Words to Birds: Text-to-Image Generation Using AttnGAN with Pre-trained Image Captioning Model and BERT

Lizhi(Gary) Yang*
lzyang@berkeley.edu

Alex Zhou
alexzhou00@berkeley.edu

Abstract

In this project we investigate the problem of text-to-image generation with GANs. Traditional text-to-image GANs encode the whole piece of text as one giant vector and pass it to the generator. However this has the problem of ignoring the specific details in a sentence, for example, attributes for the color, the shape, etc. AttnGAN [18] alleviates this problem by introducing an attention mechanism into the GAN by incorporating both sentence vectors and word vectors into the loss giving it the ability to look at these details, thus generating more accurate and vibrant images. We modify the network by utilizing the advancements in the Natural Language Processing domain, namely BERT [4], and a pre-trained image-caption model [7] to replace the text-encoders and the Deep Attentional Multimodal Similarity Model (DAMSM) reponsible for fine-grained loss in AttnGAN, thus providing a simpler, more modular model that can take advantage of pre-trained modules from other technical domains.

1. Introduction

Text-to-Image generation is an interesting problem as it has great potential in the art and design field. Just image what potentials there can be when images can be generated out of our words! The definition of this problem is as follows: given a piece of text, for example the sentence "this is a red bird.", we want to generate a photo-realistic image of a red bird (as opposed to something close but different like a green bird or a red dog).

Recent approaches to this problem use GANs to generate images from text, since GANs have the ability to encode text into feature representations and use a generator and a discriminator to do self-adversarial training in order to generate realistic images. It comes naturally to just encode the whole piece of text into a global vector and use it as the condition for image generation using GANs, such as in [10]. However, this method ignores the information at the local word level, and AttnGAN [18] addresses this problem by using word features on top of sentence features and using the Deep Attentional Multimodal Similarity Model (DAMSM) in order to compute a fine-grained loss to incorporate into the GAN.

With the recent advances in natural language processing, more and more opportunities now exist in this field as more and more powerful new tools in the natural language processing field become available to us. BERT [4] is a prime example. Thus it comes naturally that we could embed these tools within the natural language processing realm and speed up development. In this report we demonstrate how we combine a text-to-image generation model, AttnGAN [18], and advancements in natural language processing, namely BERT [4] and a pre-trained image captioning model by Luo *et al.* [7], showing satisfactory results.

2. Related Work

Image Generation

Our project requires us to be able to generate realistic looking images. Much work has been done on this already. For example, there are autoregressive models such as PixelCNN [12], PixelRNN [13], and image transformers [9], which can perform tasks like image completion, generating realistic images given a class, or super-resolution. Another method for image generation is training generative adversial networks (GANs) [6]. Rather than computing loss in a latent space, as in the aforementioned models, GANs have a second network called the discriminator network that provides feedback to the generation network. This turns image generation into a two-player mini-max game, where the generation network learns to fool the discriminator network, and the discriminator network also learns to distinguish between real and fake images. These types of networks have already found some success in applications to text-to-image synthesis ([10], [18]).

Image Captioning

For the modifications we are making to AttnGAN, we need to use an image captioning network. Of note is "Show and Tell: A Neural Image Caption Generator" [15], which

*Code available at: https://github.com/lzyang2000/cs194FinalProject

uses a CNN as an image encoder followed by an RNN decoder that generates the captions. Future works have built upon this. For example, Xu *et al.* came up with "Show, Attend and Tell" [17], which modified the RNN to include attention. This greatly improved captioning performance and allowed for better visualization of what the network understands about how captions correspond to image structure. Beyond attention, another improvement in image captioning found by Luo *et al.* [7] was to add a discriminability loss when training image captioning networks to penalize captions that are too similar for two different images, incentivising models to generate more specific and detailed captions. Their pre-trained models are imported for our work here.

Text to Features

The past decade has seen many advancements in natural language processing. In 2013, Mikolov *et al.* published word2vec [8], which could learn to represent words as vectors in a meaningful way where not only were vectors close to each other similar in meaning, but the vector difference between two words could capture their relationship. For example, the vector difference between "man" and "woman" is very similar to the vector difference between "uncle" and "aunt".

More recent works have made progress in not just representing individual words, but representing sentence structure as well. In the past, people have attempted to solve this by using RNN's and LSTM's, which have hidden states for "remembering" prior words when sequentially processing text. However, the latest powerful models like GPT-3 and BERT ([2], [14]) use transformers, which is an attention-based model that has led to breakthroughs in many language tasks like translation, summarization, and question answering.

3. AttnGAN with Pre-trained Image Captioning Model and BERT

3.1. AttnGAN

As is the case with most other Text-to-Image models, AttnGAN begins with generating an image from small to large scales one by one, similar to a image pyramid. Specifically, denoting the generators as $G_0, G_1, ..., G_{m-1}$, the hidden states in between as $h_0, h_1, ..., h_{m-1}$, and the small-to-large-scale generated images as $\hat{x}_0, \hat{x}_1, ..., \hat{x}_{m-1}$, we have:

$$\begin{aligned} h_0 &= F_0(z, F^{ca}(\bar{e})) \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{for} i = 1, 2, ..., m-1 \\ \hat{x}_i &= G_i(h_i) \end{aligned} \tag{1}$$

where z is a noise vector sampled from a standard normal distribution, $\bar{e}$ is the sentence vector and e are the word vectors. F^{ca} is the Conditioning Augmentation [19] which transforms the sentence vector $\bar{e}$ to the conditioning vector. F_i^{attn} is the attention model at the i^{th} stage of the AttnGAN. The attention model is what provides the fine-grainedness of the AttnGAN and is constructed by taking both the word features e and image features h, and computing a word-context vector for each sub-region of the image, and this is what replaces the noise z in iterations of the GAN. This highlights words that have higher importance and thus the name "Attention". For example, in the sentence "This bird has a red belly black crown and grey primaries", if we bundle everything together and simply pass that to the GAN, it may take the words "This, red, belly, bird, has" more seriously over others, losing the other colors, while if we add attention, the model would attend for words "black, red, grey, this, bird", and thus preserving the color details. The discriminators then take the various resolution images and try to tell if they match the captions. This multi-stage discriminator setup ensures that at each stage something meaningful is added. The loss for the discriminators are

$$\begin{aligned} L_{G_i} = &-\frac{1}{2}E_{\hat{x_i}\sim P_{G_i}}[log(D_i(x_i))] \\ &-\frac{1}{2}E_{\hat{x_i}\sim P_{G_i}}[log(D_i(x_i, \bar{e}))] \end{aligned} \tag{2}$$

while the loss for the generators are

$$\begin{aligned} L_{D_i} = &-\frac{1}{2}E_{x_i\sim P_{data_i}}[log(D_i(x_i))] \\ &-\frac{1}{2}E_{\hat{x_i}\sim P_{G_i}}[log(1-d_i(\hat{x}_i))] \\ &-\frac{1}{2}E_{x_i\sim P_{data_i}}[log(D_i(x_i, \bar{e}))] \\ &-\frac{1}{2}E_{\hat{x_i}\sim P_{G_i}}[log(1-d_i(\hat{x}_i, \bar{e}))] \end{aligned} \tag{3}$$

The other key piece in the model is the Deep Attentional Multimodal Similarity Model (DAMSM). DAMSM is used as an objective to check if every word in the caption is accounted for appropriately in the generated image, since the discriminator only uses whole sentence vectors. It takes an image and the bundle of word vectors and tells us how well they match. It requires pretraining since it is an expert of sorts with prior knowledge that we rely on. The total loss of the generator network is therefore:

$$L = \sum_{i=0}^{m-1} L_{G_i} + \lambda L_{DAMSM}$$

and the discriminators are disjoint, thus they have their own loss L_{D_i}.

3.2. Image Caption Model

So what if we do not want to do pretraining for the DAMSM or do not want to use it at all? We propose that

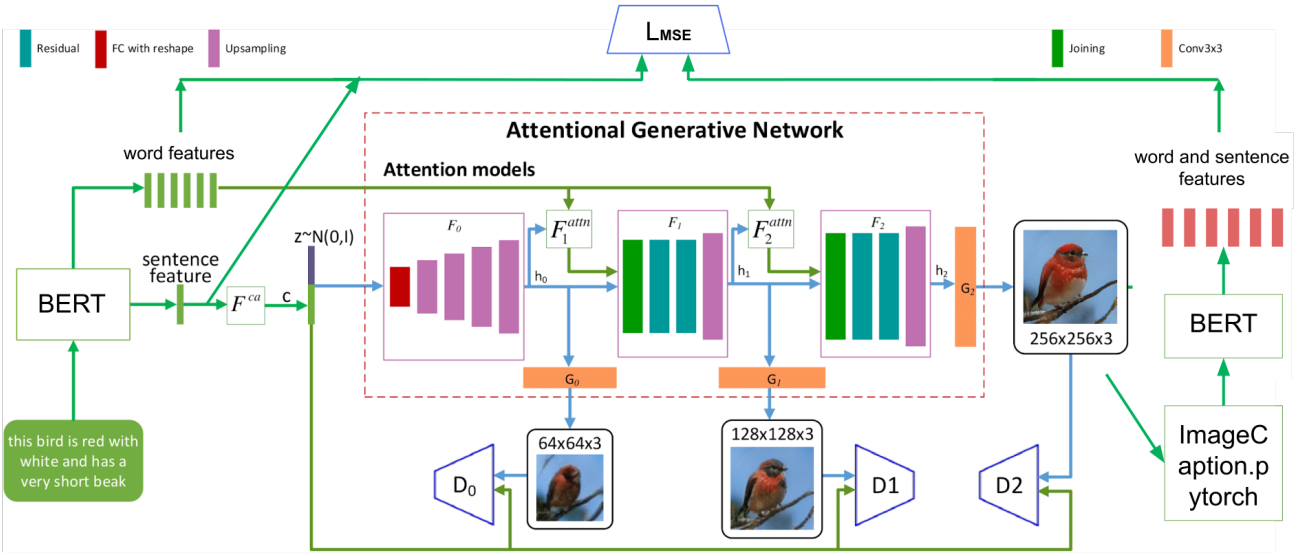

Figure 1. Modified AttnGAN: Used pre-trained BERT and image caption network to generate the word/sentence vectors used in the losses.

Figure 2. The caption reads: two zebras and a giraffe walking down a road.

instead of using DAMSM, we simply use a pre-trained image caption model, ImageCaption.pytorch [7]. It uses transformer captioning and employs self-critic [11] training and uses bottom-up features [1]. An example of image caption is shown in Figure 2.

3.3. BERT

In order to leverage the advancements in natural language processing, we utilize BERT [4] as our feature extractor for the sentence passed in in order to get the word vectors and the sentence vectors. BERT applies the bidirectional training of the Transformer [14] to language modelling and has a deeper sense of context and flow of the text than previous models that were trained either from left to right or the reverse. This gives us better understanding of the text passed in.

3.4. Our modifications

We build our model on top of the AttnGAN model by swapping out the text encoders and replacing them with pre-trained BERT weights from spacy-transformers [5], and incorporated the word vectors and sentence vectors it produced into training. It is interesting to note that the original text encoder is of output dimension 256 while BERT outputs a dimension 768 tensor as features. Instead of the DAMSM loss, we generate image captions on the output image and use a simple MSE loss to compare the sentence/word vectors of the generated captions to the ground truth captions. Our modified architecture is shown in Figure 1. Our loss would then become

$$L = \sum_{i=0}^{m-1} L_{G_i} + \lambda(L_{MSE}(\bar{e_o}, \bar{e_g}) + L_{MSE}(e_o, e_g))$$

where $\bar{e_o}$ and $\bar{e_g}$ are the sentence and word vectors generated from the image caption network with e_o and e_g being the sentence and word vectors generated from ground-truth captions.

4. Experiments

We used the Caltech-UCSD Birds 200 dataset [16] for training. The experiments were ran on a single NVIDIA GTX1080Ti GPU with batchsize of 8. The experiments were run for 45 epochs, and Figures 3 to 9 showcase some results.

Figure 3. A brown colored bird with a long tail and a very small bill in comparison to its body.

Figure 4. A small bird with yellow body black head and short sharp beak.

Figure 5. This bird has a bill and a large black eye with a yellow throat and a grey breast.

Figure 6. This bird has wings that are brown and a white belly.

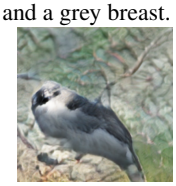

Figure 7. This bird is all white except for black wings and gray and speckled head.

Figure 8. This bird is brown and yellow in color with a stubby beak.

Figure 9. This is a bird with blue and black wings and yellow belly.

5. Conclusions

In this final project we tried to modify AttnGAN to require less specialized pretraining by replacing the AttnGAN's DAMSM with pre-existing models used for other tasks, namely the image-caption and BERT models. We acquainted ourselves with the use of attention layers in computer vision that focuses on local features for better details, and also the encoder-decoder framework used in all components, be it the GAN, the image-caption network or the BERT network. Indeed the trend of borrowing methods from each others' domain is becoming more and more popular - the recent introduction of transformers in computer vision [3] is proof of that. It would be very interesting to see and maybe do more work in this area as it does lead to a more generalized AI model. Future work extending from this project would include trying to incorporate more of the advancements in the NLP domain into the GAN structure instead of just utilizing these tools in the loss calculation, and also to search for a standardized evaluation metric to evaluate the performance of the modifications.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Explosion. spacy–transformers. https://github.com/explosion/spacy-transformers, 2019.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[7] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*, 2018.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[9] N. Parmar, A. Vaswani, J. Uszkoreit, Łukasz Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer, 2018.

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[11] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. corr abs/1612.00563, 2016.

[12] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications, 2017.

[13] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks, 2016.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator, 2015.

[16] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[18] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.