GENERATING IMAGES FROM CAPTIONS WITH ATTENTION

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba & Ruslan Salakhutdinov
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
{emansim,eparisotto,rsalakhu}@cs.toronto.edu, jimmy@psi.utoronto.ca

ABSTRACT

At the end.

1 INTRODUCTION

Statistical natural image modelling remains a fundamental problem in computer vision and image understanding. This has motivated recent approaches in generative modelling applied to natural images by employing deep neural networks for their inference and generative components. Image generative models studied previously are often restricted to learning unconditional models of images distribution or conditioned on simple structured annotations, for example classification labels. Despite the advances in generative models, learning the highly structured natural image distribution alone in the high dimensional pixel space alone proves to be a difficult task. In real world, however, images rarely appear in isolation. They often accompanied by their unstructured textual descriptions on web pages and in books. One domain presents substantial amount of relating information of the other domain. The additional information from image and unstructured text description could be used to simplify the image modelling task.

There are two directions in learning a generative model of image and text. One approach is to learn a text generative model conditioned on the images. Significant amount of recent works have been focusing on generating captions from images(Karpathy & Li, 2014), (Xu et al., 2015), (Kiros et al., 2014) and etc. The models take an image descriptor and generate unstructured texts through a recurrent decoder. By contrast, learning a generative model for image and text may also be studied by generating images correctly interpreting the text description. Generating high dimensional realistic images from their descriptions is a more difficult approach that combines two challenging components of language modelling and image generation. Namely, the model has to capture the semantic meaning expressed in the description and then use that knowledge to generate pixel intensities of the image. Although, the interesting high dimensional natural images lay on a small manifold that is difficult to capture, the additional text description cues of a target image may simplify the learning problem by focusing on the conditional distribution.

In this paper, we illustrate how simple sequential deep learning techniques can be used to build a conditional probabilistic model over natural image space effectively. By using a sequence to sequence framework to approach the problem of image generation from unstructured natural language captions, our model iteratively draws the patches on canvas, while attending to the relevant words in the description.

2 RELATED WORK

Deep Neural Networks have achieved a remarkable performance in various tasks such as image recognition (Krizhevsky et al., 2012), speech transcription (Graves et al., 2013) and etc. While most of the recent success has been achieved by discriminative models, the generative models have not yet enjoyed the same level of success. Most of the previous work in generative models has been focused on variants of Boltzmann Machines (Smolensky, 1986), (Salakhutdinov & Hinton, 2009) and Deep Belief Networks (Hinton et al., 2006). While these models are very powerful, each iteration of training requires a computationally costly step of MCMC to approximate an intractable normalization constant that makes it hard to scale them to large datasets.

Kingma & Welling (2013) have introduced the Variational Auto-Encoder (VAE) which can be seen as a neural network with continous latent variables. The encoder is used to approximate posterior distribution and the decoder is used to stochastically reconstruct the data from latent variables. The model performs an efficient inference and learning that allows it to scale to large datasets. Gregor et al. (2015) have introduced Deep Recurrent Attention Writer (DRAW), where they have incorporated a novel differentiable attention mechanism into the VAE which significantly improved its performance as well as quality of generated samples. While most of the samples of VAE and DRAW resemble a clear structure of objects, the generated images are blurry most of the time.

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are another type of generative models that use noise-contrastive estimation (Gutmann & Hyvärinen, 2010) to avoid calculating intractable normalization constant. The model consists of generator that generates samples from uniform distribution and discriminator that discriminates between real and generated images. Both networks are playing the game, where generator tries to produce samples that look real and discriminator tries not to be fooled by generator. Recently, Denton et al. (2015) have scaled those models, by training GANs at each level of Laplacian pyramid of images. While their model has generated sharp looking samples, the generated images have lacked a clear structure. Compared to other mentioned generative models, GANs are more unstable and are harder to train.

While all of the previous work has been focused on unconditional models or models conditioned on labels, to the best of our knowledge this paper is the first to introduce the generative model of images conditioned on captions.

3 Model

Our proposed model can be seen as a part of sequence-to-sequence framework (Sutskever et al., 2014), (Cho et al., 2014), (Srivastava et al., 2015) where captions are represented as a sequence of consecutive words and images are represented as a sequence of patches drawn on canvas over time $t = 1, ..., T$. Let y be the input caption, consisting of N words $y_1, y_2, ..., y_n$ and x be the image corresponding to that caption.

3.1 Encoder

The encoder is a deterministic Bidirectional LSTM that encodes the variable size sentences into the vector representation s. Bidirectional LSTM consists of one Forward LSTM and Backward LSTM which combine information from past and future respectively. The Forward LSTM computes the sequence of forward hidden states $[\overrightarrow{h}_1^{lang}, \overrightarrow{h}_2^{lang}, ..., \overrightarrow{h}_N^{lang}]$, whereas the Backward LSTM computes the sequence of backward hidden states $[\overleftarrow{h}_1^{lang}, \overleftarrow{h}_2^{lang}, ..., \overleftarrow{h}_N^{lang}]$. Then these hidden states are concatenated together into the sequence $[h_1^{lang}, h_2^{lang}, ..., h_N^{lang}]$, where $h_n^{lang} = [\overrightarrow{h}_n^{lang}, \overleftarrow{h}_n^{lang}], 1 \leq n \leq N$.

The final representation of the sentence is calculated as follows:

$$s_t = \alpha_1 h_1^{lang} + \alpha_2 h_2^{lang} + ... + \alpha_N h_N^{lang} \tag{1}$$

where h_0^{lang} is initialized to the learned bias. Setting $\alpha_{1...N}$ to $\frac{1}{N}$ turns the encoder into the vanilla model introduced in (Cho et al., 2014) without the attention. We will describe how $\alpha_{1...N}$ are calculated in the next section.

3.2 Decoder

The decoder is a DRAW model, which is a stochastic recurrent neural network that consists of Inference LSTM that infers the distribution of latent variables of image x given y and then the Generator LSTM that uses the inferred latent variables in order to reconstruct the image x given y.

Formally, the decoder iteratively computes the following equations for $t = 1, ..., T$

$$\hat{x}_t = x - \boldsymbol{\sigma}(c_{t-1}) \tag{2}$$
$$r_t = read(x_t, \hat{x}_t, h^{dec}_{t-1}) \tag{3}$$
$$h^{enc}_t = LSTM^{enc}(h^{enc}_{t-1}, [r_t, h^{dec}_{t-1}]) \tag{4}$$
$$z_t \sim Q(Z_t | h^{enc}_t) \tag{5}$$
$$h^{dec}_t = LSTM^{dec}(h^{dec}_{t-1}, z_t, s_{t-1}) \tag{6}$$
$$s_t = align(h^{dec}_{t-1}, \boldsymbol{h^{lang}}) \tag{7}$$
$$c_t = c_{t-1} + write(h^{dec}_t) \tag{8}$$

where $read$ and $write$ are the same attention operators as in (Gregor et al., 2015), $\boldsymbol{h^{lang}} = [h^{lang}_1, h^{lang}_2, ..., h^{lang}_n]$ and $c_0, h^{dec}_0, h^{enc}_0$ are initialized to the learned biases.

The $align$ function introduced by Bahdanau et al. (2015) is used to compute probabilities $\alpha_{1...n}$ by first computing the energies $e_{t,0}, e_{t,1}, ..., e_{t,n}$ and then rescaling them to the probability distribution $\alpha_1, \alpha_2, ..., \alpha_n$

$$e_{tj} = v^T tanh(U h^{lang}_j + W h^{dec}_t + b) \tag{9}$$
$$\alpha_j = softmax(e_{tj}) \tag{10}$$

where $softmax(e_{tj}) = \frac{exp(e_{tj})}{\sum_{j=1}^{N} exp(e_{tj})}$

3.3 Learning

The model is learned by the modified version of Stochastic Gradient Variation Bayes (SGVB) algorithm introduced by Kingma & Welling (2013). The model is trained to maximize the lower bound of marginal likelihood $\mathcal{L}$ of the correct image x given the input caption y. The $\mathcal{L}$ is decomposed into the latent loss $\mathcal{L}^z$ and the reconstruction loss $\mathcal{L}^x$.

The reconstruction loss $\mathcal{L}^x$ equals to $\frac{1}{L}\sum_{l=1}^{L}(log\, p(x_t|y, z)$ where L is the number of samples used during training, which was set to 1 in our experiments.

The latent loss is a negative sum of Kullback–Leibler divergence terms between distribution $Q(Z_t|h^{enc}_t)$ and some prior distribution $P(Z_t)$ over time $t = 1, ..., T$, which can be seen as a regularization term. Since the patches drawn on canvas over time are not independent of each other, naturally the sufficient statistics of the prior distribution at time t should be dependent on the sufficient statistics of the prior distribution at time $t-1$. Therefore, instead of setting $P(Z_1), ..., P(Z_T)$ to be independent unit gaussian distributions, the mean and variance of $P(Z_t)$ depends on the h^{dec}_{t-1}, as in (Bachman & Precup, 2015), where

$$\mu^{prior}_t = tanh(W_{mu} h^{dec}_{t-1}) \tag{11}$$
$$\sigma^{prior}_t = exp(tanh(W_\sigma h^{dec}_{t-1})) \tag{12}$$

Overall, the likelihood $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = -\sum_{t=1}^{T} D_{KL}(Q(Z_t|h^{enc}_t, s_{t-1}) \,||\, P(Z_t)) + \frac{1}{L}\sum_{l=1}^{L} log\, p(x_t|y, z) \tag{13}$$

$$= \frac{1}{2}\sum_{t=1}^{T}(1 - 2\, log\, \sigma^{prior}_t + 2\, log\, \sigma_t - \frac{exp(2\, log\, \sigma_t) + (\mu_t - \mu^{prior}_t)^2}{exp(2\, log\, \sigma^{prior}_t)}) + \frac{1}{L}\sum_{l=1}^{L} log\, p(x_t|y, z) \tag{14}$$

3.4 Image Generation

During the image generation step, we throw away the inference network from the decoder and instead sample from the prior distribution. Due to the blurriness of samples generated by DRAW

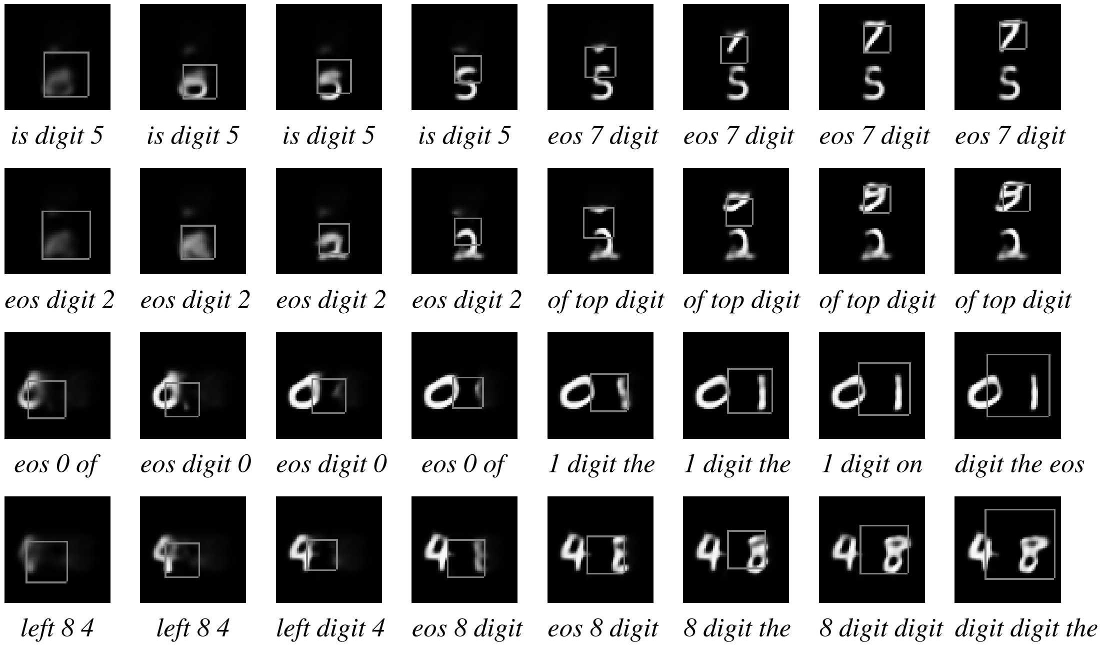


Figure 1: Four examples show the generated images unfolded over several timesteps as well as the top three words the model attends to while generating images.

model, we do an additional post processing step, where we use the generator of the adversarial network trained on residuals of laplacian pyramid to sharpen the generated images, similar to (Denton et al., 2015). By fixing the prior of adversarial generator to the mean of uniform distribution, it gets treated as a deterministic neural network which allows us to calculate the lower bound of likelihood. The reconstuction loss becomes the loss between sharpened image and correct image, whereas the latent loss stays the same. We also noticed that sampling from the mean of uniform distribution allowed us to generate much less noisy samples than by sampling from the uniform distribution itself.

4 Experiments

4.1 MNIST With Captions

As a toy experiment, we have trained our model on the two digit MNIST dataset with captions constructed on the fly. Two random digits were either placed horizontally or vertically, so that they don't overlap, on the black 60×60 background. The caption indicated the way digits were placed in the image and had the following template: *the digit (1) is (2) the (3) of the digit (4) <eos>*. (1) and (4) were the digit numbers, (2) was either on or at, (3) was one of the adjectives top, bottom, left or right. For example, the generated images corresponding to the caption *the digit seven is at the bottom of the digit five <eos>* as well as other captions are shown on **??**. While most of the generative models were trained on the version of binarized MNIST dataset (Salakhutdinov & Murray, 2008), our model was trained directly on pixel intensities with binary cross entropy cost function.

4.2 COCO

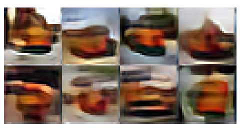
Microsoft COCO (Lin et al., 2014) is a largest dataset of images annotated with captions consisting of 83k images. The rich collection of images with variety of styles, backgrounds and objects makes the task of learning a good generative model conditioned on caption very challenging. Since some of the images have more than five captions attached to them, for consistency with related work on caption generation we disregard extra captions.

In the following subsections we make both qualitative and quantitative analysis of our model as well as compare its performance with other related generative models.

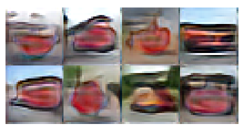
4.2.1 Analysis of Generated Images

The main goal of this work is to learn a model that can understand the semantic meaning expressed in description of the image, such as the properties of objects, the relationships between them, etc. and then use that knowledge to generate relevant images. To verify that, we generated a set of captions inspired by COCO dataset and changed some words in the captions to see whether the model made the relevant changed in the generated samples.

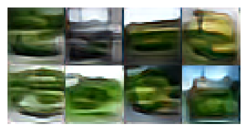
First, we wanted to see whether the model understood one of the simplest properties of any object, the color. We generated images of school buses with four different colors: yellow, red, green and blue. Although, there are images of buses with different colors in the training set, all school buses specifically are colored yellow. As you can see below, the model managed to generate images with an object that looked like a bus with the correct color.



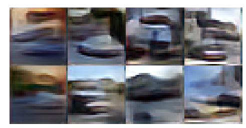
A yellow school bus parked in a parking lot.



A red school bus parked in a parking lot.

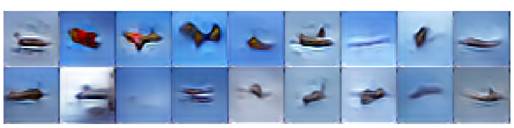


A green school bus parked in a parking lot.

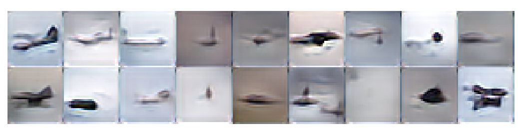


A blue school bus parked in a parking lot.

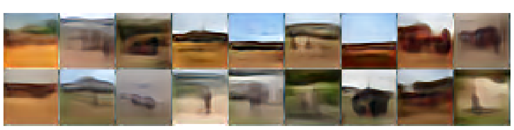
Apart from changing the colors of objects, we were curious whether changing the background of the scene described in the caption would result in the appropriate changes in the generated samples. It is a somewhat harder task for a model than changing the color, due to larger number of changes that have to be made in the generated samples. Nevertheless as shown below, changing the grass type from dry to green as well as adding clouds in the caption has resulted in the appropriate changes.



A very large commercial plane flying in the blue skies.



A very large commercial plane flying in the cloudy skies.



A heard of elephants walking across a dry grass field.

A heard of elephants walking across a green grass field.

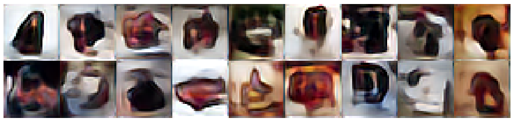
Despite an infinite number of ways of changing colors and backgrounds in descriptions, in general we have found that the model made appropriate changes as long as some similar patter was present in the training set. However, it wasn't always the case when changing an object itself in the description. In cases, when objects did't have a noticeable differences in their properties, such as shape or color, the changes in the generated samples weren't very clearly seen. This could be attributed to the limitation of the model to grasp the detailed understanding of each object.

4.2.2 Analysis of Attention

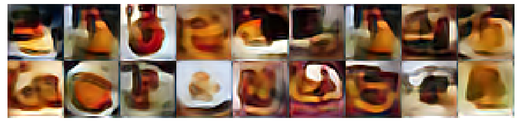
To be added soon.

4.2.3 Comparison With Other Models

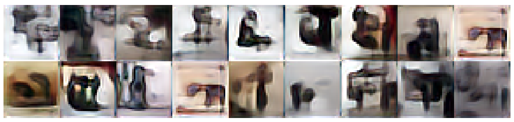
The quantitative evaluation of generative models has been a subject of ambiguity within a machine learning community. Compared to reporting classification accuracies, the measures defining generative models are intractable most of the times and might not correctly define the real quality of



The decadent chocolate dessert is on the table.



The bowl of banas is on the table.



A vintage photo of a cat.

A vintage photo of a dog.

Figure 34: Four different models displaying results from sampling caption *A group of people walk on a beach with surf boards.*

the model. To avoid ambiguity, we report results on two different metrics as well as do qualitative comparison of generative models.

As shown on Figure blah, we have sampled the images from the same caption in four of the best models.

First, to compare the the peformance of our model with other generative models trained with variational objective, we report the Recall@K, namely the mean number of sentences for which the correct image is ranked within the top-K retrieved results, based on the lower bound of likelihood $\mathcal{L}$. If you want to get a really good results build a model like Ryan did. Do deal away with easy images we take ratio with ... with mean caption. To deal with large computational complexity of looping over each image we only consider closest neighbors. The precision Recall curve is reported there...

Second,

4.3 CIFAR

REFERENCES

Bachman, Philip and Precup, Doina. Data generation as sequential decision making. *CoRR*, abs/1506.03504, 2015. URL `http://arxiv.org/abs/1506.03504`.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Cho, Kyunghyun, van Merrienboer, Bart, Gülçehre, Çaglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1724–1734, 2014.

Denton, Emily L., Chintala, Soumith, Szlam, Arthur, and Fergus, Robert. Deep generative image models using a laplacian pyramid of adversarial networks. *CoRR*, abs/1506.05751, 2015. URL `http://arxiv.org/abs/1506.05751`.

Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua. Generative adversarial nets. In *Advances in*

Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623, 2015.

Gutmann, Michael and Hyvärinen, Aapo. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pp. 297–304, 2010. URL http://www.jmlr.org/proceedings/papers/v9/gutmann10a.html.

Hinton, Geoffrey E., Osindero, Simon, and Teh, Yee Whye. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. URL http://arxiv.org/abs/1412.2306.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://arxiv.org/abs/1312.6114.

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard S. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL http://arxiv.org/abs/1411.2539.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pp. 1106–1114, 2012.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Salakhutdinov, Ruslan and Hinton, Geoffrey E. Deep boltzmann machines. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pp. 448–455, 2009. URL http://www.jmlr.org/proceedings/papers/v5/salakhutdinov09a.html.

Salakhutdinov, Ruslan and Murray, Iain. On the quantitative analysis of deep belief networks. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pp. 872–879, 2008. doi: 10.1145/1390156.1390266. URL http://doi.acm.org/10.1145/1390156.1390266.

Smolensky, Paul. Information processing in dynamical systems: foundations of harmony theory. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1986.

Srivastava, Nitish, Mansimov, Elman, and Salakhutdinov, Ruslan. Unsupervised learning of video representations using LSTMs. *ICML*, 2015.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. 2014.

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron C., Salakhutdinov, Ruslan, Zemel, Richard S., and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2048–2057, 2015. URL http://jmlr.org/proceedings/papers/v37/xuc15.html.