Bayesian Attention Modules

Xinjie Fan*, Shujian Zhang*, Bo Chen, and Mingyuan Zhou

Presented by Jiaru Zhang

SJTU Aritficial Intelligence Special Interest Group

November 12, 2021

SHANGHAI JIAO TONG UNIVERSITY

Section 1

Introduction

The paper and the authors

- This paper is published in NeurIPS 2020.

MetaReview

All reviewers recommended acceptance, pointing out that this is an interesting idea and a solid and well-executed work.

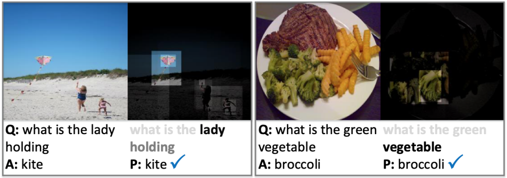
- The first authors come from The University of Texas at Austin.
- They published a series of papers these years:
  - *Bayesian Attention Belief Networks*, ICML 2021
  - *Adversarially Adaptive Normalization for Single Domain Generalization*, CVPR 2021
  - *Contextual dropout: An efficient sample-dependent dropout module*, ICLR 2021
  - ...

Introduction on attention modules

- Attention modules have been widely used in various applications.
- They are effective in not only being combined with other components, but also be used to build a stand-alone architecture.
- They also aid model visualization and enhance interpretability.

Preliminaries on attention modules

- Attention: putting more focus to important information according to current states.



- Deterministic attention:

$$\Phi = f(Q, K), \quad W_{i,j} = \exp(\Phi_{i,j}) / \sum_{j=1}^{n} \exp(\Phi_{i,j}), \quad O = WV$$

Figure source: Yu, Zhou, et al. Deep modular co-attention networks for visual question answering. In CVPR, 2019.

Advantages of stochastic attention

Advantages of making the attention weights stochastic and learning in a probabilistic manner:

- Capture more complicated dependencies
- Provide better model analysis
- Estimate uncertainties.

Related work

- Stochastic attention focus on hard attention [1] [2]
  - The weights are discrete random variables, making backpropagation difficult.
- Soft stochastic attention with normal distribution [3]
  - Weights are possibly negative and not sum to one.

[1] Xu, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[2] Deng, Yuntian, et al. Latent alignment and variational attention. In NeurIPS, 2018.

[3] Bahuleyan, Hareesh, et al. Variational Attention for Sequence-to-Sequence Models. In COLING, 2018.

Section 2

Bayesian attention modules

Main design idea

- Review on deterministic attention:

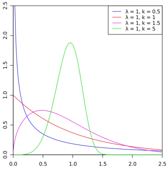
$$\Phi = f(Q, K), \quad W_{i,j} = \exp\left(\Phi_{i,j}\right) / \sum_{j=1}^{n} \exp\left(\Phi_{i,j}\right), \quad O = WV$$

- Main idea: Treat W as random variables from a distribution q_ϕ, which is parameterized by Q and K.
- Requirements:
  - $W_{i,j} > 0$
  - $\sum_j W_{i,j} = 1$
  - Amenable to gradient descent based optimization.

Two distributions

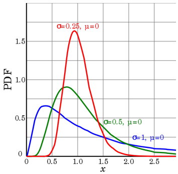
Weibull distribution

$$p(S \mid k, \lambda) = \frac{k}{\lambda^k} S^{k-1} e^{-(S/\lambda)^k}$$



Lognormal distribution

$$p(S \mid \mu, \sigma) = \frac{1}{S\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log S - \mu)^2}{2\sigma^2}\right]$$



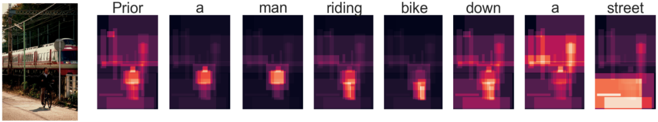
Weights W are got by applying a normalization over S:

$$W_i = \frac{S_i}{\sum_j S_{i,j}}$$

Shanghai Jiao Tong University

Contextual prior for regularization

- Motivating example:



- Idea: use keys K to construct prior attention distributions, regularizing posterior construsted for each query Q.

SHANGHAI JIAO TONG UNIVERSITY

Putting it all together

Combining all things above, the Evidence Lower BOund (ELBO) loss is given by

$$\mathcal{L}_\lambda(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\epsilon}) = \log p_{\boldsymbol{\theta}}\left(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{g}_\phi(\boldsymbol{\epsilon})\right) - \lambda \sum_{l=1}^{L} \underbrace{\mathrm{KL}\left(q_\phi\left(S_l \mid \tilde{g}_\phi\left(\boldsymbol{\epsilon}_{1:l-1}\right)\right) \| p_\eta\left(S_l \mid \tilde{g}_\phi\left(\boldsymbol{\epsilon}_{1:l-1}\right)\right)\right)}_{\text{analytic}}.$$

SHANGHAI JIAO TONG UNIVERSITY

Section 3

Experiments

Graph neural network results

- Graph attention networks (GAT) use self-attention layers to process node-features.
- Bayesian Attention Modules are used to improve the self-attention layers.
- `L` and `W` denote Lognormal and Weibull.
- `C` and `F` denote the contextual prior and the fixed prior.

Table 1: Classification accuracy for graphs.

Attention	Cora	Citeseer	PubMed
GAT	83.00	72.50	77.26
BAM (NO KL)	83.39	72.91	78.50
BAM-LF	83.24	72.86	78.30
BAM-LC	83.34	73.04	78.76
BAM-WF	83.48±0.2	73.18±0.3	78.50±0.3
BAM-WC	**83.81**±0.3	**73.52**±0.4	**78.82**±0.3

Shanghai Jiao Tong University

Visual question answering results

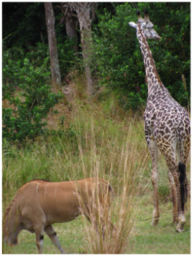
- Image features with Gaussian noise are used to evaluate the model robustness.
- Patch Accuracy vs Patch Uncertainty (PAvPU) is a metric to evaluate the quality of uncertainty estimation

Table 2: Results of different attentions on VQA.

Metric	Accuracy		PAvPU	
Data	Original	Noisy	Original	Noisy
Soft	66.95	61.25	70.04	65.34
BAM-LF	66.89	61.43	69.92	65.48
BAM-LC	66.93	61.58	70.14	65.60
BAM-WF	66.93	61.60	70.09	65.62
BAM-WC	**67.02** ±0.04	**62.89** ±0.06	**71.21** ±0.06	**66.75** ±0.08

Visual question answering results

- p-value is a metric of the model uncertainty.



Question: What animal is next to the giraffe?
Annotation set: {'wildebeest', 'horse', 'cow', 'antelope', 'gazelle', 'tapir', 'antelope', 'mountain lion', 'antelope', 'horse'}
Soft answer: deer, p-value: 0.01
BAM-WC answer: cow, p-value: 0.35

Question: What number is on the batter's shirt?
Annotation set: {'25', '25', '25', '25', '25', '25', '25','25', '25, '25'}
Soft answer: 15, p-value: 0.0
BAM-WC answer: 25, p-value: 0.0

Question: Is there mustard on the hot dog?
Annotation set: {'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes'}
Soft answer: yes, p-value: 0.48
BAM-WC answer: yes, p-value: 0.0

More experimental results

- Image captioning

Table 3: Comparing different attention modules on image captioning.

ATTENTION	BLEU-4	BLEU-3	BLEU-2	BLEU-1	CIDER	ROUGE	METEOR
SOFT[9]	24.3	34.4	49.2	70.7	-	-	23.9
HARD[9]	25.0	35.7	50.4	71.8	-	-	23.0
SOFT (OURS)	32.2	43.6	58.3	74.9	104.0	54.7	26.1
HARD (OURS)	26.5	37.2	51.9	69.8	84.4	50.7	23.3
BAM-LC	32.7	44.0	58.7	75.1	**105.0**	54.8	**26.3**
BAM-WC	**32.8**±0.1	**44.1**±0.1	**58.8**±0.1	**75.3**±0.1	104.5±0.1	**54.9**±0.1	26.2±0.1

- Neural machine translation

Table 4: Results on IWSLT.

Model	BLEU
Soft Attention	32.77
Variational Relaxed Attention	30.05
Variational Attention + Enum	33.68
Variational Attention + Sample	33.30
BAM-WC (Ours)	**33.81**±0.02

- Pretrained language models

	MRPC	COLA	RTE	MNLI	QNLI	QQP	SST	STS	SQUAD 1.1	SQUAD 2.0
ALBERT-BASE	86.5	54.5	75.8	85.1	90.9	**90.8**	92.4	90.3	80.86/88.70	78.80/82.07
ALBERT-BASE+BAM-WC	**88.5**	**55.8**	**76.2**	**85.6**	**91.5**	90.7	**92.7**	**91.1**	**81.40/88.82**	**78.97/82.23**

Section 4

Conclusion

Conclusion

Summary:

- This paper proposed a Bayesian attention module, with few modifications to standard attention.
- Experiments on a variety of tasks show its effectiveness.

Conclusion

Summary:

- This paper proposed a Bayesian attention module, with few modifications to standard attention.
- Experiments on a variety of tasks show its effectiveness.

My thoughts:

- **“Bayesian”** is all you need!

Thank You

Useful links

- Paper link (arxiv): `https://arxiv.org/pdf/2010.10604.pdf`
- Paper home page in NeurIPS: `https://papers.nips.cc/paper/2020/hash/bcff3f632fd16ff099a49c2f0932b47a-Abstract.html`
- Slides and video from the authors: `https://slideslive.ch/38937138/bayesian-attention-modules`

Details for calculating W

With the *Weibull* distribution, we treat k as a global hyperparameter and let $\lambda_{i,j}^{l,h} = \exp(\Phi_{i,j}^{l,h})/\Gamma(1 + 1/k)$, and like before $\Phi^{l,h} = f(Q^{l,h}, K^{l,h})$. Then, we sample $S_{i,j}^{l,h} \sim \text{Weibull}(k, \lambda_{i,j}^{l,h})$, which is the same as letting $S_{i,j}^{l,h} = \exp(\Phi_{i,j}^{l,h}) \frac{(-\log(1-\epsilon_{i,j}^{h,l}))^{1/k}}{\Gamma(1+1/k)}$, $\epsilon_{i,j}^{h,l} \sim \text{Uniform}(0,1)$. With the *Lognormal* distribution, we treat σ as a global hyperparameter and let $\mu_{i,j}^{l,h} = \Phi_{i,j}^{l,h} - \sigma^2/2$. Then, we sample $S_{i,j}^{l,h} \sim \text{Lognormal}(\mu_{i,j}^{l,h}, \sigma^2)$, which is the same as letting $S_{i,j}^{l,h} = \exp(\Phi_{i,j}^{l,h}) \exp(\epsilon_{i,j}^{h,l}\sigma - \sigma^2/2)$, $\epsilon_{i,j}^{h,l} \sim \mathcal{N}(0,1)$. Note our parameterizations ensure that $\mathbb{E}[S_{i,j}^{l,h}] = \exp(\Phi_{i,j}^{l,h})$. Therefore, if, instead of sampling $S_{i,j}^{l,h}$ from either distribution, we use its expectation as a substitute, then the mapping becomes equivalent to that of vanilla soft attention, whose weights are defined as in (1). In other words, if we let k of the Weibull distribution go to infinity, or σ of the Lognormal distribution go to zero, which means the variance of $S_{i,j}^{l,h}$ goes to zero and the distribution becomes a point mass concentrated at the expectation, then the proposed stochastic soft attention reduces to deterministic soft attention. Therefore, the proposed stochastic soft attention can be viewed as a generalization of vanilla deterministic soft attention.

Shanghai Jiao Tong University

Details for the two distributions

Weibull distribution: The Weibull distribution $S \sim \text{Weibull}(k, \lambda)$ has probability density function (PDF) $p(S \mid k, \lambda) = \frac{k}{\lambda^k} S^{k-1} e^{-(S/\lambda)^k}$, where $S \in \mathbb{R}_+$. Its expectation is $\lambda\Gamma(1 + 1/k)$ and variance is $\lambda^2 \left[\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2\right]$. It is reparameterizable as drawing $S \sim \text{Weibull}(k, \lambda)$ is equivalent to letting $S = \tilde{g}(\epsilon) := \lambda(-\log(1-\epsilon))^{1/k}$, $\epsilon \sim \text{Uniform}(0, 1)$. It resembles the gamma distribution, and with γ denoted as the Euler–Mascheroni constant, the KL divergence from the gamma to Weibull distributions has an analytic expression [17] as

$$\text{KL}(\text{Weibull}(k, \lambda) || \text{Gamma}(\alpha, \beta)) = \frac{\gamma\alpha}{k} - \alpha \log \lambda + \log k + \beta\lambda\Gamma(1 + \tfrac{1}{k}) - \gamma - 1 - \alpha \log \beta + \log \Gamma(\alpha).$$

Lognormal distribution: The Lognormal distribution $S \sim \text{Lognormal}(\mu, \sigma^2)$ has PDF $p(S \mid \mu, \sigma) = \frac{1}{S\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log S - \mu)^2}{2\sigma^2}\right]$, where $S \in \mathbb{R}_+$. Its expectation is $\exp(\mu + \sigma^2/2)$ and variance is $[\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$. It is also reparameterizable as drawing $S \sim \text{Lognormal}(\mu, \sigma^2)$ is equivalent to letting $S = \tilde{g}(\epsilon) = \exp(\epsilon\sigma + \mu)$, $\epsilon \sim \mathcal{N}(0, 1)$. The KL divergence is analytic as

$$\text{KL}(\text{Lognormal}(\mu_1, \sigma_1^2) || \text{Lognormal}(\mu_2, \sigma_2^2)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - 0.5.$$

Sampling $S_{i,j}^{l,h}$ from either the Weibull or Lognormal distribution, we obtain the simplex-constrained random attention weights W by applying a normalization function $\bar{g}$ over S as $W_i^{l,h} = \bar{g}(S_i^{l,h}) := S_i^{l,h} / \sum_j S_{i,j}^{l,h}$. Note W is reparameterizable but often does not have an analytic PDF.

Shanghai Jiao Tong University

Details for the contextual prior

Key-based contextual prior: Instead of treating the prior as a fixed distribution independent of the input $\boldsymbol{x}$, here we make the prior depend on the input through keys. The motivation comes from our application in image captioning. Intuitively, given an image (keys), there should be a *global* prior attention distribution over the image, indicating the importance of each part of the image even before the caption generation process. Based on the prior distribution, the attention distribution can be updated *locally* using the current state of generation (queries) at the each step (see Figure 1). This intuition can be extended to the general attention framework, where the prior distribution encodes the *global* importance of each keys shared by all queries, while the posterior encodes the *local* importance of each keys for each query. To obtain the prior parameters, we take a nonlinear transformation of the key features, followed by a softmax to obtain positive values and enable the interactions between keys. Formally, let $\Psi^{l,h} = \text{softmax}(F_2(F_{NL}(F_1(K^{l,h})))) \in \mathbb{R}^{n \times 1}$, where F_1 is linear mapping from $\mathbb{R}^{d_k}$ to a hidden dimension $\mathbb{R}^{d_{\text{mid}}}$, F_2 is linear mapping from $\mathbb{R}^{d_{\text{mid}}}$ to $\mathbb{R}$, and F_{NL} denotes a nonlinear activation function, such as ReLU [39]. With the gamma prior, we treat β as a hyperparameter and let $\alpha_{i,j}^{l,h} = \Psi_{i,1}^{l,h}$. With the Lognormal, we treat σ as a hyperparameter and let $\mu_{i,j}^{l,h} = \Psi_{i,1}^{l,h}$. Following previous work [40], we add a weight λ to the KL term and anneal it from a small value to one.