STYLIZED IMAGE CAPTIONING USING CONTROLLABLE TEXT GENERATION

Pranjal Gupta
pgupta332
Georgia Institute of Technology
Atlanta, Georgia
{pgupta332}@gatech.edu

Xueqing Li
xli958, 903639167
Georgia Institute of Technology
Atlanta, Georgia
{xli958}@gatech.edu

ABSTRACT

Image Captioning is the process of generating a textual description for a given image. It is a fundamental and important task in computer vision and natural language processing domains. However, most of the existing captions are factual descriptions of the image, without human emotions. This makes them a bit dull, since humans often rely on emotions to convey meaning. In this paper, we want to use deep learning techniques to generate stylized captions, which can be used in creative writing, plot generation, and other related fields. Previous methods that perform stylized image captioning have a specialized style module that also considers the image content. Due to this, they require paired factual and style descriptions, for a given image. To overcome these limitations, we split the Image captioning process into two parts: 1)Factual Image captioning and 2) A text-style transfer using VAEs and controlled text generation. We consider only romantic style captioning, but this work can easily be extended to generate other styles.

1 INTRODUCTION

Image captioning is the process of generating the textual description of an image. The generated output is expected to describe the contents of the image: the objects present, their properties, the actions being performed, the interaction between the objects, etc. This job seems very easy for humans: a brief glance is sufficient for the human to understand and describe what is happening in the picture. However, it is relatively harder for a machine to capture these object characteristics and inter-dependencies. Recently, many deep learning methods have been used to generate very realistic image captioning.

However, currently, the methods only focus on generating the factual description of an image, ignoring human emotion. This makes them relatively uninteresting, since humans often use emotions in their descriptions of an image, to convey meaning. For example, the factual captioning for Figure 1 is: "A skier jumps high in the air with a view of the mountains." However, a human might say: "A skiing man in a fluorescent jacket jumps very high, loving the thrill of the jump." or "A skier jumps high in the air above the mountains to show his courage." people may express their feelings, guesses, opinions or even the weather. These artistic captions could then be applied in applications like story telling, plot generation etc.

Previous methods (eg. StyleNET) that performed stylized image captioning, normally have an encoder-decoder architecture. They use specialized style modules that look at image content to generate captions of a specific style. Due to the encoder-decoder architecture, they are able to generate only a single caption per style. In addition, since each style module requires image features, they rely on paired factual and stylized caption data for training.

To overcome these limitations, we split the task of stylized image captioning into two parts: 1) Factual Image captioning and 2) A text-style transfer using VAEs and controlled text generation. The factual image captioning is generated by using the method as described in Xu et al. (2015). Once, we have the factual description, we then use the method as described in Hu et al. (2017) to perform a text style-transfer.

Figure 1: A skier jumps high in the air with a view of the mountains.

The VAE architecture allows us to produce multiple stylized captions, from a given factual caption. This is useful, since multiple versions of a stylized caption exist. For example, consider the factual sentence ”A man is standing on a beach by the ocean”. One stylized version could be ”A man is standing on the beach enjoying the beauty of nature”, while a second one could be ”A man stands on a beach dreaming of love”. This text generation model does not need any image features. Due to this, we are able to perform unpaired style transfer, i.e. for an image, we do not need corresponding factual and stylized versions of the caption. We only focus on generation of romantic styles, but this work can easily be extending to generate other styles like humorous, satirical, etc.

2 MOTIVATION

Humans often incorporate emotions in their descriptions. This allows them to convey meaning more effectively. It also makes them more interesting to other people, when compared to a dull factual description. Thus, we believe that generating a stylized caption makes the caption more realistic and interesting, as compared to factual descriptions which are robotic and uninteresting.

When given an image to describe, different people may present various descriptions, for the same image. This is based on their personal biases, guesses, opinions etc. Thus, we want our model to have the ability to generate multiple stylized sentences, to capture this variability.

There is a scarcity of paired factual and stylized caption datasets, and an abundance of unpaired factual and style caption datasets. Previous methods rely extensively on the paired datasets for training their models. We plan to leverage the unpaired datasets for this same task, due to the increased availability of this type of data.

As humans, we are able to perform this generation of stylized captions by simply modifying the factual description of an image. We don’t even need to look at the actual image to do this task (try this with some of the factual descriptions and you will see that its very easy for us). This is why we believe that stylized captioning does not require the use of image features, and should be viewed as a controlled text generation problem where we simply changing the style of the factual description.

3 DATASET

For the factual Image Captioning task, we use *MS COCO*. In this paper, you train the model on part of the data: 30,000 captions for about 20,000 images. Then we split the datasets into train, validation and test sets. The training set has 18,000 captions and 12,000 images, validation has 4,500 captions and 3,000 images and rest is the test set.

For the text style-transfer task, we borrow the dataset used in Gan et al. (2017). The dataset consists of paired factual, romantic and humorous. We only use the factual and romantic datasets. There are

7000 romantic and 8091 factual captions. We combine them and then shuffle the resulting dataset. We then split the dataset to produce a train, validation and test set. Our training set is of size 14,000, validation is 550 and rest is the test set.

Figure 2 shows some samples in the dataset.

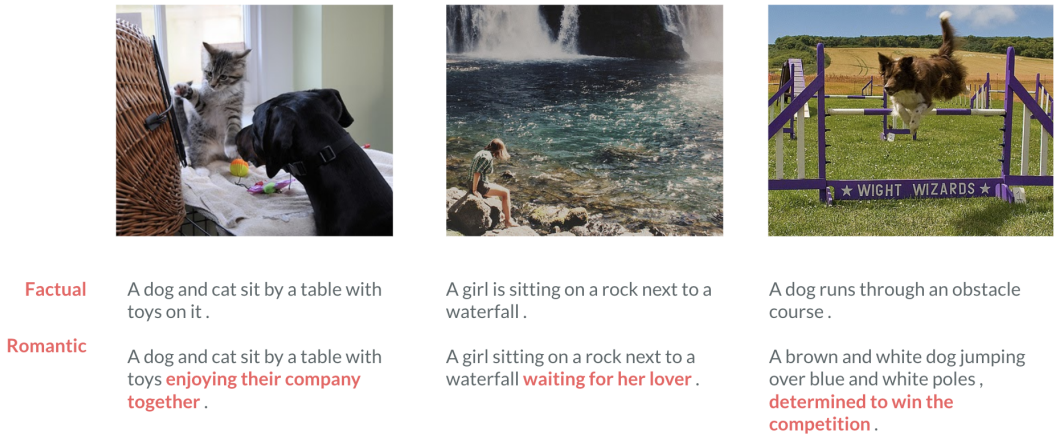


Figure 2: Dataset Samples

4 METHOD

4.1 MODEL ARCHITECTURE

We first use an Encoder-Decoder model to generate Factual Image Captioning. The encoder is a Convolutional Neural Network (CNN) network that extracts features, while the decoder is a RNN/LSTM with an Attention mechanism, which automatically learns to describe the content of images.

Then we turn Factual Image Captions into Romantic Captions using text controlled text generation with Variational Auto-Encoders(VAEs). Figure 3 is the overall architecture for the model.

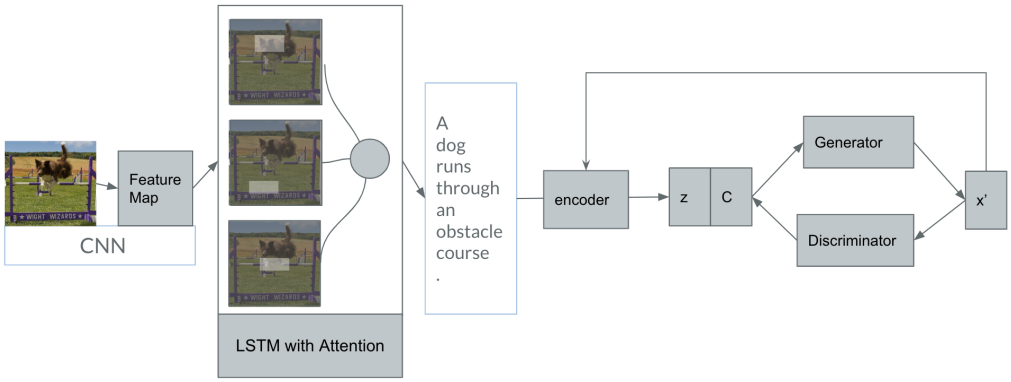


Figure 3: Model Architecture

4.2 FACTUAL IMAGE CAPTIONING

The Factual Image Captioning uses the model described in Xu et al. (2015). First, we fine-tune a pre-trained ResNet model to get the encoder. Then we train a decoder model based on LSTM and attention architecture. The input image first goes through the encoder and obtains a CNN Feature Map. The RNN with Attention attends over the image to predict the next word. This model is able to generate Factual Captions for given images.

4.3 STYLIZED IMAGE CAPTIONING

Next, we apply a text style-transfer to the factual caption obtained from the Image captioning model. To perform the stylized image captioning, we use the architecture as described in Hu et al. (2017). The model overview is shown in Figure 4. Please refer the paper for details as I will only provide a high level overview of the model.

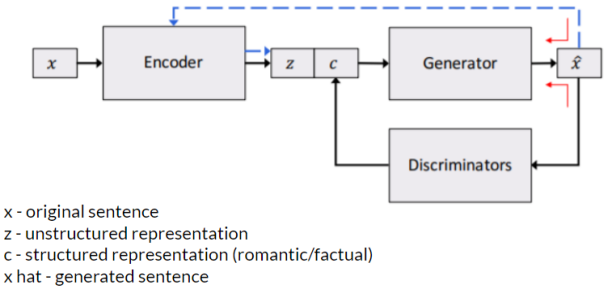


Figure 4: Controlled Generation Architecture

It consists of an Encoder, Generator and Discriminator. The encoder is trained to learn the unstructured representation of a sentence. The generator takes this unstructured representation and an input structured representation (in our case, a one-hot representation of romantic or factual) to generate sentences with the desired style; i.e. romantic or factual. To verify that the desired style is being produced by the generator, the discriminator enforces the generator to produce the correct style through feedback. To ensure that the generator only changes the desired style attribute, the encoder enforces the generator to produce sentences with the same unstructured representation as the original input sentence. The overall algorithm is presented in Figure 5

Algorithm 1 Controlled Generation of Text

Input: A large corpus of unlabeled sentences $\mathcal{X} = \{\boldsymbol{x}\}$
A few sentence attribute labels $\mathcal{X}_L = \{(\boldsymbol{x}_L, \boldsymbol{c}_L)\}$
Parameters: $\lambda_c, \lambda_z, \lambda_u, \beta$ – balancing parameters

1: Initialize the base VAE by minimizing Eq.(4) on $\mathcal{X}$ with $\boldsymbol{c}$ sampled from prior $p(\boldsymbol{c})$
2: **repeat**
3: Train the discriminator D by Eq.(11)
4: Train the generator G and the encoder E by Eq.(8) and minimizing Eq.(4), respectively.
5: **until** convergence

Output: Sentence generator G conditioned on disentangled representation $(\boldsymbol{z}, \boldsymbol{c})$

Figure 5: Controlled Generation Algorithm

5 RESULTS

5.1 HYPER-PARAMETERS

We set the hyper-parameters based on results from validation set. We select the latent vector dimension to be z_dim=100, conditioning vector c_dim=2. The encoder and generator are both LSTMs, with hidden size h_dim=300. We use Adam Optimizer with annealing learning rate, starting with lr=1e-3 and train for n_epochs=50. We use word dropout at the generator with p_word_dropout=0.3, to prevent overfitting.

All the balancing parameters ($\lambda_z, \lambda_c, \lambda_u$ and β) as 0.1 . We use GloVe word embeddings, with dimension emb_dim=300. We keep the embedding layer trainable during the initial unsupervised training of the base VAE and then freeze the embedding layer. We limit the sentences to 35 words.

5.2 FACTUAL IMAGE CAPTIONING

Figure 6 shows the loss of Factual Image Captioning training process.

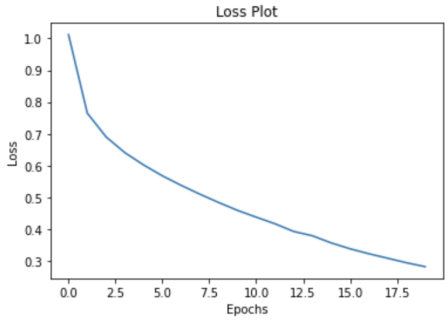


Figure 6: Loss Plot

Table 1 shows the factual image captioning results. In the first image, the factual image caption generated by the model is: "A woman in shallow water laying on their feet." For the second image, the factual image caption is: "A person skiing down a white mountain." These results are pretty good and agree with the content in the image. Table 2 shows the BLEU score for factual image captioning.

Table 1: Result

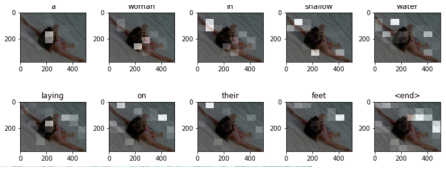
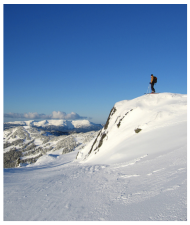
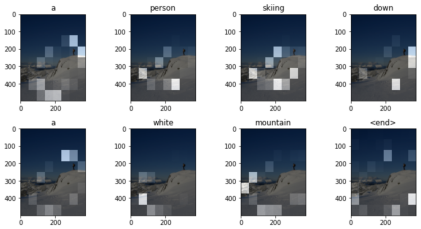
Image	**Factual Captions and Attention**
	
	

Table 2: BLEU Score

BLEU-1	BLEU-2	BLEU-3
36.84	19.23	12.40

5.3 STYLIZED IMAGE CAPTIONING

5.3.1 UNSUPERVISED MODEL(BASE VAE)

The learning curves obtained from training the base VAE, for KL loss(left) and reconstruction loss(right) are shown in Figure 7. From this, we can see that the reconstruction loss goes down, which means the model is actually learning. The KL loss is however a bit unusual, and we will talk about this in a later section. Since the base VAE is like a language model, we evaluate its text generation capabilities using perplexity, on the validation set. This result along with some sample generated sentences are shown in Figure 8.

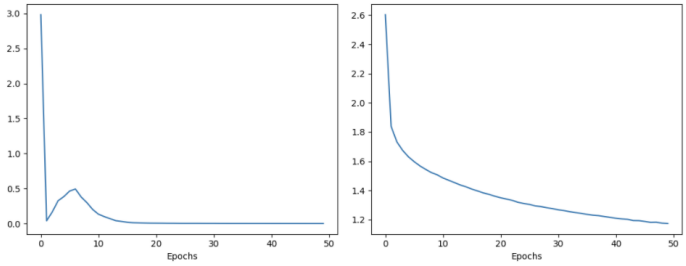


Figure 7: Unsupervised learning curves

Generated by the model:
"a girl stands on a window reading a big sign ."
"child dancing in a kitchen , looking for their lovers ."

Average perplexity on validation set: 8.3484

Figure 8: Unsupervised model perplexity and generated samples

5.3.2 SUPERVISED MODEL

The learning curves obtained from training the VAE with the discriminator using labelled data is shown in Figure 9. This plot contains the generator, encoder, and discriminator loss curves. It also contains the discriminator accuracy curve. From this figure, we can see that the discriminator is learning very fast and becomes very accurate. The generator and encoder however learn very slow. The encoder in particular does not learn well. Again, we will talk about this later.

We first test the ability of the model on the task of conditional text generation. That is, given a particular style, what percentage of samples generated are of the correct style. Since our trained discriminator is basically a very powerful classifier, we simply use this model to determine the style of the generated samples. We evaluate this over 500 generated samples. The result along with some generated samples are shown in Figure 10.

Next, we test the ability of the model on the task of controllable text generation. That is, given a sentence, generate sentences with the correct style, while retaining the original structure of the

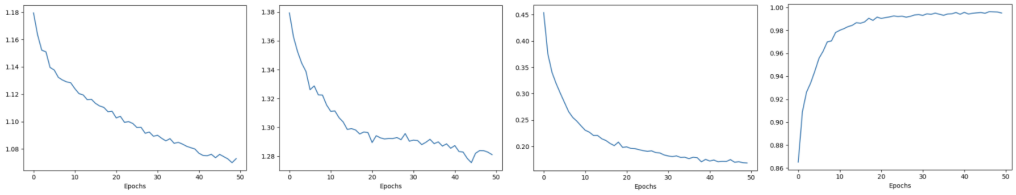


Figure 9: Supervised learning curves

Accuracy of conditional text generation: 0.7309

Romantic Samples generated by model:

*"a man and woman stand behind a tree , **enjoying their friendship** ."*
*"a dog runs along the snow in water **to meet his lover** ."*
*"a shirtless man is climbing a fence and is **enjoying the magic of being young** ."*

Factual Samples generated by model:

"a blonde girl in blue hanging costume on a nice sidewalk ."
"a child in a green shirt is laughing ."
"a dog jumps over a hurdle landing on it ."

Figure 10: Supervised model conditional text generation accuracy and generated samples

sentence. An example result we obtained is shown in Figure 11. Since, the other results we obtained are also very similar, we know that the model is not able to perform this task and hence, we do not evaluate it. Instead, we perform some further analysis and an error analysis to try to understand the underlying problem in the model.

Original sentence: an asian lady in a red jacket taking a photo, "factual"

Generated Instances
*Gen 1: two girls are kayaking in a river rapids **meeting destiny** ., "romantic"*
*Gen 2: two girls wearing florida marlins hats **proudly** between luggage in a forest ., "romantic"*
*Gen 3: a woman at a fair wearing a santa claus suit , **being silly** ., "romantic"*
Gen 4: two dogs tussle one dark object and a white uniform across the grass ., "romantic"
Gen 5: two men are standing in a red car with their backs to the camera ., "romantic"

Figure 11: Supervised model controllable text generation samples

6 Result Analysis

6.1 Latent Vector Analysis

First, we test the disentanglement between the structured and unstructured representations. To do this, we sample a random point in the latent space, and then change only the structure representation, i.e. we change the style while keeping the underlying sentence the same. From this, we can see that along with the style, other attributes are of the sentence are being changed, which means disentangled representations are not being learned.

We also test interpretability of the latent space, using latent space interpolation. That is, we select two random points in the latent space, and then 3 equally spaced points that lie on the line joining

the two points. We decode each of these points using the generator to obtain sentences. During this, we sample the style randomly.

The idea is that if the latent space is interpretable, the sentences would exhibit some sort of structure. For example, the first interpolated point when decoded should produce a sentence that has more in common with the starting point than the ending point, when they are decoded. The results of this latent space interpolation are also shown in Figure 12. We can clearly see that our latent space is not interpretable. This is probably one of the reasons why our controllable text generation does not work.

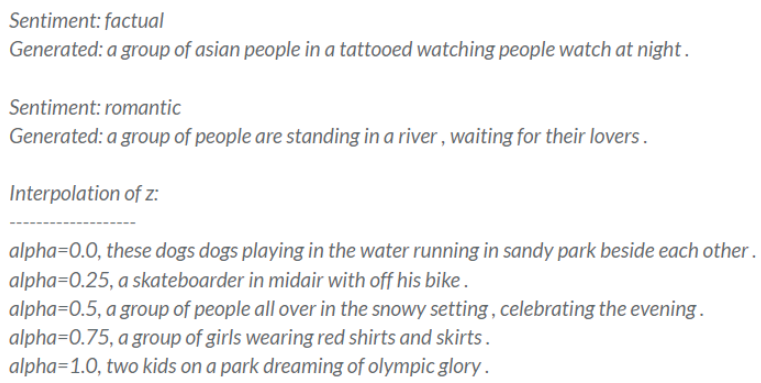


Figure 12: Latent space interpolation result

6.2 Error Analysis

The KL divergence loss curve (Figure 7) is unusual, since the posterior encoder distribution is too close to the prior. Normally, we would expect an increase in the KL divergence followed by saturation, to indicate maximum divergence from prior has been attained. But since the KL divergence is basically 0, this indicates the posterior distribution is basically random, just like the prior. This is highly undesirable.

From Figure 9, we notice that the encoder learns extremely slowly, compared to the generator or discriminator. From the previous sections, we showed that disentangled representations were not being learned, and that the latent space was not being learned effectively. All of this evidence points to poor training of the encoder. We could try to remedy this by giving a higher loss coefficient to the encoder.

Another thing we should keep in mind is that longer sentences are harder to capture (the original authors of controlled text generation used sentences of size $\leq$ 15 words). We could perhaps get better results by doing this as well.

7 Conclusions and Future Work

Overall, our model is able to generate realistic sentences, i.e. we are able to solve the text generation problem. Thus, we can use the VAE model to create augmented data for other stylized image captioning problems. To a degree, our model is also able to solve the conditional text generation problem. So we can generated augmented data of a particular style. Finally, the discriminator we train is very powerful, and achieves a nearly perfect accuracy with the efficient semi-supervised training approach. We can reuse it as a style caption classifier.

However, the VAE framework we used is unable to generate sentences based on an existing sentence, while preserving the original sentence structure. In other words, it is unable to perform the controllable text generation task. In particular, the encoder seems to be learning very poorly with this approach. In addition, the latent space is not interpretable and disentangled representations are not being learned through this approach.

Instead of using LSTMs, we could try using transformers, as described in (Vaswani et al. (2017)) since they are a much more powerful model. This would also help capture unstructured representations (particularly for longer sentences) better, due to the attention mechanism. Instead of using VAEs for text generation, we could try using GANs (Goodfellow et al. (2014)), which are more powerful generation models. More recent papers prove that disentanglement is unnecessary. We could use back-translation, as described in Prabhumoye et al. (2018). Finally, we could use relational memory recurrent networks(LMRN) as described in Santoro et al. (2018). This will allow us to perform attention during the generation process as well.

REFERENCES

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3137–3146, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pp. 1587–1596. PMLR, 2017.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. *arXiv preprint arXiv:1806.01822*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.