Latent Alignment and Variational Attention

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

{dengyuntian@seas,yoonkim@seas,justinchiu@g,dguo@college,srush@seas}.harvard.edu

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

Abstract

Neural attention has become central to many state-of-the-art models in natural language processing and related domains. Attention networks are an easy-to-train and effective method for softly simulating alignment; however, the approach does not marginalize over latent alignments in a probabilistic sense. This property makes it difficult to compare attention to other alignment approaches, to compose it with probabilistic models, and to perform posterior inference conditioned on observed data. A related latent approach, hard attention, fixes these issues, but is generally harder to train and less accurate. This work 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. We further propose methods for reducing the variance of gradients to make these approaches computationally feasible. Experiments show that for machine translation and visual question answering, inefficient exact latent variable models outperform standard neural attention, but these gains go away when using hard attention based training. On the other hand, variational attention retains most of the performance gain but with training speed comparable to neural attention.

1 Introduction

Attention networks [6] have quickly become the foundation for state-of-the-art models in natural language understanding, question answering, speech recognition, image captioning, and more [15, 78, 16, 14, 62, 77, 68, 61]. Alongside components such as residual blocks and long-short term memory networks, soft attention provides a rich neural network building block for controlling gradient flow and encoding inductive biases. However, more so than these other components, which are often treated as black-boxes, researchers use intermediate attention decisions directly as a tool for model interpretability [42, 1] or as a factor in final predictions [25, 65]. From this perspective, attention plays the role of a latent alignment variable [10, 87]. An alternative approach, hard attention [77], makes this connection explicit by introducing a latent variable for alignment and then optimizing a bound on the log marginal likelihood using policy gradients. This approach generally performs worse (aside from a few exceptions such as [27]) and is used less frequently than its soft counterpart.

Still the latent alignment approach remains appealing for several reasons: (a) latent variables facilitate reasoning about dependencies in a probabilistically principled way, e.g. allowing composition with other models, (b) posterior inference provides a better basis for model analysis and partial predictions than strictly feed-forward models, which have been shown to underperform on alignment in machine translation [38], and finally (c) directly maximizing marginal likelihood may lead to better results.

∗Equal contribution.
The aim of this work is to quantify the issues with attention and propose alternatives based on recent developments in variational inference. While the connection between variational inference and hard attention has been noted in the literature [4, 40], the space of possible bounds and optimization methods has not been fully explored and is growing quickly. These tools allow us to better quantify whether the general underperformance of hard attention models is due to modeling issues (i.e. soft attention imbues a better inductive bias) or optimization issues.

Our main contribution is a variational attention approach that can effectively fit latent alignments while remaining tractable to train. We consider two variants of variational attention: categorical and relaxed. The categorical method is fit with amortized variational inference using a learned inference network and policy gradient with a soft attention variance reduction baseline. With an appropriate inference network (which conditions on the entire source/target), it can be used at training time as a drop-in replacement for hard attention. The relaxed version assumes that the alignment is sampled from a Dirichlet distribution and hence allows attention over multiple source elements.

Experiments describe how to implement this approach for two major attention-based models: neural machine translation and visual question answering (Figure 1 gives an overview of our approach for machine translation). We first show that maximizing exact marginal likelihood can increase performance over soft attention. We further show that with variational (categorical) attention, alignment variables significantly surpass both soft and hard attention results without requiring much more difficult training. We further explore the impact of posterior inference on alignment decisions, and how latent variable models might be employed. Our code is available at https://github.com/harvardnlp/var-attn/.

Related Work Latent alignment has long been a core problem in NLP, starting with the seminal IBM models [11], HMM-based alignment models [72], and a fast log-linear reparameterization of the IBM 2 model [20]. Neural soft attention models were originally introduced as an alternative approach for neural machine translation [6], and have subsequently been successful on a wide range of tasks (see [15] for a review of applications). Recent work has combined neural attention with traditional alignment [18, 69] and induced structure/sparsity [47, 33, 43, 82, 53, 54, 48], which can be combined with the variational approaches outlined in this paper.

In contrast to soft attention models, hard attention [77, 3] approaches use a single sample at training time instead of a distribution. These models have proven much more difficult to train, and existing works typically treat hard attention as a black-box reinforcement learning problem with log-likelihood as the reward [77, 3, 52, 26, 19]. Two notable exceptions are [4, 40]: both utilize amortized variational inference to learn a sampling distribution which is used obtain importance-sampled estimates of the log marginal likelihood [12]. Our method uses uses different estimators and targets the single sample approach for efficiency, allowing the method to be employed for NMT and VQA applications.

There has also been significant work in using variational autoencoders for language and translation application. Of particular interest are those that augment an RNN with latent variables (typically Gaussian) at each time step [17, 22, 64, 23, 39] and those that incorporate latent variables into sequence-to-sequence models [81, 7, 67]. Our work differs by modeling an explicit model component (alignment) as a latent variable instead of auxiliary latent variables (e.g. topics). The term “variational attention” has been used to refer to a different component the output from attention (commonly called the context vector) as a latent variable [7], or to model both the memory and the alignment as a latent variable [9]. Finally, recent work [75] also performs exact marginalization over latent alignments for character-level transduction.
2 Background: Latent Alignment and Neural Attention

We begin by introducing notation for latent alignment, and then show how it relates to neural attention. For clarity, we are careful to use alignment to refer to this probabilistic model (Section 2.1), and soft and hard attention to refer to two particular inference approaches used in the literature to estimate alignment models (Section 2.2).

2.1 Latent Alignment

Figure 2(a) shows a latent alignment model. Let $x$ be an observed set with associated members $\{x_1, \ldots, x_i, \ldots, x_T\}$. Assume these are vector-valued (i.e. $x_i \in \mathbb{R}^d$) and can be stacked to form a matrix $X \in \mathbb{R}^{d \times T}$. Let the observed $\hat{x}$ be 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 $y$. The generative process we consider is:

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

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 $\theta$ by maximizing the log marginal likelihood of training examples $(\hat{x}, y)^T$:

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

Directly maximizing this log marginal likelihood in the presence of the latent variable $z$ is often difficult due to the expectation (though tractable in certain cases).

For this to represent an alignment, we restrict the variable $z$ to be in the simplex $\Delta^{T-1}$ over source indices $\{1, \ldots, T\}$. We consider two distributions for this variable: first, let $\mathcal{D}$ be a categorical where $z$ is a one-hot vector with $z_i = 1$ if $x_i$ is selected. For example, $f(x, z)$ could use $z$ to pick from $x$ and apply a softmax layer to predict $y$, i.e. $f(x, z) = \text{softmax}(WXz)$ and $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$,

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

This computation requires a factor of $O(T)$ additional runtime, and introduces a major computational factor into already expensive deep learning models.\(^2\)

Second we consider a relaxed alignment where $z$ is a mixture taken from the interior of the simplex by letting $\mathcal{D}$ be a Dirichlet. This objective looks similar to the categorical case, i.e. $\log p(y = \hat{y} \mid x, \hat{x}) = \log \mathbb{E}_z[\text{softmax}(WXz)_{\hat{y}}]$, but the resulting expectation is intractable to compute exactly.

2.2 Attention Models: Soft and Hard

When training deep learning models with gradient methods, it can be difficult to use latent alignment directly. As such, two alignment-like approaches are popular: soft attention replaces the probabilistic model with a deterministic soft function and hard attention trains a latent alignment model by maximizing a lower bound on the log marginal likelihood (obtained from Jensen’s inequality) with policy gradient-style training. We briefly describe how these methods fit into this notation.

---

\(^2\)When clear from context, the random variable is dropped from $\mathbb{E}[\cdot]$. We also interchangeably use $p(\hat{y} \mid x, \hat{x})$ and $f(x, z ; \theta)$ to denote $p(y = \hat{y} \mid x, \hat{x})$.

\(^3\)Although not our main focus, explicit marginalization is sometimes tractable with efficient matrix operations on modern hardware, and we compare the variational approach to explicit enumeration in the experiments. In some cases it is also possible to efficiently perform exact marginalization with dynamic programming if one imposes additional constraints (e.g. monotonicity) on the alignment distribution.\(^8\)\(^9\)\(^3\).
**Soft Attention**  Soft attention networks use an altered model shown in Figure 2b. Instead of using a latent variable, they employ a deterministic network to compute an expectation over the alignment variable. We can write this model using the same functions $f$ and $a$ from above,

$$\log p_{\text{soft}}(y \mid x, \hat{x}) = \log f(x, E_z[z]; \theta) = \log \text{softmax}(W X E_z[z])$$

A major benefit of soft attention is efficiency. Instead of paying a multiplicative penalty of $O(T)$ or requiring integration, the soft attention model can compute the expectation before $f$. While formally a different model, soft attention has been described as an approximation of alignment [77]. Since $E[z] \in \Delta^{T-1}$, soft attention uses a convex combination of the input representations $X E[z]$ (the context vector) to obtain a distribution over the output. While also a “relaxed” decision, this expression differs from both the latent alignment models above. Depending on $f$, the gap between $E[f(x, z)]$ and $f(x, E[z])$ may be large.

However there are some important special cases. In the case where $p(z \mid x, \hat{x})$ is deterministic, we have $E[f(x, z)] = f(x, E[z])$, and $p(y \mid x, \hat{x}) = p_{\text{soft}}(y \mid x, \hat{x})$. In general we can bound the absolute difference based on the maximum curvature of $f$, as shown by the following proposition.

**Proposition 1.** Define $g_{x,\hat{y}} : \Delta^{T-1} \rightarrow [0, 1]$ to be the function given by $g_{x,\hat{y}}(z) = f(x, z)\hat{y}$ (i.e. $g_{x,\hat{y}}(z) = p(y = \hat{y} \mid z, x, \hat{x})$) for a twice differentiable function $f$. Let $H_{g_{x,\hat{y}}}(z)$ be the Hessian of $g_{x,\hat{y}}(z)$ evaluated at $z$, and further suppose $\|H_{g_{x,\hat{y}}}(z)\|_2 \leq c$ for all $z \in \Delta^{T-1}$, $\hat{y} \in Y$, and $x$, where $\| \cdot \|_2$ is the spectral norm. Then for all $\hat{y} \in Y$,

$$|p(y = \hat{y} \mid x, \hat{x}) - p_{\text{soft}}(y = \hat{y} \mid x, \hat{x})| \leq c$$

The proof is given in Appendix A.\footnote{Empirically the soft approximation works remarkably well, and often moves towards a sharper distribution with training. Alignment distributions learned this way often correlate with human intuition (e.g. word alignment in machine translation) [38].}\footnote{It is also possible to study the gap in finer detail by considering distributions over the inputs of $f$ that have high probability under approximately linear regions of $f$, leading to the notion of approximately expectation-linear functions, which was originally proposed and studied in the context of dropout [45].}

**Hard Attention**  Hard attention is an approximate inference approach for latent alignment (Figure 2b) [77, 4, 52, 26]. Hard attention takes a single hard sample of $z$ (as opposed to a soft mixture) 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, $\log E_z[p(y \mid x, z)] \geq E_z[\log p(y \mid x, z)]$, then maximize this lower-bound with policy gradients/REINFORCE [73] to obtain unbiased gradient estimates,

$$\nabla_{\theta}\mathbb{E}_z[\log f(x, z)] = \mathbb{E}_z[\nabla_{\theta} \log f(x, z) + (\log f(x, z) - B)\nabla_{\theta} \log p(z \mid x, \hat{x})]$$

where $B$ is a baseline that can be used to reduce the variance of this estimator. To implement this approach efficiently, hard attention uses Monte Carlo sampling to estimate the expectation in the gradient computation. For efficiency, a single sample from $p(z \mid x, \hat{x})$ is used, in conjunction with other tricks to reduce the variance of the gradient estimator (discussed more below) [77, 49, 50].

### 3 Variational Attention for Latent Alignment Models

Amortized variational inference (AVI, closely related to variational auto-encoders) [36, 60, 49] is a class of methods for efficient approximate latent variable inference, using learned inference networks. In this section we explore this technique for deep latent alignment models, and propose methods for variational attention that combine the benefits of soft and hard attention.

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.\footnote{Another way of viewing soft attention is as simply a non-probabilistic learned function. While it is possible that such models encode better inductive biases, our experiments show that when properly optimized, latent alignment attention with explicit latent variables do outperform soft attention.}\footnote{Prior works on hard attention have generally approached the problem as a black-box reinforcement learning problem where the rewards are given by $\log f(x, z)$. Ba et al. (2015) [4] and Lawson et al. (2017) [40] are the notable exceptions, and both works utilize the framework from [50] which obtains multiple samples from a learned sampling distribution to optimize the IWAE bound [12] or a reweighted wake-sleep objective.}
inference methods directly aim to tighten this gap. In particular, the evidence lower bound (ELBO) is a parameterized bound over a family of distributions \( q(z) \in \mathcal{Q} \) (with the constraint that the support \( q(z) \subseteq \text{supp} p(z | x, \hat{x}, y) \)).

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

This allows us to search over variational distributions \( q \) to improve the bound. It is tight when the variational distribution is equal to the posterior, i.e. \( q(z) = p(z | x, \hat{x}, y) \). Hard attention is a special case of the ELBO with \( q(z) = p(z | x, \hat{x}) \).

There are many ways to optimize the evidence lower bound; an effective choice for deep learning applications is to use amortized variational inference. 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. \( \lambda = \text{enc}(x, \hat{x}, y; \phi) \). The objective aims to reduce the gap with the inference network \( \phi \) while also training the generative model \( \theta \):

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

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. In the rest of this section, we consider strategies for accurately and efficiently computing this objective; in the next section, we describe instantiations of \( \text{enc} \) for specific domains.

### Algorithm 1: Categorical Alignments

First consider the case where \( D \), the alignment distribution, and \( \mathcal{Q} \), the variational family, are categorical distributions. Here the generative assumption is that \( y \) is generated from a single index of \( x \). Under this setup, a low-variance estimator of \( \nabla_{\theta} \text{ELBO} \) is easily obtained through a single sample from \( q(z) \). For \( \nabla_{\theta} \text{ELBO} \), the gradient with respect to the KL portion is easily computable, but there is an optimization issue with the gradient with respect to the first term \( \mathbb{E}_{z \sim q(z)} [\log f(x, z)] \).

Many recent methods target this issue, including neural estimates of baselines [49, 50], Rao-Blackwellization [53], reparameterizable relaxations [31, 46], and a mix of various techniques [70, 24]. We found that an approach using REINFORCE [73] along with a specialized baseline was effective. However, note that REINFORCE is only one of the inference choices we can select, and as we will show later, alternative approaches such as reparameterizable relaxations work as well. Formally, we first apply the likelihood-ratio trick to obtain an expression for the gradient with respect to the inference network parameters \( \phi \),

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

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

\[
\mathbb{E}_{z \sim q(z)} \left[ \left( \frac{f(x, z)}{f(x, \mathbb{E}[z] \mid x, \hat{x})} \right) \nabla_{\phi} \log q(z | x, \hat{x}) \right]
\]

Effectively this weights gradients to \( q \) based on the ratio of the inference network alignment approach to a soft attention baseline. Notably the expectation in the soft attention is over \( p \) (and not over \( q \)), and therefore the baseline is constant with respect to \( \phi \). Note that a similar baseline can also be used for hard attention, and we apply it to both variational/hard attention models in our experiments.
We experiment with variational attention in two different domains where attention-based models are

The inference network applies a bidirectional LSTM over the source and

Algorithm 2: Relaxed Alignments Next consider treating both D and Q as Dirichlets, where z represents a mixture of indices. This model is in some sense closer to the soft attention formulation which assigns mass to multiple indices, though fundamentally different in that we still formally treat alignment as a latent variable. Again the aim is to find a low variance gradient estimator. Instead of using REINFORCE, certain continuous distributions allow the use reparameterization [36], where sampling z ∼ q(z) can be done by first sampling from a simple unparameterized distribution U, and then applying a transformation gα(·), yielding an unbiased estimator,

\[ \mathbb{E}_{u \sim U}[\nabla_\phi \log p(y|x, g_\phi(u))] - \nabla_\phi \text{KL}[q(z) \parallel p(z \mid x, \tilde{x})] \]

The Dirichlet distribution is not directly reparameterizable. While transforming the standard uniform distribution with the inverse CDF of Dirichlet would result in a Dirichlet distribution, the inverse CDF does not have an analytical solution. However, we can use rejection based sampling to get a sample, and employ implicit differentiation to estimate the gradient of the CDF [32].

Empirically, we found the random initialization would result in convergence to uniform Dirichlet parameters for λ. (We suspect that it is easier to find low KL local optima towards the center of the simplex). In experiments, we therefore initialize the latent alignment model by first minimizing the Jensen bound, \[ \mathbb{E}_{z \sim p(z \mid x, \tilde{x})}[\log p(y \mid x, z)] \], and then introducing the inference network.

4 Models and Methods

We experiment with variational attention in two different domains where attention-based models are essential and widely-used: neural machine translation and visual question answering.

Neural Machine Translation Neural machine translation (NMT) takes in a source sentence and predicts each word of a target sentence yj in an auto-regressive manner. The model first contextually embeds each source word using a bidirectional LSTM to produce the vectors x1...xT. The query \( \tilde{x} \) consists of an LSTM-based representation of the previous target words y1:j−1. Attention is used to identify which source positions should be used to predict the target. The parameters of D are generated from an MLP between the query and source [6], and f concatenates the selected x_i with the query \( \tilde{x} \) and passes it to an MLP to produce the distribution over the next target word y_j.

For variational attention, the inference network applies a bidirectional LSTM over the source and the target to obtain the hidden states x_1,...,x_T and h_1,...,h_S, and produces the alignment scores at the j-th time step via a bilinear map, s_i(j) = exp(h_j^T U x_i). For the categorical case, the scores are normalized, q(s_i(j) = 1) ∝ s_i(j); in the relaxed case the parameters of the Dirichlet are α_i(j) = s_i(j). Note, the inference network sees the entire target (through bidirectional LSTMs). The word embeddings are shared between the generative/inference networks, but other parameters are separate.

Visual Question Answering Visual question answering (VQA) uses attention to locate the parts of an image that are necessary to answer a textual question. We follow the recently-proposed “bottom-up top-down” attention approach [2], which uses Faster R-CNN [59] to obtain object bounding boxes and performs mean-pooling over the convolutional features (from a pretrained ResNet-101 [27]) in each bounding box to obtain object representations x_1,...,x_T. 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 is also similar to the NMT case: we 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,

\[ q(z_i = 1) ∝ \exp(u^T \tanh(U_1(x_i \odot \text{ReLU}(V_1 h_y))) + U_2(\tilde{x} \odot \text{ReLU}(V_2 h_y)))) \]

where \( \odot \) is the element-wise product. This parameterization worked better than alternatives. We did not experiment with the relaxed case in VQA, as the object bounding boxes already give us the ability to attend to larger portions of the image.

Inference Alternatives For categorical alignments we described maximizing a particular variational lower bound with REINFORCE. Note that other alternatives exist, and we briefly discuss them...
with variational attention improve on soft attention and hard attention. However, the task-specific nature of VQA makes it challenging to evaluate soft and hard attention. For VQA, the trend is largely similar, and results for NLL are often better with variational attention.

Table 2 (left) considers test inference for variational attention, comparing enumeration to sampling. When training hard attention, we sample both at training and test time. However, when training variational attention, we sample at training time but use the exact same baseline at test time.

5 Experiments

Setup For NMT we mainly use the IWSLT dataset. This dataset is relatively small, but has become a standard benchmark for experimental NMT models. We follow the same preprocessing as in [21] with the same Byte Pair Encoding vocabulary of 14k tokens. To show that variational attention scales to large datasets, we also experiment on the WMT 2017 English-German dataset, following the preprocessing in [71] except that we use newstest2017 as our test set.

Results and Discussion Table 1 shows the main results. We first note that hard attention underperforms soft attention, even when its expectation is enumerated. This indicates that Jensen’s inequality alone is a poor bound. On the other hand, on both experiments, exact marginal likelihood outperforms soft attention, indicating that when possible it is better to have latent alignments.

For NMT, on the IWSLT 2014 German-English task, variational attention with enumeration and sampling performs comparably to optimizing the log marginal likelihood, despite the fact that it is optimizing a lower bound. We believe that this is due to the use of $q(z)$, which conditions on the entire source/target and therefore potentially provides better training signal to $p(z | x, \tilde{x})$ through the KL term. Note that it is also possible to have $q(z)$ come from a pretrained external model, such as a traditional alignment model. Table 2 (left) shows these results in context compared to the best reported values for this task. Even with sampling, our system improves on the state-of-the-art.

On the larger WMT 2017 English-German task, the superior performance of variational attention persists: our baseline soft attention reaches 24.10 BLEU score, while variational attention reaches 24.98. Note that this only reflects a reasonable setting without exhaustive tuning, yet we show that we can train variational attention at scale. For VQA the trend is largely similar, and results for NLL with variational attention improve on soft attention and hard attention. However, the task-specific evaluation metrics are slightly worse.

Table 2 (left) considers test inference for variational attention, comparing enumeration to K-max with $K = 5$. For all methods exact enumeration is better, however K-max is a reasonable approximation.

Note that enumeration does not imply exact if we are enumerating an expectation on a lower bound.
where soft attention and (right)). This possibly indicates that soft attention models are approximating latent alignment models.

We evaluate intrinsically on negative log-likelihood (NLL) (lower is better) and extrinsically on VQA evaluation (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.

Table 2 shows the PPL of different models as we increase $K > 1$, but we only get marginal benefits for $K > 5$. Finally, we observe that it is possible to train with soft attention and test using $K$-Max with a small performance drop (Soft $K$Max in Table 2 (right)). This possibly indicates that soft attention models are approximating latent alignment models.

On the other hand, training with latent alignments and testing with soft attention performed badly.

Table 3 (upper right) compares inference alternatives for variational attention. REINFORCE reaches a comparable performance as REINFORCE, but at a higher memory cost as it requires multiple samples. Gumbel-Softmax reaches nearly the same performance and seems like a viable alternative; although we found its performance is sensitive to its temperature parameter. We also trained a non-amortized SVI model, but found that at similar runtime it was not able to produce satisfactory results, likely due to insufficient updates of the local variational parameters. A hybrid method such as semi-amortized inference might be a potential future direction worth exploring.

Despite extensive experiments, we found that variational relaxed attention performed worse than other methods. In particular we found that when training with a Dirichlet KL, it is hard to reach low-entropy regions of the simplex, and the attentions are more uniform than either soft or variational categorical attention. Table 3 (lower right) quantifies this issue. We experimented with other distributions such as Logistic-Normal and Gumbel-Softmax but neither fixed this issue. Others have also noted difficulty in training Dirichlet models with amortized inference.

Besides performance, an advantage of these models is the ability to perform posterior inference, since the $q$ function can be used directly to obtain posterior alignments. Contrast this with hard attention where $q = p(z | x, \tilde{x})$, i.e. the variational posterior is independent of the future information. Figure 3 shows the alignments of $p$ and $q$ for variational attention over a fixed sentence (see Appendix C for more examples). We see that $q$ is able to use future information to correct alignments. We note that the inability of soft and hard attention to produce good alignments has been noted as a major issue in NMT. While $q$ is not used directly in left-to-right NMT decoding, it could be employed for other applications such as in an iterative refinement approach.

Table 1: Evaluation on NMT and VQA for the various models. $\mathcal{E}$ 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).

Table 2: (Left) Performance change on NMT from exact decoding to $K$-Max decoding with $K = 5$. (see section 5 for definition of $K$-max decoding). (Right) Test perplexity of different approaches while varying $K$ to estimate $\mathbb{E}_e[p(y|x, \tilde{x})]$. Dotted lines compare soft baseline and variational with full enumeration.
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 has lower entropy.

<table>
<thead>
<tr>
<th>Model</th>
<th>IWSLT BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam Search Optimization [74]</td>
<td>26.36</td>
</tr>
<tr>
<td>Neural PBMT + LM [29]</td>
<td>30.08</td>
</tr>
<tr>
<td>Minimum Risk Training [21]</td>
<td>32.84</td>
</tr>
<tr>
<td>Soft Attention</td>
<td>32.77</td>
</tr>
<tr>
<td>Marginal Likelihood</td>
<td>33.29</td>
</tr>
<tr>
<td>Hard Attention + Enum</td>
<td>31.40</td>
</tr>
<tr>
<td>Hard Attention + Sample</td>
<td>30.42</td>
</tr>
<tr>
<td>Variational Relaxed Attention</td>
<td>30.05</td>
</tr>
<tr>
<td>Variational Attention + Enum</td>
<td>33.69</td>
</tr>
<tr>
<td>Variational Attention + Sample</td>
<td>33.30</td>
</tr>
</tbody>
</table>

Table 3: (Left) Comparison against the best prior work for NMT on the IWSLT 2014 German-English test set. (Upper Right) Comparison of inference alternatives of variational attention on IWSLT 2014. (Lower Right) Comparison of different models in terms of implied discrete entropy (lower = more certain alignment).

Potential Limitations While this technique is a promising alternative to soft attention, there are some practical limitations: (a) Variational/hard attention needs a good baseline estimator in the form of soft attention. We found this to be a necessary component for adequately training the system. This may prevent this technique from working when T is intractably large and soft attention is not an option. (b) For some applications, the model relies heavily on having a good posterior estimator. In VQA we had to utilize domain structure for the inference network construction. (c) Recent models such as the Transformer [71], utilize many repeated attention models. For instance the current best translation models have the equivalent of 150 different attention queries per word translated. It is unclear if this approach can be used at that scale as predictive inference becomes combinatorial.

6 Conclusion

Attention methods are ubiquitous tool for areas like natural language processing; however they are difficult to use as latent variable models. This work explores alternative approaches to latent alignment, through variational attention with promising result. Future work will experiment with scaling the method on larger-scale tasks and in more complex models, such as multi-hop attention models, transformer models, and structured models, as well as utilizing these latent variables for interpretability and as a way to incorporate prior knowledge.
Acknowledgements

We are grateful to Sam Wiseman and Rachit Singh for insightful comments and discussion, as well as Christian Puhrsch for help with translations. This project was supported by a Facebook Research Award (Low Resource NMT). YK is supported by a Google AI PhD Fellowship. YD is supported by a Bloomberg Research Award. AMR gratefully acknowledges the support of NSF CCF-1704834 and an Amazon AWS Research award.

References


