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 has been focused 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. Overall, the main contributions of this work are the following:

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

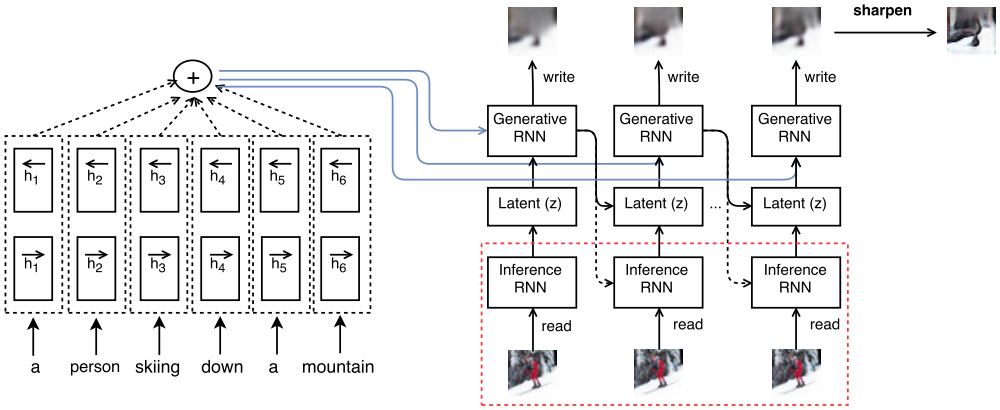


Figure 1: The image of the proposed model

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 viewed 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 $\mathbf{y}$ be the input caption, consisting of N words $y_1, y_2, ..., y_n$ and $\mathbf{x}$ be the image corresponding to that caption.

3.1 Language Model: the Bidirectional Attention RNN

The input caption sentences are fed into 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 $h_n^{lang} = [\overrightarrow{h}_n^{lang}, \overleftarrow{h}_n^{lang}], 1 \leq n \leq N$.

3.2 IMAGE MODEL: THE CONDITIONAL DRAW NETWORK

The DRAW network Gregor et al. (2015) is a sequential probabilistic model generating images by accumulating the output at each iterative step. While the original DRAW network assumes the latent variables are independent, it has shown in (Bachman & Precup, 2015) the model performance is improved by including the dependencies of latent variables.

We extended the architecture the DRAW network generative process to include additional input caption from the language model described in Sec. (3.1). Similarly to the original DRAW network, the conditional DRAW network 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 Generative LSTM that uses the inferred latent variables in order to reconstruct the image x given y. The $align$ function is used to compute the alignment between the input caption and intermediate image generative steps as in Bahdanau et al. (2015):

Formally, the image is generated by iteratively computing the following equations for $t = 1, ..., T$

$$\hat{x}_t = x - \boldsymbol{\sigma}(c_{t-1}) \tag{1}$$

$$r_t = read(x_t, \hat{x}_t, h_{t-1}^{dec}) \tag{2}$$

$$h_t^{enc} = LSTM^{enc}(h_{t-1}^{enc}, [r_t, h_{t-1}^{dec}]) \tag{3}$$

$$z_t \sim Q(Z_t | h_t^{enc}) \tag{4}$$

$$h_t^{dec} = LSTM^{dec}(h_{t-1}^{dec}, z_t, s_{t-1}) \tag{5}$$

$$s_t = align(h_{t-1}^{dec}, \boldsymbol{h^{lang}}) \tag{6}$$

$$c_t = c_{t-1} + write(h_t^{dec}) \tag{7}$$

where $read$ and $write$ are the same attention operators as in (Gregor et al., 2015). Given the caption representation from the language model, $\boldsymbol{h^{lang}} = [h_1^{lang}, h_2^{lang}, ..., h_N^{lang}]$, the $align$ operator computes the final sentence representation s_t through a weighted sum using alignment probabilities $\alpha_{1...N}$:

$$s_t = align(h_{t-1}^{dec}, \boldsymbol{h^{lang}}) = \alpha_1 h_1^{lang} + \alpha_2 h_2^{lang} + ... + \alpha_N h_N^{lang}. \tag{8}$$

The corresponding alignment probabilities $\alpha_{1...n}$ at each step are obtained by:

$$e_{tj} = v^T tanh(U h_j^{lang} + W h_t^{dec} + b) \tag{9}$$

$$\alpha_j = \frac{exp(e_{tj})}{\sum_{j=1}^{N} exp(e_{tj})}. \tag{10}$$

Here 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.

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 $\mathcal{L}^z$ is a negative sum of Kullback–Leibler divergence terms between distribution $Q(Z_t|h_t^{enc})$ 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_{t-1}^{dec}, which forms a Markov chain $P(Z_1), P(Z_2 \mid Z_1), ..., P(Z_T \mid Z_{T-1})$ as in (Bachman & Precup, 2015), where

$$\mu_t^{prior} = tanh(W_\mu h_{t-1}^{dec}) \tag{11}$$

$$\sigma_t^{prior} = exp(tanh(W_\sigma h_{t-1}^{dec})) \tag{12}$$

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

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{Q(Z_{1:t} \mid \mathbf{y}, \mathbf{x})} \left[-\log p(\mathbf{x} \mid \mathbf{y}, Z_{1:T}) + \sum_{t=2}^{T} \mathrm{D}_{\mathrm{KL}} \left(Q(Z_t \mid Z_{t-1}, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_t \mid Z_{t-1}, \mathbf{y}) \right) \right] \\ &+ \mathrm{D}_{\mathrm{KL}} \left(Q(Z_1 \mid \mathbf{y}, \mathbf{x}) \,\|\, P(Z_1 \mid \mathbf{y}) \right). \end{aligned} \tag{13}$$

The expectation can be approximate by L Monte Carlo samples $\tilde{Z}_{1:T}$ from $Q(Z_{1:T} \mid \mathbf{y}, \mathbf{x})$:

$$\begin{aligned}\mathcal{L} \approx& \frac{1}{L} \sum_{l=1}^{L} \left[-\log p(\mathbf{x} \mid \mathbf{y}, \tilde{Z}_{1:T}^l) + \sum_{t=2}^{T} \mathrm{D}_{\mathrm{KL}} \left(Q(Z_t \mid \tilde{Z}_{t-1}^l, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_t \mid \tilde{Z}_{t-1}^l, \mathbf{y}) \right) \right] \\ &+ \mathrm{D}_{\mathrm{KL}} \left(Q(Z_1 \mid \mathbf{y}, \mathbf{x}) \,\|\, P(Z_1 \mid \mathbf{y}) \right). \end{aligned} \tag{14}$$

3.4 Generating Images from Captions

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 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 COCO

Microsoft COCO (Lin et al., 2014) is the largest dataset of images annotated with captions consisting of roughly 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 the performances of other related generative models.

4.1.1 Analysis of Generated Images

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

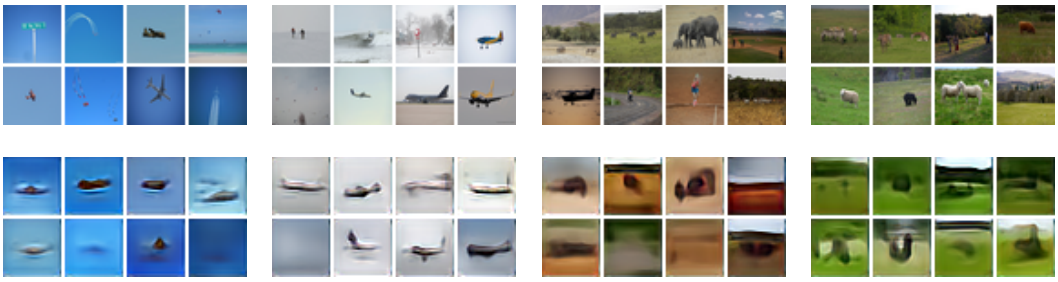
First, we wanted to see whether the model understood one of the most basic properties of any object, the color. As shown in Figure 2, we generated images of school buses with four different colors: yellow, red, green and blue. Although, there are images of buses with different colors in the training

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.

Figure 2: Examples of changing the color while keeping the caption fixed.

set, all school buses specifically are colored yellow. Despite that, the model managed to generate images of an object that is reminiscent of the bus pained with the relevant color.

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. Changing the background in images is a somewhat harder task for a model that changing the color, due to the larger visual area in images that is taken by the background. Nevertheless, as shown in Figure 3 changing the skies from blue to rainy as well as changing the grass type from dry to green resulted in the appropriate changes. The nearest images from the training set also indicate that the model was not simply copying the patterns it observed during the learning phase.



A very large commercial plane flying in blue skies. | A very large commercial plane flying in rainy skies. | A herd of elephants walking across a dry grass field. | A herd of elephants walking across a green grass field.

Figure 3: Examples of changing the background while keeping the caption fixed. The respective nearest training images based on pixelwise distance are displayed on top.

Despite an infinite number of ways of changing colors and backgrounds in descriptions, in general we 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. As you can see in Figure 4, when objects didn't have a clear fine grained differences, such as different shape or color, the relevant changes in the generated samples weren't very clearly seen. This highlighted the limitation of the model to grasp the detailed understanding of each object.

The decadent chocolate desert is on the table. | A bowl of banas is on the table. | A vintage photo of a cat. | A vintage photo of a dog.

Figure 4: Examples of changing the object while keeping the caption fixed.

4.1.2 Analysis of Attention

Unfortunately, we found that there was no connection between the patches drawn on canvas and most attended words at each timestep. During the image generation, the model mostly focused on

several words that carried the semantic meaning of caption. The words which were mostly attended during all generation steps indicated the kind of scene the model would generate. For example, as shown in Figure 5 by equally looking at words deset and forest allowed the model to make relevant changes in the scene. Whereas in the second example, the model completely ignored word sun and didn't make any changes.

A rider on a blue motorcycle in the desert.

A rider on a blue motorcycle in the forest.

A surfer, a woman, and a child walk on the beach.

A surfer, a woman, and a child walk on the sun.

Figure 5: Examples of most attended words while changing the background in the caption.

4.1.3 Comparison With Other Models

The quantitative evaluation of generative models has been a subject of ambiguity in a machine learning community. Compared to reporting classification accuracies in discriminative models, the measures defining generative models are intractable most of the times and might not correctly define the real quality of the model. To get a better comparisions between performances of generative models, we report results on two different metrics as well as do some qualitative comparison of different generative models.

As you can see on Figure 6, we generated several samples from prior of each of the current state-of-the-art generative models corresponding to the caption "A group of people walk on a beach with surf boards". While all of the samples look sharp, the images generated by LAPGAN look more noisy and don't have a very clear structure in them, whereas the images generated by variational models trained with L2 cost function have a watercolor effect on the images.

As for the quantitative comparison of different models, we first compare the performances of the model trained with variational methods. We rank the images in test set conditioned on the captions based on the variational lower bound of likelihood and then report the Precision-Recall metric to evaluate the quality of the generative model. As we expected, the quality of image retrieval using generative models is worse compared to the disriminative models that were specifically build for retrieval. To deal with large computational complexity of looping through each test image, we create a shortlist of hundred images including the correct one, based on the convolutional features of VGG like model trained on CIFAR dataset. Since there are "easy" images for which the model assigns high likelihood independent of the query caption, we look at the ratio of likehood of image conditioned on the sentence to likelihood of image conditioned on the mean sentence representation in the training set. We found that the reconstruction error $\mathcal{L}^x$ increased for the sharpened images that considerably hurt the retrieval results. Since sharpening changes the statistics of images, computing reconstruction error for each pixel is not necessarily a good metric.

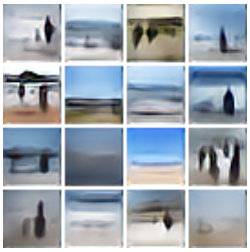
Instead of calculating error per each pixel, we turn to the smarter metric, such as Structural Similarity Index (SSI), which incorporates luminace and contrast masking into the error calculation. Due to strong inter-dependencies of closer pixels, the metric is calculated on the small windows of the image. To calculate SSI, we sampled fifty images from prior of each generative model per each caption in the test set.

References

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

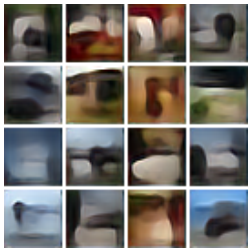
Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Cho, Kyunghyun, van Merrienboer, Bart, Gülçehre, Çaglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using RNN encoder-



Align DRAW

LAPGAN



Conv. VAE

Figure 6: Four different models displaying results from sampling caption *A group of people walk on a beach with surf boards.*

COCO (before sharpening)						
	Image Search					Image Similarity
Model	**R@1**	**R@5**	**R@10**	**R@50**	**Med r**	**SSI**
LAPGAN	-	-	-	-	-	0.08
Fully-conn. VAE (L2 cost)	1.0	5.6	10.4	51.1	51	0.156
Conv. VAE (L2 cost)	1.0	5.9	10.9	50.8	50	**0.164**
Skipthought DRAW	2.0	11.2	18.9	63.3	36	0.157
Noalign DRAW	2.8	**14.1**	**23.1**	68.0	**31**	0.155
Align DRAW	**3.0**	14.0	22.9	**68.5**	**31**	0.156

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.