Attend, tell and ground: Weakly-supervised Object Grounding with Attention-based Conditional Variational Autoencoders

Effrosyni Mavroudi and René Vidal
Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD
{emavrou1, rvidal}@jhu.edu

Abstract

We address the problem of weakly-supervised object grounding, i.e., learning how to align multiple words in a sentence with visual regions using only image-caption pairs. Recent approaches use captioning as a downstream task to guide object grounding, i.e., extract region proposals, attend over them to predict the next word and then ground the word by selecting the most attended regions. However, attention coefficients are computed without knowing the word that needs to be localized. To address this shortcoming we propose a novel grounded captioning framework based on Conditional Variational Autoencoders (CVAEs). In particular, we introduce a discrete random variable modeling the alignment of a word and a region, and learn its posterior distribution conditioned on the word to be grounded. Furthermore, to ensure our latent variables capture meaningful alignments, we propose a modified CVAE training objective to mitigate the issue of the posterior alignment distribution collapsing to the prior, which often arises when training CVAEs with language models. Experiments on the challenging Flickr30k Entities dataset validate the effectiveness of the different components of our framework and show that it can substantially outperform soft-attention-based baselines in grounding.

1 Introduction

This paper studies the *visual object grounding problem*, where given an input image and a sentence describing it, the goal is to find where the referred entities (actors, objects) appear in the image. Addressing this task, i.e., linking words to regions, is critical for developing autonomous systems that can effectively interact with humans by comprehending instructions (Alomari et al., 2017; Hu et al., 2020). Training visual grounding systems typically requires annotations of textual descriptions combined with bounding boxes for each entity. Since constructing datasets with such fine-grained bounding box annotations is rather time-consuming and costly, we focus on weakly-supervised visual grounding which requires only textual descriptions.

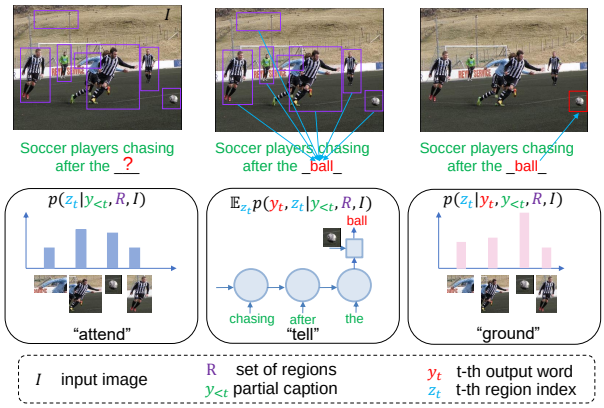

Figure 1: Overview of our proposed GVD-CVAE framework for weakly-supervised object grounding in images. We model word-to-region alignments as a sequence of discrete latent variables Z in a generative model of sentences. Given a partial caption, our model is able to *attend* over the candidate region proposals, *tell* the next word by marginalizing out the latent word-to-region alignments from the joint distribution and *ground* the word by leveraging the learned prior and approximate posterior alignment distributions.

Weakly-supervised learning from textual descriptions by using grounded captioning as a surrogate supervision task is a possible way to alleviate the need for bounding box annotations. The idea is to learn how to ground words by learning how to generate sentences based on detected regions. To do so, we can leverage soft-attention mechanisms used in grounded captioning encoder-decoder models and select regions with maximum attention coefficients (Zhou et al., 2019; Ma et al., 2020; Liu et al., 2020). Given a partially generated sentence, these attention coefficients are computed for each input region and the attention-weighted sum of region features is utilized as relevant visual context for predicting the next word.

However, exploiting soft-attention as a grounding mechanism is restricted by two major limitations. First, despite being an effective, end-to-end learnable pooling mechanism for summarizing variable-length inputs to guide sequential generation tasks, attention is not encouraged to capture meaningful alignments, unless it is supervised (Liu et al., 2017; Zhou et al., 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019). More importantly, attention coefficients are computed **before** predicting the word to be grounded. As a motivating example, consider the task of grounding the words *'hat'* and *'jacket'* given the two sentences: *"A man is wearing a hat"* and *"A man is wearing a jacket"*. Soft-attention-based grounding will predict the same box for 'hat' and 'jacket', given that the language context is the same ("A man is wearing a"). We address these limitations by learning the posterior distribution of word-to-region alignments given the word to be grounded using our proposed Grounded Visual Description CVAE (GVD-CVAE), as shown in Fig. 1.

2 Method

Our approach builds upon an earlier image description paradigm (Xu et al., 2015; Pedersoli et al., 2017), which uses a latent-variable probabilistic model, $p(Y \mid R, I) = \sum_Z p_\theta(Y, Z \mid R, I)$, for sentence (sequence of words) $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ given an image I and region proposals R. In this model, the sequence of word-to-region alignments is modeled as a sequence of discrete latent variables $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$, where $z_{t,i} = 1$ when the i- th region proposal is used for generating the t-th word $\mathbf{y}_t$ and $z_{t,i} = 0$ otherwise. In particular, assuming that the t-th word depends only on the region $\mathbf{z}_t$ given the partial caption $\mathbf{y}_{1:t-1}$ and that the region-to-word alignments $\mathbf{z}_t$ for each word are independent with each other conditioned on the partial caption, our joint probability distribution $p_\theta(Y, Z \mid R, I)$ takes the form:

$$\prod_{t=1}^{T} \overbrace{p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_t, R, I)}^{\text{language decoder}} \overbrace{p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, R, I)}^{\text{region prior}}. \quad (1)$$

Encoder. We adopt the neural network models used in GVD (Zhou et al., 2019) to encode the image in a global feature vector $\mathbf{v}$, represent words in the vocabulary $\mathcal{V}$ with learnable embeddings $\text{emb}(\mathbf{y}_t)$, and represent the i-th region with grounding-aware region encoding $\mathbf{x}_i$.

Decoder model. The distribution over words in the vocabulary is conditioned on the partial caption, the aligned region and the input image. The dependence of the current word on the sentence generated so far ($\mathbf{y}_{<t}$) and on the input image I is captured in an LSTM hidden state $\mathbf{s}_t$. The dependence of the word $\mathbf{y}_t$ on the aligned region with $z_{t,j} = 1$ is modeled by feeding $\mathbf{x}_j$ concatenated with $\mathbf{s}_t$ to a classifier. The output of the decoder network yields the parameters $g_\theta(\mathbf{s}_t, \mathbf{z}_t, \mathbf{x}) \in \mathbb{R}^{|\mathcal{V}|}$ of the multinomial word distribution $p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_t, R, I)$:

$$\mathbf{s}_t = \text{RNN}_\theta(\mathbf{s}_{t-1}, [\mathbf{v}; \text{emb}(\mathbf{y}_{t-1})]), \quad (2)$$

$$g_\theta(\mathbf{s}_t, \mathbf{z}_t, \mathbf{x}) = \text{softmax}\left(W_c \left[\sum_{i=1}^{M} z_{t,i}\mathbf{x}_i; \mathbf{s}_t\right]\right). \quad (3)$$

Prior model. The prior $p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, R, I)$ is a multinomial distribution over possible word-to-region alignments. We parameterize it using a standard attention mechanism that computes coefficients $\alpha_\theta(\mathbf{s}_t, \mathbf{x}) \in \mathbb{R}^M$ given the query $\mathbf{s}_t$, encoding the partial caption and image, and the regions $\mathbf{x}$:

$$\alpha_\theta^{(i)}(\mathbf{s}_t, \mathbf{x}) = \frac{\exp\left(\mathbf{u}^T \tanh(W_a[\mathbf{s}_t; \mathbf{x}_i])\right)}{\sum_{j=1}^{M} \exp\left(\mathbf{u}^T \tanh(W_a[\mathbf{s}_t; \mathbf{x}_j])\right)}. \quad (4)$$

To learn the parameters of our conditional generative model we leverage Amortized Variational Inference (AVI). Therefore, our model becomes a Conditional Variational Autoencoder (Sohn et al., 2015) (CVAE) with sequential discrete latent space and sentences as observations. In the CVAE framework, a variational distribution $q_\phi(Z|Y, R, I)$ is introduced to approximate the true posterior and is parameterized via a neural network with parameters ϕ, also known as the "inference network".

Inference model. We choose to approximate the true posterior with a filtering approximate posterior: $q_\phi(Z \mid Y, R, I) = \prod_{t=1}^{T} q_\phi(\mathbf{z}_t \mid \mathbf{y}_{\leq t}, R, I)$. The dependence of the word-to-region alignment $\mathbf{z}_t$ on the caption up to and **including the current word** $\mathbf{y}_{\leq t}$ is encoded in the hidden state of a separate LSTM network that takes as input the current word $\mathbf{y}_t$ at each timestep:

$$\mathbf{h}_t = \text{RNN}_\phi(\mathbf{h}_{t-1}, [\mathbf{v}; \text{emb}(\mathbf{y}_t)]). \quad (5)$$

Then, the parameters $\alpha_\phi(\mathbf{h}_t, \mathbf{x}) \in \mathbb{R}^M$ of the categorical approximate posterior distribution can be obtained by another learnable attention module.

Training. During training, we are given a dataset consisting of N image-sentence pairs. To train our GVD-CVAE we minimize the following hybrid objective w.r.t. the parameters θ and ϕ (omitting the conditioning of all distributions on $I^{(n)}$ for readability):

$$\mathcal{L} = \frac{1}{N}\sum_{n,t} \lambda \mathcal{L}_{CVAE}(n,t) \tag{6}$$
$$+ (1-\lambda)\log p_\theta(\mathbf{y}_t^{(n)} \mid \mathop{\mathbb{E}}_{\mathbf{z}_t \sim p_\theta}[\mathbf{z}_t], \mathbf{y}_{<t}^{(n)}, R^{(n)}),$$

where

$$\mathcal{L}_{CVAE} = \overbrace{\mathop{\mathbb{E}}_{\mathbf{z}_t \sim q_\phi}\Big[-\log p_\theta(\mathbf{y}_t^{(n)} \mid \mathbf{y}_{<t}^{(n)}, \mathbf{z}_t, R^{(n)})\Big]}^{\text{word reconstruction loss}}$$
$$+\beta \mathrm{KL}(q_\phi(\mathbf{z}_t \mid \mathbf{y}_{\leq t}^{(n)}, R^{(n)}) \,||\, p_\theta(\mathbf{z}_t \mid \mathbf{y}_{<t}^{(n)}, R^{(n)}). \tag{7}$$

For $\lambda = 1$ and $\beta = 1$, we recover the Evidence Lower Bound Objective (ELBO) resulting from our factorization of the joint probability distribution and our choice of the approximate posterior. Similar to prior work, we observe that optimizing the ELBO often results in an inference model that produces posteriors almost identical to the prior, which translates to word-to-region alignments that do not take into account the word to be grounded. To mitigate this phenomenon of "posterior collapse", we propose to re-weight the KL loss term with a scalar factor β, which starting from 0 is gradually increased up to a value $\beta_{clip} < 1$ during training.

For $\lambda = 0$, only the decoder and prior networks are trained and region samples are replaced with their expected value according to the prior distribution, i.e. a word is predicted based on the language context and the region context $\sum_{i=1}^{M} \alpha_\theta^{(i)}(\mathbf{s}_t, \mathbf{x})\mathbf{x}_i$. Thus, we end up training a discriminative, attention-based encoder-decoder captioning model, where the attention module of the prior model plays the role of a soft-attention mechanism. This serves as our baseline. Moreover, jointly optimizing the CVAE and baseline loss facilitates training by regularizing the prior attention coefficients.

To optimize the final hybrid objective using Stochastic Gradient Descent, we approximate the reconstruction loss term using Monte-Carlo samples, with S region samples drawn from the Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) continuous approximation of the categorical distribution q_ϕ with temperature τ.

3 Related Work

Early attempts for weakly-supervised grounding of multiple entities in images and videos relied on graphical models (Yu and Siskind, 2013; Ramanathan et al., 2014). Another line of work formulates grounding as a Multiple Instance Learning problem, using image retrieval as supervision (Karpathy and Fei-Fei, 2017), but is not able to generate descriptions. Recently, researchers have proposed grounding based on the attention coefficients of grounded captioning models (Zhou et al., 2019; Ma et al., 2020; Liu et al., 2020). Instead, we propose treating word-to-region alignments as discrete latent variables and parameterize prior/posterior alignment distributions with attention mechanisms, inspired by discrete latent-variable models for image captioning/neural machine translation (Xu et al., 2015; Pedersoli et al., 2017; Deng et al., 2018; Shankar and Sarawagi, 2019). Our proposed CVAE-based captioning model is also related to CVAEs developed for modeling sequential data and particularly those with sequential latent variables (Chung et al., 2015; Goyal et al., 2017; Serban et al., 2017; Chu et al., 2018; Aneja et al., 2019; Graber and Schwing, 2020), instead of a single latent variable initiating the sequence generation (Wang et al., 2017; Bhattacharyya et al., 2018; Pagnoni et al., 2018; Zhao et al., 2017). However, the majority of those models have non-interpretable, continuous latent random variables, are not parameterized with attention mechanisms and generally differ in their choice of prior, approximate posterior and decoder. In addition, their goal is to model the sequence likelihood, while we propose exploiting the latent variables for grounding, and thus need to overcome additional challenges, such as posterior collapse (Alemi et al., 2018; Fu et al., 2019; Dieng et al., 2020).

4 Results and Discussion

Dataset. We evaluate our method on the Flickr30k Entities dataset, which provides captions describing images and bounding boxes associated with noun phrases in the captions. Since we are operating on the weakly-supervised grounding regime, we ignore bounding box annotations during training. We evaluate word grounding (instead of phrase grounding (Rohrbach et al., 2016)), following Zhou et al. (2019). Specifically, we evaluate based on word-to-region alignments of 480 groundable words out of the 8639 vocabulary words.

	Inputs	Captioning				Grounding		
						GT	Generated	
		B@4	M	C	S	Acc.	$F1_{all}$	$F1_{loc}$
GVD (Sup.)	G	27.3	22.5	62.3	16.5	41.4	7.55	22.2
GVD	G	26.9	22.1	60.1	16.1	21.4	3.88	11.7
GVD (Grd)	G	26.9	22.1	60.1	16.1	25.5	3.88	11.7
Cyclical	G	26.6	22.3	60.9	16.3	-	4.85	13.4
DPA	G	27.6	22.6	62.7	16.7	-	4.79	15.5
BUTD	U	27.3	21.7	56.6	16.0	24.2	-	-
DPA	U	27.2	22.3	60.8	16.3	-	5.45	15.3
Sub-GC	S	28.5	22.3	61.9	16.4	-	5.98	16.53
POS-SCAN†	GP	28.0	**22.6**	66.2	**17.0**	-	6.53	15.79
POS-SCAN†	UP	**30.1**	22.6	**69.3**	16.8	-	**7.17**	**17.49**
Baseline ($\lambda = 0$)	G	24.3	20.9	54.4	15.5	21.5	4.25	13.37
GVD-CVAE	G	25.4	21.6	56.1	15.8	**27.7**	5.42	15.30

Table 1: **Results on the Flickr30k Entities test set**. The performance of the fully-supervised GVD model (Sup.) is reported as an upper-bound to the weakly-supervised approaches. Types of model inputs: regions encoded following **G**VD or B**U**TD, **P**OS (part-of-speech) tokens, **S**cene-graphs. † denotes models trained with an additional captioning reinforcement learning loss. B: Bleu, M: METEOR, C: CIDEr, S: SPICE. GVD-CVAE results are averaged across 5 runs.

Metrics. Predicted boxes with IoU above 0.5 with a ground-truth box are considered correct. Grounding performance given GT sentences is then measured using box accuracy averaged over object classes. Additionally, to compare with state-of-the-art methods, we also evaluate our model on the downstream task (grounded captioning) using our decoder and prior. Grounding on generated sentences is evaluated using F1-score metrics (Zhou et al., 2019), to account for existing objects not included in the generated caption. Captioning is evaluated using standard metrics, such as CIDEr.

Results. We use the region proposals and pre-trained region/image features from (Zhou et al., 2019), with 100 region proposals per image. We train our model for 40 epochs with the ADAM optimizer, having an initial learning rate of $5e{-}4$, decayed by 0.8 every 3 epochs. Our batch size is 80 and $S = 10$, $\tau = 0.8$, $\beta_{clip} = 0.2$.

In Table 1, we compare our baseline and GVD-CVAE to state-of-the-art methods (GVD (Zhou et al., 2019), BUTD (Anderson et al., 2018), Cyclical (Ma et al., 2020), DPA (Liu et al., 2020), Sub-GC (Zhong et al., 2020), POS-SCAN (Zhou et al., 2020)). Exploiting our learned approximate posterior alignment, our method achieves the state of the art in weakly-supervised object grounding. It yields a relative improvement of 29% upon the soft-attention baseline (21.5% to 27.7% $\pm$ 1.12%). Our model also outperforms both Cyclical Attention and DPA - despite us having a weaker, single LSTM language model - in generated sentence grounding (5.4% $\pm$ 0.27% vs 4.8%). Hence, modeling word-to-region alignments as latent variables in our deep conditional generative model leads to better grounding than adopting attention regularization techniques during training.

In Table 2 we study the impact of the training objective and motivate the need for our proposed hybrid objective ($\lambda = 0.5$), instead of the vanilla CVAE objective (ELBO). Finally, Fig. 2 illustrates two representative object grounding cases, where selecting regions based on our learned prior or approximate posterior alignment distributions yields better grounding results than using the soft-attention coefficients.

Training objective	CVAE-p	CVAE-q
Cross-entropy (CE)	21.5	-
ELBO	3.29	3.16
CE + ELBO	25.22	23.99
CE + ELBO + β anneal	26.07	25.61
CE + ELBO + β anneal + clip	**26.31**	**28.88**

Table 2: Impact of various training objectives on weakly-supervised object grounding. Performance measured via Box accuracy (%) on the Flickr30k Entities validation set. *CVAE-p* denotes box accuracy obtained using the learned prior alignment distribution, while *CVAE-q* using the approximate posterior.

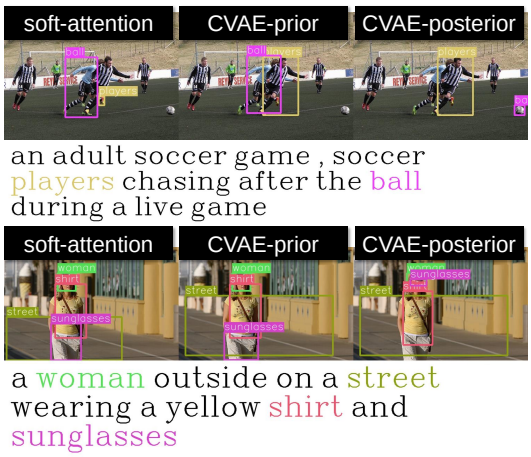

Figure 2: Qualitative comparison of weakly-supervised object grounding results obtained by the baseline and our GVD-CVAE on images from Flickr30k Entities. For each caption, we show three copies of the image with grounding results obtained by the soft-attention baseline, our prior and posterior alignment distributions, respectively. Best viewed zoomed in and in color.

5 Acknowledgements

The authors thank Carolina Pacheco, Ambar Pal, Benjamin Béjar Haro for helpful discussions in early stages of this work. Research supported by IARPA DIVA contract D17PC00345.

References

Alexander A. Alemi, Ben Poole, Ian Fische, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. 2018. Fixing a broken elbo. In *International Conference on Machine Learning*, volume 1.

Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *AAAI Conference on Artificial Intelligence*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019. Sequential latent spaces for modeling the intention during diverse image captioning. In *IEEE International Conference on Computer Vision*, pages 4260–4269.

Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. 2018. Accurate and diverse sampling of sequences based on a 'best of many' sample objective. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Hong Min Chu, Chih Kuan Yeh, and Yu Chiang Frank Wang. 2018. Deep generative models for weakly-supervised multi-label classification. In *European Conference on Computer Vision*, pages 409–425. Springer Verlag.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Neural Information Processing Systems*, volume 2015-January.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M. Rush. 2018. Latent alignment and variational attention. In *Neural Information Processing Systems*, volume 2018-December.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and David M. Blei. 2020. Avoiding latent variable collapse with generative skip models. In *International Conference on Artificial Intelligence and Statistics*.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Anirudh Goyal, Alessandro Sordoni, Marc Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. 2017. Z-forcing: Training stochastic recurrent networks. In *Neural Information Processing Systems*, volume 2017-December.

Colin Graber and Alexander G. Schwing. 2020. Dynamic neural relational inference. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2020. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Annual Meeting of the Association for Computational Linguistics*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *AAAI Conference on Artificial Intelligence*.

Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. 2020. Prophet attention: Predicting attention with future attention. In *Neural Information Processing Systems*.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded visual captions without localization supervision. In *European Conference on Computer Vision*.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Artidoro Pagnoni, Kevin Liu, and Shangyan Li. 2018. Conditional variational autoencoder for neural machine translation.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *IEEE International Conference on Computer Vision*.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with "their" names using coreference resolution. In *European Conference on Computer Vision*, volume 8689 LNCS.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI Conference on Artificial Intelligence*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Shiv Shankar and Sunita Sarawagi. 2019. Posterior attention models for sequence to sequence learning. In *International Conference on Learning Representations*.

Kihyuk Sohn, Xinchen Yan, and Honglak Lee. 2015. Learning structured output representation using deep conditional generative models. In *Neural Information Processing Systems*.

Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Neural Information Processing Systems*, volume 2017-December.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Annual Meeting of the Association for Computational Linguistics*, volume 1.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Annual Meeting of the Association for Computational Linguistics*.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J. Corso, and Marcus Rohrbach. 2019. Grounded video description. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6571–6580. IEEE Computer Society.

Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4776–4785.