Variational Image Captioning Using Deterministic Attention

Paola Cascante Bonilla
Computer Science
University of Virginia
pc9za@virginia.edu

Hyun Jae Cho
Computer Science
University of Virginia
hc2kc@virginia.edu

Alphonse N Akakpo
Systems Engineering
University of Virginia
ana2cy@virginia.edu

Abstract

Generating descriptions from images has been a challenging research topic that intersects Computer Vision and Natural Language Processing (NLP). Image captioning is a difficult, yet important task that has recently been drawing much attention. However, generating diverse and context-rich captions is not widely studied, which could represent a limitation for a broader range of applications. In this project, we present a novel approach to generate diverse captions using Conditional Variational Auto-Encoders and deterministic attention.

1 Problem Definition

Generating images descriptions is a challenging task, it requires to detect which objects are in the image, and also capture the representation and relations between these objects using natural language sentences. Both tasks are considered as difficult problems; even descriptive language is sometimes challenging for common understanding. Popular image datasets such as the MS-COCO dataset [8] include image captions, and a popular task is to generate descriptions based on those images, but generating context-rich captions requires understanding of the semantics in a given image.

Traditional models has successfully employed a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to generate captions from images [12, 10, 7], but they are generally incapable of generating novel, more descriptive and diverse captions. In traditional image captioning with deep neural networks, a CNN is used to extract a dense feature representation a_t that represents the input image. This vector is used as the initial state of a RNN. At each iteration, the RNN generates a predicted next word in a sentence given its previous state. Therefore, the result from the CNN influences the following word predictions.

The ability to generate more descriptive, novel captions is valuable for suiting systems with the ability to perceive semantic variations on the image compositions. The challenge is to be more accurate about the visual alterations, instead of generating the same short caption for many different images. The true challenge resides on getting diverse syntactic language representations from images, instead of reproducing the ground-truth captions already seen on the sample collections. Some images can be hard to interpret, the purpose of this project is to present a novel end-to-end solution to this problem.

2 Related Work

Attention mechanism has been employed to improve the state-of-the-art performance in many neural machine translation tasks including [7, 1] by selectively referring back to the corresponding source text. Xu *et al*. [12] also showed that visual attention is capable of improving caption generation quality by assigning weights to the feature vector to focus on certain areas at every iteration of the decoder network. Since then, visual attention has been widely adopted for generating more accurate captions.

More recently, Variational Autoencoders (VAEs) have been explored for generating sentences. Standard recurrent neural network language models generate one word at a time based on the previous word, but this approach lacks global context over the sentence. Using VAEs for language modeling has proven to exploit the characteristics of this generative approach to incorporate distributed latent representations of entire sentences. Doing this, the system is capable of modeling more general properties, such as style and high-level syntactic features [2, 5].

Additionally, prior work on exploiting VAEs and CVAEs for language models that include image related tasks has shown successful results [6, 11]. In [6] the authors take advantage of this latent variable model for Visual Question Answering (VQA) tasks. VQA is a research area about answering questions based on an image and also involves both image recognition and natural language processing. In [11] the authors explore image caption generation using CVAEs, augmenting the representation with an additional data dependent latent variable, in which they use the annotated objects in a given image. The main difference between this work and ours is that instead of using the data augmentation and exploiting a Gaussian Mixture Model, we instead encourage the model to maintain the fixed Gaussian prior and force the captioning decoder to rely mainly on the attention weights learned when iterating over the whole captioning model.

3 Data

We use the MS COCO dataset [8] which is a widely used dataset for caption generation tasks. COCO dataset contains 80 labels, 1.5 million object instances, and 330K images, of which more than 200K are labeled. Each image contains 5 human-generated captions, which makes it an ideal dataset for our caption generation task. We used 113.286 images for training, 5000 images for validation, and 5000 images for evaluation.

a couple of horses standing in a river next to an island.
two horses are walking through a river together
two horses standing in a stream next to a wooded area.
horses standing in shallow water in a wooded area
two horses that are standing in some water.

woman on the sidewalk putting her umbrella up.
a woman's black umbrella blew inside out in the rain.
a very cute lady holding an indie out umbrella.
a woman holding an inverted black umbrella in her hands.
a woman's umbrella has flipped backwards in the rain.

Figure 1: Examples of images in the COCO dataset. On the left hand side, we see captions explaining two horses in water in a wooded area. However, no caption mentions the white stripes the horse on the right or the fallen trees in the background. Moreover, the words *brown* or *green* never appear in any of the captions. On the right hand side, 4 of the 5 captions mention the inside out umbrella, but none mentions her clothes, the darkness of the image, or the puddle of water on the street.

4 Proposed Method

4.1 Overview

We propose an end-to-end model to generate diverse image descriptions for each input image using a Conditional Variational Autoencoder (CVAE), where the encoder compresses data into a latent space, and the decoder uses attention mechanism and generates captions given an image feature representation and the latent representation. A diagram of the model is shown in Figure 1 for visual illustration.

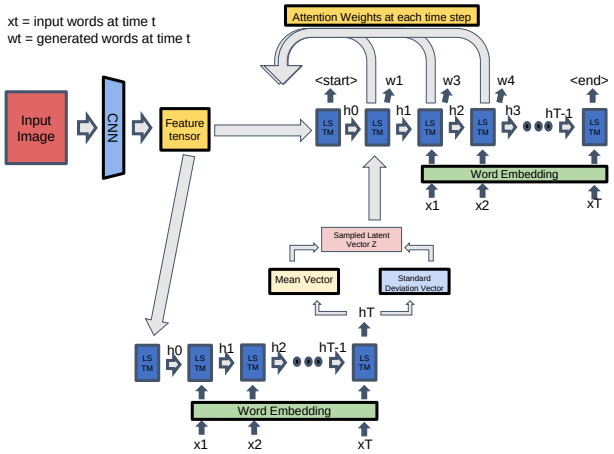

Figure 2: Model diagram. Given an input image, a CNN is used to extract out a feature representation. This feature representation acts as the initial hidden state of the encoder LSTM (bottom LSTM). After a forward propagation, the last hidden state h_T is obtained, but it is represented by a mean vector and a standard deviation vector. Latent vector z is sampled from them via the reparametrization trick, and it is given to the LSTM decoder network (above) as an initial input, along with the original image feature representation. At each timestep t, the decoder outputs α_{ti}, which is multiplied to the feature representation to become the context vector that tells the network which regions of the image to "attend", or focus on, when generating w_t. During training, teacher forcing is used. At test time, beam search is used.

4.2 Encoder - LSTM

We first used a CNN–a pretrained ResNet [3] on Imagenet with 101 layers–that extracts out a feature vector of size 2048, a_i, given an input image. This representation is fed into the initial state of the encoder, which is a Long-Short Term Memory (LSTM) [4]. The encoder additionally takes as input one embedded word during each timestep, and at the last iteration, the hidden state h_T is produced. This hidden state contains rich information about the whole input (image and caption). It is then processed through a set of linear layers to extract two separate vectors, one representing the mean and the other representing the standard deviation of h_T. Using the reparametrization trick, the latent vector z is sampled from those two vectors.

4.3 Decoder - LSTM

The decoder gets two inputs before it starts getting the embedded words as inputs. First, z is fed into the decoder along with the feature representation a_i extracted from the pretrained Resnet. Note that z is learned to model a Gaussian distribution to give a variable condition to the decoder. Lastly, the decoder takes as input one embedded input word per timestep using the attention mechanism.

Attention mechanism works as follows: at each iteration t, the decoder uses a multi-layer perceptron (MLP) f_{att} that is conditioned on the previous hidden state h_{t-1} and the a_i to produce the attention weight vector α_t. To be mathematically precise,

$$e_{ti} = f_{att}(h_{t-1}, a_t)$$

$$\alpha_{ti} = \frac{exp(e_{ti})}{\Sigma_k exp(e_{tk})}.$$

Finally, the weights are multiplied to the feature representation a_i to become the context vector c_t. In mathematical terms,

$$c_t = \Sigma_i \alpha_{ti} \cdot a_i.$$

These context vectors become the input to the LSTM in the next timestep, together with an embedded input word. The logic behind attention mechanism is that by focusing on the areas of the input that

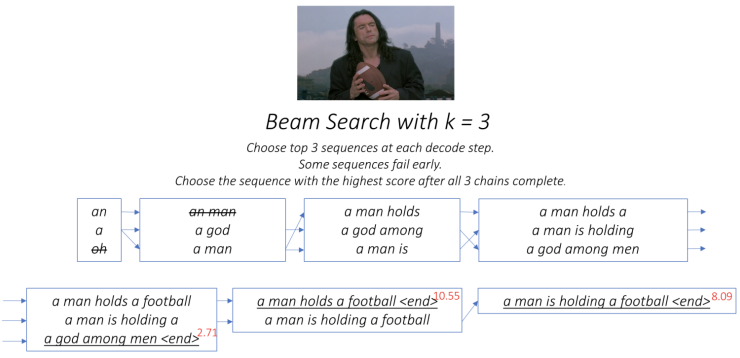

Figure 3: Example of beam search with $k = 3$. At each iteration, top k words are chosen, and from them the next k words are generated, and so on until the <end> tag is encountered. source: [9]

are highly correlated with w_t, which is the word that is generated at time t, the more accurate w_t will be. Therefore, each w_t is conditioned on the initial feature representation a_i, the latent vector z, attention context vector c_t, and the previous words $x_1...x_{t-1}$. However, instead of feeding in the prediction word w_{t-1} at timestep t, which can be incorrect, we feed the actual word x_t. We used this techniche (teacher forcing) due to its efficiency and accuracy demonstrated when training recurrent neural network models.

4.4 Objective Loss Function

We use KL-Divergence to model the distribution of z as close to the Gaussian distribution as possible:

$$D_{KL}[q(z|x_i)||p(z)] = -\Sigma_i q(z|x_i) \cdot log(\frac{p(z)}{q(z|x_i)}).$$

In addition, we use cross entropy of the generated words and target words when training the decoder:

$$\Sigma_{j=1} A_j \cdot log(P_j),$$

where A_j represents the embedding of the actual next word and P_i represents the softmax of the prediction.

Joining the two loss functions, we obtain out objective function:

$$L = \Sigma_j A_j \cdot log(P_j) - \Sigma_i q(z|x_i) \cdot log(\frac{p(z)}{q(z|x_i)}).$$

4.5 Inference

At test time, we only use the decoder to generate captions. When generating the captions instead of directly choosing the word with the highest probability, we utilize beam search. Beam search works as follows: when generating w_0, we consider k words with the highest probability. Then for each of the following timestep, choose k words with the highest probability given the k words in the previous timestep. An example of beam search with $k = 3$ is illustrated in Figure 3 [9].

5 Experiments

We experimented by alternating the initial state order of the Decoder, introducing first the image feature representations and then the latent vector z during and after training. Our conclusion was that the LSTM initial vectors order had a considerable impact on the caption generation. We decided to give priority to the latent vector z since it yielded better results. We also experimented changing the dimensions of z, but that added to much variance to our model and the results were diverse but extremely inaccurate.

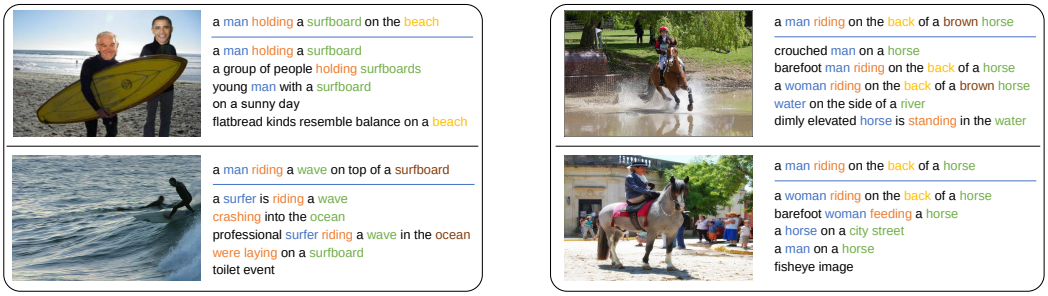

Figure 4: Example outcomes. The first caption is generated by the baseline model[2]. The other captions are generated using our model. Note that our model also generates the baseline caption, but in addition is able to figure out more fine grained and general characteristics of the image for its corresponding description. For example, using the first image the model is able to figure out that it is a sunny day. The last sentence demonstrates an example of poorly generated caption due to the diversity enforced by the fixed Gaussian condition, even with the attention mechanism.

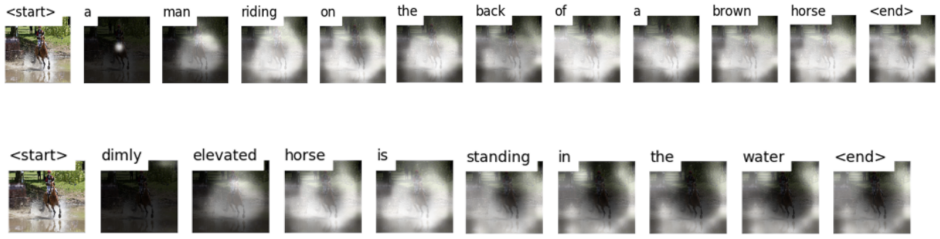

Figure 5: Attention mechanism in action is shown clearly in the caption "*dimly elevated horse is standing in the water*". When the word "*horse*" is being generated, we can see that the model is focusing its attention in the center of the image, which is where the horse is. On the other hand, when the word "*water*" is being generated, the model is attending the periphery of the image.

We also experimented using different beam sizes, getting better results with $k = 5$. Figure 4 demonstrates a few examples of generated sentences given an input image. Figure 5 shows attention mechanism in work.

6 Evaluation

We use the BLEU and the METEOR scores to evaluate the performance of generated caption in Table 1. The BLEU score measures the similarity between sentences, and its main purpose is to evaluate the exact word-to-word comparison between the ground truth and the generated sentence. The METEOR score finds the optimal semantic alignment, by looking up for synonyms in WordNet [9]. We believe that the two metrics, however, are not ideal for evaluating the task of generating diverse captions, for diverse captions require the use of words that may not exist in ground truth. As a result, one potential future work is to have human evaluation, such as through Amazon Mechanical Turk, to generate scores for captions that aim to generate diverse captions.

7 Conclusion

We conclude by reiterating the importance of producing diverse captions in caption generation tasks. Diverse captioning can explain the image more fully and therefore is widely applicable to many tasks.

[2]The baseline is a open-sourced implementation of the attention mechanism available here: https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

	BLEU4	METEOR
LSTM	0.413	0.285
CVAE	0.261	0.246
GMM-CVAE	0.371	0.274
A-CVAE (ours)	0.352	0.198

Table 1: Quantitative results. CVAE and GMM-CVAE are benchmark models from [11]. Note that our model, A-CVAE (Attention-Conditional Variational Auto-Encoder), seems to perform poorer than the LSTM and the benchmark models according to the metrics. However, diverse captions are prone to result in lower scores for these metrics, and only human evaluation can be accurate measurements for the performance of our model.

In this project, we have shown an end-to-end solution to the traditionally limited image captioning task. The fixed Gaussian introduced by the latent vector allows the word prediction model to be more diverse than the original caption generation models. Similarly, attention mechanism helps the model to focus only on the parts of the image that matter the most when generating each word during each timestep. As a result, by combining a conditional variational autoencoder and attention mechanism, we are able to produce captions that are both accurate and diverse. The most accurate evaluation will be human evaluation since the existing metrics fail to accurately evaluate diversity in captions.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. *CoRR*, abs/1511.06349, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[5] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Controllable text generation. *CoRR*, abs/1703.00955, 2017.

[6] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. Creativity: Generating diverse questions using variational autoencoders. *CoRR*, abs/1704.03493, 2017.

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017.

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[9] George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164, 2015.

[11] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5758–5768, 2017.

[12] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 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*, pages 2048–2057, 2015.