Latent Alignment and Variational Attention

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

*Equal Contribution.

Introduction

- Attention is a core building block in many applications
  - ▷ makes the model more interpretable
  - ▷ can be used for other prediction tasks
- The prevalent Soft Attention is *deterministic*
- However, *Latent* Alignment is still attractive
  - ▷ facilitates composibility in a principled probabilistic manner
  - ▷ posterior inference provides better post-hoc interpretability
  - ▷ modeling uncertainties might lead to better performance
- We propose Variational Attention to learn Latent Alignments
  - ▷ efficient training based on amortized variational inference
  - ▷ tighter approximation bound than Hard Attention
- We experiment with translation and visual question answering
  - ▷ inefficient exact Latent Alignment outperforms Soft Attention
  - ▷ gains with Latent Alignments go away with Hard Attention
  - ▷ Variational Attention reaches similar performance as exact

Attention versus Alignment

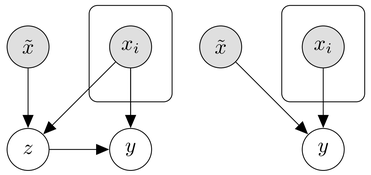


Figure 1: (a) Latent Alignment (b) Soft Attention with deterministic z absorbed

- Notations
  - ▷ $x = x_1, \cdots, x_T$: observed set, e.g. encoded source words
  - ▷ $\tilde{x}$: query, such as decoder hidden state at a timestep
  - ▷ y: output, such as the current target word
  - ▷ z: the latent alignment, indicating which member (or mixture of members) of x generates y.
  - ▷ $\mathcal{D}$: the prior of z
  - ▷ $f(x, z; \theta)$: the distribution of y given z and x
- Generative Process

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

- Training Objective (maximizing marginal log-likelihood)

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

- Direct optimization is computationally expensive
  - ▷ Discrete $z \sim \mathcal{D}$: $O(T)$ additional runtime
  - ▷ Continuous $z \sim \mathcal{D}$: intractable
- Traditional Workarounds:
  - ▷ Soft Attention [Bahdanau et al 2014]: nested expectation

$$\log \mathbb{E}_z[f(x, z; \theta)] \approx \log f(x, \mathbb{E}_z[z]; \theta)$$

  - ▷ Hard Attention [Xu et al 2015]: Jensen inequality and REINFORCE

$$\log \mathbb{E}_z[f(x, z; \theta)] \geq \mathbb{E}_z \log[f(x, z; \theta)]$$

Proposal: Variational Attention to learn Latent Alignment Models

- For any distribution $q(z) \in \mathcal{Q}$ ($\mathcal{Q} \subseteq \operatorname{supp} p(z \mid y, x, \tilde{x})$)

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

- The gap between log marginal likelihood and ELBO is $\mathrm{KL}[q(z) \,\|\, p(z \mid y, x, \tilde{x})]$
- Hard Attention uses prior as $q(z)$, which may result in a large gap hence a poor bound
- Instead we use an inference network to learn $q(z \mid y, x, \tilde{x})$ and maximize ELBO
  - ▷ $\mathbb{E}_{z \sim q(z)}[\log p(y \mid x, z)]$: approximate posterior attentions need to explain the target y
  - ▷ $-\mathrm{KL}[q(z) \,\|\, p(z \mid x, \tilde{x})]$: Prior attentions should follow posterior attentions

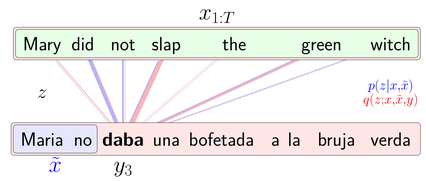


Figure 2: The blue prior p is restricted to past info, while the red variational posterior q may take into account future observations

Variational Categorical Attention

- $z \sim \mathcal{D}$ and $q(z)$: categorical
- Estimate gradients with REINFORCE

$$\mathbb{E}_{z \sim q(z)}[\nabla_\theta \log f(x, z) + (\log f(x, z) - B) \nabla_\phi \log q(z)]$$

- B is baseline for reducing variance, we use soft attention:

$$B = \log f(x, \mathbb{E}[z])$$

Variational Relaxed Attention

- $z \sim \mathcal{D}$ and $q(z)$: Dirichlet
- Use reparameterization trick [Kingma et al 2013]
  - ▷ Sample u from a simple distribution $\mathcal{U}$
  - ▷ Apply a transformation $g_\phi(\cdot)$ to obtain $z = g_\phi(u)$
- The gradient estimator takes the form

$$\mathbb{E}_{u \sim \mathcal{U}}\left[\nabla_{\theta,\phi} \log f(x, g_\phi(u))\right]$$

Experiments

Model	Objective	$\mathbb{E}$	NMT (IWSLT) PPL	NMT (IWSLT) BLEU	VQA (VQA 2.0) NLL	VQA (VQA 2.0) Eval
Soft Attention	$\log p(y \mid \mathbb{E}[z])$	-	7.17	32.77	1.76	58.93
Marginal Likelihood	$\log \mathbb{E}[p]$	Enum	6.34	33.29	1.69	60.33
Hard Attention	$\mathbb{E}_p[\log p]$	Enum	7.37	31.40	1.78	57.60
Variational Relaxed Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	7.58	30.05	-	-
Variational Categorical Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Enum	6.08	33.68	1.69	58.44
Variational Categorical Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	6.17	33.30	1.75	57.52

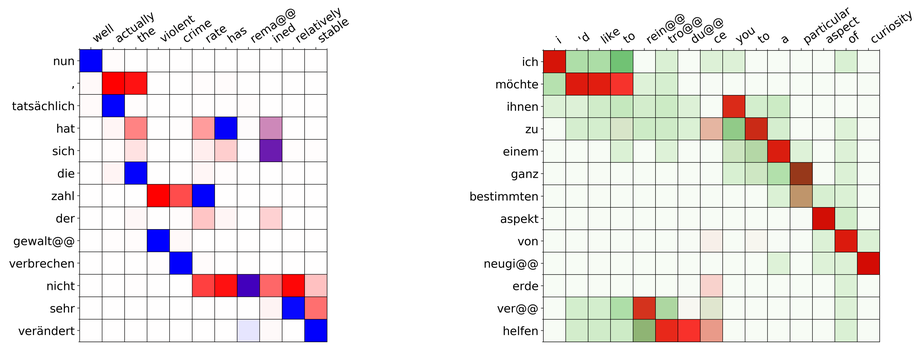


Figure 3: DE-EN attentions. (Left) red: prior; blue: variational posterior. (Right) red: prior; green: Soft Attention.

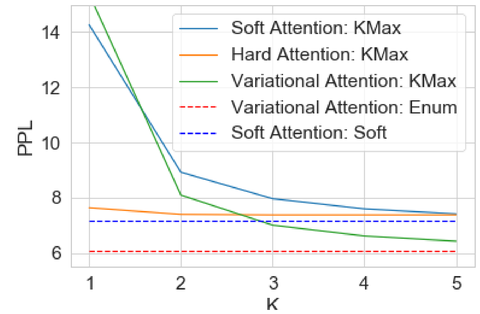
Alternative Inference Methods

Inference Method	#Samples	PPL	BLEU
REINFORCE	1	6.17	33.30
Reweighted Wake-Sleep	5	6.41	32.96
Gumbel-Softmax	1	6.51	33.08

- Gumbel-Softmax is a viable alternative
- Reweighted Wake-Sleep incurs higher memory cost

Approximate Decoding

- At test time we need to marginalize out z
- K-max decoding uses the top K entries of $p(z)$ to estimate



- Performance gain after $K = 5$ is marginal

Low Resource MT Preliminary Results†

Training Size	10k	25k	50k	Full (160k)
Soft Attn	12.10	22.88	26.65	32.77
Marginal Likelihood	16.79	23.44	26.99	33.29
Var Attn Enum	14.90	23.50	27.26	33.68
Var Attn Sample	12.20	23.35	27.87	33.30

†Results provided by Tim Lee

- Latent Alignments look promising at low resource scenarios
- Performance gain is lost when using Variational Attention with a single sample

Conclusions

- Exact Latent Alignment outperforms Soft Attention but is computationally expensive
- Hard Attention loses the performance gain of latent modeling even with exact enumeration of the expectation
- Variational Attention reaches similar performance as exact Latent Alignment Modeling