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.
{*nkvyas, srallaba, lmorishe, hovy, awb*} *@cs.cmu.edu*

ABSTRACT

Multimodal tasks require learning joint representation across modalities. In this paper, we present an approach to employ latent stochastic models for a multimodal task - image captioning. Encoder Decoder models with stochastic latent variables are often faced with optimization issues such as latent collapse preventing them from realizing their full potential of rich representation learning and disentanglement. 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 such as Audio Visual Speech Recognition [1], Visual Question Answering [2], Video Transcription [3], Translation [4], etc are AI complete in some capacity and therefore need to deal with the challenges of Representation Learning, Translation, Alignment, Fusion and Co-learning [5] of the modalities present. Such tasks are also deceptively non trivial - they tend to give a false illusion of having learnt visually grounded representations [6]. Traditional encoder-decoder architectures for such tasks have shown to learn biases present in the data [7, 8]. Such models fail to learn robust representations, and do not generalize to unseen compositions of the seen objects [9]. In addition, such models are easily prone to adversarial attacks [10, 11, 12, 13]. In this paper, we present an initial approach to incorporate and learn latent stochastic random variables in encoder decoder models [14, 15, 16] for such multimodal tasks using image captioning as a case study.

Specifically, we investigate the ability of latent stochastic encoder decoder models to learn disentangled representations. Disentangled representations are defined as ones where a change in a single unit of the representation corresponds to a change in single factor of variation of the data while being invariant to others [17]. Such representations are attractive from the perspective of generalizabilty across tasks [18], zero shot learning[19], transfer learning 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 are more interpretible [20].

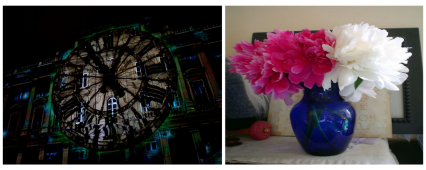

Fig. 1. (a) **Ground Truth**: A Gigantic clock is displayed on the side of a building. **Proposed Model**: a very tall clock with roman numerals on a wall. (b) **Ground Truth**: A small blue glass vase on a table. **Proposed Model**: A vase filled with pink roses on top of a table.

For image captioning, the deployed models are first expected to summarize both global information like objects and their positions in an image and local information like attributes and relation with other objects. Further, the models are required to generate factual and grammatically meaningful text descriptions. We hypothesize that latent stochastic models provide a flexible framework for address the challenges involved in such generative tasks. These models provide a mechanism to jointly train both the latent representations as well as the downstream inference network. They are expected to both discover and disentangle causal factors of variation present in the distribution of original data, so as to generalize at inference time. We believe that disentanglement is an important property for such tasks as it can improve the ability of models to generate new concepts by combining different global and local properties (see Figure 1). Due to the nature of challenges involved and the flexible framework of deep-latent models, we employ image captioning using latent stochastic models as the testbed for our experiments in this study.

*These two authors contributed equally

While training latent stochastic models, optimizing the exact log likelihood can be intractable. To address this, a recognition network is employed to approximate the posterior probability using reparameterization [21]. When deployed in encoder decoder models, this approach is often subject to an optimization challenge referred to as KL-collapse [22] - wherein the generator (usually an RNN) marginalizes the learnt latent representation. Typical approaches to dealing this issue involve annealing the KL divergence loss[22, 23], weakening the generator [24] and ensuring the recall using bag of words loss.

In this work, we present a method to incorporate inductive bias into latent stochastic models by forcing the prior distribution to be slightly more complex compared to the univariate Gaussian distribution typically employed. Specifically, we propose to split the latent prior space used for approximating the posterior distribution into continuous and discrete counterparts. This is motivated by the observation that tasks involving multiple modalities usually inherently contain both continuous and discrete factors that are responsible for the generation of observed data. In the context of caption generation both the involved modalities - textual even though primarily symbolic and visual even though primarily spatial - are characterized by distinct discrete and continuous factors of variation. For instance, distinct objects or entities would intuitively perhaps be better represented by discrete variables, while their spatial location and relationships between them might be represented by continuous variables. Based on this hypothesis, we constrain the latent prior space to include both continuous as well as discrete variables, thus forcing the model to encode important information into the latent representation, and subsequently forcing the generator to use this information during inference. 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.

2. PROPOSED APPROACH

In this section, we first present a brief analysis of the relationship between disentanglement of causal factors of variation and the optimization in latent stochastic models. Based on this analysis, we next present our approach to split the prior space into continuous and discrete components in such models.

2.1. Analysis of optimization and disentanglement

Latent stochastic models have shown promising results in unsupervised, unimodal settings and are the preferred models for learning disentangled representations of the causal factors of variation in the data. However, when they are combined with powerful generative models as decoders, optimization becomes harder due to KL-vanishing [22, 16]. To illustrate this, consider the following decomposition of the unregularized Variational Lowerbound (ELBO) being optimized by a vanilla Conditional Variational Encoder Decoder(CVED) framework:

$$logp_\theta(y|c,z) = logp_\theta(y|z,c) + logp_\theta(z|c) \quad (1)$$

where $logp_\theta(y|c,z)$ is approximated by an RNN, z is the latent representation approximated by the posterior network and c is the conditioning. In the context of image captioning, c typically corresponds to the spatial feature representation learnt from the image and z is approximated by performing reparameterization on the posterior estimates from the encoder. It can be observed from the decomposition though that the optimal value of this lowerbound estimate can be conditionally independent of the latent representation(z) and therefore, there is a possibility for the model to marginalize it entirely[25]. This becomes even more apparent if we consider the KL divergence between the approximate posterior and the assumed prior distribution. The divergence is expected to act as a regularizer thus forcing the model place information into the latent representation. When optimization is performed in expectation over minibatches, the KL divergence is the upper bound on the mutual information that can be encoded into the latent representation [26]. Therefore, reaching the global optimum for the divergence term, which is 0, effectively translates to limiting the amount of information the posterior network can encode into the latent representations. Thus, on the one hand while deep latest stochastic models are flexible and powerful, they are also incentivized to trivially ignore the latent representation.

Models such as β VAE[27] and the subsequently proposed channel capacity based approaches [28] aim to address this issue by gradually increasing the channel capacity, thus resulting in pressurizing the posterior distribution to match the prior closely. However, following such an approach translates to an unreasonable constraint in scenarios that have categorical output distributions. In other words, it seems impractical to assume that the true prior that generates latent distribution is a continuous Gaussian when the likelihood is based on discrete sequential data as in the contexts of language modeling, machine translation and image captioning. We hypothesize that a more reasonable constraint is to assume that the prior distribution is a mixture of discrete and continuous variables. This implicitly makes the prior space more complex compared to a univariate Gaussian distribution. The decoder is naturally weakened and is forced to encode only the information required to generate a coherent grammatical structure, while the remaining specific information such as the identity of objects and their relationships are encoded in the latent representation. This leads to, as a byproduct, disentanglement of

independent factors of variation in the original data. However, the weakened decoder model might not be able to accurately re-compose the learnt low dimensional simplistic representations into high dimensional and rich structured natural data with both independent as well as interdependent causal factors of variation. Therefore, optimization in latent stochastic models follows a compromise between the capability for reconstruction and the potential for disentanglement.

2.2. Concrete Conditional Variational Encoder Decoder

In this subsection we present our proposed model - Concrete Conditional Variational Encoder Decoder (CCVED). The objective of CCVED is to model the probabilistic generative process of discrete sequential data (here, captions), conditioned on spatial information (here, image) and the prior space (z). Based on the analysis in Section 2.1, we assume the z to be a mixture of continuous (z_c) and discrete random variables (z_d).

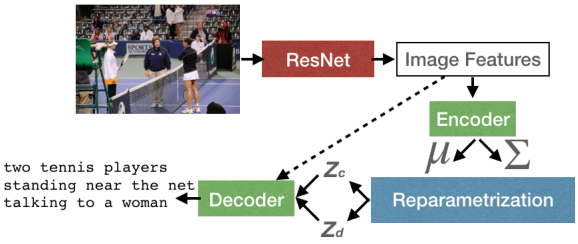

Fig. 2. The latent representation space in the proposed model is split into continuous (z_c) and discrete (z_d) prior space.

The procedure is as follows:

- Extract features from images using pretrained ResNet [29].
- Encode the extracted features to generate mean (μ) and diagonal variance (Σ) for the posterior distribution
- Sample N *i.i.d* latent continuous variables $z_c = \{z^n\}_{n=1}^N$ from μ and Σ via reparameterization trick [21].
- Sample N *i.i.d* latent discrete variables $z_d = \{z^n\}_{n=1}^N$ from μ and Σ via Gumbel argmax trick [30].
- Encoded image features are fed into the decoder in the first timestep. We then combine z_c, z_d with captions using a gated combination and feed to the decoder in the subsequent timesteps.

At inference, we use μ as z_c and argmax of posterior as z_d. Image features are fed at the first time step. Then z_c and $z-d$ are combined with Start token and fed as second timestep, Subsequent words for the caption are generated at each timestep, until end token is generated. As we are interested in learning the conditional distribution $P(caption|z, image)$, we factorize the encoder to learn $P(z|image)$, and not the joint distribution $P(z|image, captions)$.

Fig. 3. Examples of generated captions across models (Blue words represent generated concepts that are factual, but not in the gold caption. Green words represent generated concepts that are present in the gold. Red words represent non-factual concepts.)

2.3. Baseline Models

Base System (RNN): This model uses an RNN based encoder and decoder. The input to the encoder are pretrained ResNet features which are converted into a representation vector of fixed size. The input to decoder at each timestep is a stacked vector of the caption-word embedding and encoded image representation. The model is trained using teacher forcing and cross-entropy loss.

Latent Stochastic Baseline Model (VED): This model uses a similar RNN based model as described above. Specifically, we designed our encoder model to output the mean (μ) and log variance (Σ) of the latent distribution. We then sample a latent representation (z) using reparameterization trick [21]. The input to decoder at each timestep is a stacked vector of the caption-word embedding and encoded latent image representation. The model is trained with scheduled annealing using logistic function (step size of 2500) for KL divergence [22, 23].

3. EXPERIMENTAL SETUP

Dataset: We conduct our experiments using the challenging MS COCO (2014) dataset [31], which has 82,783 images and was generated using human subjects on the Amazon Mechanical Turk (AMT). We used the NLTK tokenizer for the captions and limit the vocabulary to include words that occur atleast 10 times. The final vocabulary size was 8855. We do not

System	BLEU 4	METEOR	CIDER	ROUGE L
RNN Baseline	12	0.15	0.32	0.38
VED Baseline	13	0.15	0.33	0.40
CCVED (Ours)	**16**	**0.18**	**0.49**	**0.43**

Table 1. Performance comparison across models

N-grams	Gold	RNN Baseline	VED Baseline	CCVED (Ours)
a man sitting	23	3716	40	**19**
a dog	400	**694**	1010	714
a woman sitting	11	2508	51	**21**

Table 2. Count of n-grams that appear at start of caption

repartition the training and validation sets for MS COCO to increase the training data since we wanted to test the ability of the models to generalize to novel combinations.

Evaluation Metrics: We report the performance of our proposed approach as well as the baseline models using BLEU, a measure that loosely corresponds to precision of word n-grams between hypothesis and reference sentences. Additionally we also report the results based on METEOR, ROUGE and CIDEr.

Hyperparameters: Hyperparameters across all our experiments were kept constant. z, z_d and z_c were fixed to 128 dimensions. 512 dimensions were used for hidden. Adam was used for optimization with a learning rate of 0.001. Epsilon value of 1e-12 was used for Gumbel argmax. We use minimal KL-annealing with logistic function between 300 and 2500 steps. All our models use greedy decoding (beam size=1).

4. ANALYSIS

We observe that our proposed approach of using both continuous and discrete variables for representing latent space has consistent gains across different metrics, as compared to the baseline models (see Table 1). RNN based models optimize the likelihood objective via cross entropy loss. This biases the decoder to over-generate n-gram patterns that occur more frequency in training data, leading to non-factual captions. On the contrary, our proposed approach optimizes KL-divergence that outperforms the baseline models in estimating the prior n-gram distribution (see Table 2).

Captions generated by our proposed model capture more details than the baseline models (see Figure 3). The model is evidently able to disentangle the learnt properties and create new abstract concepts at inference time. As a result, our proposed model generates more diverse and relevant captions compared to the baseline models. For example, the model generates *stone structure* for describing the *building* in the image. The model is also able to map similar properties to each other. For example, the model learns *leaning* and *sitting* fall into the same semantic space. However, our proposed

Fig. 4. Counting errors in generated captions (a) a plate with a sandwich and three sandwiches (b) a number of horses on a beach near the water (c) four guys relaxing on a narrow sofa

Fig. 5. Common sense errors in generated captions. (a) a man in a giraffe has a branch pinned between his ear (b) a black man unk a fish under a framed view of the unk

model is also prone to errors. We observed that our model is weak at counting (see Figure 4(a) and (c)). Sometimes, it produces factually correct, but more general words like *many* and *number of* to denote multiple objects in the image (see Figure 4(b)). Unfortunately, this is penalized by the evaluation metric. Another shortcoming of our proposed model is its lack of common sense knowledge. This leads to generation of bizarre captions. For example in Figure 5(a), the branch in background is visible from in between the giraffes ear, and is not pinned between his ear. In Figure 5(b), the model assumes the reflection of a man in black-suit on the window is a black-man standing. Nevertheless the model is able to create novel concepts like *pinned in between something*.

5. CONCLUSION

Multimodal problems like caption generation require learning representation across modalities. In this work, we proposed an approach to incorporate joint continuous and discrete representation in the prior distribution. Our model learns better representations, and generalizes well on unseen data. It outperforms baseline models on a multitude of metrics, and is able to generate more detailed, relevant and diverse captions. In future, we would like study this module in other zero-shot learning tasks.

Acknowledgements

We would like to thank Louis-Philippe Morency, Ying Shen, and Bhavya Karki for their valuable comments.

6. REFERENCES

[1] T. Afouras, J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *CoRR*, vol. abs/1809.02108, 2018.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," *CoRR*, vol. abs/1505.00468, 2015.

[3] S. Chen, J. Chen, and Q. Jin, "Generating video descriptions with topic guidance," *CoRR*, vol. abs/1708.09666, 2017.

[4] Y. Su, K. Fan, N. Bach, C. J. Kuo, and F. Huang, "Unsupervised multi-modal neural machine translation," *CoRR*, vol. abs/1811.11365, 2018.

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

[6] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi, "FOIL it! find one mismatch between image and language caption," *CoRR*, vol. abs/1705.01359, 2017.

[7] 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.

[8] 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.

[9] A. Agrawal, A. Kembhavi, D. Batra, and D. Parikh, "C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset," *arXiv preprint arXiv:1704.08243*, 2017.

[10] H. Chen, H. Zhang, P. Chen, J. Yi, and C. Hsieh, "Show-and-fool: Crafting adversarial examples for neural image captioning," *CoRR*, vol. abs/1712.02051, 2017.

[11] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *CoRR*, vol. abs/1810.00069, 2018.

[12] Y. Zhao, H. Zhu, Q. Shen, R. Liang, K. Chen, and S. Zhang, "Practical adversarial attack against object detector," *CoRR*, vol. abs/1812.10217, 2018.

[13] X. Yuan, P. He, Q. Zhu, R. R. Bhat, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *CoRR*, vol. abs/1712.07107, 2017.

[14] 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.

[15] 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.

[16] 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.

[17] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," *CoRR*, vol. abs/1206.5538, 2012.

[18] B. Esmaeili, H. Wu, S. Jain, S. Narayanaswamy, B. Paige, and J.-W. Van de Meent, "Hierarchical disentangled representations," *arXiv preprint arXiv:1804.02086*, 2018.

[19] I. Higgins, A. Pal, A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner, "Darla: Improving zero-shot transfer in reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1480–1490.

[20] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

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

[22] 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.

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

[24] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," *arXiv preprint arXiv:1703.10960*, 2017.

[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] "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[28] 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.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv preprint arXiv:1611.00712*, 2016.

[31] 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.