Repulsive Bayesian Sampling for Diversified Attention Modeling

Bang An
State University of New York at Buffalo
anbang@buffalo.edu

Xuannan Dong
University of Science and Technology of China
dxn0714@mail.ustc.edu.cn

Changyou Chen
State University of New York at Buffalo
changyou@buffalo.edu

Abstract

Attention mechanism stands as the key technique to the success of many deep-learning models. The multi-head attention extends single-head attention by allowing a model to jointly focus on information from different perspectives. Without explicit constraints, however, multi-head attention may suffer from attention collapse in the sense that several heads might attend to the same information, thus losing representation power. In this paper, for the first time, we provide a novel understanding of multi-head attention from a Bayesian-sampling perspective. Based on particle-optimization sampling methods, we further propose non-parametric approaches that explicitly improve the diversity of multi-head attention, which could strengthen a model's expression ability. We apply our framework to four representative models with multi-head attention, including the Transformer and Graph Attention Networks, and evaluate it on six different tasks. Experimental results show that our framework can significantly improve the diversity of multi-head attention, leading to performance improvement on all the tasks considered.

1 Introduction

Attention is one of the most popular and effective modules in deep learning neural networks, with impressive performances gains in many tasks. By extending a single head to multiple paralleled heads of attention, multi-head attention is widely used to capture different attentive information and strengthen the expressive ability of a model. The key point of multi-head attention is its ability to jointly attend to information from different representation subspaces at different positions. However, there are no explicit mechanisms guaranteeing this desired property, thus it could potentially lead to attention redundancy or collapse. Although there exist works by directly adding regularization on loss functions to encourage diversity of multi-head attention [1, 2], the underlying working principle has not been well-validated, and improvement has not been significant.

In this paper, we provide a principled and more interpretable solution for this problem from a Bayesian perspective. In order to incorporate uncertainty in attention, we propose to adapt the deterministic attention to a stochastic setting. Consequently, multi-head attention could be understood as multiple Bayesian samples that approximate a posterior attention distribution. To further introduce repulsiveness into attention heads, we adopt the particle-optimization sampling methods by viewing each head as a particle and jointly updating their weights to approximate a multimodal posterior distribution of the attention. With this, multi heads are enforced to move to different modes in the parameter space, thus improving the diversity in multi-head attention and enhancing their expressiveness power.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

We call our framework repulsive multi-head attention. Experiments on various attention models demonstrate the effectiveness of our approaches.

2 Repulsive Bayesian Attention

This section describes our framework on Bayesian modeling of diversified attention. For completeness, an introduction of standard attention mechanisms is presented in Appendix A, including the two most commonly used attention functions, the additive attention and dot-product attention.

2.1 Understanding Multi-Head Attention from a Bayesian-Inference Perspective

Consider the simplest case of self attention with a single head. As described in Appendix A, the attention mechanism can be represented as a deterministic mapping, f_{att}, from an input space to an output attention feature space, *e.g.*, $\boldsymbol{z} = f_{att}(\boldsymbol{x}; \theta)$ with θ the parameter of the mapping. In this example, f_{att} could represent an additive attention or a dot-product attention. Since $\boldsymbol{z}$ is deterministic, it lacks the ability to model and propagate uncertainty in attention modeling. To overcome this problem, we propose to adapt the deterministic attention to a stochastic version with Bayesian modeling.

Multi-head attention as hierarchical Bayesian modeling In our framework, instead of modeling attention as a deterministic transformation $\boldsymbol{z} = f_{att}(\boldsymbol{x}; \theta)$, we consider it as a hierarchical stochastic generative process: $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}; \theta)$, where the posterior for the parameter θ, $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$, serves as the prior for the parameter of the attention feature $\boldsymbol{z}$. Bayesian inference for attention then computes the predictive distribution $p(\boldsymbol{z}|\boldsymbol{x}, \mathcal{D})$ of the attentive latent representation $\boldsymbol{z}$ for an input $\boldsymbol{x}$ given the training data $\mathcal{D}$ by:

$$p(\boldsymbol{z}|\boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{z}|\boldsymbol{x}; \theta)p(\theta|\mathcal{D})d\theta \ . \tag{1}$$

Note a caveat of such a generalization is that it makes backpropagation used in standard attention mechanism difficult, due to the sampling operator $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}; \theta)$ in the internal node. To overcome this problem, instead of adopting the potentially complicated variational inference technique, we propose a simple workaround by reformulating Equation (1) with a set of M samples from $p(\theta|\mathcal{D})$, *i.e.*, we define the sampling process for $\boldsymbol{z}$ as the following generative process:

$$\begin{aligned} &\boldsymbol{z} = g(\boldsymbol{z}_1, ..., \boldsymbol{z}_M), \text{ where} \\ &\theta_i \sim p(\theta|\mathcal{D}), \ \ \boldsymbol{z}_i = f_{att}(\boldsymbol{x}) \ . \end{aligned} \tag{2}$$

Here $g(\cdot)$ is an aggregate function such as a linear projection. We can see that the above equation reduces to the multi-head attention when the parameters θ_i's are independent instead of being drawn from a shared prior $p(\theta|\mathcal{D})$. Thus, multi-head attention could be understood as a special case of the stochastic attention with independent parameter priors. In the next section, we introduce Bayesian sampling techniques to enforce repulsiveness to the parameters θ_i's. The repulsiveness could then be propagated to make the attention features $\boldsymbol{z}_i$'s diverse.

2.2 Repulsive Multi-Head Attention

We adopt the recently proposed particle-optimization sampling techniques [3, 4] to generate repulsive samples of the attention parameters. Generally speaking, particle-optimization sampling interactively updates a set of particles to approximate a target distribution by leveraging the optimal transport theory. To apply particle-optimization to multi-head attention, the parameter of every head θ_i is considered as a particle, which, according to (2), is a sample from the posterior distribution $p(\cdot|\mathcal{D})$. With a total of M heads $\{\theta_i\}_{i=1}^M$, we are able to well approximate the distribution $p(\theta|\mathcal{D})$, which are updated interactively according to some particle-optimization rules described later. In this way, multi-head attention modeling is equivalently transformed to a Bayesian inference problem. In this paper, we utilize two representative particle-optimization sampling algorithms: Stein Variational Gradient Descent (SVGD) [5] and Stochastic Particle-Optimization Sampling (SPOS) [4], for posterior inference. Details of these two methods are described in Appendix B.

To achieve repulsive multi-head attention, we only need to modify the learning process of attention parameters with particle-optimization sampling methods as shown in Algorithm 1, while keeping

Algorithm 1 Diversified Multi-Head Attention

Input: Initialized M-head attention model $\mathcal{A}$ with attention parameters $\Theta_0 = \{\theta_i\}_{i=1}^M$ and other parameters Ω_0; Training data $\mathcal{D} = \{D_k\}_{k=1}^N = \{(x_k, y_k)\}_{k=1}^N$;
Output: Optimized attention model with learned parameters $\hat{\Theta}$ and $\hat{\Omega}$;
Train:

```
for iteration ℓ do
  forward: ŷ_k = A(x_k; Θ_ℓ, Ω_ℓ), ∀k;
  calculate loss: L({ŷ_k}, {y_k});
  backward and calculate gradients:
  gradient of Ω_ℓ: φ(Ω_ℓ) ← ∇_{Ω_ℓ} L
  for attention head i do
    calculate φ(θ_ℓ^(i)) with SVGD (Eq (7)) or SPOS (Eq (8));
    gradient of θ_ℓ^(i): φ(θ_ℓ^(i)) ← ε_ℓ φ(θ_ℓ^(i));
  end for
  update parameters:
  Ω_{ℓ+1} ← Optimizer(Ω_ℓ, φ(Ω_ℓ))
  Θ_{ℓ+1} ← Optimizer(Θ_ℓ, φ(Θ_ℓ))
end for
```

model architectures unchanged. To be specific, in standard multi-head attention, the parameter of every head is updated independently according to the corresponding gradient of the loss function. To achieve repulsive multi-head attention, we follow the particle-optimization sampling update rule (*e.g.* Equation 7 or Equation 8) to update the parameter of every head interactively, while updating other parameters in the same way as standard multi-head attention. Equation 7 and 8 can be seen as a modified gradient with explicit repulsive intention and can be applied with any optimizer, *e.g.*, Adam [6]. Note that $\nabla_{\theta_\ell^{(i)}} U(\theta_\ell^{(i)})$ equals to the gradient of the $\theta_\ell^{(i)}$ in standard multi-head attention when negative log-likelihood is defined as the loss function and the prior of $\theta^{(i)}$ is assumed to be uniform. In practice, the update of M heads is parallel conducted with matrix operations.

3 Experimental Results

We apply our framework to four representative attention-based models. Experiments are conducted on six different tasks including author profiling, sentiment classification, textual entailment, translation, scientific publication classification and text generation. Extensive experimental results show that our approach can significantly improve the diversity in multi-head attention and strengthen the expression ability of original models, leading to consistent performance improvement in all the tasks considered.

3.1 Self-attentive Sentence Embedding Model

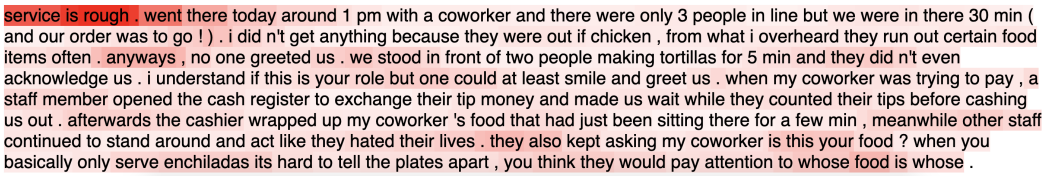
Self-attentive sentence embedding model [1] combines BiLSTM with multi-head attention to generate the sentence embedding matrix for specific tasks. We build a model containing 30-head self-attention following [1] and apply particle-optimization approach (SVGD and SPOS) to parameters of multi-head attention. All three tasks in [1] including author profiling, sentiment analysis and textual entailment are evaluated with the Age, Yelp, and SNLI datasets respectively. Author profiling is to predict the age range of the user by giving their tweets. Sentiment analysis is to predict the number of stars the user who wrote that review assigned to by analysis their reviews. Textual entailment is to tell whether the semantics in the two sentences are entailment or contradiction or neutral.

Results are presented in Table 1. With the proposed repulsive multi-head attention, the model achieves much higher accuracy on all three tasks, especially on the sentiment analysis task, yielding a 2.3% improvement. Our approaches also outperform the regularization method in [1], which penalize Frobenius norm of multi-head attention to introduce diversity. For particle-optimization rules, comparing with SVGD, with the help of additional noise, SPOS appears to get better performance.

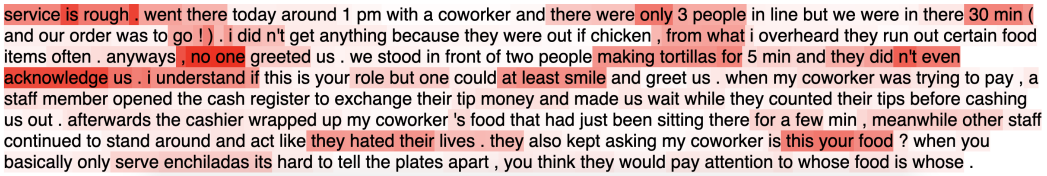
Visualization of attention is shown in Figure 1 which serves as an interpretation of the learned sentence embedding. All 30 heads are depicted in one heatmap yielding a general view of what the

Table 1: Performance (accuracy) comparison on Yelp, Age and SNLI dataset

Models	Yelp	Age	SNLI
BiLSTM + Multi-head Attention [1]	69.3%	81.47%	83.79%
BiLSTM + Multi-head Attention + Penalization [1]	70.2%	81.30%	84.55%
BiLSTM + Repulsive Multi-head Attention (SVGD)	71.2%	81.82%	84.58%
BiLSTM + Repulsive Multi-head Attention (SPOS)	**71.7%**	**82.55%**	**84.76%**



(a) Multi-head attention



(b) Repulsive multi-head attention

Figure 1: Heatmap of a 1 star Yelp review. The red mark indicates the weight of corresponding word in sentence embedding. The deeper the red, the more important the word is for sentiment analysis. Perspectives of 30 heads are depicted in one heatmap.

sentence embedding mostly focuses on. The text is a 1-star review in the Yelp dataset. As shown in Figure 1, the original multi-head attention tends to incur mode collapse, where almost all heads focus on one single factor. On the contrary, repulsive multi-head attention is able to capture multiple key factors in the review that indicate strongly on the sentiment behind the sentence. More examples can be found in Appendix D.

3.2 Transformer

Transformer is a representative model entirely relying on the multi-head attention. We evaluate our approach on Transformer with two standard translation datasets: IWSLT14 De-En and WMT14 En-De. The details of the Transformer model are given in Appendix C. To apply repulsive multi-head attention on Transformer, we modify all the multi-head self-attention in encoder and decoder as well as multi-head inter-attention between encoder and decoder to be repulsive. Parameters of every head are together seen as a particle, and we apply SVGD particle-optimization to all heads in all layers. Results are presented in Table 2, which show that, with the repulsive multi-head attention, Transformer models achieve a remarkable improvement on the BLEU score in both datasets. Furthermore, it is encouraging to see that Transformer-`base` with repulsive multi-head attention achieves comparable performance with Transformer-`big`, while the parameters are much less.

Table 2: Translation performance (BLEU) comparison

Models	IWSLT14 De-En	WMT14 En-De
Transformer-`small` [7]	34.4	/
Transformer-`base` [7]	/	27.3
Transformer-`big` [7]	/	**28.4**
Transformer-`small` + Repulsive Multi-head Attention	**35.2**	/
Transformer-`base` + Repulsive Multi-head Attention	/	**28.4**

Table 3: Performance (accuracy) comparison on Cora dataset

Models	**Cora**
Graph Attention Network [8]	$83.0 \pm 0.7\%$
Graph Attention Network + Repulsive Multi-head Attention(SVGD)	$85.1 \pm 0.8\%$
Graph Attention Network + Repulsive Multi-head Attention(SPOS)	$\mathbf{85.3} \pm \mathbf{0.8}\%$

Table 4: Automatic evaluations of generation systems

Models	**BLEU**	**METEOR**
GAT [9]	12.2 ± 0.44	17.2 ± 0.63
GraphWriter [9]	14.3 ± 1.01	18.8 ± 0.28
GraphWriter + Repulsive Multi-head Attention	$\mathbf{15.1} \pm \mathbf{0.97}$	$\mathbf{19.5} \pm \mathbf{0.29}$

3.3 Graph Attention Networks

Graph data are particularly useful in real world to represent irregular structure such as 3D meshes, social networks and brain connectomes. To learn the representation of graph-based data, attention mechanism is also widely used. A typical model is Graph Attention Networks (GAT) [8] which leverage masked multi-head self-attention to pass messages along graphs. By stacking multi-head attention layers in which nodes are able to attend over their neighborhoods' features, the model enables implicitly specifying different weights to different nodes in different subspaces. The attention in GAT is in the form of dot-product attention. We follow the model configuration in [8] and adapt the multi-head attention to the repulsive attention. Experiments are conducted on the Cora dataset, which represents a citation network consisting of 2708 scientific publications classified into seven classes. The task is node classification. Results are shown in Table 3, which shows that the GAT model with repulsive multi-head attention outperforms the original model by 2% in accuracy, demonstrating the effectiveness of our framework.

3.4 GraphWriter

Finally, we evaluate our framework on a more complicated task of generating coherent multi-sentence texts from a knowledge graph. To be specific, the task is to generate a text abstract given the title of a scientific article and a knowledge graph encoding annotations. We build on the graph-to-text model named GraphWriter [9], which contains dot-product multi-head attention. We modify the model with repulsive multi-head attention, and get results as shown in Table 4. Similarly, the GraphWriter model with repulsive multi-head attention outperforms the original model in terms of both BLEU and METER scores.

4 Conclusion

In this paper, motivated by uncertainty modeling of attention, we propose a principled way of understanding multi-head attention from a Bayesian-modeling perspective. Based on existing particle-optimization sampling techniques, we propose a simple yet efficient way to modify multi-head attention to be repulsive without additional parameters nor regularizers. Experimental results on six tasks demonstrate that our framework can significantly improve the diversity of multi-head attention, leading to performance improvement on all the tasks considered.

Acknowledgement

CC would like to thank the support of the Yahoo! FREP program.

References

[1] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.

[2] Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. In *EMNLP*, 2018.

[3] Changyou Chen, Ruiyi Zhang, Wenlin Wang, Bai Li, and Liqun Chen. A unified particle-optimization framework for scalable bayesian sampling. In *UAI*, 2018.

[4] Jianyi Zhang, Ruiyi Zhang, and Changyou Chen. Stochastic particle-optimization sampling and the non-asymptotic convergence theory. *arXiv preprint arXiv:1809.01293*, 2018.

[5] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, 2016.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[8] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[9] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text generation from knowledge graphs with graph transformers. In *NAACL-HLT*, 2019.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[11] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

Appendix A Multi-head Attention

Attention mechanism aims at modelling dependencies of representation pairs on different positions without regard to their distance. The attention could be between two different sequences such as two sentences of different languages, or inside a sequence which is called self-attention (also called intra-attention). The two most commonly used attention functions are additive attention [1, 10] and dot-product attention (also called multiplicative attention) [7]. We specify multi-head attention in these two kinds of attention.

Additive Attention First proposed by [10], additive attention uses a one-hidden layer feed-forward network to calculate the attention alignment. We use the attention function in [1] which is also a self-attention as an example.

$$\boldsymbol{a} = \mathsf{Softmax}(\boldsymbol{v}^T \tanh(WH^T)), \quad \boldsymbol{z} = \boldsymbol{a}H \tag{3}$$

$H \in \mathbb{R}^{n\times d}$ is the hidden state matrix of a sentence with n words. $\boldsymbol{a} \in \mathbb{R}^{1\times n}$ is the normalized alignment score vector for each word. $W \in \mathbb{R}^{d_a\times d}$ and $\boldsymbol{v} \in \mathbb{R}^{d_a\times 1}$ are attention parameters. The final sentence representation vector $\boldsymbol{z}$ is a weighted sum of words' hidden states weighted by attention vector. In order to capture overall semantics of the sentence instead of a specific component, multi-head attention could be applied.

$$A = \mathsf{Softmax}(V^T \tanh(WH^T)), \quad Z = AH \tag{4}$$

where $V \in \mathbb{R}^{d_a\times r}$ is the matrix performs r heads, $A \in \mathbb{R}^{r\times n}$ is the r-head attention matrix and $Z \in \mathbb{R}^{r\times d}$ is the resulting sentence representation matrix contains semantics from multiple aspects.

Dot-product Attention Transformer [7] is an architecture based on multi-head scaled dot-product attention. The attention function for a single head can be described as mapping a query and a set of key-value pairs to an output as

$$A_i = \mathsf{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}}), \quad Z_i = A_i V_i \tag{5}$$
$$\text{where} \quad Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V$$

$\{W_i^Q, W_i^K, W_i^V\}$ are parameters of i-th head. r-head attention projects the queries, keys and values into r subspaces with different, learned linear projections. Attention functions of all heads are performed in parallel and are concatenated and once again projected, resulting in the final values.

$$Z = \mathsf{Concate}(Z_1, ..., Z_i, ..., Z_r), \quad \mathsf{MultiHead}(Q, K, V) = ZW^O \tag{6}$$

Appendix B Particle-optimization Sampling Methods

In general, particle-optimization sampling is to interactively updates a set of particles to approximates a distribution. In this paper, we utilize the following two representative particle-optimization sampling algorithms.

Stein Variational Gradient Descent (SVGD) Considering $\mathcal{D} = \{D_k\}_{k=1}^N$ is a set of i.i.d. observation, $\boldsymbol{x} \in \mathbb{R}^d$ is a continuous random variable or parameter of interest with prior $p_0(\boldsymbol{x})$. In Bayesian sampling, we aims to generate random samples from the posterior distribution $p(\boldsymbol{x}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{x})p_0(\boldsymbol{x})$. Define the potential energy as $U(\boldsymbol{x}) \triangleq -\log p(\mathcal{D}|\boldsymbol{x}) - \log p_0(\boldsymbol{x}) = -\sum_{k=1}^N \log p(D_k|\boldsymbol{x}) - \log p_0(\boldsymbol{x})$, the posterior distribution is $p(\boldsymbol{x}|\mathcal{D}) \propto \exp(-U(\boldsymbol{x}))$.

SVGD [5] iteratively and interactively transports a set of particles $\{\boldsymbol{x}^{(i)}\}_{i=1}^M$ to match the target distribution, by applying a form of functional gradient descent that minimizes the KL divergence. The update rule for particles $\{\boldsymbol{x}_\ell^{(i)}\}_{i=1}^M$at the ℓ-th iteration with stepsize ϵ_ℓ is

$$\boldsymbol{x}_{\ell+1}^{(i)} = \boldsymbol{x}_\ell^{(i)} + \epsilon_\ell \phi(\boldsymbol{x}_\ell^{(i)}) \tag{7}$$
$$\text{where} \quad \phi(\boldsymbol{x}_\ell^{(i)}) = \frac{1}{M}\sum_{j=1}^M [-\kappa(\boldsymbol{x}_\ell^{(j)}, \boldsymbol{x}_\ell^{(i)})\nabla_{\boldsymbol{x}_\ell^{(j)}} U(\boldsymbol{x}_\ell^{(j)}) + \nabla_{\boldsymbol{x}_\ell^{(j)}} \kappa(\boldsymbol{x}_\ell^{(j)}, \boldsymbol{x}_\ell^{(i)})]$$

$\kappa(\cdot,\cdot)$ is a positive definite kernel e.g. the RBF kernel. The two terms in ϕ play different roles: the first term drives the particles towards the high probability areas of $p(\boldsymbol{x}|\mathcal{D})$ by following a smoothed gradient direction, which is the weighted sum of the gradients of all the points weighted by the kernel function. The second term acts as a repulsive force that prevents all the points to collapse together into local modes of $p(\boldsymbol{x}|\mathcal{D})$.

Stochastic Particle-Optimization Sampling (SPOS) Though obtaining significant empirical success, under certain conditions, SVGD experiences a theoretical pitfall, where particles tend to collapse. To overcome it, based on unified particle-optimization framework [3], [4] generalize POS to a stochastic setting by injecting random noise into particle updates. The update rule for particles $\{\boldsymbol{x}_\ell^{(i)}\}_{i=1}^M$ in (7) change to

$$\begin{aligned}\phi(\boldsymbol{x}_\ell^{(i)}) = \frac{1}{M}\sum_{j=1}^{M}[-\kappa(\boldsymbol{x}_\ell^{(j)},\boldsymbol{x}_\ell^{(i)})\nabla_{\boldsymbol{x}_\ell^{(j)}}U(\boldsymbol{x}_\ell^{(j)}) + \nabla_{\boldsymbol{x}_\ell^{(j)}}\kappa(\boldsymbol{x}_\ell^{(j)},\boldsymbol{x}_\ell^{(i)})] \\ - \beta^{-1}\nabla_{\boldsymbol{x}_\ell^{(i)}}U(\boldsymbol{x}_\ell^{(i)}) + \sqrt{2\beta^{-1}\epsilon_\ell^{-1}}\xi_\ell^{(i)}\end{aligned} \tag{8}$$

where $\xi_\ell^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the random Gaussian noise that would enhance the ability of the algorithm to jump out of local modes, leading to better ergodic properties compared to standard SVGD. The non-asymptotic convergence of SPOS is proved by [4].

Appendix C Experiment Details

For all our experiments, RBF kernel $\kappa(x,y) = \exp(-\frac{1}{h}\|x-y\|_2^2)$ with the bandwidth $h = med^2/\log M$ is used as the kernel function, where med denotes the median of the pairwise distance between current particles. Please find details of the datasets and experiments in Appendix C.

Author Profiling The Author Profiling task is to use tweets from Twitter as input to predict the age range (18-24, 25-34, 35-49, 50-64, 65+) of the user. We use the same Age dataset as in [1] which contains 68485 tweets as training set, 4000 as development set, and 4000 as test set.

Sentiment Analysis Yelp dataset consists of 2.7M yelp reviews is used for sentiment analysis task. We take the review as input and predict the number of stars the user who wrote that review assigned to the corresponding business store. As in [1], we randomly select 500K review-star pairs as training set, and 2000 for development set, 2000 for test set. Different from [1], we use Spacy toolkit as tokenizer and GloVe (GloVe 840B 300D) as pretrained word embedding.

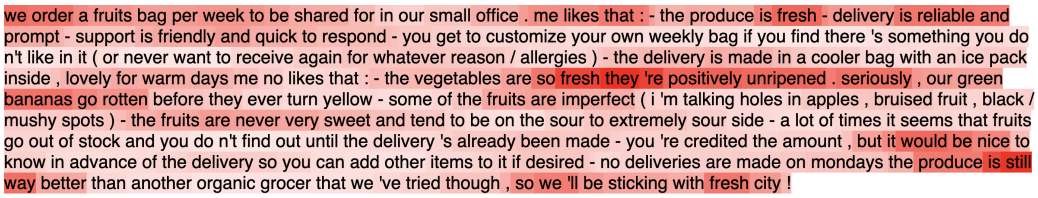
Textual Entailment The SNLI corpus [11] is used for this task. It is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels {*entailment*, *contradiction*, *neutral*}. The model will be given a pair of sentences, called hypothesis and premise respectively, and asked to tell if the semantics in the two sentences are entailment or contradiction or neutral. We applied the standard train (550k)/ validation (10k) / test (10k) split in this paper.

Transformer Experiments on Transformer are conducted on two widely used public datasets: IWSLT14 German-to-English (De-En) and WMT14 English-to-German (En-De) dataset. For the WMT14 dataset, we follow the `base` setting of Transformer in [7] which consistes a 6-layer encoder and a 6-layer decoder. The size of the hidden nodes and embeddings is 512 and the number of heads is 8. IWSLT14 dataset is much smaller than the WMT14 dataset, so we use the small setting of Transformer, whose size of hidden states and embeddings is set to 512 and the number of heads is set to 4. Our implementation is based on the open-sourced `fairseq` code base and follows the hyperparameter settings in [7].

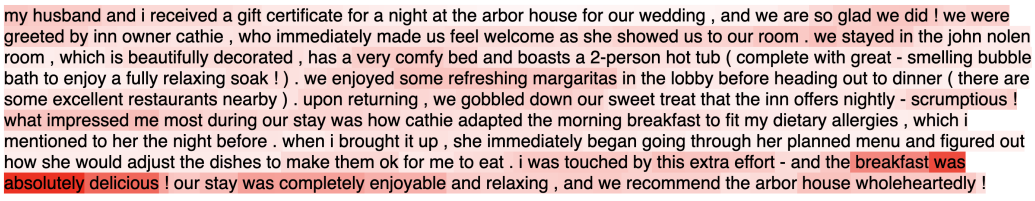
Graph Attention Networks We completely follow the model structure in [8] and evaluate our approach on Cora dataset. The Cora dataset contains 2708 nodes, 5429 edges, 7 classes and 1433 features per node. As in [8], we use 140 nodes for training, 500 for validation and 1000 for testing.

GraphWriter Experiments are conducted on the Abstract GENeration DAtaset (AGENDA), a dataset of knowledge graphs paired with scientific abstracts. Our dataset consists of 40k paper titles and abstracts from the Semantic Scholar Corpus taken from the proceedings of 12 top AI conferences. We use the standard split of AGENDA dataset in our experiments: 38,720 for training, 1000 for validation, and 1000 for test.

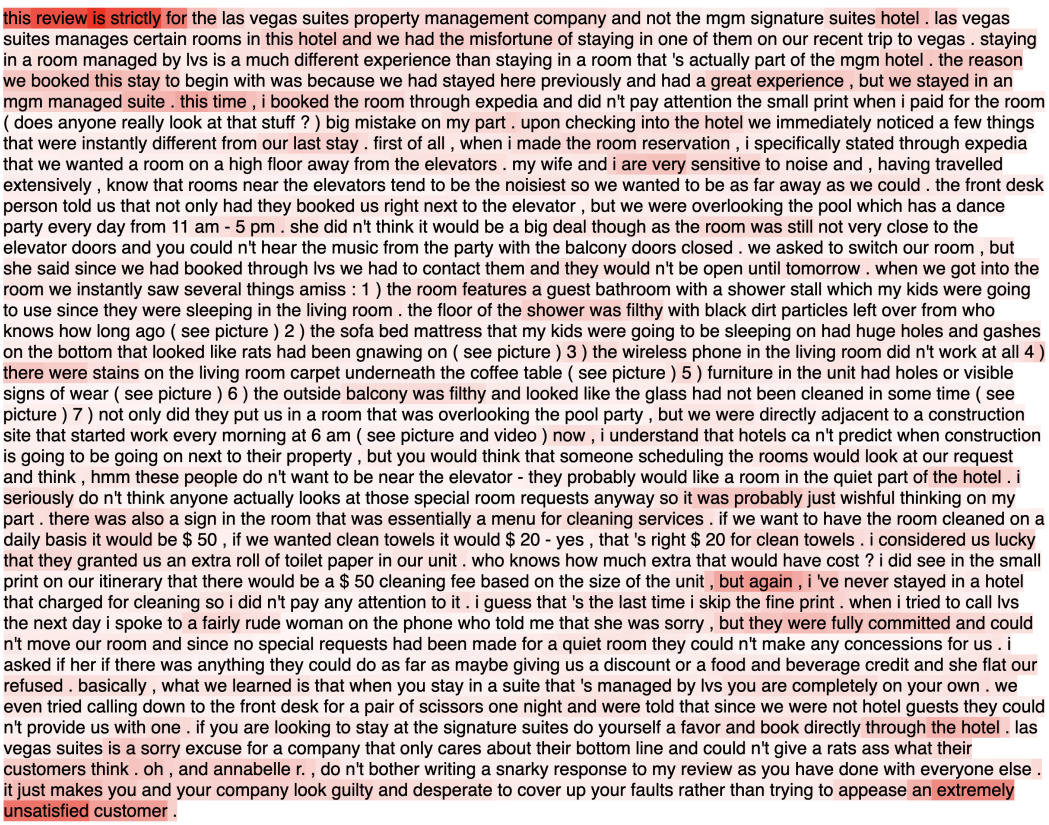
Appendix D Visualization of Multi-head Attention



(a) Example 1



(b) Example 2



(c) Example 3

Figure 2: Heatmap of Yelp reviews with original multi-head attention

we order a fruits bag per week to be shared for in our small office . me likes that : - the produce is fresh - delivery is reliable and prompt - support is friendly and quick to respond - you get to customize your own weekly bag if you find there 's something you do n't like in it (or never want to receive again for whatever reason / allergies) - the delivery is made in a cooler bag with an ice pack inside , lovely for warm days me no likes that : - the vegetables are so fresh they 're positively unripened . seriously , our green bananas go rotten before they ever turn yellow - some of the fruits are imperfect (i 'm talking holes in apples , bruised fruit , black / mushy spots) - the fruits are never very sweet and tend to be on the sour to extremely sour side - a lot of times it seems that fruits go out of stock and you do n't find out until the delivery 's already been made - you 're credited the amount , but it would be nice to know in advance of the delivery so you can add other items to it if desired - no deliveries are made on mondays the produce is still way better than another organic grocer that we 've tried though , so we 'll be sticking with fresh city !

(a) Example 1

my husband and i received a gift certificate for a night at the arbor house for our wedding , and we are so glad we did ! we were greeted by inn owner cathie , who immediately made us feel welcome as she showed us to our room . we stayed in the john nolen room , which is beautifully decorated , has a very comfy bed and boasts a 2-person hot tub (complete with great - smelling bubble bath to enjoy a fully relaxing soak !) . we enjoyed some refreshing margaritas in the lobby before heading out to dinner (there are some excellent restaurants nearby) . upon returning , we gobbled down our sweet treat that the inn offers nightly - scrumptious ! what impressed me most during our stay was how cathie adapted the morning breakfast to fit my dietary allergies , which i mentioned to her the night before . when i brought it up , she immediately began going through her planned menu and figured out how she would adjust the dishes to make them ok for me to eat . i was touched by this extra effort - and the breakfast was absolutely delicious ! our stay was completely enjoyable and relaxing , and we recommend the arbor house wholeheartedly !

(b) Example 2

this review is strictly for the las vegas suites property management company and not the mgm signature suites hotel . las vegas suites manages certain rooms in this hotel and we had the misfortune of staying in one of them on our recent trip to vegas . staying in a room managed by lvs is a much different experience than staying in a room that 's actually part of the mgm hotel . the reason we booked this stay to begin with was because we had stayed here previously and had a great experience , but we stayed in an mgm managed suite . this time , i booked the room through expedia and did n't pay attention the small print when i paid for the room (does anyone really look at that stuff ?) big mistake on my part . upon checking into the hotel we immediately noticed a few things that were instantly different from our last stay . first of all , when i made the room reservation , i specifically stated through expedia that we wanted a room on a high floor away from the elevators . my wife and i are very sensitive to noise and , having travelled extensively , know that rooms near the elevators tend to be the noisiest so we wanted to be as far away as we could . the front desk person told us that not only had they booked us right next to the elevator , but we were overlooking the pool which has a dance party every day from 11 am - 5 pm . she did n't think it would be a big deal though as the room was still not very close to the elevator doors and you could n't hear the music from the party with the balcony doors closed . we asked to switch our room , but she said since we had booked through lvs we had to contact them and they would n't be open until tomorrow . when we got into the room we instantly saw several things amiss : 1) the room features a guest bathroom with a shower stall which my kids were going to use since they were sleeping in the living room . the floor of the shower was filthy with black dirt particles left over from who knows how long ago (see picture) 2) the sofa bed mattress that my kids were going to be sleeping on had huge holes and gashes on the bottom that looked like rats had been gnawing on (see picture) 3) the wireless phone in the living room did n't work at all 4) there were stains on the living room carpet underneath the coffee table (see picture) 5) furniture in the unit had holes or visible signs of wear (see picture) 6) the outside balcony was filthy and looked like the glass had not been cleaned in some time (see picture) 7) not only did they put us in a room that was overlooking the pool party , but we were directly adjacent to a construction site that started work every morning at 6 am (see picture and video) now , i understand that hotels ca n't predict when construction is going to be going on next to their property , but you would think that someone scheduling the rooms would look at our request and think , hmm these people do n't want to be near the elevator - they probably would like a room in the quiet part of the hotel . i seriously do n't think anyone actually looks at those special room requests anyway so it was probably just wishful thinking on my part . there was also a sign in the room that was essentially a menu for cleaning services . if we want to have the room cleaned on a daily basis it would be $ 50 , if we wanted clean towels it would $ 20 - yes , that 's right $ 20 for clean towels . i considered us lucky that they granted us an extra roll of toilet paper in our unit . who knows how much extra that would have cost ? i did see in the small print on our itinerary that there would be a $ 50 cleaning fee based on the size of the unit , but again , i 've never stayed in a hotel that charged for cleaning so i did n't pay any attention to it . i guess that 's the last time i skip the fine print . when i tried to call lvs the next day i spoke to a fairly rude woman on the phone who told me that she was sorry , but they were fully committed and could n't move our room and since no special requests had been made for a quiet room they could n't make any concessions for us . i asked if her if there was anything they could do as far as maybe giving us a discount or a food and beverage credit and she flat our refused . basically , what we learned is that when you stay in a suite that 's managed by lvs you are completely on your own . we even tried calling down to the front desk for a pair of scissors one night and were told that since we were not hotel guests they could n't provide us with one . if you are looking to stay at the signature suites do yourself a favor and book directly through the hotel . las vegas suites is a sorry excuse for a company that only cares about their bottom line and could n't give a rats ass what their customers think . oh , and annabelle r. , do n't bother writing a snarky response to my review as you have done with everyone else . it just makes you and your company look guilty and desperate to cover up your faults rather than trying to appease an extremely unsatisfied customer .

(c) Example 3

Figure 3: Heatmap of Yelp reviews with repulsive multi-head attention