Interspeech 2022
18-22 September 2022, Incheon, Korea

Bayesian Transformer Using Disentangled Mask Attention

Jen-Tzung Chien, Yu-Han Huang

Dept of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan
{jtchien,yhhuang.ee08}@nycu.edu.tw

Abstract

Transformer conducts self attention which has achieved state-of-the-art performance in many applications. Multi-head attention in transformer basically gathers the features from individual tokens in input sequence to form the mapping to output sequence. There are twofold weaknesses in transformer. First, due to the natural property that attention mechanism would mix up the features of different tokens in input and output sequences, it is likely that the representation of input tokens contains redundant information. Second, the patterns of attention weights between different heads tend to be similar, the model capacity is bounded. To strengthen the sequential learning, this paper presents a variational disentangled mask attention in transformer where the redundant features are enhanced with semantic information. Latent disentanglement in multi-head attention is learned. The attention weights are filtered by a mask which is optimized by semantic clustering. The proposed attention mechanism is then implemented according to a Bayesian learning for clustered disentanglement. The experiments on machine translation show the merit of the disentangled mask attention.

Index Terms: sequential learning, disentangled representation, mask attention, Bayesian clustering, transformer

1. Introduction

Attention mechanism has been achieving the promising performance in different sequence-to-sequence learning tasks. In recent years, transformer [1, 2] has obtained state-of-the-art results on sequential learning in the applications of speech recognition, machine translation, dialogue generation [3], language understanding [4] to name a few. In spite of the success of self attention [5] in transformer, there are still some issues which restrict the learning performance such as the inference speed, computational complexity and representation redundancy, etc. To deal with theses challenges, several variants of transformer were proposed by using mask attention schemes [6]. Accordingly, the sparse transformer [7] was proposed to calculate the dot-product in attention by only using a small portion of tokens. The computational complexity was reduced by applying binary mask on attention weights. Contrarily, the adaptively sparse transformer [8, 9, 10] was proposed to construct the real-valued mask where the α-entmax function [11] was presented to filter out redundancy features in attention. In [12], a mask attention network was built to carry out a dynamic attention mask based on the distance between two tokens in a sequence. In addition, the transformer [13] with the Gaussian-weighted self-attention [14] was exploited by constructing an attention mask where the relations for the pairs of tokens were measured. In addition, it was shown that some of attention heads were redundant [8] or the attention weights lacked the semantic interpretation. In [15], it was found that the similarity of attention patterns between individual heads was high in vanilla transformer so that similar performance was obtained after pruning some attention heads. In [16], the adversarial training was applied to estimate the attention weights of transformer, and the resulting transformer received similar outputs even the attention weights were considerably different. This phenomenon revealed that attention weights might not contain sufficient semantics.

This paper aims to increase the semantic meaning as well as reduce the redundancy of attention weights within each head and across different heads. A new disentangled transform is constructed through two stages. The first stage is to disentangle the representation of attention weights within individual heads. The semantics of these heads are represented via a latent topic model through a variational sequence-to-sequence learning [17, 18, 19] based on the mixture of Gaussians as the prior model. Bayesian clustering is performed to construct a semantic mask and the mask is applied over the attention weights in latent space for those semantically-close tokens. The real-valued clusters of attention mask are implemented to strengthen the attention mechanism by a variational inference procedure. The second stage is to reduce the redundancy of attention weights and disentangle the multi-head attention across various heads. The mutual information of query vectors between two heads is calculated as the disentanglement objective which is minimized to reduce the redundancy of attention patterns in attention-based representation. The contributions and the novelties of this work are summarized. First, a semantic mask on attention weights is proposed to reduce the attention redundancy and enhance the weights for semantically similar tokens. Second, a stochastic clustering is incorporated to implement latent disentanglement for variational attention. Third, a variational sequence-to-sequence model is carried out for a probabilistic transformer. Four, the experiments on machine translation illustrate the performance improvement in terms of BLEU score and model size.

2. Bayesian Disentangled Learning

2.1. Disentangled representation

Disentangled representation is a line of researches [20, 21, 22] which aims to factorize the latent representation into several independent low-dimensional representations by optimizing a specialized objective function [23, 24]. This study pursues the mutually independent latent variables in accordance with the variation of information (VI). VI between variables $\mathbf{z}_i$ and $\mathbf{z}_j$ acts as the metric to measure the degree of independence which is nonnegative and is defined through the notations of entropy $\mathrm{H}(\cdot)$ and mutual information $\mathrm{I}(\cdot,\cdot)$ as $\mathrm{VI}(\mathbf{z}_i,\mathbf{z}_j) = \mathrm{H}(\mathbf{z}_i)+\mathrm{H}(\mathbf{z}_j)-2\mathrm{I}(\mathbf{z}_i,\mathbf{z}_j)$. Considering the disentanglement of variables $\mathbf{z}_i$ and $\mathbf{z}_j$ from original variable $\mathbf{x}$, the triangular inequality for these variables is held by $\mathrm{VI}(\mathbf{x},\mathbf{z}_i)+\mathrm{VI}(\mathbf{x},\mathbf{z}_j) \geq \mathrm{VI}(\mathbf{z}_i,\mathbf{z}_j)$. The equality only holds when $\mathbf{z}_i$ and $\mathbf{z}_j$ are statistically independent. VI is closely related to mutual information (MI). Minimizing MI is comparable to maximizing VI to learn independent components. More specifically, the objective of disentanglement for independence between $\mathbf{z}_i$ and $\mathbf{z}_j$ from $\mathbf{x}$ is

10.21437/Interspeech.2022-10457

measured by the difference $D(\cdot)$ between two sides of triangular inequality as

$$\begin{aligned} D(\mathbf{x};\mathbf{z}_i,\mathbf{z}_j) &= \mathrm{VI}(\mathbf{x},\mathbf{z}_i)+\mathrm{VI}(\mathbf{x},\mathbf{z}_j)-\mathrm{VI}(\mathbf{z}_i,\mathbf{z}_j) \\ &= 2(\mathrm{H}(\mathbf{x})+\mathrm{I}(\mathbf{z}_i,\mathbf{z}_j)-\mathrm{I}(\mathbf{x},\mathbf{z}_i)-\mathrm{I}(\mathbf{x},\mathbf{z}_j)) \end{aligned} \tag{1}$$

where $\mathrm{H}(\mathbf{x})$ is seen as a constant and the remaining MI terms are required during model optimization. However, direct calculation of MI is intractable. There are different upper and lower bounds of MI provided in [25]. These bounds are feasible to estimate MI and build information-theoretic objectives without the calculation of true value of MI. Latent disentanglement is performed by minimizing $D(\mathbf{x};\mathbf{z}_i,\mathbf{z}_j)$, or equivalently minimizing the upper bound of $\mathrm{I}(\mathbf{z}_i,\mathbf{z}_j)$ and simultaneously maximizing the lower bounds of $\mathrm{I}(\mathbf{x},\mathbf{z}_i)$ and $\mathrm{I}(\mathbf{x},\mathbf{z}_j)$.

2.2. Bayesian clustering with GMM prior

In addition, this paper presents the semantic mask attention where the Bayesian clustering in neural network (NN) [26, 27] is performed. In [28], a Gaussian mixture model (GMM) was introduced to carry out the variational deep embedding where the distribution of latent embedding in NN was characterized. Each latent sample $\mathbf{z}$ of observation $\mathbf{x}$ belongs to a cluster c according to a GMM $p(\mathbf{z}) = \sum_c p(c)p(\mathbf{z}|c) = \sum_c \pi_c^z \mathcal{N}(\boldsymbol{\mu}_c^z, \mathrm{diag}\{(\boldsymbol{\sigma}_c^z)^2\})$ where $\boldsymbol{\pi}^z = \{\pi_c^z\} \in \mathbb{R}^{n_c}$ denotes the weights of n_c clusters, $\boldsymbol{\mu}^z = \{\boldsymbol{\mu}_c^z\} \in \mathbb{R}^{n_c \times d}$ denotes the mean vectors, and $(\boldsymbol{\sigma}^z)^2 = \{(\boldsymbol{\sigma}_c^z)^2\} \in \mathbb{R}^{n_c \times d}$ denotes the variance entries of a diagonal matrix. A Bayesian clustering method is developed via variational inference based on a specialized variational autoencoder (VAE) [29] by maximizing the likelihood of training data $\mathbf{x}$ which are encoded as embedding $\mathbf{z}$ in latent space where $\mathbf{z}$ is modeled by a GMM for reconstruction of $\mathbf{x}$ in decoder. The loss is derived as a negative evidence lower bound (ELBO) in the right-hand-side of the following

$$\begin{aligned} &\log p(\mathbf{x}) = \log \int \sum_c p(\mathbf{x},\mathbf{z},c)d\mathbf{z} \geq \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{x}|\mathbf{z})] \\ &- \mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|c)) - \mathrm{KL}(q(c|\mathbf{x})\|p(c)) \triangleq -\mathcal{L}_{\mathrm{ELBO}} \end{aligned} \tag{2}$$

where the first term represents a reconstruction objective for $\mathbf{x}$ under a latent variable model and the remaining terms imply the Kullback-Leibler (KL) divergence due to latent variables $\mathbf{z}$ and c driven by a variational distribution $q(\mathbf{z},\mathbf{c}|\mathbf{x})$ which is close to a prior of GMM using $p(\mathbf{z}|c)$ and $p(c)$ through KL minimization. Here, c denotes a latent cluster of $\mathbf{z}$ corresponding to an input $\mathbf{x}$. The Bayesian clustering is implemented as a new type of VAE where a mixture of Gaussians is adopted as the variational distribution $q(\mathbf{z},\mathbf{c}|\mathbf{x})$. A deep clustering and embedding model is constructed to represent semantic topics which can be employed in semantic mask attention based on transformer.

3. Disentangled Mask Attention

This paper presents the disentangled mask attention (DMA) to build a disentangled transformer. DMA is implemented to replace the attention module in the encoder and decoder of a vanilla transformer. This DMA is constructed with three schemes. First, the semantic mask attention is presented to enhance the semantic meaning of attention weights [30] via latent clustering. Second, the disentangled attention heads are calculated by maximizing the independence among attention heads for the queries $\mathbf{q}$ of word sequences $\mathbf{x}$. Third, the information-theoretic disentanglement is implemented to carry out a variational learning procedure of semantic-aware transformer. The detailed solution is addressed in what follows.

3.1. Semantic mask attention

Given a source sequence $\mathbf{x}$ and a target sequence $\mathbf{y}$, a new Bayesian transformer for sequence-to-sequence (S2S) learning is presented by merging GMM to express the prior distribution of latent variable $\mathbf{z}$. By extending the Bayesian clustering in Eq. (2), this paper presents a novel latent variable model for transformer by minimizing S2S classifier loss or the negative ELBO of conditional likelihood $p(\mathbf{y}|\mathbf{x})$ of training data $\{\mathbf{x},\mathbf{y}\}$

$$\begin{aligned} \mathcal{L}_{\mathrm{s2s}} = &-\sum_n \mathbb{E}_{\mathbf{z}^{n,h}\sim q(\mathbf{z}^{n,h}|\mathbf{x})}\left[\log p(\mathbf{y}|\mathbf{z}^{n,h},\mathbf{x})\right] \\ &+ \mathrm{KL}(q(\mathbf{z}^{n,h},\mathbf{c}_z^{n,h}|\mathbf{x})\|p(\mathbf{z}^{n,h},\mathbf{c}_z^{n,h})) \end{aligned} \tag{3}$$

where $\mathbf{z}^{n,h}$ is the feed-forward network output of the nth transformer layer after applying the head separation for multi-head attention, h is the head index, and $\mathbf{c}_z^{n,h}$ denotes the semantic clusters of layer n and head h. In this latent variable model, there are two latent variables $\mathbf{z}^{n,h}$ and $\mathbf{c}_z^{n,h}$. The stochastic gradient variational Bayes estimator [29] is applied to draw latent sample $\mathbf{z}^{n,h}$ from the variational distribution $q(\mathbf{z}^{n,h}|\mathbf{x})$ where the Gaussian parameters are calculated by the transformer layer. Alternatively, the probability of latent sample $\mathbf{c}_z^{n,h}$ corresponding to $\mathbf{z}_i^{n,h}$ of token $\mathbf{x}_i$ for cluster c is calculated by

$$p(\mathbf{c}_z^h = c|\mathbf{z}_i^h) = \frac{\pi_c^h \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_c^h, \mathrm{diag}\{(\boldsymbol{\sigma}_c^h)^2\})}{\sum_{c'} \pi_{c'}^h \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_{c'}^h, \mathrm{diag}\{(\boldsymbol{\sigma}_c^h)^2\})}. \tag{4}$$

Overall, the parameters of transformer layer and GMM $\{\pi_c^h, \boldsymbol{\mu}_c^h, (\boldsymbol{\sigma}_c^h)^2\}$ are trained by minimizing $\mathcal{L}_{\mathrm{s2s}}$ in Eq. (3). For ease of expression, the layer index n is ignored hereafter. Through the estimated prior of latent variable c using GMM, the semantic relation between two variables $\mathbf{z}_i^h$ and $\mathbf{z}_j^h$ associated with word tokens $\mathbf{x}_i$ and $\mathbf{x}_j$ is characterized by constructing the semantic mask $M = \{M_{ij}^h\}$ based on the clustering probability of tokens belonging to the same cluster c

$$M_{ij}^h = \frac{\sum_c p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)p(\mathbf{c}_z^h = c|\mathbf{z}_j^h)}{\sum_j \sum_{c'} p(\mathbf{c}_z^h = c'|\mathbf{z}_i^h)p(\mathbf{c}_z^h = c'|\mathbf{z}_j^h)}. \tag{5}$$

This calculation measures a probabilistic correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ driven and integrated by different semantic clusters c. This semantic mask is helpful to enrich the semantic information of attention weight as $\bar{A}_{ij}^h$ by using M_{ij}^h as the soft mask of original attention weight A_{ij} in a form of

$$\bar{A}_{ij}^h = \frac{M_{ij}^h A_{ij}^h}{\sum_{j'} M_{ij'}^h A_{ij'}^h}, A_{ij}^h = \mathrm{Softmax}\left(((\mathbf{k}_{1:J}^h)^T \mathbf{q}_i^h)/\sqrt{d_k}\right)_j$$

where attention weight A_{ij}^h is calculated by dot-product of query $\mathbf{q}_i^h = W_q^h \mathbf{x}_i + \mathbf{b}_q^h$ and key $\mathbf{k}_j^h = W_k^h \mathbf{x}_j + \mathbf{b}_k^h$ transformed by $\{W_q^h, \mathbf{b}_q^h\}$ and $\{W_k^h, \mathbf{b}_k^h\}$, respectively. A softmax function is calculated over $\mathbf{k}_{1:J}^h$ for a set of J key tokens of head h, where $\mathbf{k}_j^h \in \mathbb{R}^{d_k}$, and then retrieved by jth entry.

3.2. Disentangled attention heads

Semantic mask attention enhances the semantic meaning of attention weights in each head by using the estimated $M = \{M_{ij}^h\}$. However, different heads are likely similar such that the redundancy in multi-head representation does exist. To alleviate this issue, this paper implements a new disentangled

mask attention (DMA) as shown in Figure 1 which focuses on the disentanglement of query vectors $\mathbf{q}^h$ in multi-head attention. The disentangled queries are used to implement the semantic mask attention. The disentanglement objective $\mathcal{L}_D = \sum_{h=1}^{n_h} \sum_{h' \neq h}^{n_h} \mathrm{I}(\mathbf{q}^h, \mathbf{q}^{h'}) - \mathrm{I}(\mathbf{x}, \mathbf{q}^h)$ is constructed by extending the measure of independence over two latent vectors in Eq. (1) to that over multiple query vectors for n_h heads where n_h is the number of attention heads and $\mathbf{q}^h = \{\mathbf{q}_i^h\}$ denotes all queries of tokens $\mathbf{x} = \{\mathbf{x}_i\}$ in head h. Latent disentanglement is performed by minimizing $\mathcal{L}_D$ which correspondingly pursues the *semantic independence* for individual queries $\{\mathbf{q}_i^h\}$ across different heads h. This is because that the semantic mask attention is applied by using those queries driven by the semantic clusters or latent topics via GMM. The independence of queries $\mathbf{q}^h = \{\mathbf{q}_i^h\}$ is enhanced, and the redundancy of attention weights across different heads is accordingly reduced.

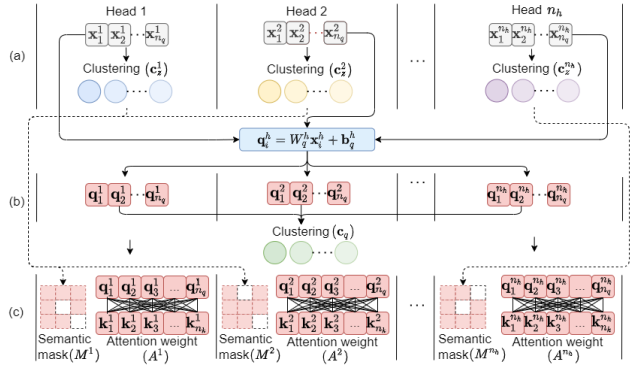

Figure 1: *Implementation for disentangled mask attention where n_q and n_k denotes the number of query and key tokens, respectively. First in* ***(a) semantic clustering****, the input $\mathbf{x}_i^h$ of token i in head h is clustered as $\mathbf{c}_z^h$. The clustering probability is used to construct the semantic mask M for mask attention indicated by dotted line. Then in* ***(b) disentangled heads****, $\mathbf{x}_i^h$ is transformed to query $\mathbf{q}_i^h$ by $\{W_q^h, \mathbf{b}_q^h\}$, and is clustered by a set of additional clusters $\mathbf{c}_q$ for disentanglement of $\mathbf{q}_i^h$ over different heads. Finally in* ***(c) mask attention****, the semantic mask M is employed in attention weight A^h calculated by $\mathbf{q}^h$ and $\mathbf{k}^h$ to construct a new attention weight. This procedure is run over n_h heads. Calculating key $\mathbf{k}$ is omitted in this figure.*

The exact computation of MI terms $\mathrm{I}(\mathbf{q}^h, \mathbf{q}^{h'})$ and $\mathrm{I}(\mathbf{q}^h, \mathbf{x})$ is intractable. To pursue the independence, the minimization of disentanglement objective $\mathcal{L}_D$ is performed by minimizing the upper bound $\mathcal{L}_{D_{qq}}$ of the first MI term $\mathrm{I}(\mathbf{q}^h, \mathbf{q}^{h'})$ derived as

$$\mathcal{L}_{D_{qq}} = \mathbb{E}\left[\sum_{i=1}^{n_q}\left(\frac{\log p(\mathbf{q}_i^h|\mathbf{q}_i^{h'})}{n_q} - \frac{\sum_{j\neq i}^{n_q}\log p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})}{n_q(n_q-1)}\right)\right] \quad (6)$$

and simultaneously maximizing the lower bound $\mathcal{L}_{D_{xq}}$ of the second MI term $\mathrm{I}(\mathbf{x}, \mathbf{q}^h)$ obtained by [31]

$$\mathcal{L}_{D_{xq}} = \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n_q}\left(\log\frac{\exp(f(\mathbf{q}_i^h, \mathbf{x}_i))}{\frac{1}{n_q}\sum_{j=1}^{n_q}\exp(f(\mathbf{q}_i^h, \mathbf{x}_j))}\right)\right] \quad (7)$$

where n_q is the number of samples in $\mathbf{q}^h$ or $\mathbf{q}^{h'}$. In Eq. (6), a variational leave-one-out upper bound of MI between $\mathbf{q}^h$ and $\mathbf{q}^{h'}$ is calculated. The conditional probabilities $p(\mathbf{q}_i^h|\mathbf{q}_i^{h'})$ and $p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})$ are estimated by a trainable NN. In Eq. (7), a variational lower bound is measured by using a critic function $f(\mathbf{q}_i^h, \mathbf{x}_j)$ which identifies the semantic relation between $\mathbf{q}_i^h$ and $\mathbf{x}_j$. The critic function $f(\mathbf{q}_i^h, \mathbf{x}_j) \triangleq \sum_c p(\mathbf{c}_q = c|\mathbf{q}_i^h)p(\mathbf{c}_q = c|\mathbf{q}_j^h)$ reflects an integrated correlation over different semantic clusters c. The probabilistic correlation between $\mathbf{q}_i^h$ and $\mathbf{x}_j$ under the same cluster or latent topic $\mathbf{c}_q = c$ is measured. Importantly, a new GMM is introduced to express the prior of latent query as $p(\mathbf{q}_i^h) = \sum_c \pi_c^q \mathcal{N}(\boldsymbol{\mu}_c^q, \mathrm{diag}\{(\boldsymbol{\sigma}_c^q)^2\})$ with parameters $\{\pi_c^q, \boldsymbol{\mu}_c^q, (\boldsymbol{\sigma}_c^q)^2\}$. The posterior probability of the cluster of query $p(\mathbf{c}_q = c|\mathbf{q}_i^h)$ is computed similar to that of transformer layer output $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ as shown in Eq. (4).

3.3. Bayesian hierarchical learning

This study presents the semantic mask attention based on the disentangled attention heads for sequence-to-sequence (S2S) learning. There are four latent variables in this Bayesian hierarchical model. The first level involves latent variables of transformer layer outputs and semantic clusters in different heads and layers $\{\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h}\}$ while the second level consists of queries and query clusters in different layers $\{\mathbf{q}^{n,h}, \mathbf{c}_q^n\}$. Attention heads are disentangled over queries $\mathbf{q}^h$ across different heads h in a layer n. By extending Eq. (3), the latent variable model is trained by minimizing the negative ELBO or S2S classifier loss $\mathcal{L}_{\mathrm{s2s}} = -\sum_n \mathbb{E}_{\mathbf{z}^{n,h}, \mathbf{q}^{n,h}}\left[\log p(\mathbf{y}|\mathbf{z}^{n,h}, \mathbf{q}^{n,h}, \mathbf{x})\right] + \mathrm{KL}(q(\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h}|\mathbf{x})\|p(\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h})) + \mathrm{KL}(q(\mathbf{q}^{n,h}, \mathbf{c}_q^n|\mathbf{x})\| p(\mathbf{q}^{n,h}, \mathbf{c}_q^n))$. An additional KL term is derived to regularize the estimated variational distribution $q(\mathbf{q}_i^h|\mathbf{x})$ which is close to its shared GMM prior $p(\mathbf{q}_i^h)$. The disentanglement of attention weights A_{ij} is performed via that of queries $\mathbf{q}^h$ over the groups of queries across different heads h. A new type of *disentangled transformer* is fulfilled by minimizing an information-theoretic objective by using the estimators of MI or the upper bound $\mathcal{L}_{D_{qq}}$ and lower bound $\mathcal{L}_{D_{xq}}$ of MI. The disentanglement objective is then formed by $\mathcal{L}_D = \mathcal{L}_{D_{qq}} - \mathcal{L}_{D_{xq}}$ where $p(\mathbf{q}_i^h|\mathbf{q}_i^{h'})$ in $\mathcal{L}_{D_{qq}}$ is calculated by integrating the probabilities of queries $\{\mathbf{q}_i^h, \mathbf{q}_i^{h'}\}$ under the same cluster c as $p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) = \sum_c p(\mathbf{q}_i^h|\mathbf{c}_q = c)p(\mathbf{c}_q = c|\mathbf{q}_i^{h'})$ where $p(\mathbf{c}_q = c|\mathbf{q}_i^{h'})$ is computed by referring Eq. (4) and the first term is computed via GMM as $p(\mathbf{q}_i^h|\mathbf{c}_q = c) = \frac{p(\mathbf{c}_q=c|\mathbf{q}_i^h)p(\mathbf{q}_i^h)}{\sum_{i'} p(\mathbf{c}_q=c|\mathbf{q}_{i'}^h)p(\mathbf{q}_{i'}^h)}$. In addition, the loss functions $\mathcal{L}_{D_{qq}}$ and $\mathcal{L}_{D_{xq}}$ in Eqs. (6) and (7) are calculated over all queries over n_h heads including n_q samples in each head. The overall loss $\mathcal{L}$ for the transformer with disentangled mask attention is composed of classification loss, disentanglement loss and diversity loss as $\mathcal{L} = \mathcal{L}_{\mathrm{s2s}} + \mathcal{L}_D$ with the regularization parameters in two objectives. The parameters of the encoder based on transformation of query, key and value $\{W_q^h, \mathbf{b}_q^h, W_k^h, \mathbf{b}_k^h, W_v^h, \mathbf{b}_v^h\}$ and the GMMs $\{\pi_c^z, \boldsymbol{\mu}_c^z, (\boldsymbol{\sigma}_c^z)^2, \pi_c^q, \boldsymbol{\mu}_c^q, (\boldsymbol{\sigma}_c^q)^2\}$ are estimated by finding the corresponding gradients over $\mathcal{L}$.

4. Experiments

In the experiments, three machine translation tasks including IWSLT'14 De-En, WMT'14 En-De and WMT'17 Zh-En were used. All models were evaluated by using Fairseq [32] toolkit.

4.1. Experimental setup

IWSLT'14 De-En contained 167K of German and English sentence pairs, and WMT'14 En-De contained 4.5M of English and

German sentence pairs. For WMT'17 Zh-En, only the data in 'news-commentary-v12' set were collected for training. There were 212K of Chinese and English sentence pairs. The byte-pair-encoding (BPE) dictionary of the models in WMT'14 En-De were shared between encoder and decoder. For the other two tasks, encoder and decoder have independent BPE dictionary. The performance of translation models were evaluated in terms of BLEU score similar to the setting in [33]. In addition, two metrics were introduced to measure the redundancy of attention weights based on the layer redundancy (LR) and the head redundancy (HR) using the Jensen-Shannon (JS) distance [8, 15]. LR measures the similarity of attention weights between different attention heads in the same layer while HR measures the similarity between each attention head in whole model. The lower the redundancy metric, the larger the difference of attention weights between each head.

The vanilla transformer was used as baseline model. All settings in transformer and the proposed DMA transformer were identical for fair comparison. All models in the experiments were composed of six layers of encoder and decoder. The semantic mask was disregarded in the masked DMA block of transformer decoder in test phase. DMA transformer empirically worked well under this setting. In order to rapid the convergence of learning GMMs, the gradient of cluster centroid was multiplied by a constant $c_{\text{grad}} = 10$ during training, except that the models on WMT'17 Zh-En was trained with $c_{\text{grad}} = 5$. To evaluate the proposed model under different model sizes, DMA transformer was built by three configuration types with different size of attention and feed-forward layers, which were *base*, *small*, and *tiny*. In loss calculation, there were three regularization parameters c_{qq}, c_{xq}, and c_{kl} which were used to control the contributions of loss terms $\mathcal{L}_{D_{qq}}$ and $\mathcal{L}_{D_{xq}}$, and KL terms in $\mathcal{L}_{s2s}$. There were 4 clusters in all configurations of DMA transformer, except that DMA transformer (*base*) used 8 clusters on IWSLT'14 De-En.

4.2. Experimental results

The results on three tasks are illustrated. The proposed DMA transformer achieved higher BLEU score than baseline transformer (*base*) [34] in various tasks. In IWSLT'14 De-En (Table 1), DMA transformer absolutely improved 0.8 BLEU score over transformer. In WMT'14 En-De (Table 2), DMA transformer outperformed transformer by 0.6 BLEU score. In WMT'17 Zh-En (Table 3), DMA transformer achieved 0.37 BLEU score higher than transformer. These results show that DMA transformer performs better than the other models for translation with grammatically similar and dissimilar languages.

Model	Params	LR/HR	BLEU
Transformer (*base*) [34]	42.0M	- /-	34.30
Transformer (*base*, ours)	39.5M	0.74/0.65	34.50
DMA transformer (*base*)	39.7M	0.62/0.53	**35.31**
DeLighT [34]	14.0M	-/-	33.80
DMA transformer (*small*)	26.1M	0.64/0.57	35.00
DMA transformer (*tiny*)	11.9M	0.63/0.58	34.96

Table 1: *Results on IWSLT'14 De-En translation task. 'Ours' denotes the model trained by ourselves. Model size is evaluated.*

In addition, the number of additional parameters using the proposed methods was only 1% more compared to vanilla transformer. In IWSLT'14 De-En (Table 1), DMA transformer (*tiny*) achieved 0.46 BLEU score higher than transformer by using only 30% of parameters. Compared to another state-of-the-art model, DeLighT [34], with smallest model size, DMA transformer (*tiny*) obtained about 1.16 BLEU score higher with using 5% less of parameters. In WMT14'En-De, DMA transformer (*small*) achieved 0.41 BLEU score higher than transformer with using 30% less of parameters (Table 2). Furthermore, the performance of DMA transformer (*base*) is close to transformer (*big*) with slightly 0.05 drop in terms of BLEU score, but using only one third of the size of parameters (Table 2). In conclusion, introducing the objectives of Bayesian clustering and disentangled attention heads into transformer does strengthen the latent representation of sequences so that similar or even higher performance can be achieved by using much smaller model.

Model	Params	LR/HR	BLEU
Transformer (*base*) [34]	67.0M	-/-	27.70
Transformer (*base*, ours)	66.5M	0.79/0.69	27.75
DMA transformer (*base*)	66.6M	0.73/0.58	**28.35**
DeLighT [34]	54.0M	-/-	28.00
DMA transformer (*small*)	46.4M	0.70/0.57	28.16
Transformer (*big*)	213.0M	-/-	28.40

Table 2: *Results on WMT'14 En-De translation task.*

Model	Params	LR/HR	BLEU
Transformer (ours)	55.0M	0.71/0.60	12.76
DMA transformer	55.1M	0.62/0.55	**13.13**

Table 3: *Results on WMT'17 Zh-En translation task.*

Furthermore, LR and HR metrics of DMA transformer in three translation tasks (Tables 1-3) are lower than transformer by 0.05 to 0.1, respectively, while the performance is improved. These results illustrate that the proposed methods could encourage different attention heads identifying different semantic relations between individual tokens. The model performance is benefited from the reduction of attention redundancy.

5. Conclusions

This paper presented a variant of transformer with the disentangled mask attention, which replaced the vanilla attention with the disentangled mask attention. The disentangled and variational learning was developed to enhance the compactness and robustness in attention-based representation. This variational attention represented the prior probability distribution of feed-forward output and query as the mixture of Gaussians, constructed the semantic mask based on the corresponding clustering probability, and optimized the model with the objective of disentangled attention heads. Experimental results showed the redundancy of attention weight was reduced, and the semantic diversity of query within the same head and across different heads was increased. The encoder and decoder in DMA transformer were examined. By applying the proposed methods for sequence-to-sequence learning, the compact DMA transformer outperformed the other transformers in different translation tasks either with identical or smaller model size.

6. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[2] W. Zhou, T. Ge, F. Wei, M. Zhou, and K. Xu, "Scheduled DropHead: A regularization method for transformer models," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 1971–1980.

[3] T.-C. Luo and J.-T. Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[4] C.-T. Chu, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Augmentation strategy optimization for language understanding," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7952–7956.

[5] J.-T. Chien and C.-W. Wang, "Hierarchical and self-attended sequence autoencoder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] H. Lio, S.-E. Li, and J.-T. Chien, "Adversarial mask transformer for sequential learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4178–4182.

[7] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[8] G. M. Correia, V. Niculae, and A. F. T. Martins, "Adaptively sparse transformers," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 2174–2184.

[9] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online compressive transformer for end-to-end speech recognition," *Proc. of Annual Conference of International Speech Communication Association*, pp. 2082–2086, 2021.

[10] S. Wu, X. Xiao, Q. Ding, P. Zhao, Y. Wei, and J. Huang, "Adversarial sparse transformer for time series forecasting," in *Proc. of International Conference on Neural Information Processing Systems*, 2020, pp. 17 105–17 115.

[11] B. Peters, V. Niculae, and A. F. T. Martins, "Sparse sequence-to-sequence models," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2019, pp. 1504–1519.

[12] Z. Fan, Y. Gong, D. Liu, Z. Wei, S. Wang, J. Jiao, N. Duan, R. Zhang, and X. Huang, "Mask attention networks: Rethinking and strengthen transformer," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1692–1701.

[13] M. Guo, Y. Zhang, and T. Liu, "Gaussian transformer: A lightweight approach for natural language inference," *Proc. of AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6489–6496, 2019.

[14] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6649–6653.

[15] Y. Bian, J. Huang, X. Cai, J. Yuan, and K. Church, "On attention redundancy: A comprehensive study," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 930–945.

[16] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 3543–3556.

[17] H. Bahuleyan, L. Mou, O. Vechtomova, and P. Poupart, "Variational attention for sequence-to-sequence models," in *Proc. of International Conference on Computational Linguistics*, 2018, pp. 1672–1682.

[18] J.-T. Chien and W.-H. Chang, "Dualformer: a unified bidirectional sequence-to-sequence learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7718–7722.

[19] J.-T. Chien and C.-J. Tsai, "Variational sequential modeling, learning and understanding," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 480–486.

[20] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. of International Conference on Machine Learning*, 2019, pp. 4114–4124.

[21] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *Proc. of International Conference on Machine Learning*, 2018, pp. 5670–5679.

[22] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," *Advances in Neural Information Processing Systems*, pp. 4417–4426, 2017.

[23] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," in *Proc. of International Conference on Learning Representations*, 2021.

[24] S.-J. Huang and J.-T. Chien, "Attribute decomposition for flow-based domain mapping," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 1710–1714.

[25] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," in *Proc. of International Conference on Learning Representations*, 2020.

[26] J.-T. Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[27] S. Watanabe and J.-T. Chien, *Bayesian speech and language processing*. Cambridge University Press, 2015.

[28] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: an unsupervised and generative approach to clustering," in *Proc. of International Joint Conference on Artificial Intelligence*, 2017, pp. 1965–1972.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. of International Conference on Learning Representations*, 2014.

[30] J.-T. Chien and Y.-H. Chen, "Learning continuous-time dynamics with attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[31] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. of the International Conference on Machine Learning*, 2019, pp. 5171–5180.

[32] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Demonstrations*, 2019, pp. 48–53.

[33] F. Wu, A. Fan, A. Baevski, Y. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," in *Proc. of International Conference on Learning Representations*, 2019.

[34] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "DeLighT: Deep and light-weight transformer," in *Proc. of International Conference on Learning Representations*, 2021.