Summer Review 10

Latent Alignment and Variational Attention

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, Alexander M. Rush

Paper Link

School of Engineering and Applied Sciences Harvard University
Cambridge, MA, USA

Reviewed by : Arshdeep Sekhon

[1]Department of Computer Science, University of Virginia
https://qdata.github.io/deep2Read/

Attention and Latent alignment

- attention decisions directly as a tool for model interpretability
- or as a factor in final predictions
- attention plays the role of a latent alignment variable
- hard attention: explicit by introducing a latent variable for alignment and then optimizing a bound on the log marginal likelihood using policy gradients
- VQA and NMT: alignment

latent alignment approach

appealing for several reasons:

- latent variables facilitate reasoning about dependencies in a probabilistically principled way, e.g. allowing composition with other models
- posterior inference provides a better basis for model analysis and partial predictions than strictly feed-forward models
- directly maximizing marginal likelihood may lead to better results

This paper

- quantify the issues with attention
- propose alternatives based on variational inference.
- variational attention approach that can effectively fit latent variable alignments while remaining tractable to train.
- two variants: categorical and relaxed

Background: Latent Alignment

- $x = \{x_1, \ldots, x_i, \ldots, x_T\}$
- $X \in D \times T$
- $\tilde{x}$ is query
- discrete output variable : $y \in Y$
- process is mediated through a latent alignment variable z
- which indicates which member (or mixture of members) of x generates y.
- $z \sim D(a(x, \tilde{x}; \theta))$ $y \sim f(x, z; \theta)$
- a produces the parameters for an alignment distribution D

Background: Latent Alignment

- a produces the parameters for an alignment distribution D.
- $max_\theta logp(y = y\hat{y}|x, \tilde{x}) = max_\theta logE_z[f(x, z; \theta)_{\hat{y}}]$
- Directly maximizing this log marginal likelihood in the presence of the latent variable z is often difficult due to the expectation
-

Background: Latent Alignment

- For this to represent an alignment, restrict the variable z to be in the simplex over source indices $\{1, \ldots, T\}$
- let D be a categorical where z is a one-hot vector
- $z_i = 1$ if x_i is selected
- Example: $f(x; z)$ could use z to pick from x and apply a softmax layer to predict y, i.e. $f(x; z) = softmax(WXz)$ and $W \in R^{|Y| \times d}$
- Second : a relaxed alignment where z is a mixture taken from the interior of the simplex by letting D be a Dirichlet.

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

Background: Soft Attention

- deterministic
- expectation over the alignment variable
- an approximation of alignment
- soft attention uses a convex combination of the input representations XE[z] (the context vector) to obtain a distribution over the output

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

Hard Attention

- approximate inference approach for latent alignment
- takes a single hard sample of z (as opposed to a soft mixture) and then backpropagates through the model.
- First apply Jensens inequality to get a lower bound on the log marginal likelihood (KEY STEP)
- then maximize this lower-bound with policy gradients/REINFORCE

Variational Attention for Latent Alignment Models

- Key step in hard attention: could be large, poor performance
- variational inference methods directly aim to tighten this gap
- ELBO parameterized bound over a family of distributions $q(z) in Q$
- search over variational distributions q to improve the bound. tight when the variational distribution is equal to the posterior

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

optimize the evidence lower bound

- tight when $q(z) = p(z|x, \tilde{x}, y)$
- Hard attention is a special case of the ELBO with $q(z) = p(z|x; \tilde{x})$.
- many ways to optimize the evidence lower bound
- amortized variational inference

Amortized Variational Inference

- AVI uses an inference network to produce the parameters of the variational distribution $q(z; \lambda)$
- $\lambda = enc(x, \tilde{x}, y; \phi)$
- objective aims to reduce the gap with the inference network ϕ while also training the generative model θ
-

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

Categorical Alignments

- generative assumption is that y is generated from a single index of x.
- a low-variance estimator of $\nabla_\theta ELBO$, is easily obtained through a single sample from q(z).
- For $\nabla_\phi ELBO$, the gradient with respect to the KL portion is easily computable,
- but optimization issue with the gradient with respect to the first term
-

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

Categorical Alignment

- Variance reduction of this estimate falls to the baseline term B.
- The ideal baseline would be $E_{z\sim q(z)}[logf(x;z)]$, 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. $logf(x;E[z])$).

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

Relaxed Alignment

- both D and Q as Dirichlets
- z represents a mixture of indices.
- closer to the soft attention formulation which assigns mass to multiple indices
- fundamentally different in that we still formally treat alignment as a latent variable.
- certain continuous distributions allow the use reparameterization
- sampling $z \sim q(z)$ can be done by first sampling from a simple unparameterized distribution U,
- and then applying a transformation $g\phi(\dot{})$, yielding an unbiased estimator

$$\mathbb{E}_{u\sim\mathcal{U}}\left[\nabla_\phi \log p(y|x, g_\phi(u))\right] - \nabla_\phi \,\mathrm{KL}\left[q(z) \,\|\, p(z \,|\, x, \tilde{x})\right]$$

Neural machine Translation

- Attention is used to identify which source positions should be used to predict the target.
- For variational attention, the inference network enc applies a bidirectional LSTM over the source and the target to obtain the hidden states $x_1, \ldots, x_T and h_1, \ldots, h_S$,
- produces the alignment scores at the j-th time step via a bilinear map, $s^(j)_i = exp(h_j^T U x_i)$
- For the categorical case, the scores are normalized, $q(z^(j)_i = 1) \propto s^(j)_i$
- in the relaxed case the parameters of the Dirichlet are $\alpha^(j)_i = s^(j)_i$

Visual Question Answering

- The query $\tilde{x}$ is obtained by running an LSTM over the question
- the attention function *a* passes the query and the object representation through an MLP.
- The prediction function f: concatenate the chosen x_i with the query $\tilde{x}$ to use as input to an MLP which produces a distribution over the output.
- The inference network *enc* uses the answer embedding h_y and combines it with x_i and $\tilde{x}$ to produce the variational (categorical) distribution

$$q(z_i = 1) \propto \exp(u^\top \tanh(\mathbf{U}_1(x_i \odot \mathrm{ReLU}(\mathbf{V}_1 h_y)) + \mathbf{U}_2(\tilde{x} \odot \mathrm{ReLU}(\mathbf{V}_2 h_y))))$$

Results

Model	Objective	$\mathbb{E}$	NMT PPL	NMT BLEU	VQA NLL	VQA Eval
Soft Attention	$\log p(y \mid \mathbb{E}[z])$	-	7.03	32.31	1.76	58.93
Marginal Likelihood	$\log \mathbb{E}[p]$	Enum	6.33	33.08	1.69	60.33
Hard Attention	$\mathbb{E}_p[\log p]$	Enum	7.37	31.40	1.78	57.60
Hard Attention	$\mathbb{E}_p[\log p]$	Sample	7.38	31.00	1.82	56.30
Variational Relaxed Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	7.58	30.05	-	-
Variational Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Enum	6.03	33.10	1.69	58.44
Variational Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	6.13	33.09	1.75	57.52

Table 1: Evaluation on neural machine translation (NMT) and visual question answering (VQA) for the various models. $\mathbb{E}$ column indicates whether the expectation is calculated via enumeration (Enum) or a single sample (Sample) during training. For NMT we evaluate intrinsically on perplexity (PPL) (lower is better) and extrinsically on BLEU (higher is better), where for BLEU we perform beam search with beam size 10 and length penalty (see Appendix B for further details). For VQA we evaluate intrinsically on negative log-likelihood (NLL) (lower is better) and extrinsically on VQA evaluation metric (higher is better). All results except for relaxed attention use enumeration at test time.

Results

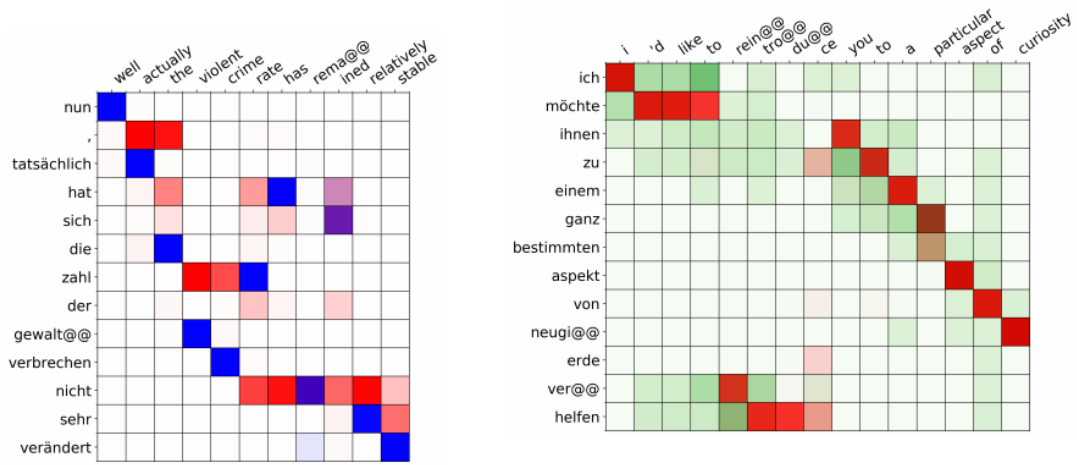

Figure 3: (Left) An example demonstrating the difference between the prior alignment (red) and the variational posterior (blue) when translating from DE-EN (left-to-right). Note the improved blue alignments for `actually` and `violent` which benefit from seeing the next word. (Right) Comparison of soft attention (green) with the p of variational attention (red). Both models imply a similar alignment, but variational attention is lower entropy.