Out of Context - Fixing Attention in AttnGAN: Final Report

G112 (s1870915, s1866666, s1876426)

Abstract

In this work we propose two solutions to the AttnGANs problem with attention accuracy. The AttnGAN for text-to-image synthesis, which uses an auxiliary attentional model, the deep attentional multimodal similarity model (DAMSM), often fails to correctly match text to the corresponding image sub-regions. We propose to augment the attention mechanism to incorporate lexical context information via a weighted average of attention scores, called AttnGAN+lex (lexical-aware attentional generative adversarial network). Furthermore, we propose similar changes to the DAMSM, called LASM.

1. Introduction

Despite the great progress made in past years in the field of automatic image generation, generating images from natural language description remains a hard problem. The state-of-the-art attentional generative adversarial network (AttnGAN) has been proved capable of producing high-resolution images with plausible contextual accuracy by using an attentional mechanism which enables the network to draw sub-regions of an image with the focus on different words in a descriptiom(Xu et al., 2017).

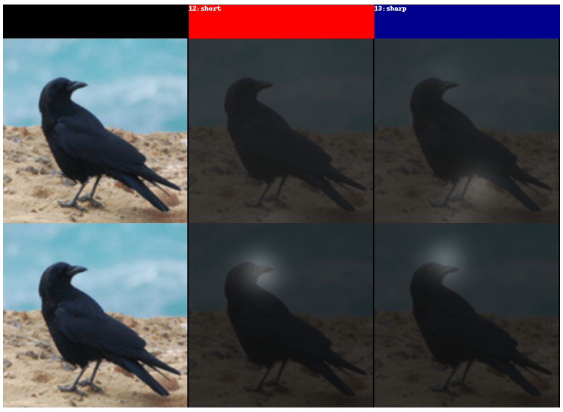

Figure 1. The attention of the "short" and "sharp" in "This medium sized all black bird has long tail feathers and features a beak that is short, sharp pointed and looks very sturdy." The top row is produced using the original model (Xu et al., 2017), while the bottom row is produced by ours.

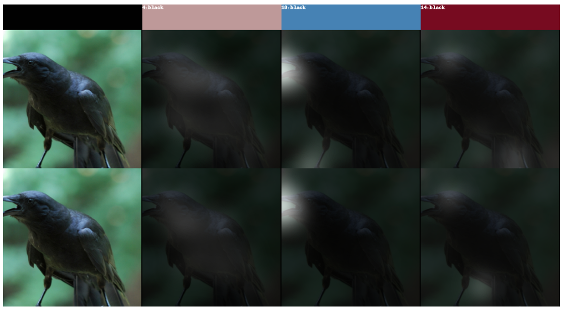

Figure 2. The attention of the three "black"s in "This bird is shiny black in color, and has a black beak with a black eye ring." produced by the original model (top) and our model (bottom). The three "black"s are correctly attended by our model on the corresponding part of the bird: body, beak, and eye. Although the first and third one are not quite confident.

Nevertheless, the attention obtained by the model is often not so accurate as to be useful in the image generation. We attribute this inaccuracy to the text encoder used to extract word and sentence vectors in the attention model.

The text encoder in the AttnGAN is a recurrent neural network (RNN), (more precisely a single-layer bi-directional long short term memory (LSTM) network), which takes a sequence of word embedding vectors w and outputs a sequence of word encoding vectors e. In addition, the last output of the sequence is also used as the sentence vector $\bar{e}$. Because both the global (sentence-level) and local (word-level) features have to be extracted using the same RNN, the model is forced to settle at a middle-ground between the two scales by making the word encoding vectors e more 'contextual', and hence degrade the local attention accuracy.

To ameliorate this problem, we propose a lexical attentional similarity model (LASM) which includes additional lexical information in the computation of the attention, whilst preserving sufficient contextual information. Shown in the figure 1 is the attention map produced by our lexical attentional model and the original model. It shows a fail case where the original model does not capture the selected words "short" and "sharp" from the caption "this medium sized all black bird has long tail feathers and features a beak that is short, sharp pointed and looks very sturdy." We believe the problem here is that the RNN has to encode the verbose contextual information in this long sentence that each output has little information related to the very word it is processing. By using the word embedding vector l in the decision process, our model is able to attend correctly to the corresponding part of the image. Figure 2 shows another example of our model where the same word in a sentence has different attention. This indicates that our model is also able to determine the attention according to the context.

To evaluate our model quantitatively, we build an LASM-adapted version of AttnGAN and compute the inception score of the generated images for both, the original AttnGAN and our lexical AttnGAN. We found that this novel attentional generative adversarial network with lexical attentional similarity model (AttnGAN+lex) produced higher scores than the original AttnGAN, thus proving the improvement of our model.

1.1. Change of project (objectives)

Our original project idea focused on using the AttnGAN for Style Transfer, relying heavily on the attention mechanism. Unfortunately, our initial experiments did not produce any informative results related to our objective.

But while trying out the AttnGAN and testing it on the CUB data set 3.1, we observed that the attention masks on the image pixels often did not match the corresponding input words. This gave us the idea to include the lexical vectors into the computation of attention, which should direct the attention during the image generation to the correct subregions of the image.

The remaining contents of this paper are organized in the following way. Section 2 introduces related work in this area of study. Section 3 describes the data set we used in our experiments. Section 4 elaborates on the AttnGAN and the lexical attentional model. Section 5 provides the experiment details and section 6 concludes the paper and provides directions for future work.

2. Related work

GANs were initially proposed to produce novel images based on a zero-sum competition between a discriminator and a generator. The discriminator attempts to distinguish the fake images produced by the generator from the real ones, while the generator attempts to fool the discriminator into classifying the fake images as real (Im et al., 2016). Conditional GANs, as described in (Mirza & Osindero, 2014; Springenberg, 2015) can be used to generate images of a specific class by conditioning on class labels. Despite the difficulties in training such model, e.g. modal collapse, they has been able to achieve high class accuracy. The additional challenge, when generating images from text descriptions, is that we need to learn a multimodal translation from text to image (Gorti & Ma, 2018; Xu et al., 2017).

Some examples for other GAN architectures producing photo-realistic images based on text are the StackGAN (Zhang et al., 2016) and the TAC-GAN (Dash et al., 2017).

The StackGAN improves the inclusion of necessary image details and the image overall quality by stacking multiple GANs on top of another. Each GAN receives the descriptive text as input. The subsequent GAN additionally takes in the output of the previous GAN. Starting with a rough sketch, the image is refined in a consecutive GAN, correcting mistakes and adding realistic details (Zhang et al., 2016).

The TAC-GAN (Text Conditioned Auxiliary Classifier Generative Adversarial Network) combines descriptive text information with class information, increasing diversity and structural coherence of the produced images. They do this by conditioning on the input text and also feeding in the text information to the classifier before classification. This GAN has been shown to perform even better than the before mentioned StackGAN (Dash et al., 2017).

Still, none of these models make use of an attention mechanism to improve text-image correspondence. While attention is broadly used in natural language processing for machine translation, its applications for text-to-image translation have not yet been widely explored, with the AttnGAN being the first to do so (Vaswani et al., 2017).

Additionally, the inclusion of context into the attention model for natural language processing has been shown to improve machine translation significantly(Nguyen & Chiang, 2017). Even though text-to-image translation is a different translation task, depending on other networks and architectures, the underlying (theoretical) ideas to improve machine translation might still be applicable or transferable to this kind of translation.

3. Data set and task

To improve the choice of image pixels where the different words should attend to, we decided to include additional lexical information into the attention mechanism. Our hypothesis is that this can help to put the attention on the correct image patch and thereby improve text-to-image synthesis. To prove this, we train on a class specific data set, the CUB-200-2011.

3.1. The data

The Caltech-UCSD Birds-200-2011 (CUB-200-2011)(Wah et al., 2011) data set consists of 11,788 images of birds in 200 categories. Each image is annotated with 10 captions describing the image, as well as one bounding box and its corresponding label. Additionally, for each image 15 names of parts and their locations are given, as well as the certainty for each of the 312 binary attributes (with labels) to be present in the image. Only the annotated captions will be used in our task.

Two example images from the data set with two example captions each are shown in Figure 3 and 4.

The results of our experiments were quantitatively evaluated using the inception score(Salimans et al., 2016). To obtain the inception score, the Inception model(Szegedy et al., 2015) is applied to all the generated images, and the image label distribution is computed. The score then measures the entropy of the label distribution, which should be low whenever the contents are of high enough quality. Additionally, the diversity between the different images is measured and incorporated into the score. The final score gives a measure of the KL-divergence between the two properties.

Furthermore, we evaluated and compared a set of images generated by our improved model and the original model manually, focusing on differences in the attention masks. Due to the difficulty of finding an objective qualitative measurement, we compared the attention masks on images produced for the captions "yellow bird with red wings and a round beak" and "white bird with black wings and large beak" only. The overall image quality was assessed on a variety of bird classes (or rather their corresponding captions).

Figure 3. A photo of a Cardinal from the data set.
"the bird has small beak when compared to its body, the beak is sharp and pointed, and it has red breast and belly."
"bird has red body feathers, red breast feather, and red beak"

Figure 4. A photo of a Painted Bunting from the data set.
"colorful bird! the crown is bright blue, the breast and belly are a pinkish orange. the wings are yellow."
"a small bird with a red belly and breast, green wings and a blue head and neck."

4. Methodology

This section describes the AttnGAN+lex, an attentional generative adversarial network with lexical attentional similarity model (LASM). We first briefly introduce the relevant aspects of the baseline model AttnGAN, and then go on to the contributions of the AttnGAN+lex.

4.1. The Attentional Generative Adversarial Network

The Attentional Generative Adversarial Network (AttnGAN)(Xu et al., 2017) consists of two components: an attentional generative network and a deep attentional multimodal similarity model(DAMSM). The architecture is shown in figure 5.

The attentional generative network generates images in several consecutive stages. Following the notations in the original paper, the attentional generative network has m generators $(G_0, G_1, ..., G_{m-1})$, which take the hidden states $(h_0, h_1, ..., h_{m-1})$ as input and generate images $(\hat{x}_0, \hat{x}_1, ..., x_{\widehat{m-1}})$. The generative pipline can be defined as:

$$h_0 = F_0(z, F^{ca}(\bar{e})); \tag{1}$$

$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})); \tag{2}$$

$$\hat{x}_i = G_i(h_i) \tag{3}$$

In this, z is a random noise sampled from a standard normal distribution, $\bar{e}$ is the global sentence encoding, and e is the matrix of word encodings. F^{ca} is the Conditioning Augmentation which converts $\bar{e}$ to a conditional vector. F_i^{attn} is the attentional at i^{th} stage and F_i are the hidden convolutional networks which refines and up-scales the hidden states stage by stage.

In each stage, $F_i^{attn}(e, h)$ takes a image feature state $h \in \mathbb{R}^{\hat{D}\times N}$ and a word encoding matrix $e \in \mathbb{R}^{D\times T}$ to determine the conditional vectors of the next stage. Here, each column of e represents a word feature and each column of h represents the context features of a sub-region of the image. The word encoding matrix e is first transformed into the common semantic space of the image features by a perceptron layer, i.e $e' = Ue$, $U \in \mathbb{R}^{\hat{D}\times D}$. Let $\beta_{j,i}$ denotes the weights the model attends to the i^{th} word when generating the j^{th} sub-region of the image, then $\beta_{j,i}$ can be computed by:

$$\beta_{j,i} = \frac{exp(s'_{j,i})}{\sum_{k=0}^{T-1} exp(s'_{j,k})} \tag{4}$$

$$s'_{j,i} = h_j^T e'_i \tag{5}$$

The word-context matrix $F_i^{attn}(e, h) = (c_0, c_1, ..., c_{N-1}) \in \mathbb{R}^{\hat{D}\times N}$ for image feature matrix h is computed by:

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i \tag{6}$$

Finally, the word-context matrix is concatenated with the image feature h to generate images at the next stage.

The training object of the AttnGAN is to minimize the generator loss function L:

$$L = L_G + \lambda L_{DAMSM}, \text{ where } L_G = \sum_{i=0}^{m-1} L_{G_i}. \tag{7}$$

$$L_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[log(D_i(\hat{x}_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[log(D_i(\hat{x}_i, \bar{e})]}_{\text{conditional loss}} \tag{8}$$

Here, λ is a hyper-parameter λ is used to balance the two terms, and L_{DAMSM} is the attentional loss function, which measures how well the local and global text-image attention fits the data. To learn an accurate text-image alignment, the attentional model must be pre-trained on the real-image dataset before the training of the GAN (generators and discriminators) with the objective of minimizing only the attentional loss function L_{DAMSM}. This term is also included in the GAN loss function because the generator has

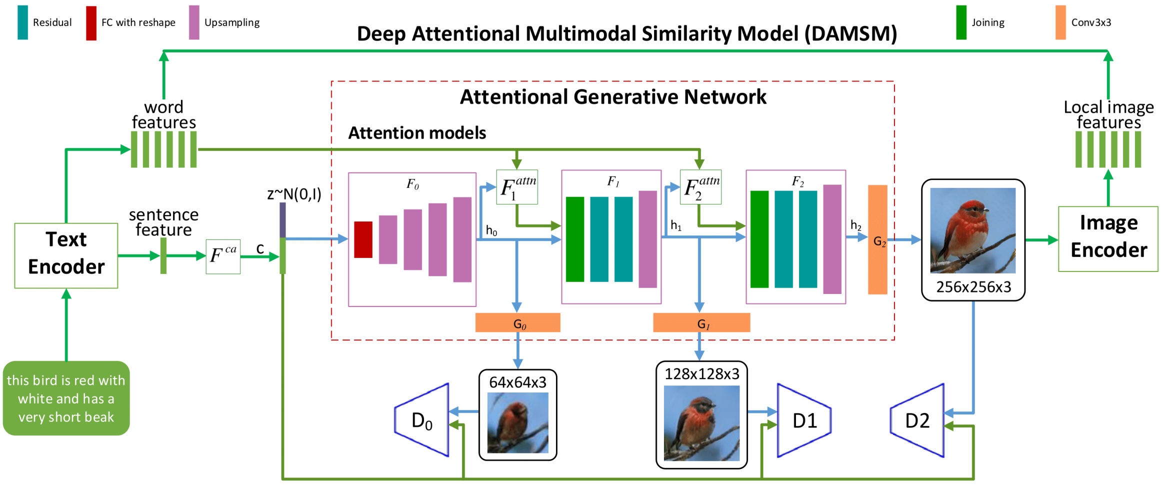

Figure 5. The architecture of AttGAN. Figure adopted from (Xu et al., 2017)

to work under the same semantic space of the attentional model so as to meaningfully utilize the attention produced by the attentional model. As LASM uses the same loss function as DAMSM, we will not elaborate the term in this paper. Please go to the original paper for more details (Xu et al., 2017).

To achieve the training objective, we also have to minimize the loss function for discriminators $(D_1, ..., D_{m-1})$:

$$L_{D_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{data_i}}[log(D_i(x_i)] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{data_i}}[log(1 - D_i(\hat{x}_i)]}_{\text{unconditional loss}} + \tag{9}$$

$$\underbrace{\left\{-\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{data_i}}[log(D_i(x_i, \bar{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[log(1 - D_i(\hat{x}_i, \bar{e})]\right\}}_{\text{conditional loss}} \tag{10}$$

4.2. Lexical Attentional Similarity Model

Our approach to incorporate the lexical context into the attention scores is inspired by a paper proposing a lexical module to reduce wrong translations of rare words in neural machine translation (Nguyen & Chiang, 2017).

In this paper, the neural network used for machine translation is an RNN based encoder-decoder with attention. The authors use a simple feed-forward network to compute a weighted average of the word embedding vectors at each time step (t) weighted by the attention weights. This weighted lexical embedding vector is then combined with the decoder output and included in the computation of the final adapted scores. This proves the importance of lexical information in the text-text alignment. Therefore, we assume that it is also the case in the text-image alignment.

Based on this assumption, we made a simple modification to DAMSM to make it 'lexical'. Instead of just using the RNN outputs e to represent the word features, we add a skip connection and two perceptron layers to produce a word embedding vector e_l combining both context-level features $e \in \mathbb{R}^{D \times T}$ and lexical-level features $w \in \mathbb{R}^{\bar{D} \times T}$ which is obtained from the one-hot encoding before the RNN. The final word embedding vector $e_l \in \mathbb{R}^{D \times T}$ is computed by

$$e_l = \tanh(W_r e + b_r + W_l w + b_l) \tag{11}$$

where $W_r \in \mathbb{R}^{D \times D}$, $W_l \in \mathbb{R}^{D \times \bar{D}}$ and $b_r, b_l \in \mathbb{R}^{D \times 1}$ are biases. We then use this new word embedding vector e_l instead of e to train the attentional model. Meanwhile, we adapt the attentional generative network accordingly. Denoting $\beta'^{l}_{j,i}$ as the new attentional weights the model attends to the i^{th} word when generating j^{th} sub-region of the image. The new word-context matrix $F_i^{attn+l}(e^l, h) = (c_0^l, c_1^l, ..., c_{N-1}^l) \in \mathbb{R}^{\hat{D} \times N}$ is computed by

$$c_i^l = \sum_{i=0}^{T-1} \beta'^{l}_{j,i} e'^{l}_i \tag{12}$$

5. Experiments

To test the performance of the lexical model, we conducted a set of experiments, described in the remainder of this section.

Following Xu et al. (2017), we split the CUB dataset into balanced training (8855 images) and validation (2933 images) sets. We first pre-train a baseline attentional model (DAMSM) and a lexical attentional model (LASM) to learn the text-image alignment.

Each model is trained for 200 epochs, among which the model with best validation loss is selected. The images are resized into 299 × 299 so as to fit the inception model. The 10 captions for each image are selected randomly during training and only the fist 18 words in each caption are taken as the input of the text encoder for simplicity and generalization purposes.

We use ADAM(Kingma & Ba, 2014) optimizer on both text and image encoder with a geometric learning rate annealing: denoting the learning rate at the i^{th} epoch as α_i, we have $\alpha_i = 0.98 * \alpha_{i-1}$. Other hyper-parameters used in the DAMSM and LASM pre-training are listed in table 1.

batch size	48	**RNN hidden size**	256
ADAM α_0	0.002	**RNN grad clip**	0.25
ADAM β_1	0.5	γ_1	4.0
ADAM β_2	0.999	γ_2	5.0
RNN dropout	0.5	γ_3	10.0

Table 1. The hyper-parameters used in the attentional models, the γs are the parameters in the computation of the loss function L_{DAMSM}

Using the pre-trained attentional models, we continue to train two attentional generative networks with the attentional model weights frozen. For each model, we deploy 64 filters in the generators, and 64 filters in the discriminators. The initial random noise is 100 in size, and the conditional and hidden states in each stage have 32 channels (together 64). Following Xu et al. (2017), we use 3-stage generative models and the size of final image output is 256x256. More detailed hyper-parameter information are listed in table 2.

After training each model for 600 epochs, we evaluate their performance using the inception score. Because computing the inception score is expensive, we only perform the evaluation once per 50 epochs. The result of the evaluation is shown in the figure 6 and table 3. It can be seen that AttnGAN+lex outperforms the AttnGAN in terms of inception score. We also tried to unfreeze the attentional model and continue to train it jointly with the GAN after a certain number of epochs (550) of GAN training. We perform 50 epochs of joint training and the result turned out to be worse than the model it starts from: the inception score decreases to merely 4.37 ± 0.15. This is because of a shift in the training objective. The loss function of the GAN does not converge, while the loss of the attentional model does. The gradients w.r.t the GAN losses overwhelm the gradients w.r.t. the attentional losses, and gradually alter the attentional model in favor of the GAN, causing an undesirable degradation.

batch size	20	γ_1	4.0
ADAM α_0	0.0002	γ_2	5.0
ADAM β_1	0.5	γ_3	10.0
ADAM β_2	0.999	λ	5.0

Table 2. The hyper-parameters used in the training of AttnGAN/AttnGAN+lex.

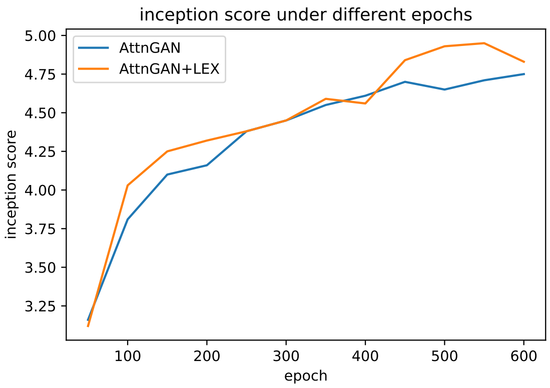

Figure 6. The inception score of the baseline and lexical model, under different epochs.

AttnGAN	4.75 ± 0.22
AttnGAN+lex	**4.95 ± 0.19**

Table 3. The best inception score obtained by the baseline and lexical model

5.1. Interpretation and discussion

While visually inspecting and comparing the images generated by the baseline and the lexical for the two captions "a yellow bird with red wings and a round beak" and "a white bird with black wings and a large beak" (20 images each), we observed that actually the original AttnGAN did get the color for the bird and for the wings right more often than the AttnGAN+lex.

The red wings were also better portrayed by the original model. It produced clearly visible red wings 11 out of 20 times, compared to only 2 for AttnGAN+lex. Both models still seem to have trouble drawing the beak right. Overall, the birds produced by AttnGAN+lex have a higher chance of actually having a beak, even though not necessarily of the right form, than the ones produced by the original.

When only comparing the numbers of photo-realistic images produced for a variety of captions, there seem to be more realistic ones produced by the AttnGAN+lex, even though it did not draw all the details correctly in the before mentioned examples.

Since these visible qualities were assessed rather subjectively on a far too small sample, they might not represent the truth. Another difficulty in the assessment is the fact that we can not confidently say which effects were caused by the changes to the model and which only depend on the random initialisation (or image generation in general, randomness), making it harder to determine which of the observations were actually caused by the LASM.

In the two examples of attention heatmaps in Figures 8 and 10, it becomes visible that the attention mask did change to include the context of the current word. Especially for the colors, red and yellow, attention to a larger image subregion is indeed desired (attention scattered over a larger area). The attended regions in the original model are generally smaller and noticeably outlined, but this is only useful for small and clearly delimited features, like "beak". For these kind of details, the original model seems to have an advantage.

A few images, like the ones presented, do furthermore indicate that the lexical sometimes has an undesired side effect. In Figure 7, it seems as if the bird has become rounder instead of the break. We also observed some examples for "long beak", where instead of the beak, the body of the

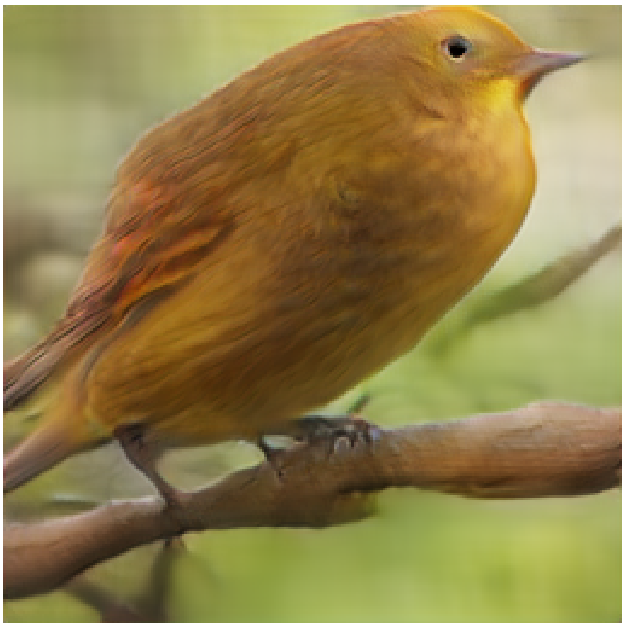

Figure 7. The best image produced by AttnGAN+lex for "a yellow bird with red wings and a round beak".

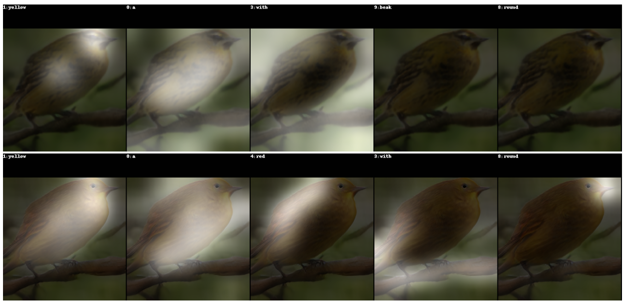

Figure 8. The attention heat maps for the generation of the image in Figure 7.
"yellow", "a", "with", "beak", "round"
"yellow", "a", "red", "with", "round"

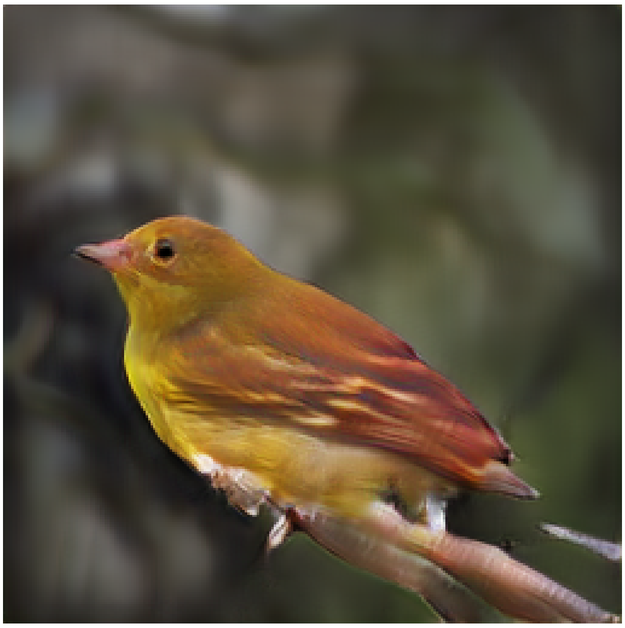

Figure 9. The best image produced by the original AttnGAN for "a yellow bird with red wings and a round beak".

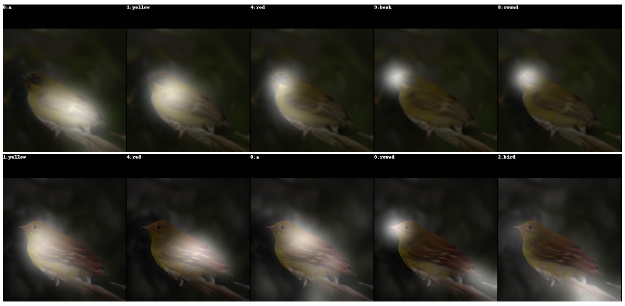

Figure 10. The attention heat maps for the generation of the image in Figure 9.
"a", "yellow", "red", "beak", "round"
"yellow", "red", "a", "round", "bird"

bird was elongated. We assume that this is due to the fact that the words "a bird" occur more often than "beak", and therefore the impact of the attention score of those words in the weighted average are higher, biasing the final attention masks to the wrong sub-region.

As shown in Figure 11, these shortcomings do not occur all the time. Instead, in this case, the colors have been matched nicely to the image sub-regions, with attention ranging from larger areas to smaller details corresponding to the text input.

6. Conclusions

As the inception score shows, our lexical model does indeed improve the overall image quality of the AttnGAN. We were also able to improve the accuracy of the attention when changing the DAMSM to LASM.

The experiments have also shown, where our model needs further improvement. Especially for finer details such as beaks, our model scatters the attention to broadly, and the more common words, steal too much attention from the less common words.

For future work, it might be worthwhile to incorporate more advanced lexical or rather syntax models. I.e., using dependency parsing to determine to which noun an adjective belongs would enable the model to correctly identify and use only the context needed for the current word. Thereby, it should further improve the accuracy of the attention for attributive words, like "round" or "long" and the overall image quality.

References

Dash, Ayushman, Gamboa, John Cristian Borges, Ahmed, Sheraz, Liwicki, Marcus, and Afzal, Muhammad Zeshan. TAC-GAN - text conditioned auxiliary classifier generative adversarial network. *CoRR*, abs/1703.06412, 2017. URL http://arxiv.org/abs/1703.06412.

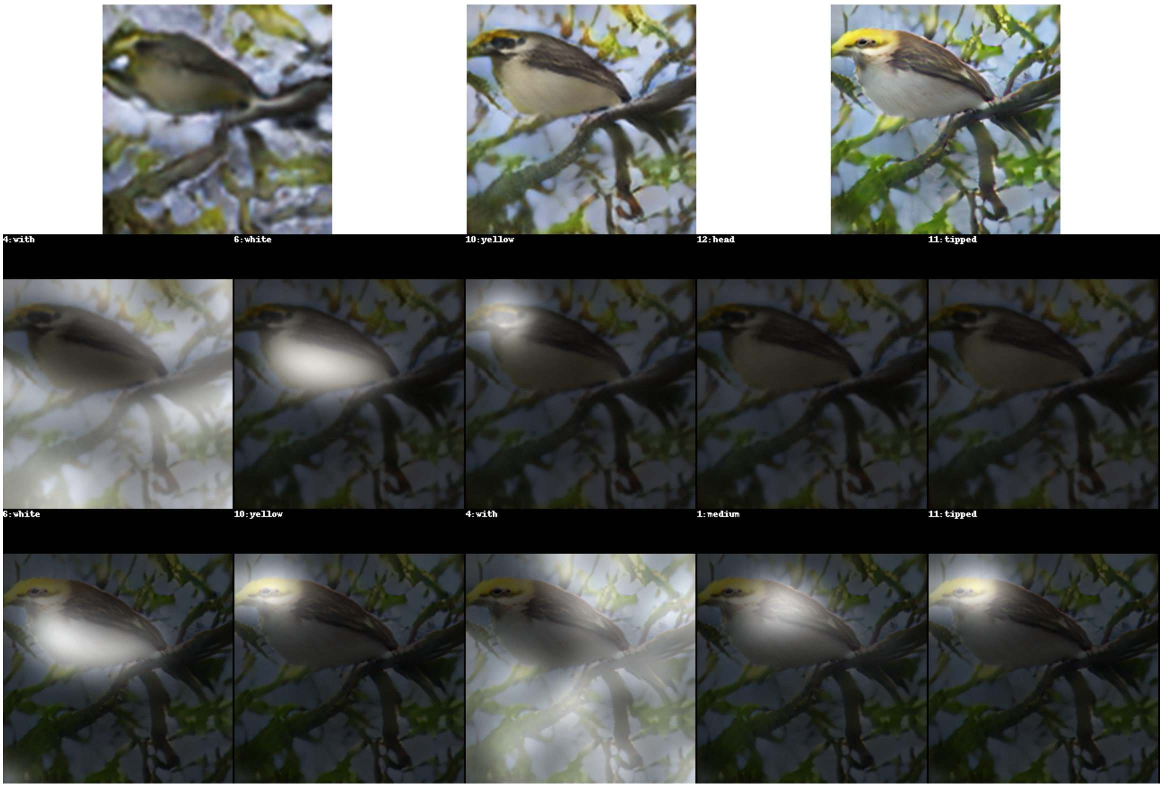

Figure 11. The attention heat maps for the generation of an image by AttnGAN+lex. The top row is the image output in the 3 stages of the model scaled to the sames size. The bottom two rows are the most attended heatmaps in the last 2 stages.
"with", "white", "yellow", "head", "tipped"
"white", "yellow", "with", "medium", "tipped"

Gorti, Satya Krishna and Ma, Jeremy. Text-to-image-to-text translation using cycle consistent adversarial networks. *CoRR*, abs/1808.04538, 2018. URL http://arxiv.org/abs/1808.04538.

Im, Daniel Jiwoong, Kim, Chris Dongjoo, Jiang, Hui, and Memisevic, Roland. Generating images with recurrent adversarial networks. *CoRR*, abs/1602.05110, 2016. URL http://arxiv.org/abs/1602.05110.

Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. dec 2014. URL http://arxiv.org/abs/1412.6980.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Nguyen, Toan Q. and Chiang, David. Improving lexical choice in neural machine translation. *CoRR*, abs/1710.01329, 2017. URL http://arxiv.org/abs/1710.01329.

Salimans, Tim, Goodfellow, Ian J., Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. URL http://arxiv.org/abs/1606.03498.

Springenberg, Jost Tobias. Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*, 2015.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jonathon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Polosukhin, Illia. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xu, Tao, Zhang, Pengchuan, Huang, Qiuyuan, Zhang, Han, Gan, Zhe, Huang, Xiaolei, and He, Xiaodong. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017. URL http://arxiv.org/abs/1711.10485.

Zhang, Han, Xu, Tao, Li, Hongsheng, Zhang, Shaoting, Huang, Xiaolei, Wang, Xiaogang, and Metaxas, Dimitris N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016. URL http://arxiv.org/abs/1612.03242.