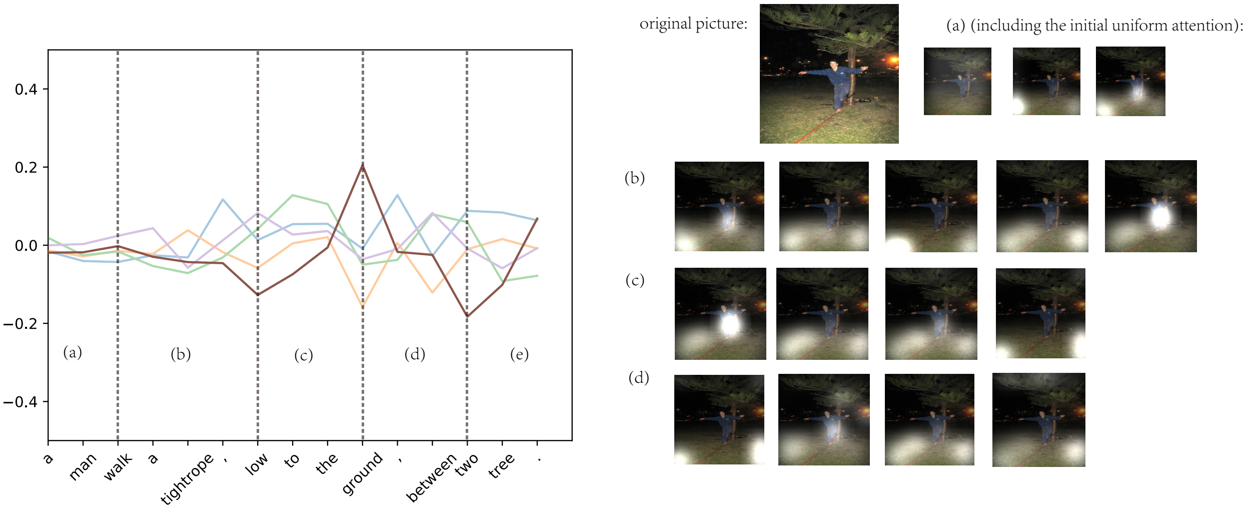

Figure 3. Examples of semantic alignment in ground truth. Left is the subgoal vector during text generation. Right is attention over time segment by segment. Notice that the picture in the end of the last segment is the same as the first one in the next segment.

4.4. Initialization and Other Settings

For the initialization, we use orthogonal method for those RNN models and a scheme called Xavier initialization proposed by[24] for the others.

To start up, the initial context are processed to be an average of the annotation vectors. Before being fed into the generator, we concatenate it to a start-up token $\mathbf{x}_0$, which serves as the initial word and will not appear in the final sentence output.

Concretely, this is an end-to-end model that minimizes the following penalized negative log-likelihood:

$$L_d = -\log(P(\mathbf{y}|\mathbf{x}))$$

5. Experiments

5.1. Data

We measure the performance of this architecture on the popular Flickr8k dataset which has 8,000 images. Each image is accompanied by at least five captions of varying length. We use official regularizations of the captions, which simplify the tense and plural form of words. The vocabulary size is around 5,000 adapted to the dataset, and the maximum sentence length is restricted to 20, which covers 97.5% of the data.

5.2. Semantic Alignment in Ground Truth

to be finished

5.3. Semantic Alignment in Generated Captions

to be finished

5.4. Analysis

We investigate the quality of the inferred attention and subgoal vector alignment with visualization. The result of these experiments can be found in Fig**??**.

More descriptions about specific details in figures to be finished, the monotonic segment of each dimension of the subgoal vector illustrates sub-structures of the captions*i.e.* semantic segments

We notice that our model learned the segment-wise alignments while the prior work AttendCap focus on point-wise ones, which can be explained as a result of highway connections. SRU used in our model and LSTM in AttendCap differ in how the hidden information is passed. Adding highway connections results in better generalization in time horizon by incorporating extra information.

We also observe that there is something that soft attention is potential to handle while single-head hard attention can hardly handle. As is shown in Figure

However, the time cost of our training and inference procedures is higher than that of simple RNN language model. The extra spend is due to the calculation of subgoal (MANAGER) and sub-feature (WORKER) proposals. And in some cases, we find that the subgoal vectors suffer collapse with all vectors representing the similar goal. We can try to implement normalization to alleviate the problem of goal collapse furthermore.

6. Conclusion

We propose a hierarchical, adversarial, attention based approach for image captioning that has more potential to fit the pixel-to-text distribution in reality. By simply testing

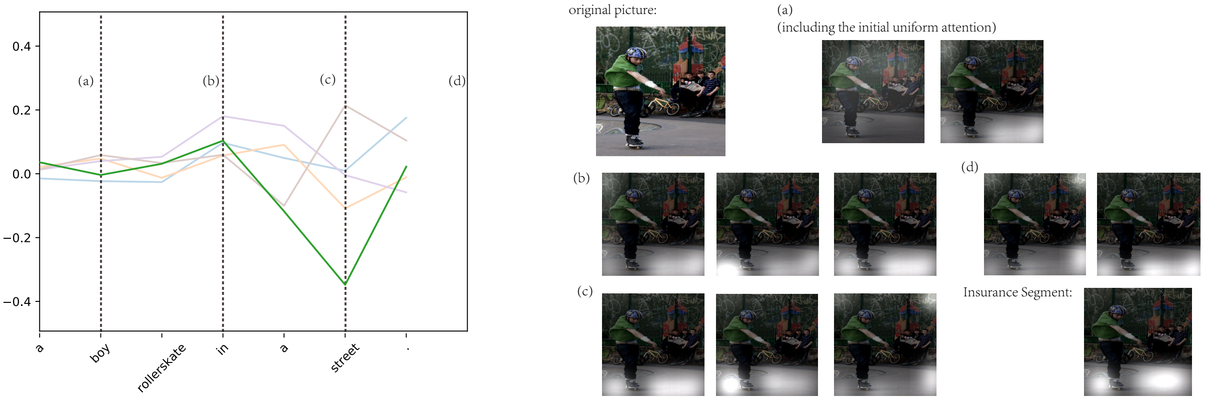

Figure 4. Examples of semantic alignment in generated captions.

the framework on Flickr8k dataset, an appealing observation of our model is that it learns to modulate the attention to align with the subgoal vector. That means it has some interesting finding about the inner correspondences between attention-goal pair. We also visualize the attention and its corresponding subgoal produced by the MANAGER to verify if the graphical semantic and linguistic semantic is well aligned.

7. Future Works

- Although our finding is encouraging, the time cost of training process is too high. It remains thinking whether there is a more sensitive implementation of our framework for a more effective procedure.
- As is described above, our model has learned some intuitively connected pattern between graphical semantic and linguistic semantic information. However, the final quantitive results such as matrix like BLEU are not so satisfying. The optimization of network structure and hyperparameters may be one of the directions to assist the interpretable attention-goal correspondences to leverage the strengths for further improvement.
- Just like most image captioning models, our feature extractor pre-trained without fine tuning. In this case, there is an orientation for framework optimization to apply the fine tuning into the network structure.
- We notice that although some of our outputs are very fluent and natural due to the guidance of MANAGER. However, not all natural text is strongly related to original images, this may be due to the mode collapse caused by the hierarchical architecture since discriminator is only trained to discriminate between real and fake text. It is not, under our current settings, explicitly evaluate whether a caption is practically appropriate for an image. In future work, we will try to introduce context information into the discriminative model to better guide the WORKER.
- We modify the decoder to improve the performance of caption generation by utilizing the linguistic guidance from discriminator to generator. Furthermore, we can implement supervised semantic alignment method to set another kind of information from content comparison to provide semantic guidance for generator.

References

[1] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. 2015.

[2] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

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

[4] Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

[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] 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.

[7] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.

[8] 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.

[9] Jiaxian Guo, Sidi Lu, Weinan Zhang, Jun Wang, and Yong Yu. Long text generation via adversarial training with leaked information. 2017.

[10] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1703.01161*, 2017.

[11] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[12] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603, 2014.

[13] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015.

[14] Xinlei Chen and C Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.

[15] 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.

[16] 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.

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

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[19] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[20] Tao Lei and Yu Zhang. Training rnns as fast as cnns. *CoRR*, abs/1709.02755, 2017.

[21] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. *arXiv preprint arXiv:1507.06228*, 2015.

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

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.