JMLR: Workshop and Conference Proceedings 1–6

Latent Alignment and Variational Attention

Elena Orlova, Vadim Kuzmin, Roman Tsukanov

1. Introduction

Attention mechanism has became central to many state-of-the-art models in natural language processing and computer vision. Neural networks with attention may be considered as a method for softly simulating alignment; however, the approach does not marginalize over latent alignments in a probabilistic sense. A related latent approach, hard attention, fixes these issues, but is generally harder to train and less accurate. In a paper Deng et al. (2018), the authors considers variational attention networks, alternatives to soft and hard attention for learning latent variable alignment models, with tighter approximation bounds based on amortized variational inference.

By reproducing the key experiments of the paper we show that variational attention is a successful method for learning latent variable alignment models, based on amortized variational inference. It outperforms soft attention and using policy-based/REINFORCE optimization allows to train the model in reasonable time.

2. Problem statement and definitions

Firstly, let us introduce some definitions. Let x be an observed set, $\hat{x}$ an arbitrary "query". These generate a discrete output variable $y \in \mathcal{Y}$. This process is mediated through a latent alignment variable z, which indicates which member (or mixture of members) of x generates y. We consider the generative process

$$z \sim \mathcal{D}(a(x, \tilde{x}; \theta)) \quad y \sim f(x, z; \theta),$$

where a produces the parameters for an alignment distribution $\mathcal{D}$. The function f gives a distribution over the output, e.g. an exponential family. To fit this model to data, we set the model parameters θ by maximizing the log marginal likelihood of training examples

$$\max_{\theta} \log p(y = \hat{y}|x, \tilde{x}) = \max_{\theta} \log \mathbb{E}_z \left[f(x, z; \theta)_{\hat{y}} \right].$$

Directly maximizing this log marginal likeli- hood in the presence of the latent variable z is often difficult due to the expectation.

In order to overcome this issue, we will consider $\mathcal{D}$ as categorical distribution, i.e z is one-hot encoded vector. So, f can be represented as:

$$f(x, z) = \text{softmax}(\text{WXz})$$

Then, marginal log-likelihood is computed by this formula:

$$\log p(y = \hat{y}|x, \tilde{x}) = \log \sum_{i=1}^{T} p\left(z_i = 1|x, \tilde{x}\right) p\left(y = \hat{y}|x, z_i = 1\right) = \log \mathbb{E}_z \left[\text{softmax}(\mathbf{W}Xz)_{\hat{y}}\right]$$

Soft Attention Instead of using a latent variable, they employ a deterministic network to compute an expectation over the alignment variable:

$$\log p_{\text{soft}}\left(y|x, \tilde{x}\right) = \log f\left(x, \mathbb{E}_z[z]; \theta\right) = \log \text{softmax}\left(\mathbf{W}X\mathbb{E}_z[z]\right).$$

Hard Attention Hard attention is an approximate inference approach for latent alignment Ba et al. (2015), Mnih et al. (2014). Hard attention takes a single hard sample of z and then backpropagates through the model. The approach is derived by two choices: first apply Jensen's inequality to get a lower bound on the log marginal likelihood, then maximize this lower-bound with policy gradients/reinforce to obtain unbiased gradient estimates,

$$\nabla_\theta \mathbb{E}_z[\log f(x, z))] = \mathbb{E}_z\left[\nabla_\theta \log f(x, z) + (\log f(x, z) - B)\nabla_\theta \log p(z|x, \tilde{x})\right],$$

where B is a baseline that can be used to reduce the variance of this estimator. First note that the key approximation step in hard attention is to optimize a lower bound derived from Jensen's inequality. This gap could be quite large, contributing to poor performance. Variational inference methods directly aim to tighten this gap.

To approximate latent variable inference, amortized variational inference (AVI) Rezende et al. (2014). Mnih et al. (2014), that exploits learned inference networks, is used. The key approximation step in hard attention is to optimize a lower bound derived from Jensen's inequality. In AVI an inference network produces the parameters of the variational distribution $q(z; \lambda)$. With the right choice of optimization strategy and inference network this form of variational attention can provide a general method for learning latent alignment models. And the authors propose propose two algorithms: for categorical and relaxed alignments. We worked only with categorical attention.

Variational attention Amortized variational inference (AVI, closely related to variational auto-encoders) is a class of methods to efficiently approximate latent variable inference, using learned inference networks. We need to sse ELBO (Evidence Lower-Bound) over a family of distributions $q(z)$:

$$\log \mathbb{E}_{z\sim p(z|x,\tilde{x})}[p(y|x, z)] \geq \mathbb{E}_{z\sim q(z)}[\log p(y|x, z)] - \text{KL}[q(z)\|p(z|x, \tilde{x})]$$

Then, we use amortized variational inference to optimize ELBO. And, finally, AVI uses an inference network to produce the parameters of the variational distribution $q(z; \lambda)$. The inference network takes in the input, query, and the output, i.e. $\lambda = enc(x, \tilde{x}, y; \phi)$. The objective aims to reduce the gap with the inference network ϕ while also training the generative model θ

$$\max_{\phi,\theta} \mathbb{E}_{z\sim q(z;\lambda)}[\log p(y|x, z)] - \text{KL}[q(z; \lambda)\|p(z|x, \tilde{x})].$$

So, to accurately and efficiently compute this objective it is possible to use categorical alignments. Note that low-variance estimator of $\nabla_\theta ELBO$, is easily obtained through a

single sample from $q(z)$ and for $\nabla_\phi ELBO$, the gradient with respect to the KL portion is easily computable. Then, use REINFORCE Williams (1992) along with a specialized baseline to estimate gradient:

$$\nabla_\phi \mathbb{E}_{z\sim q(z)}[\log p(y|x,z)] = \mathbb{E}_{z\sim q(z)}\left[(\log f(x,z) - B)\nabla_\phi \log q(z)\right].$$

As with hard attention, we take a single Monte Carlo sample (now drawn from the variational distribution). Variance reduction of this estimate falls to the baseline term B. The ideal (and intuitive) baseline would be $\mathbb{E}_{z\sim q(z)}[\log f(x,z)]$, analogous to the value function in reinforcement learning. While this term cannot be easily computed, there is a natural, cheap approximation: soft attention (i.e. $\log f(x, \mathbb{E}[z])$). Then the gradient is

$$\mathbb{E}_{z\sim q(z)}\left[\left(\log \frac{f(x,z)}{f\left(x, \mathbb{E}_{z'\sim p(z'|x,\bar{x})}\left[z'\right]\right)}\right)\nabla_\phi \log q(z|x,\tilde{x})\right]$$

3. Experiments

We focus on Neural Machine Translation task and reproduce the key experiments to evaluate the presented results. We work with IWSLT German-English dataset. We implemented soft attention, variational categorical attention with exact ELBO and variational categorical attention with REINFORCE.

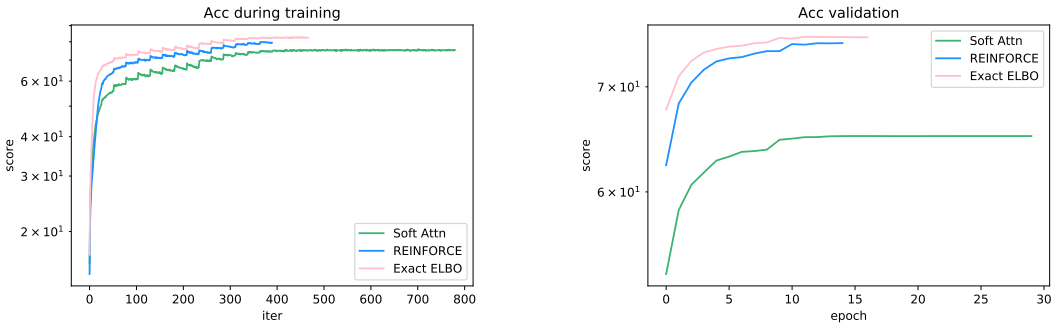
As for the network architecture, the encoder is a two-layer bi-directional LSTM with 512 units in each direction, and the decoder as a two-layer LSTM with with 768 units. For the decoder, the convex combination of source hidden states at each time step from the attention distribution is used as additional input at the next time step. Word embedding is 512-dimensional.

The inference network consists of two bi-directional LSTMs (also two-layer and 512-dimensional each) which is run over the source/target to obtain the hidden states at each time step. These hidden states are combined using bilinear attention to produce the variational parameters. Only the word embedding is shared between the inference network and the generative model.

4. Results

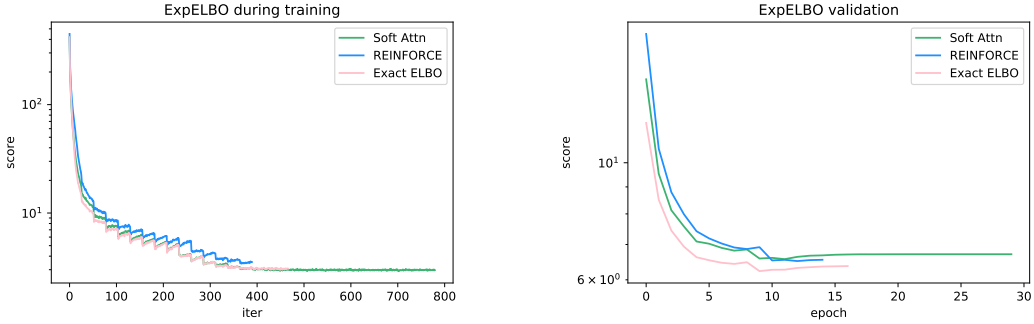
We performed set of experiments on translation task on IWSLT dataset (German to English). Due to computing resources limitation, we run soft mode for 30 epochs, mode with exact ELBO for 18, the mode with REINFORCE for 15. Also, note that model with attention with exact ELBO is the most computationally expensive and the model with REINFORCE is more expensive then soft attention, so, we run less epochs for these epoch. However, even in such case these models outperfoms the model with soft attention.

The accuracy scores are presented in Fig. 1, ELBO expectation are in Fig.2 and traditional NLP-metric perplexity is shown in Fig. 3. The plots illustrate that model with categorical attention with exact ELBO outperfoms other approaches.



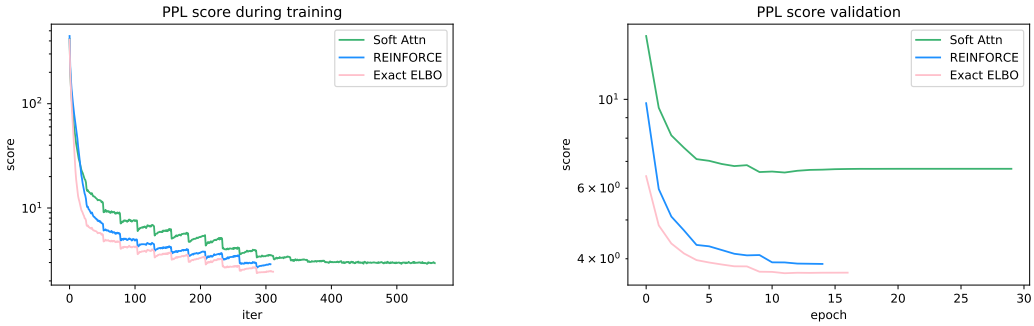
(*a*) Training (*b*) Validation

Figure 1: Accuracy of translation during training and validation.



(*a*) Training (*b*) Validation

Figure 2: Values of ExpELBO during training and validation.



(*a*) Training (*b*) Validation

Figure 3: Perplexity during training and validation.

Model	Best accuracy score on validation	Best perplexity score on validation
Soft attention	65.1	6.7
REINFORCE	74.5	3.9
Exact ELBO	75.1	3.7

As for the testing stage, we calculated BLEU score and provide translation examples in the Table below. Scores are of the same order, and, from the example, seems that all models produce similar texts.

Model	Text	BLEU score
Original (German)	wissen sie , eines der großen vernügen beim reisen und eine der freuden bei der ethnographischen forschung ist , gemeinsam mit den menschen zu leben , die sich noch an die alten tage erinnern können . die ihre vergangenheit noch immer im wind spüren , sie auf vom regen geglätteten steinen berühren , sie in den bitteren blättern der pflanzen schmecken .	-
English translation	you know , one of the intense pleasures of travel and one of the delights of ethnographic research is the opportunity to live amongst those who have not forgotten the old ways , who still feel their past in the wind , touch it in stones polished by rain , taste it in the bitter leaves of plants .	-
Soft mode	you know , one of the great people in the travel , and one of the joys in ethnographic research is to live together with the people who can still remember the old days , who still feel their past in the wind , touch them on the rain , taste them in the bitter leaves of the plants .	33.23
Exact ELBO	you know , one of the great chunks of travel and one of the pleasures in ethnographic research is to live together with the people who can remember the old days , who still feel their past still in the wind , touch them on the rains , taste them in the bitter flights of the plants .	33.63
REINFORCE	you know , one of the great reason in traveling , and one of the joys of ethnographic research is to live together with the people who are able to remember the old days , and that their past feels still in the wind , touching them on the rain , touching them on the biter leaves of the plants , taste them in the bitter pets of the plants .	33.41

5. Conclusions

Model with variational attention and exact ELBO optimization shows the best scores but is more time-consuming than REINFORCE inference method. Soft attention is less time-consuming, but even with more epochs the performance is worse. Variational attention with REINFORCE inference method is faster to train and the scores are sufficient. While this technique is a promising alternative to soft attention, there are some practical limitations, for example, popular models such as the Transformer, utilize many repeated attention models such as 100-150. It is unclear if this approach can be used at that scale as predictive inference becomes combinatorial.

6. Contribution

- Elena Orlova – metrics visualization, reproducing variational attention model with exact ELBO;
- Vadim Kuzmin – data preprocessing, reproducing variational attention model with REINFORCE;
- Roman Tsukanov – translation examples, reproducing soft attention model.

References

Jimmy Ba, Ruslan R Salakhutdinov, Roger B Grosse, and Brendan J Frey. Learning wake-sleep recurrent attention models. In *Advances in Neural Information Processing Systems*, pages 2593–2601, 2015.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2018.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.