Text-to-Image synthesis via visual-memory creative adversarial network

Shengyu Zhang[1], Hao Dong[2], Wei Hu[3], Yike Guo[2], Chao Wu[1], Di Xie[4], Fei Wu[1] ⋆

[1] Zhejiang University, Hangzhou, China
[2] Imperial College London, London, UK
[3] Baidu Research
[4] Hikvision Research Institute
{light.e.gal, elweihu}@gmail.com
{wufei, chao.wu}@zju.edu.cn
{hao.dong11, y.guo}@imperial.ac.uk
xiedi@hikvision.com

Abstract. Despite recent advances, text-to-image generation on complex datasets like MSCOCO, where each image contains varied objects, is still a challenging task. In this paper, we propose a method named visual-memory Creative Adversarial Network (vmCAN) to generate images depending on their corresponding narrative sentences. vmCAN appropriately leverages an external visual knowledge memory in both multi-modal fusion and image synthesis. By conditioning synthesis on both internally textual description and externally triggered "visual proposals ", our method boosts the inception score of the baseline method by 17.6% on the challenging COCO dataset.

Keywords: Text-to-Image · Visual Knowledge · Adversarial Model.

1 Introduction

Realistic image generation from natural language descriptions is an active research task. The technique is applicable to many practical applications such as image editing and sketch or game designing. Models based on Generative Adversarial Networks (GAN) [6] have achieved promising results on datasets merely consisting of single category objects in images like CUB [35] and Oxford Flower [20]. However existing methods are far from promising on complex dataset like MSCOCO [16], in which generally one image contains varied objects and objects are rarely centered in the image [24, 38]. In order to generate complex scenes, existing approaches attempt to utilize word level attention to fine-grain image [37], establish hierarchical text-to-image mapping [8] and enhance the text description in a manner of dialog [5]. However, little work has been carried out using auxiliary visual knowledge. According to the human painting process, real-world

⋆ Corresponding author

scenes or some references may help a painter learn quickly during training and improve the generation quality during inference. That is to say, one sophisticate painter in general triggered many of the relevant visual cues during his/her painting.

Based on these intuitions, we suggest using sub-images as the visual cues to enhance text-to-image generation. More specifically, we use proposals extracted by Region Proposal Network [26] as visual cues which are stored in the external visual-knowledge memory. A proposal feature vector can be viewed as a visual summary of a meaningful sub-image, especially the one with the highest probability containing a real-world object.

The extracted proposals (i.e., visual cues) bears many of visual details such as texture, shape, color, size, etc., and they can potentially be inspired together to synthesize images after they are triggered by corresponding textual descriptions.

In this paper, given one textual sentence and the external visual-knowledge memory, we first utilize the multi-modal Encoder to encode the textual sentence into a multi-modal hidden vector. The multi-modal encoder is similar to the Memory Network model proposed by [31]. However, this paper uses semantic embeddings instead of bag-of-words representation.

The key contributions of our work are listed as following: we propose a model named visual-memory Creative Adversarial Network (vmCAN) for generating complex real-world images in a synthesis manner via the appropriate integration of an external visual-knowledge memory (i.e., visual cues). We employ a multi-modal encoder to encode visual cues and textual description into a multi-modal hidden vector to trigger the relevant visual counterparts of sentence descriptions. Knowledge retrieval process is stacked along with the stacked image generation process. We conduct experiment and evaluations on MSCOCO and our proposed approach boosts the inception score by 17.6% than the baseline.

2 Related Work

2.1 Memory Network

First proposed by [32, 31], Memory Network has been utilized to augment neural networks for different tasks, such as algorithm inference [7], conversational systems [29, 34, 5] and question answering [36, 31]. Memory helps extend the capability to capture long-term dependencies and provides a way to model relevant information inside their surroundings. As for unconditioned image generation area, [14] presents a deep generative model (DGMs) with memory and attention to capture the local detail information and [12] successfully applied a life-long memory network [11] to adversarial models.

Compared with these memory augmented networks, we propose to employ an external visual-knowledge and memory network to model and leverage the correlations between visual images and textual sentences. The uniqueness of our model will be specified in the next section.

2.2 Generative Adversarial Networks

Recently, Generative Adversarial Networks have shown the ability to generate appealing images with conditions. Generated images are required not only realistic but also well aligned with the condition constraints. The condition variable can be simple discrete class labels [19, 22, 3, 21] and language sequence [24, 4, 38] which is complex in structure and plentiful in expression. Constrained on visual domain, GAN model has been applied to domain transfer [9, 17], image editing [2, 39], super-resolution [13, 10] and style transfer [10]. [25] managed to draw pictures conditioned on object location. [4] proposed a method to edit a given image with specific textual description. Compared with these Conditional GAN models, our proposed vmCAN attempts to synthesize images conditioned on the textual description and multiple relevant sub-images, which can be seen as an appropriate extension to CGAN framework.

3 Knowledge Grounded Synthesis

We formulate the *sentence-to-image* problem as following: given an image description t, which may remark objects, properties of objects and relations between objects, we aim to learn a series of stacked multi-modal encoders $ME_0, ...,$ ME_{gn} and stacked generators $G_0, ...G_{gn}$. The final output is one corresponding image $s = G_{gn}(ME_{gn}(...(G_0(ME_0(t, P_0))...), P_{gn}))$, where P_i is one group of m proposals sampled from Kownledge Proposal Memory and gn is the number of stacked processes. We set $gn = 2$ in our experiment.

Compared with textual-knowledge based system [5, 28], one primary challenge in leveraging visual knowledge is that images relevant to the target synthesized image still contain much irrelevant information. In our opinion, objects are typically the most important part of an image and they can be easily extracted using Region Proposal Network.

Another problem is that relevant sub-images cannot be directly applied to the target image. For example, virtual viewpoint synthesis [30] requires large viewpoint inputs like video sequence which can provide important structure and texture information. As a result, we use semantic vectors to represent these sub-images and employ attention mechanism to leverage them.

Our model will be demonstrated in three parts: 1. Proposal Extraction for knowledge preparation. 2. Multi-modal Encoder to encode text and relevant proposals (i.e. visual cues) into a multi-modal hidden vector. 3. Stacked Adversarial Generation to generate the realistic image in a stacked manner conditioned on the multi-modal hidden vector.

3.1 Proposal Extraction

Region Proposal Network proposed by [26] ranks and refines region boxes called anchors to generate high-quality region proposals which most likely contain an object. After RPN, a Region of Interest Pooling layer is used to normalize different sized CNN feature map into the same size. The output of ROI pooling is used as visual cues in our visual-memory knowledge.

We extract about 320000 proposals from MSCOCO **training** dataset images to build our Proposal Knowledge Memory. Each proposal is a semantic vector of dimension 1024. When there are more than 5 proposals in one image, we just keep top 5 proposals with the highest predicted objectness score. Extraction can be finished in an offline fashion.

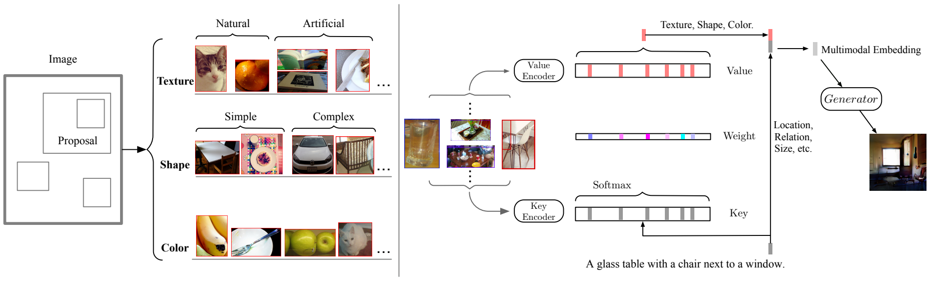

Fig. 1. We extract proposals which may provide texture, shape and color cues to build our visual-knowledge memory. After given a sentence "a glass table with a chair next to a window", the multi-modal encoder triggers some of useful visual cues (i.e. proposals in visual-knowledge memory) to generate a multi-modal hidden vector w.r.t. the given sentence.

3.2 Multi-modal Encoder

Based on memory networks, our Multi-modal Encoder ME uses two encoders and attention mechanism to model proposals (i.e., visual cues) and textual descriptions.

We first encode text description t into a continuous representation $\varphi(t)$ using a pre-trained text encoder φ [24]. We further augment the text embedding using a method proposed by [38]. This augmentation helps generate a large number of additional text embeddings for adversarial training [4]. More formally, a fully connected layer is applied over the input text embedding to generate μ and σ. The augmented text embedding is computed as $\varphi(t)' = \mu + \sigma \bigodot \varepsilon$ where ς is sampled from $\mathcal{N}(0, I)$ and $\bigodot$ is the element-wise multiplication operator. Augmented text vector is of dimension d. We randomly sample m proposals $P = \{p_1, p_2, p_3, ..., p_m\}$ from Proposal Knowledge Memory. Based on [31], these proposals are encoded into key representations and value representations respectively:

$$\begin{aligned} k_i &= \kappa_0(p_i) \\ v_i &= \nu_0(p_i) \end{aligned} \tag{1}$$

Where p_i is a proposal feature vector of dimension 1024, k_i and v_i are of dimension d. Key representation is used to attend and weight retrieved knowledge. Value representation contains useful guidance necessary for generation. Both key and value encoders are neural networks which are simple fully connected layers with ReLU activation function in our model. The multi-modal hidden vector is produced as follows:

$$
\begin{gathered}
a_i = Softmax(\varphi(t)'^T k_i) \\
o = \sum_i v_i a_i \\
c = (o, \varphi(t)') \\
\hat{c} = ReLU(Wc + b)
\end{gathered} \tag{2}
$$

$\hat{c} = ME(t, P)$ is the final multi-modal hidden vector of dimension d.

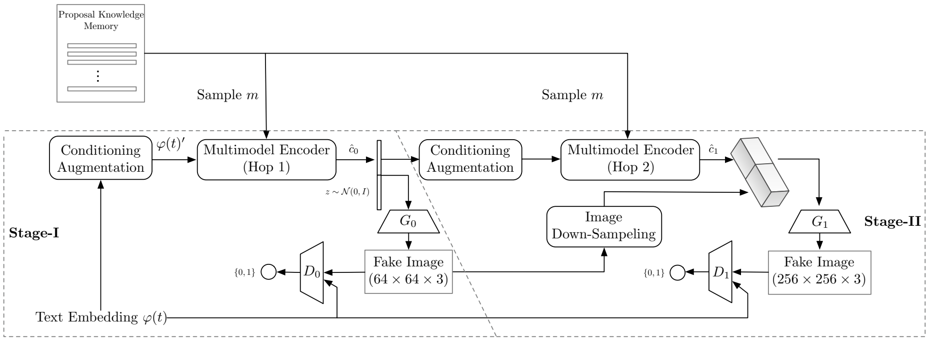
Proposal Knowledge Memory
Sample m
Sample m
Conditioning Augmentation
$\varphi(t)'$
Multimodel Encoder (Hop 1)
$\hat{c}_0$
$z \sim \mathcal{N}(0, I)$
Conditioning Augmentation
Multimodel Encoder (Hop 2)
$\hat{c}_1$
G_0
Image Down-Sampeling
G_1
Stage-I
Stage-II
$\{0,1\}$
D_0
Fake Image ($64 \times 64 \times 3$)
$\{0,1\}$
D_1
Fake Image ($256 \times 256 \times 3$)
Text Embedding $\varphi(t)$

Fig. 2. Pipeline of the proposed vmCAN. At the first stage of generation, augmented text embedding $\varphi(t)'$ and sampled m proposals will be encoded to a multi-modal represenatation $\hat{c}_0$. This feature vector will guide the first sketch-like image generation. For stage-II, mutimodal tensor $\hat{c}_0$ from the first stage will make induction on newly sampled m propoals. The output $\hat{c}_1$ and the downsampled first-stage generated image will be concatenated and used for real-world image generation.

3.3 Stacked Adversarial Generation

Building upon StackGAN-v1 [38], the whole pipeline model is defined in figure 2. **Stage-I Generation** For stage-I Generation, more sketch-like information like shapes from proposals (i.e. visual cues) will be used. The multimodel condition vector $\hat{c}_0$ helps generator produce a low-resolution sketch-like image. A noise vector is sampled from a normal distribution p_z. Concatenated by this noise vector, the multi-modal hidden vector goes through several upsampling blocks to generate a $W_0 \times H_0$ color image. Then Discriminator D_0 downsamples this image to $M_d \times M_d \times N_d$ feature map. Meanwhile, the augmented text embedding

$\varphi(t)'$ is spatially replicated to $M_d \times M_d \times d$ and then concatenated with above image feature map. The concatenated tensor is further downsampled to a fake-real score whose range is between 0 and 1. The loss functions of G_0 and D_0 are defined as follows:

$$\begin{aligned}\mathcal{L}_{D_0} = & \mathop{\mathbb{E}}_{(I,t)\sim p_{data}}[logD_0(I, \varphi(t)')]+ \\ & \mathop{\mathbb{E}}_{z\sim p_z, t\sim p_{data}}[log(1 - D_0(G_0(z, \hat{c}_0), \varphi(t)'))] \\ \mathcal{L}_{G_0} = & \mathop{\mathbb{E}}_{z\sim p_z, t\sim p_{data}}[log(1 - D_0(G_0(z, \hat{c}_0), \varphi(t)'))] \\ & + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi(t)')), \Sigma_0(\varphi(t)')||\mathcal{N}(0, I))\end{aligned} \tag{3}$$

where I is the real image, t is the pre-trained text embedding, z is a noise vector sampled from a given distribution such as Gaussian distribution in our experiment and $\hat{c}_0 = ME_0(t, P_0)$ is the multi-modal hidden vector. In our model, the discriminator is not conditioned on proposals. That is to say, the generated image doesn't need to be well aligned with proposals. Only some useful visual cues inside these proposals are used. The Stage-I model is trained by alternating between maximizing $\mathcal{L}_{D_0}$ and minimizing $\mathcal{L}_{G_0}$.

Stage-II Generation For stage-II, more comprehensive information from proposals (i.e. visual cues) will be used. Such information helps to rectify the imperfection in Stage-I results and add appealing details to them. Stage-I multi-modal hidden vector $\hat{c}_0$ makes induction on newly sampled m proposals to produce $\hat{c}_1 = ME_1(G_0, P_1)$. Encoders κ_1 and ν_1 inside ME_1 in stage-II are trained from scratch in this paper but they can reuse weights from κ_0 and ν_0 inside ME_0 to ease training since they reduce the number of parameters. The image generated by G_0 is downsampled to $M_g \times M_g \times N_g$. $\hat{c}_1$ is spatially replicated to $M_g \times M_g \times d$ and concatenated to the image feature map. Generator G_1 upsamples the concatenated feature map to a $W_1 \times H_1$ image. Discriminator D_1 downsampling process is the same as Stage-I except that the input image is larger and downsampling networks are more complex. Similar to stage-II, the loss functions of G_1 and D_1 are:

$$\begin{aligned}\mathcal{L}_{D_1} = & \mathop{\mathbb{E}}_{(I,t)\sim p_{data}}[logD_1(I, \varphi(t)')]+ \\ & \mathop{\mathbb{E}}_{s_0\sim p_{G_0}, t\sim p_{data}}[log(1 - D_1(G_1(s_0, \hat{c}_1), \varphi(t)'))] \\ \mathcal{L}_{G_1} = & \mathop{\mathbb{E}}_{s_0\sim p_{G_0}, t\sim p_{data}}[log(1 - D_1(G_1(s_0, \hat{c}_1), \varphi(t)'))] \\ & + \lambda D_{KL}(\mathcal{N}(\mu_0(\varphi(t)')), \Sigma_0(\varphi(t)')||\mathcal{N}(0, I))\end{aligned} \tag{4}$$

where s_0 is the generated image by Stage-I G_0. To make it more directly comparable to baseline StackGAN model, we set the model parameters as $N_z = 100$, W_0 =64,H_0 =64,M_d =4,N_d =512,N_d =128,M_g =16,N_g =512,d =128,W_1 =256, $H_1 = 256$, and $\lambda = 1$.

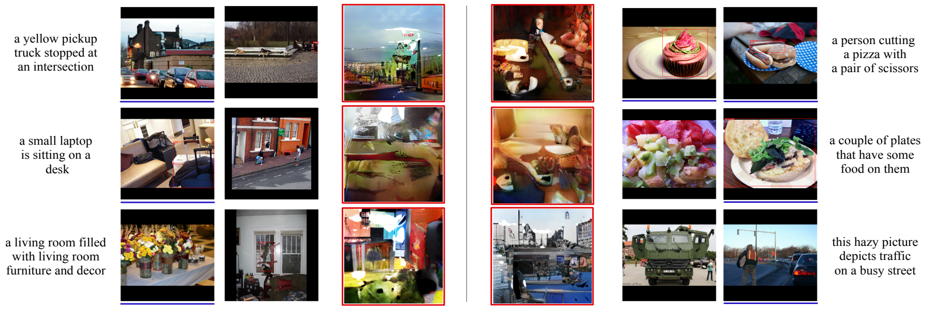
a yellow pickup truck stopped at an intersection
a small laptop is sitting on a desk
a living room filled with living room furniture and decor
a person cutting a pizza with a pair of scissors
a couple of plates that have some food on them
this hazy picture depicts traffic on a busy street

Fig. 3. Some samples of descriptions, top 2 relevant proposals and generated images. Generated images are highlighted using red boxes and relevant proposals are underlined in blue.

4 Experiments

4.1 Experimental Setup

Datasets Our model is evaluated on COCO captioning 2015 dataset. By default, it contains 80k images for training and 40k images for validation with 5 captions per image. There are over 80 semantic object categories in total and each image contains varied objects which are rarely centered in the image.

Evaluation In order to measure images generation recognizability and generation diversity, we use inception score. Moreover, we generate captions to quantitatively measure how well the generated images are conditioned on the textual descriptions.

Inception Score - Inception score is first proposed by [27] and has been acknowledged to be well correlated with human evaluation on the quality and diversity of generated images.

Caption Quality[8] - Since the inception score cannot reflect whether the generated images are well conditioned on the given text descriptions, we generate captions using a pre-trained caption model [18] trained on MS-COCO. Then we measure how similar these generated captions are to textual input using four standard language similarity metrics: BLEU [23], METEOR [1], ROGUE_L [15] and CIDEr [33].

In addition to quantitative evaluations above, we also conduct qualitative evaluations in terms of visualization.

4.2 quantitative Results

Ablative Analysis In order to better understand the impact of visual knowledge, we conduct ablative analysis by using Ground Truth knowledge for each textual description. Given a text-image training pair, proposals extracted from the paired image are considered as Ground Truth knowledge in terms of the

Table 1. Inception Score by different models on MSCOCO test sets. Higher is better.

	Inception Score
StackGAN [38]	8.35 ± 0.03
vmCAN-R	9.94 ± 0.12
vmCAN-GT	$\mathbf{10.36 \pm 0.17}$
chatPainter [28]	9.74 ± 0.02
Hong et al. (2018) [8]	11.46 ± 0.09
AttnGAN [37]	$\mathbf{25.89 \pm 0.47}$

corresponding textual description. Instead of training a multi-modal encoder to encode proposals and text to a meaningful multi-modal vector, we simply average this group of proposal vectors, linearly transform the averaged vector to a tensor of dimension d, concatenate this tensor with augmented text embedding $\varphi(t)'$ and finally non-linearly encode them into a multi-modal hidden vector of dimension d. This vector will be used for further generation. This is, in fact, a weak upper-bound because we simply average these proposal feature vectors. Even though, we get substantial improvement both on the Inception Score and Caption Generation BLEU. This weak upper-bound model is noted as vmCAN-GT and the previous model is named vmCAN-R.

Table 1 shows the image recognisability analysis. In detail, we compare the test set image generation Inception Score between our method, the baseline method, and some other approaches based on conditional GANs. Our model boosts the baseline method by 17.6%. This quantitatively shows that proposals (i.e. visual cues) can help enhance the generation quality and potentially increase the generation variety. We will point out that the proposed vmCAN does not achieve a better performance compared to attnGAN[37]. However, attnGAN, which also has a stacked generation process, can be enhanced by incorporating the visual-knowledge memory and the multi-modal encoders.

Table 2. Evaluation metrics based on caption generation to measure whether the generated images are well conditioned on the given text descriptions. Higher is better in all columns.

	Caption Generation						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEROR	ROUGE_L	CIDEr
StackGAN[38]	0.400	0.188	0.078	0.037	0.092	0.267	0.039
vmCAN-R	0.399	0.187	0.079	0.038	0.093	0.266	0.039
vmCAN-GT	**0.467**	**0.261**	**0.137**	**0.075**	**0.124**	**0.307**	**0.145**
Real Image	0.743	0.577	0.427	0.313	0.273	0.488	0.946

Table 2 shows the caption generation result. By conditioning both on text description and additional visual knowledge, our method yields little loss on text-image(generated) relevance with randomly sampled knowledge and an improvement with Ground Truth knowledge. This result further shows the effectiveness of the utilization of proposals (i.e. visual cues) and the necessity of building an efficient and accurate text-proposal retrieval system.

4.3 Qualitative Results

Figure 3 shows some examples of generated image and Top 2 relevant proposals. In detail, we compute the cosine similarity between the text embedding $\varphi(t)'$ and key encodings of K sampled proposals and visualize the top 2 relevant proposals represented using bounding box. This result shows that our multi-modal encoder is able to activate relevant visual knowledge although some of them are irrelevant from the human perspective.

Figure 4 shows some creative generation examples generated by our proposed method. These results illustrate that our method is able to generate novel images which do not exist in the source dataset.

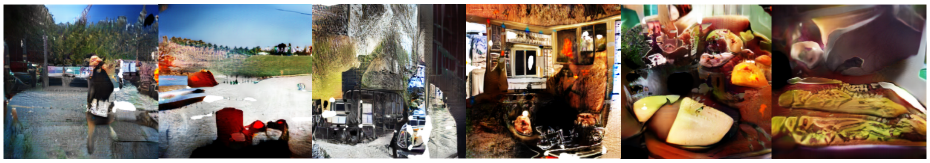

Fig. 4. Some novel images generated by our model.

5 Conclusions

In this paper, we propose a visual-memory augmented approach named vm-CAN for *sentence-to-image* synthesis. Our model obtains substantial improvement over the baseline method on the challenging MSCOCO dataset.

6 Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Nos.61751209,U1611461), the Key Program of Zhejiang Province, China (No. 2015C01027), and Artificial Intelligence Research Foundation of Baidu Inc.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Brock, A., Lim, T., Ritchie, J.M., Weston, N.: Neural photo editing with introspective adversarial networks. In: ICLR (2017)
3. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)
4. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: ICCV (2017)
5. Ghazvininejad, M., Brockett, C., Chang, M.W., Dolan, B., Gao, J., Yih, W.t., Galley, M.: A knowledge-grounded neural conversation model. In: AAAI (2017)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
7. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
8. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. arXiv preprint arXiv:1801.05091 (2018)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
10. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. Springer (2016)
11. Kaiser, L., Nachum, O., Roy, A., Bengio, S.: Learning to remember rare events. In: ICLR (2017)
12. Kim, Y., Kim, M., Kim, G.: Memorization precedes generation: Learning unsupervised gans with memory networks. arXiv preprint arXiv:1803.01500 (2018)
13. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
14. Li, C., Zhu, J., Zhang, B.: Learning to generate with memory. In: ICML (2016)
15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Marie-Francine Moens, S.S. (ed.) Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (July 2004)
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. In: ECCV. Springer (2014)
17. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
18. Luo, R.: An image captioning codebase in pytorch. https://github.com/ruotianluo/ImageCaptioning.pytorch (2017)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
20. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on. pp. 722–729. IEEE (2008)

21. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML (2017)
22. van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: NIPS (2016)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. Association for Computational Linguistics (2002)
24. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
25. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS (2016)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015)
27. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS (2016)
28. Sharma, S., Suhubdy, D., Michalski, V., Kahou, S.E., Bengio, Y.: Chatpainter: Improving text to image generation using dialogue. arXiv preprint arXiv:1802.08216 (2018)
29. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B.: A neural network approach to context-sensitive generation of conversational responses. In: NAACL-HLT (2015)
30. Starck, J., Hilton, A.: Virtual view synthesis of people from multiple view video sequences. Graphical Models **67**(6), 600–620 (2005)
31. Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. In: NIPS (2015)
32. Sukhbaatar, S., Weston, J., Fergus, R., et al.: Memory networks. In: ICLR (2015)
33. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
34. Vinyals, O., Le, Q.: A neural conversational model. In: ICML (2015)
35. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
36. Weston, J., Bordes, A., Chopra, S., Rush, A.M., van Merriënboer, B., Joulin, A., Mikolov, T.: Towards ai-complete question answering: A set of prerequisite toy tasks. In: ICLR (2016)
37. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. arXiv preprint arXiv:1711.10485 (2017)
38. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
39. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV. Springer (2016)