Investigating Diversity in Responses to Artwork through Latent Representations of Language Captions

Ananya Devarakonda Samuel Showalter

MAHE, Manipal UC, Irvine

Abstract

The field of natural language processing (NLP) has seen immense growth in applications ranging from machine translation to text generation. There has been a rise in the use of generative models in NLP applications. The variational autoencoder (VAE) is a powerful generative model that learns probabilistic representations of latent space encodings. There are many versions of the VAE, including the βVAE, which introduces a hyperparameter β to encourage disentangled representations in the latent space. This paper examines the latent space encodings of an LSTM β-VAE trained on the ArtEmis dataset to generate emotion-driven text and draw conclusions about specific works of art and genres in the dataset. The ArtEmis dataset is a collection of subjective captions of artwork based on the WikiArt dataset. We quantitatively evaluate the diversity of responses elicited by a work of art by calculating the Euclidean distance between points in the learned latent space representation of captions in the dataset and their respective centroid. We use the average distance as a metric to evaluate the diversity of responses elicited by a particular work of art. In addition, we qualitatively examine the latent representations of captions in the dataset using PCA. We found that analytical cubism, synthetic cubism, pointillism, contemporary realism, and action painting were the genres of art that elicited the most diverse responses. The source code for the project can be found here: https://github.com/ananya/Artemis VAE.git

1 Introduction

Natural language processing (NLP), a subfield of artificial intelligence, deals with computational linguistics with applications in areas such as document classification [1,2], machine translation [3, 4], and sentiment analysis [5, 6]. Typically, applications in NLP use autoregressive models such as recurrent neural networks (RNNs) and their variants-gated recurrent units (GRUs) and long short-term memory (LSTM) models. An essential and well-studied application of NLP is text generation, which deals with the automatic production of readable and coherent text. Recently, the use of generative models has shown promising results in the application of text generation. Among the generative methods used, the variational autoencoder (VAE), as introduced by Kingma and Welling [7], has proven to be a powerful self-supervised model for text generation [8]. Variational autoencoders (VAEs) are distinct from standard autoencoders as they take a Bayesian approach to learn continuous latent representations of features in the dataset. One version of the VAE, known as the β-VAE [9], enforces the learning of disentangled representations in the latent space of the VAE by adding a hyperparameter β. It is valuable to learn disentangled or statistically independent representations to coherently infer properties of a particular dataset to draw conclusions about fundamental relationships between features in the dataset.

This paper outlines an emotion-driven text generation model using a β-VAE trained on the ArtEmis dataset [10]. The ArtEmis dataset is an image captioning dataset that contains around 455,000 descriptions of 80,000 works of art. Unlike traditional image captioning datasets, such as the Flickr 8k [11], Flickr 30k [11], and MS-COCO datasets [12] that focus on entirely objective descriptions of images, the ArtEmis dataset contains captions of artwork that highlight personal interpretation and emotional response.

This paper attempts to quantify a topic that is considered subjective: ranking artwork based on the diversity of the responses elicited. Quantification is done by examining the similarities in the latent encodings learned by the β-VAE. The paper also draws conclusions about artwork mentioned in the ArtEmis dataset and comments on the diversity of responses elicited by a particular work of art. Thus, our contributions are as follows:

1. Introducing an emotion-driven text generation method using an LSTM β-VAE model.
2. Exploring latent space representations of the captions in the ArtEmis dataset to quantitatively analyze features in the dataset.
3. Qualitatively analyzing diversity in response to artwork using PCA.

2 Literature Review

Several subfields of artificial intelligence, including computer vision and natural language processing, use the variational autoencoder (VAE) for various applications. In computer vision, the VAE is popularly used in image generation [13–16] by interpolating between the learned disentangled latent representations of features in the dataset. In addition to image generation, Pu et al. [14] developed a novel VAE to generate accurate labels and captions along with images. The versatility of the VAE as a generative model is further highlighted by Yan et al. [13], who introduced a VAE model to generate images from high-level textual descriptions.

The VAE has shown promising results in the field of natural language processing (NLP) as well. Hayashi et al. [17] used a VAE for an end-to-end speech-to-text model and stated that they obtained state-of-the-art results. Sheng et al. [18] proposed a novel VAE model for accurate neural machine translation (NMT). Other applications of VAEs in NLP include document modelling [19], dialog generation [20], and text summarization [21]. Text generation, an essential application of natural language generation (NLG), is a subfield of NLP that involves the automatic generation of grammatically correct and coherent text. Probabilistic models used for sentence generation trained to predict words based on sequential input data are dubbed language models (LMs). Traditionally, language models used autoregressive models, including recurrent neural networks (RNNs), long short-term memory (LSTM) models, and gated recurrent units (GRUs). Such models are called recurrent neural network language models or RNNLMs [22]. Vinyals et al. [23] used an LSTM-based model for image caption generation and obtained realistic results. More recently, Liu et al. [24] used an RNN decoder for poem generation in an image-to-poem model.

Although RNNLMs are shown to generate coherent sentences, they do not learn interpretable representations of dataset attributes; in fact, the behavior of an RNNLM remains black-box. Unlike RNNLMs, models based on the variational autoencoder (VAE) introduce an element of interpretability by regularizing for a smooth latent representation of features in a dataset. Thus, VAEs have been extensively used in text generation models to not only generate accurate sentences but also to reveal properties of the dataset by analyzing learned latent representations. In particular, the β-VAE encourages the learning of disentangled representations of features in the latent space. This quality is conducive to learning continuous hidden representations of captions from the ArtEmis dataset. These latent representations can then be analyzed quantitatively to help sort the works of art mentioned in the ArtEmis dataset by the diversity of response and interpretation- a topic that is inherently subjective.

Bowman et al. [8] introduced an LSTM-based VAE model to generate sentences by interpolating within the learned latent space. They found that the VAE model, without optimization, behaved like an RNNLM. This problem is dubbed "posterior collapse." Pelsmaeker et al. [25] describe and compare methods of overcoming the problem of posterior collapse. Bowman et al. [8] tackled this problem by optimizing the model using KL-cost annealing as well as word dropout and history-less decoding. Post-optimization, the model produced coherent and diverse sentences by interpolating between learned latent representations. Yang et al. [26] proposed a variational autoencoder model with an LSTM encoder and a dilated convolutional neural network (CNN) decoder for conditional text generation.

In our implementation, we give priority to interpretability over the general text generation model and place more emphasis on the posterior as our objective is to rank the diversity in the response of artwork using captions. The papers mentioned primarily focus on generating sentences as accurately as possible and do not provide much information on inferring properties of the datasets using learned latent representations. This paper provides information on the model used for emotion-driven text generation and analyzes the features in the ArtEmis dataset by examining the learned latent encodings to sort works of art based on the diversity of their interpretations.

3 Experimental Design

3.1 Dataset

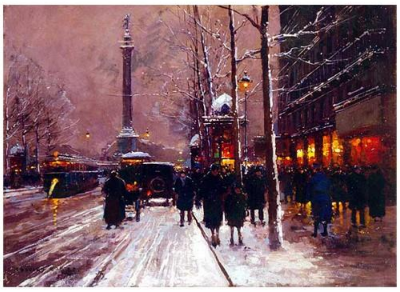
The ArtEmis dataset [10] is a large-scale image captioning dataset that contains over 455,000 emotion attributes and subjective captions of over 80,000 works of art. Image captioning datasets such as the Flickr 8k [11], Flickr 30k [11], and MS-COCO datasets [12] contain only objective descriptions of real-world images. Unlike real-world image data, art is intrinsically subjective. Thus, for effective image captioning of art, it is essential to retain elements of personal interpretation. Recognizing that subjectivity is fundamental to artwork, the ArtEmis dataset contains captions highlighting the emotions elicited by a particular work of art. Much like real-world text data, the captions from the ArtEmis dataset contain grammatical errors, spelling errors, and show human bias. Captions are categorized by art style ("art style") and name of the painting ("painting"). The authors also specify a column to assign a general emotional attribute to each artwork ("emotion"). An example from the dataset, along with the WikiArt image, is given in Figure 1 [10].

It is essential to recognize that this emotional attribute is chosen based on eight emotion options (amusement, anger, awe, fear, sadness, contentment, disgust, excitement) and one default option (something else). Around 53,000 records categorize artwork as "something else." Furthermore, much like real-world data, there are instances in the dataset where the chosen emotion attribute does not align with the sentence. Thus, using only emotional attributes to judge the diversity of emotional responses to an artwork would be inaccurate. Keeping this in mind, we use raw captions from the ArtEmis dataset so that the model learns appropriate latent representations for further analysis.

3.2 Preprocessing

The dataset used to train the model consists of all the captions in the ArtEmis dataset. First, we tokenize the captions. Tokenization refers to the splitting of each sentence in the dataset into its respective words and symbols. Next, we split the dataset containing all captions into a training (99%) and validation (1%) dataset. The validation dataset ensures that the model has not overfit and performs well with unseen data. We then build a vocabulary of all words and symbols in the tokenized training dataset using pre-trained GloVe vectors [27]. While building the vocabulary, we add special tokens

Edouard Cortes
place-de-la-bastille-4



Contentment
Cold, dark, and snowing, but people still left their homes and socialized.

Figure 1: An example from the ArtEmis dataset

such as the <*sos*>, <*eos*>, and <*pad*> tokens that refer to start of sentence, end of sentence, and padding respectively. A few sentences in the ArtEmis dataset have no spaces between words. For such instances, we have added spaces wherever required.

3.3 Model

3.3.1 Variational Autoencoder (VAE)

Autoencoders are self-supervised deep generative neural network models that feature an encoder-decoder architecture with a bottleneck layer having reduced dimensions to learn latent representations of given data. They are trained to reconstruct a given input. The primary difference between a variational autoencoder (VAE) [7] and a standard autoencoder is that the VAE encodes features as a continuous distribution over the latent space. Thus, instead of encoding singular points for each feature, the variational autoencoder learns a “soft” representation of the data in the latent space.

Fundamentally, for an input x the VAE learns a latent representation z and generates a reconstruction of x, say, x̂. In such a case, p(z|x) is given as:

$$\frac{P(x|z)\,P(z)}{P(x)} \quad (1)$$

Usually, p(x) is an intractable distribution, and thus, variational inference is used to estimate the value of p(z|x). The process involves selecting another distribution, say q(z|x), that we know is tractable (such as the Gaussian distribution) and making it as similar to p(z|x) as possible. This is done by minimizing the KL divergence between the two distributions. Thus, the loss function used by the VAE is given in equation (2):

$$L(x,\hat{x}) = l(x,\hat{x}) + KL(q(z|x)\,||\,p(z|x)) \quad (2)$$

Where $l(x, \hat{x})$ refers to the reconstruction loss. Reconstruction loss represents a metric such as mean squared error (MSE) or categorical cross-entropy and is chosen according to the assumed prior. Data generation using the VAE is done by sampling from the learned distribution in the latent space. As the output from sampling is discrete, backpropagation is not directly possible. Thus, Kingma et al. [7] proposed the reparameterization trick that made backpropagation possible in this circumstance. Assuming a Gaussian prior, sampling can be represented by:

$$z \sim q(z|x) = \mathrm{N}(z\,;\mu\,,\sigma)$$

Then, the reparameterization trick is as follows:

$$z = \mu + \epsilon \cdot \sigma, \text{ where } \sigma \sim \mathrm{N}(0, I)$$

This trick helps to train the VAE model using gradient descent and backpropagation.

The β-VAE [9] is a version of the original VAE model that encourages the learning of continuous disentangled representations by adding a Lagrangian multiplier β to equation (2), as represented by:

$$L(x, \hat{x}) = l(x, \hat{x}) + \beta KL(q(z|x) \,||\, p(z|x))$$

When $\beta = 1$, the model behaves equivalently to the original VAE. For values of $\beta > 1$, the model encourages the learning of disentangled feature representations in the latent space.

provides an illustration of the VAE model with an LSTM encoder and decoder.

The main advantage of using the β-VAE is that it forces the model to learn continuous disentangled representations of features in the dataset. The model we define uses categorical cross-entropy as the reconstruction loss and Adam optimization. After training over 60 epochs, the model achieved a final training loss of 0.046. The LSTM β-VAE model was optimized using KL cost annealing by gradually increasing the hyperparameter β during training. The process uses sigmoid annealing where β was increased from a value of one to eight. KL annealing helps tackle the problem of "posterior collapse" and ensures that the value of the KL term does not diminish. Thus, KL annealing encourages the LSTM β-VAE model to learn continuous disentangled latent representations of captions in the ArtEmis dataset.

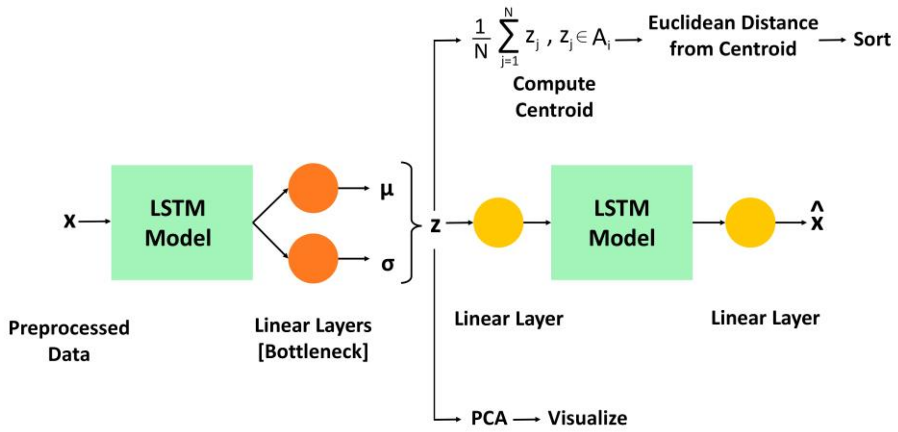


Figure 2: An illustration of the experimental design

3.3.2 LSTM β-VAE

This paper uses a β-variational autoencoder (β-VAE) [9] with an LSTM encoder and decoder as introduced by Bowmann et al. [8]. The model architecture resembles that of the sequence-to-sequence (Seq2Seq) model but has fully connected layers to encode latent representations after the LSTM encoder. Figure 2

3.4 Latent Space Analysis

To analyze the latent space of captions from the ArtEmis dataset, after sorting by artwork, we compute the centroid by calculating the mean of the latent representation z learned by the LSTM β-VAE for a particular work of art or art style. We then compute the Euclidean distance of all latent encodings of captions from the centroid and save the average distance. We use the average distance as the metric to sort artwork and art styles mentioned in the ArtEmis dataset based on the diversity of the response it elicits. Thus, the higher the average distance from the centroid, the greater the diversity of response and vice-versa.

Finally, for qualitative analysis, we use principal component analysis (PCA) to linearly project the latent space from a 100-dimensional representation to a two-dimensional representation of captions ($\mathsf{R}^{100} \rightarrow \mathsf{R}^{2}$) for each work of art. After PCA, we plot the points onto a two-dimensional plane and qualitatively infer properties of the latent representations.

4 Results

4.1 Text Generation

The LSTM β-VAE is trained to reconstruct captions from the ArtEmis dataset to accurately learn their latent space representations.

this young lady appears to be very pleased to sit for her portrait .

this young lady appears to be very pleased to sit for her portrait .

the person walking in the woods look peaceful and at ease .

the person walking in the woods look peaceful and at ease .

Table 1: Examples of reconstructed sentences from training data

beautiful painting of a tree with bright yellow and red leaves .

beautiful painting of a tree with bright yellow and red leaves .

it looks like a beautiful evening near the sea.

it looks like a beautiful evening behind the sea .

Table 2: Examples of reconstructed sentences from validation data

To measure the accuracy of reconstruction, we use cross-entropy and found that the validation data had a final cross-entropy of **0.068**. Tables 1 and 2 present a few examples of the reconstructed captions from the train and validation datasets, respectively. In both tables, the sentence in bold represents the source sentence, and the sentence below it represents the reconstructed sentence. As illustrated in tables 1 and 2, the model accurately learns to reconstruct the input data. The results also show that the model replicates grammatical errors present in the dataset. Hence, there is scope for future research in building a model that is capable of generating grammatically correct text even with errors in the dataset.

4.2 Sentence Interpolation

As the LSTM β-VAE model is trained to learn continuous latent representations of the captions in the ArtEmis dataset, it is possible to generate sentences by sampling from this continuous latent space. This property of the model makes it possible to choose two captions from the dataset and generate intermediate sentences by sampling from the smooth learned latent space. This process is called sentence interpolation. Table 3 provides an example of interpolation from a positive sentence expressing awe to a negative sentence expressing disgust. The two sentences selected from the dataset for interpolation are in bold. Although these sentences contain grammatical errors, they demonstrate a gradual shift in the overall emotion and meaning of each sentence. This gradual change in emotion through each step in interpolation is representative of the continuous learned latent space.

the scene is quite beautiful and reminds me of fairy fairy tales .

the scene is majestic it and reminds me of magic queen .

the scene looks looks and seems exciting of and evil .

the scene made look and is serious and very detailed .

the man on the right looks very mean looking.

Table 3: Sentence interpolation

Painting	Average Distance
alphonse-mucha holy-mount-athos-1926	3.130717
andrea-mantegna madonna-with-saints-st-johnthebaptist-st-gregory-i-the-greatst-benedict-1506	3.106543
brice-marden suicide-note-1973	3.051492
el-greco portrait-of-a-man-2	3.033126

Table 4: Artwork with the greatest average distances

Painting	Average Distance
thomas-eakins photograph-1910-8	0.001139
edwin-henry-landseer a-distinguished-member-ofthe-humane-society	0.000937
louay-kayyali motherhood-1974	0.000897
jacob-jordaens bust-of-satyr-1621	0.000661
nikolay-bogdanov-belsky the-former-defender-of-the -homeland	0.000575

Table 5: Artwork with the least average distances

4.3 Ranking Based on Diversity of Response

4.3.1 Artwork

Tables 4 and 5 contain the results of the latent space analysis done on captions from each work of art from the ArtEmis dataset. Table 4 lists the top four most diverse works of art after analysis of the learned latent representations of the LSTM β-VAE. Similarly, table 5 presents the five works of art with the most similar (least diverse) responses to art from the ArtEmis dataset. Alphonse Mucha's "Holy Mount Athos" was recognized as the work of art with the most diverse responses, while Nikolay Bogdanov-Belsky's "The former Defender of the Homeland" was recognized as the artwork with the most similar interpretations.

To qualitatively analyze the diversity of the responses, figure 3 provides a plot of the latent space of a work of art after PCA. Clearly, captions labeled with a particular emotion in the ArtEmis dataset seem to form distinct clusters. This is, however, not always the case, as in a few instances, there exists variation in emotion within clusters. This variation can be attributed to the inherent subjectivity in the topic of emotions. It is well known that there is no quantitative distinction between one emotion and another. By nature, emotions are defined by personal experience and this subjectivity can often lead to a blur between one emotion and another.

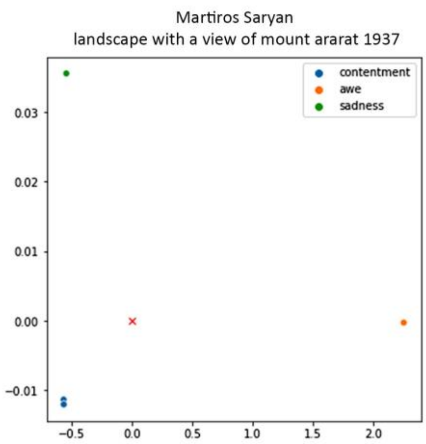


Figure 3: A scatter plot of latent representations after PCA

Art Style	Average Distance
Analytical Cubism	1.632765
Synthetic Cubism	1.562646
Pointillism	1.557388
Contemporary Realism	1.521055
Action painting	1.47832

Table 6: Art styles with the greatest average distances

4.3.2 Art Style

Along with individual works of art, we used the LSTM β-VAE model to rank art styles based on the general diversity of responses to works of art within that particular art style. As humans easily relate to typical shapes and objects, we hypothesized that nonrepresentational designs would elicit more diverse responses. The results validated this hypothesis as we found that art in the styles of analytical and synthetic cubism had the most diversity in response while art in the style of expressionism had the most similar responses (least variation in response). Table 6 shows the top five styles of art that elicit the most diverse responses, along with their average Euclidean distance. Table 7 shows the five art styles with the most similar responses. Our hypothesis is further supported by the result that action painting, a style of art that is by definition abstract and open to interpretation, is among the top five art styles that elicit the most diverse responses.

Art Style	Average Distance
Impressionism	1.254654
Art Nouveau Modern	1.254276
Baroque	1.251281
Post Impressionism	1.250021
Expressionism	1.247783

Table 7: Art styles with the least average distances

5 Discussion

The results obtained from the LSTM β-VAE fall into three main categories- text generation, sentence interpolation, and ranking of art based on the diversity of response elicited by that work of art. This paper primarily focuses on the subjective topic of ranking artwork based on the different responses people have by viewing a particular work of art or art style. The result that art styles such as analytical cubism, synthetic cubism, pointillism, contemporary realism, and action paintings elicit the most diverse responses in people aligns with the historical context and type of art.

It is important to note that while evaluating our model, we have made the assumption that the data contains grammatically correct and accurate captions for artwork. However, as the dataset represents real-world data, this is not always the case. Thus, there is scope for future work in building a model that captures the essence of real-world data without its limitations, such as grammatical errors and human bias.

6 Conclusion

While the generation of artwork using artificial intelligence is a popular topic today in the domain of computer science, analyzing the inherent subjectivity and variation in the interpretation of works of art remains unexplored. In this paper, we use an LSTM β-VAE model to learn latent representations of the captions in the ArtEmis dataset and calculate the average Euclidean distance between points in these latent representations for each work of art and art style. We use the average distance as a metric to sort individual works of art and art styles based on how differently human beings respond to them.

We found that our model produced accurate and coherent text with unseen data. Further, we demonstrated that it is possible to generate sentences by sentence interpolation, and as illustrated in the results, there is scope for building a model with grammatically correct interpolations and higher accuracy. Finally, we found that certain works of art had a much greater average distance as compared to others, and they primarily fall in the style of analytical cubism. In contrast, works of art with the least average distances mostly were in the style of expressionism.

Possible directions for future work include testing variants of the LSTM β-VAE model and accurately representing the captions from the ArtEmis dataset by considering them as characteristic of real-world data while taking into account inaccuracies, human bias, and grammatical errors.

References

[1] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[2] A. Adhikari, A. Ram, R. Tang, and J. Lin, "Docbert: Bert for document classification," 2019.

[3] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5039–5049.

[4] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1810–1822.

[5] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 375–384.

[6] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30–38.

[7] D. P. Kingma and M. Welling, "Autoencoding variational bayes," 2014.

[8] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio,"Generating sentences from a continuous space," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 10–21. [Online]. Available: https://aclanthology.org/K16-1002

[9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[10] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "Artemis: Affective language for visual art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11569–11579.

[11] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. European Conference on Computer Vision, September 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/

[13] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[14] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," *Advances in neural information processing systems*, vol. 29, pp. 2352–2360, 2016.

[15] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[16] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *stat*, vol. 1050, p. 8, 2021.

[17] T. Hayashi and S. Watanabe, "Discretalk: Text-to-speech as a machine translation problem," *arXiv e-prints*, pp. arXiv–2005, 2020.

[18] X. Sheng, L. Xu, J. Guo, J. Liu, R. Zhao, and Y. Xu, "Introvnmt: An introspective model for variational neural machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8830–8837.

[19] Y. Miao, L. Yu, and P. Blunsom, "Neural variational inference for text processing," in *International conference on machine learning*. PMLR, 2016, pp. 1727–1736.

[20] T.-H. Wen, Y. Miao, P. Blunsom, and S. Young, "Latent intention dialogue models," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3732–3741.

[21] Y. Miao and P. Blunsom, "Language as a latent variable: Discrete generative models for sentence compression," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics,

Nov. 2016, pp. 319–328. [Online]. Available: https: //aclanthology.org/D16-1031

[22] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 234–239.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[24] B. Liu, J. Fu, M. P. Kato, and M. Yoshikawa, "Beyond narrative description: Generating poetry from images by multi-adversarial training," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 783–791.

[25] T. Pelsmaeker and W. Aziz, "Effective estimation of deep generative language models," in *ACL*, 2020.

[26] Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *International conference on machine learning*. PMLR, 2017, pp. 3881–3890.

[27] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http: //www.aclweb.org/anthology/D14-1162