LEARNING DISENTANGLED REPRESENTATION IN LATENT STOCHASTIC MODELS: A CASE STUDY WITH IMAGE CAPTIONING

Nidhi Vyas, SaiKrishna Rallabandi, Lalitesh Morishetti, Eduard Hovy and Alan W Black

Language Technologies Institute, Carnegie Mellon University, PA, USA

ABSTRACT

Latent stochastic encoder decoder models often are faced with optimization issues such as latent collapse preventing them from realizing their full potential of rich representation learning and disentanglement. In this paper we present an approach to train such models by incorporating joint continuous and discrete representation in the prior distribution. We evaluate the performance of proposed approach on a multitude of metrics against vanilla latent stochastic models. We also perform a qualitative assessment and observe that the proposed approach indeed has the potential to learn composite information and explain novel combinations not seen in the training data.

Index Terms: disentanglement, latent representation, captioning, composition, multimodal, continuous, discrete

1. INTRODUCTION

Tasks involving multiple modalities are challenging due to two main reasons: Firstly, the model needs to represent and summarize multimodal data so as to exploit the complementarity and redundancy of the involved modalities [1]. Secondly the model has to be able to translate relevant information to one of the modalities without loss of generality. Consider image captioning[2, 3] as an example to demonstrate this: Models aimed at captioning are required to generate factual and grammatically meaningful image descriptions. For accomplishing this, the model needs to learn a joint representation that can capture global information about an image such as objects, their attributes, relations between the objects while discarding local information such as texture, etc. The model relies on the learnt representation to simultaneously reason about the inferred relationships between objects in a different domain, natural language.

Attempting to tackle such tasks using encoder decoder models leads to learning biases present in the data [4, 5]. Such models usually fail to recognize previously unseen compositions of the existing objects [6] since while they are universal function approximators, they lack the hierarchy to learn robust representations in their formulation. Deep latent stochastic models [7, 8] on the other hand, provide a flexible framework that promises to address such concerns [9]. These models provide a mechanism to jointly train both the latent representations as well as the generator network. Therefore, such models are expected to both discover and disentangle causal factors of variation present in the distribution of original data. Disentanglement is an attractive property from the perspective of zero shot, transfer learning [10] and low resource scenarios. Moreover, disentangled representations are usually aligned with the attributes of original data and are conditionally dependent on variance in the original data, hence interpretible.

However, while training latent stochastic models, the exact log likelihood is usually intractable and current approaches use a recognition network to approximate the posterior probability using reparameterization [11]. In practise, such models often are subject to an optimization challenge referred to as KL-collapse [12] - wherein the generator network which is usually an RNN ignores the learnt latent representation. Typical approaches to address this issue involve weakening the generator network. However, this is not desired when the downstream task is generative in nature. In this paper we take a different approach: We first present an analysis of the optimization performed in latent stochastic models. From this, we show that latent collapse happens since a trivial solution for generative models is to ignore the latent representation under certain constraints. We then present an approach to make the prior distribution more complex thereby forcing the model to encode information into the latent representation.

Our contributions are as follows (1) We propose a simple yet effective architecture that splits the latent space into continuous and discrete factors that better capture the relations between entities. (2) We perform quantitative and qualitative analysis on MSCOCO dataset and observe that the model is able to not only generate diverse captions but also makes less mistakes in terms of entity attributes.

This paper is outlined as follows: In section 2 we present relevant previous works in the context of image captioning. This is followed by an analysis of the optimization and disentanglement in latent stochastic models (3). In section 4 we present an approach to split the latent representation into joint continuous and discrete components. This is followed by experiments in section 5 and conclusion.

2. RELATED WORK

The common approach for generating image descriptions conditioned on the input image involve using a feature extractor as the encoder and then a language-model decoder to generate the captions. [13] use a Convolutional neural network (CNN) trained on images to extract the features from the pen-ultimate layer, and then use a maximum-entropy language model to generate the captions. [14, 15] replace the decoder with a Recurrent Neural Network (RNN) decoder, whereas [16] use Log-bilinear models. However, since these models are trained to maximize the likelihood with reference descriptions, they generate good short captions, but fail to generate longer diverse captions that are conditioned on the images.

[17] used Generative Adverserial Network (GAN) style caption generation model that produces captions that are indistinguishable from reference captions. [18] used a conditioned GAN to produce the image description and also evaluate the description with respect to the image content. [19] [20] [21] use Conditional Variational Auto-Encoder (CVAE) style generative modelling to add diversity in the decoder. All these models first learn a latent vector distribution at training time for the images and text. At test time, the input image, along with a latent vector z that is sampled from this distribution are fed into the decoder. The different sampled z vectors result in diversity at the decoder end. [21] used a fixed Gaussian prior to model z for all images, which resulted in collapsed conditional posterior probability at the decoder end. To account for the different objects present in an image, and to generate diverse captions around these objects, [20] substituted this single prior with different Gaussian priors for each object class.

However, these models do not capture all the causal factors of variation in the data (e.g. colour, count etc) reliably (see Figure 5. in [20]). We hypothesize that disentangling representations to capture both the properties and the objects in the image separately might help the model to overcome the sparseness issue, and generate more factual captions for the input image. Particularly, we would like to see if the model can learn different dimensions for each of these properties and objects, and hence learn to combine them to generate / comprehend novel concepts. Our work is similar in spirit to [22, 23]. Both these works attempt to learn a joint continuous and discrete latent representation in a single modality while we deal with a task that involves learning and inferring from a shared space across modalities.

3. ANALYSIS OF OPTIMIZATION AND DISENTANGLEMENT IN LATENT STOCHASTIC MODELS

Latent Stochastic Models has shown promising results in unsupervised, uni modal settings and are the preferred models for representation learning. However, when we apply these models in an encoder decoder framework, optimization becomes harder due to KL-vanishing[12]. This is mainly because the latent variable distributions are usually approximated by simpler networks compared to the powerful RNNs used in the encoders and decoder [9].

The problem becomes apparent by looking at the Variational Lowerbound (ELBO) such models try to optimize. For instance, consider the ELBO being optimized by Beta CVAE:

$$E_{q_\phi(z|x,c)}[logp_\theta(x|c,z)] - \beta|D_{KL}(q_\phi(z|x,c)||p_\theta(z|c)) - C_z| \tag{1}$$

where C_z is the channel capacity term [24]. The first term in ELBO is the reconstruction error while the second is the divergence between approximate and true posteriors. Rewriting the first term as

$$logp_\theta(x|c,z) = logp_\theta(x|z,c) + logp_\theta(z|c) \tag{2}$$

It can be seen that the optimal value of this likelihood estimate can be conditionally independent of the latent representation (z) if the recognition network is complex enough [25]. In other words, if the decoder network employs powerful universal approximators, the model is incentivized to ignore the latent representation. The second term in the expression acts as a regularizer to penalize such behavior. However, a trivial solution for the model is to force this posterior distribution to closely follow the Gaussian prior distribution [9].

The second term, KL divergence between the true and approximate posterior distributions obtains the global minimum 0 only when both the distributions match each other. From Bayes rules,

$$p_\theta(z|x,c) = \frac{p_\theta(x|z,c)p_\theta(z|c)}{p_\theta(x|c)} \tag{3}$$

It can be seen that a trivial solution to reach global minimum again is by ignoring the latent variable. Models such as β VAE and the subsequently proposed channel capacity based approaches [24] address this issue by gradually increasing the channel capacity. This would effectively result in pressurizing the posterior distribution to match the prior closely. However, following such an approach translates to an unrealistic constraint in scenarios that have categorical distribution as their output (tasks such as language modeling, machine translation, image captioning among others). In addition, it is unintuitive to assume that the true prior that generates latent distribution is a Gaussian when the likelihood is based on discrete sequential data. In such cases the decoder is implicitly weakened and the model is forced to encode information into the latent dimensions. At this stage, any local information is encoded within the hidden states of the decoder while remaining information is encoded in the latent space [25]. Thus, we ob-

tain disentanglement of independent factors of variation in the original data.

However, it has to be noted that during training optimization is performed in expectation over minibatches. The KL term can then be written as

$$E_{p(x)}[D_{KL}(q_\phi(Z|x)||p(z))] = I(x;z) + D_{KL}(q(z)||p(z)) \tag{4}$$

In other words, the KL term is the upperbound on the mutual information that can be encoded into the latent dimensions [26]. Penalizing this mutual information results in an increased reconstruction loss. Therefore, optimization in latent stochastic models follows a compromise between the capability for reconstruction and the potential for disentanglement.

4. PROPOSED APPROACH

In this work, we present an approach to improve the training of multimodal variational encoder decoders by incorporating a joint discrete and continuous prior in the latent space. Based on the analysis presented in the previous section, we hypothesize that using a more flexible prior distribution helps accomplish decent disentanglement without losing the reconstruction loss. In other words, if the prior distribution is flexible, it increases the pressure on model to match the prior more closely thereby improving disentanglement. At the same time, it gives more flexibility to the model to encode information into the latent representation thereby improving the reconstruction loss as well.

We make an observation that the tasks such as image question answering and image captioning require learning representations from both image and textual modalities. It has to be noted that while the textual modality is primarily symbolic, the visual modality is spatial. However, both these modalities can be explained by distinct discrete and continuous factors of variation. In the context of images, argument for discrete representation refers to individual objects while the continuous counterpart corresponds to the spatial relationships between objects in the image. Based on this intuition, we split the latent representation to include both continuous as well as discrete variables.

Let $\{ z_c, z_d \}$ denote set of continuous and discrete latent random variables respectively. We define joint posterior $q_\phi(z_c, z_d|x)$, prior $p(z, c)$ and likelihood $p_\theta(x|z, c)$. The objective for β variational encoder decoder with both continuous and discrete latent variables becomes:

$$E_{q_\phi(z_c,z_d|x)}[logp_\theta(x|z_c, z_d)] - \beta * K \tag{5}$$

$$K = |D_{KL}(q(z_c, z_d|x)||p(z_c, z_d)) - C_j| \tag{6}$$

where C_j is denotes joint channel capacity for both continuous and discrete latent spaces. Assuming that the continuous and discrete latent representations are independent of each other, the divergence between the true prior and approximate prior becomes:

$$\begin{aligned} &D_{KL}(q_\phi(z_c, z_d|x)||p(z_c, z_d)) = \\ &\quad E_{q_\phi(z_c|x)}[logq_\phi(z_c|x)] - E_{q_\phi(z_c|x)}[logp(z_c)] \\ &\quad + E_{q_\phi(z_d|x)}[logq_\phi(z_d|x)] - E_{q_\phi(z_d|x)}[logp(z_d)] \end{aligned} \tag{7}$$

Following [23], we further split the channel capacity into continuous and discrete latent channels and force the model to encode relevant information in both channels.

5. EXPERIMENTAL SETUP

5.1. Dataset

We conduct our experiments using the challenging MS COCO (2014) dataset [27], which has 82,783 images and was generated using human subjects on the Amazon Mechanical Turk (AMT). We have only applied trivial tokenization to the captions. We have used a threshold of 10 and every word with lower frequency was replaced by UNK. The final vocabulary size was 8855.

5.2. Evaluation Metrics

We report the performance of our systems with the frequently used BLEU metric, a measure that loosely corresponds to precision of word n-grams between hypothesis and reference sentences. However, there has been a criticism regarding interpreting BLEU scores. Hence we also present METEOR, ROUGE and CiDer [28].

5.3. Systems built

We have built the following systems for our task.

- *Base System - CNN + RNN*: As the base for our latent stochastic models we used a simple but powerful Encoder Decoder architecture. In our encoder framework we have used pretrained ResNet features. The decoder is trained in a teacher forcing fashion by stacking together encoder output and caption embedding.

- *Latent Stochastic Baseline Model:* This system is a modification of our base system to include variational inference. Specifically, we designed our encoder model to output the mean and log variance of the latent distribution. We then sample a latent representation using reparameterization trick [11]. Decoder is same as our baseline. The input to decoder is stacked vector of latent vector and caption embedding. For training this model, we have used scheduled annealing using logistic function for KL divergence as pointed out in [12, 22]. The step size for logistic function was fixed at 2500.

- *Multi Space Latent Stochastic Model*: In this system we have incorporated joint continuous and discrete latent representation as the prior distribution being modeled by the latent stochastic model. Since there are around 80 unique objects in MS COCO dataset, it might be intuitive to allow atleast so many dimensions in the discrete space. Following this intuition, we have used 128 dimensions each for the discrete and continuous components.

5.4. Quantitative Analysis

The results from quantitative evaluation of the systems is presented in table 1. As it can be seen, using both continuous and discrete variables for representing the latent space does seem to have consistent gains across different metrics.

Table 1. Quantitative Evaluation of proposed approaches

Dataset			[%]
MS COCO	Latent Stochastic Baseline	BLEU 4	0.13
		METEOR	0.15
		CIDEr	0.33
		ROUGE L	0.4
	Multi space latent stochastic model	BLEU 4	0.16
		METEOR	0.18
		CIDEr	0.49
		ROUGE L	0.43

5.5. Qualitative Analysis

Observing the captions generated by the multispace model (an example shown in figure 1 it appears that the proposed model is better able to disentangle the individual objects from the image. On the other hand, vanilla latent stochastic model tends to use the bias that a clock appears together with tower.

Fig. 1. **Baseline Latent Stochastic Model**: a clock tower with a weather vane and a clock on top of it. **Multi Space Latent Stochastic Model**: A clock on a cycle

6. FUTURE WORK

In this preliminary study, we have presented an approach to use joint continuous and discrete latent variables in latent stochastic models. However, the presented work has to be evaluated against the state of the art approaches. We also believe that a much detailed qualitative analysis has to be performed. We would like to use this module in a subsequent task, visual question answering.

7. CONCLUSION

Caption generation is an AI complete task requiring representation learning and translation across modalities. In this paper we have presented an approach to train latent stochastic encoder decoder models for such tasks by incorporating joint continuous and discrete representation in the prior distribution. We evaluate the performance of proposed approach on a multitude of metrics with vanilla latent stochastic models. We also perform a qualitative assessment and observe that the proposed approach indeed has the potential to learn composite information and explain novel combinations not seen in the training data.

8. REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[3] 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.

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[5] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Dont just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4971–4980.

[6] A. Agrawal, A. Kembhavi, D. Batra, and D. Parikh, "C-vqa: A compositional split of the visual ques-

tion answering (vqa) v1. 0 dataset," *arXiv preprint arXiv:1704.08243*, 2017.

[7] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[8] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems*, 2016, pp. 4743–4751.

[9] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[10] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3712–3722.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[13] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.

[14] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[15] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," *arXiv preprint arXiv:1505.01809*, 2015.

[16] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, 2014, pp. 595–603.

[17] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4155–4164.

[18] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2989–2998.

[19] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3483–3491. [Online]. Available: http://papers.nips.cc/paper/5775-learning-structured-output-representation-using-deep-conditional-generative-models.pdf

[20] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *Advances in Neural Information Processing Systems*, 2017, pp. 5756–5766.

[21] U. Jain, Z. Zhang, and A. G. Schwing, "Creativity: Generating diverse questions using variational autoencoders." in *CVPR*, 2017, pp. 5415–5424.

[22] C. Zhou and G. Neubig, "Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction," *arXiv preprint arXiv:1704.01691*, 2017.

[23] E. Dupont, "Learning disentangled joint continuous and discrete representations," *stat*, vol. 1050, p. 11, 2018.

[24] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in beta-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[25] X. Shen, H. Su, S. Niu, and V. Demberg, "Improving variational encoder-decoders in dialogue generation," *arXiv preprint arXiv:1802.02032*, 2018.

[26] A. Makhzani and B. J. Frey, "Pixelgan autoencoders," in *Advances in Neural Information Processing Systems*, 2017, pp. 1975–1985.

[27] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.