DISENTANGLED MASK ATTENTION IN TRANSFORMER

Anonymous authors
Paper under double-blind review

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 learning representation using 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 representation capacity of the model might be bounded. To strengthen the sequential learning representation, this paper presents a new disentangled mask attention in transformer where the redundant features are reduced and the semantic information is enriched. 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 implemented for sequential learning according to the clustered disentanglement objective. The experiments on machine translation show the merit of this disentangled transformer in sequence-to-sequence learning tasks.

1 INTRODUCTION

Attention mechanism has been achieving the promising performance in different sequence-to-sequence learning tasks. In recent years, transformer (Vaswani et al., 2017) has obtained state-of-the-art results on sequential learning in the applications of speech recognition, machine translation, question answering, reading comprehension, to name a few. In spite of the success of self attention in transformer, there are still some issues which restrict the learning performance such as the inference speed, computational complexity and representation redundancy, etc. In order to deal with theses challenges, several variants of transformer were proposed by using mask attention schemes. Accordingly, the sparse transformer (Child et al., 2019), routing transformer (Roy et al., 2021) and reformer (Kitaev et al., 2020) were 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. In contrast to the binary mask, the adaptively sparse transformer (Correia et al., 2019) and the adversarial sparse transformer (Wu et al., 2020) were proposed to construct the real-valued mask where the α-entmax function (Peters et al., 2019) was presented to filter out redundancy features in attention. In (Fan et al., 2021), 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 (Guo et al., 2019) with Gaussian-weighted self-attention (Kim et al., 2020) was exploited by constructing an attention mask where the relations for the pairs of tokens were measured. In addition, there were a number of works pointing out that some of attention heads were redundant (Correia et al., 2019) or the attention weights lacking of semantic interpretation. In (Bian et al., 2021), 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 (Jain & Wallace, 2019), 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. To cope with these challenges, 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 (Bahuleyan et al., 2018) based on the mixture of Gaussians as the prior model. The probabilistic clustering is performed to construct a semantic mask and the mask is applied on 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. There are threefold novelties compared with vanilla transformer. 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. The experiments on machine translation tasks under different scenarios illustrate the performance of the proposed semantic mask with disentanglement objective.

2 UNSUPERVISED REPRESENTATION LEARNING

The unsupervised learning of disentangled and semantic features is surveyed for model construction.

2.1 DISENTANGLED LEARNING

Disentangled representation is a line of researches (Locatello et al., 2019; Yingzhen & Mandt, 2018; Denton & Birodkar, 2017) which aims to factorize the latent representation into several independent low-dimensional representations by optimizing a specialized objective function. These researches have been widely employed in various technical data such as image (Higgins et al., 2017; Chen et al., 2018), text (John et al., 2019; Cheng et al., 2020) and voice (Yuan et al., 2021). 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

$$\mathrm{VI}(\mathbf{z}_i,\mathbf{z}_j)=\mathrm{H}(\mathbf{z}_i)+\mathrm{H}(\mathbf{z}_j)-2\mathrm{I}(\mathbf{z}_i,\mathbf{z}_j). \tag{1}$$

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) as shown in Eq. (1). 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 measured by the difference $D(\cdot)$ between two sides of triangular inequality as

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

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 (Tschannen et al., 2020). These bounds are feasible to find the estimators of MI to 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 VARIATIONAL CLUSTERING

Based on the latent disentanglement, this paper presents the semantic mask attention where variational clustering in neural network is performed. In (Jiang et al., 2017), an Gaussian mixture model (GMM) was introduced to carry out the variational deep embedding where the distribution of latent embedding in neural network was characterized. Each latent sample $\mathbf{z}$ of observation $\mathbf{x}$ belongs to a cluster c according to an 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 set of variance entries of diagonal matrices. A variational clustering method is developed via variational inference based on a special type of

variational autoencoder (Kingma & Welling, 2014) by maximizing the likelihood of training data $\mathbf{x}$ which is encoded as feature vector $\mathbf{z}$ in latent space where $\mathbf{z}$ is modeled by an GMM and used for reconstruction of $\mathbf{x}$ in decoder. The loss is derived as a negative evidence lower bound (ELBO)

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int \sum_c p(\mathbf{x}, \mathbf{z}, c) d\mathbf{z} \\ &\geq \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}|c)) - \mathrm{KL}(q(c|\mathbf{x}) \| p(c)) \triangleq -\mathcal{L}_{\mathrm{ELBO}}.\end{aligned} \tag{3}$$

In Eq. (3), 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 sequence $\mathbf{x}$. Overall, the variational clustering is implemented as a new type of VAE where a mixture of Gaussians is adopted as the variational distribution $q(\mathbf{z}, \mathbf{c}|\mathbf{x})$. This method is feasible to build a deep clustering and embedding model to represent similar samples with the same semantic topics which can be developed and employed in semantic mask attention based on transformer.

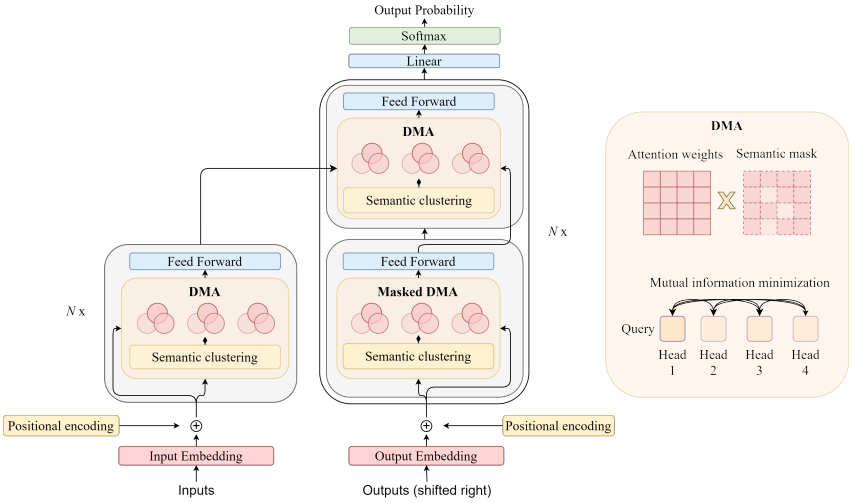

Figure 1: Architecture of transformer driven by the disentangled mask attention.

3 Disentangled Mask Attention

This paper presents the disentangled mask attention (DMA) to build a disentangled transformer as shown in Figure 1 where DMA is implemented to replace the attention module in encoder and decoder of vanilla transformer. This DMA is constructed with three schemes. First, the semantic mask attention is presented to enhance the semantic meaning of attention weights via latent clustering. Second, the disentangled attention heads are calculated by maximizing the independence among attention heads for the queries $\mathbf{q}$ of words $\mathbf{x}$. Third, the information-theoretic disentanglement is implemented to carry out a variational learning procedure of transformer.

3.1 Semantic mask attention

Given a source sequence $\mathbf{x}$ and a target sequence $\mathbf{y}$, a new variational transformer for sequence-to-sequence learning is presented by merging GMM to express the prior distribution of latent variables $\mathbf{z}$. By extending the unsupervised learning from data $\mathbf{x}$ using $p(\mathbf{x})$ based on VAE with GMM for $\mathbf{z}$ in Eq. (3), this paper presents a novel latent variable model for transformer based on the supervised learning by minimizing the sequence-to-sequence (S2S) classifier loss or the negative ELBO of conditional likelihood $p(\mathbf{y}|\mathbf{x})$ of training data $\{\mathbf{x}, \mathbf{y}\}$

$$\mathcal{L}_{\mathrm{s2s}} = -\sum_n \mathbb{E}_{\mathbf{z}^{n,h} \sim q(\mathbf{z}^{n,h}|\mathbf{x})} \left[\log p(\mathbf{y}|\mathbf{z}^{n,h}, \mathbf{x})\right] + \mathrm{KL}(q(\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h}|\mathbf{x}) \| p(\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h})) \tag{4}$$

where $\mathbf{z}^{n,h}$ is the feed-forward network output of the nth transformer block or 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 block 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 (Kingma & Welling, 2014) is applied by using the reparameterization trick 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 as

$$p(\mathbf{c}_z^h = c|\mathbf{z}_i^h) = \frac{\pi_c^h \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_c^h, \text{diag}\{(\boldsymbol{\sigma}_c^h)^2\})}{\sum_{c'} \pi_{c'}^h \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_{c'}^h, \text{diag}\{(\boldsymbol{\sigma}_c^h)^2\})}. \tag{5}$$

Overall, the parameters of transformer layer and GMM $\{\pi_c^h, \boldsymbol{\mu}_c^h, (\boldsymbol{\sigma}_c^h)^2\}$ are trained by minimizing $\mathcal{L}_{\text{s2s}}$ in Eq. (4). For ease of expression, the block or layer index n is ignored hereafter. Through the estimated prior of latent variable c using GMM, the semantic relation between two latent vectors $\mathbf{z}_i^h$ and $\mathbf{z}_j^h$ associated with word tokens $\mathbf{x}_i$ and token $\mathbf{x}_j$ is characterized by constructing the semantic mask $M = \{M_{ij}^h\}$ based on the clustering probability of tokens belonging to the same cluster c

$$M_{ij}^h = \frac{\sum_c p(\mathbf{c}_z^h = c|\mathbf{z}_i^h) p(\mathbf{c}_z^h = c|\mathbf{z}_j^h)}{\sum_j \sum_{c'} p(\mathbf{c}_z^h = c'|\mathbf{z}_i^h) p(\mathbf{c}_z^h = c'|\mathbf{z}_j^h)} \tag{6}$$

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

$$\bar{A}_{ij}^h = \frac{M_{ij}^h A_{ij}^h}{\sum_{j'} M_{ij'}^h A_{ij'}^h} \quad \text{where } A_{ij}^h = \text{Softmax}\left(((\mathbf{k}_{1:J}^h)^T \mathbf{q}_i^h)/\sqrt{d_k}\right)_j \tag{7}$$

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 the parameters $\{W_q^h, \mathbf{b}_q^h\}$ and $\{W_k^h, \mathbf{b}_k^h\}$. A softmax function is calculated over $\mathbf{k}_{1:J}^h$ for the set of J key tokens of head h, where $\mathbf{k}_j^h \in \mathbb{R}^{d_k}$, and then retrieved by jth entry. Empirically, the training of this semantic mask attention can be improved by an interpolation scheme to find the final attention for each word pair (i, j) as $\widehat{A}_{ij}^h = (1-\gamma)A_{ij}^h + \gamma \bar{A}_{ij}^h$ where $\gamma = 1 - \max(1-\alpha, e^{5\times 10^{-4} t})$ is an annealing mixing rate which is controlled a hyperparameter α at each learning iteration t. New weights $\{\widehat{A}_{ij}^h\}$ are used to implement a semantic-aware transformer.

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. Model capacity is restricted accordingly. To alleviate this issue, this paper implements a new variant of attention mechanism, called the disentangled mask attention (DMA), as shown in Figure 2 which focuses on the disentanglement of query vectors $\mathbf{q}^h$ in multi-head attention method. The disentangled queries are used to implement the semantic mask attention. The disentanglement objective is constructed by extending the measure of independence over two latent vectors in Eq. (2) to that over multiple query vectors for n_h heads

$$D(\mathbf{x}; \mathbf{q}^{h=1}, ..., \mathbf{q}^{h=n_h}) = \sum_{h=1}^{n_h} \sum_{h' \neq h}^{n_h} \text{I}(\mathbf{q}^h, \mathbf{q}^{h'}) - \text{I}(\mathbf{x}, \mathbf{q}^h) \triangleq \mathcal{L}_D \tag{8}$$

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 Eq. (8) 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 topics of the words 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.

In the implementation, the exact computation of MI terms $\text{I}(\mathbf{q}^h, \mathbf{q}^{h'})$ and $\text{I}(\mathbf{q}^h, \mathbf{x})$ is intractable. Finding the estimator of MI is required. To pursue independence, the minimization of disentangle-

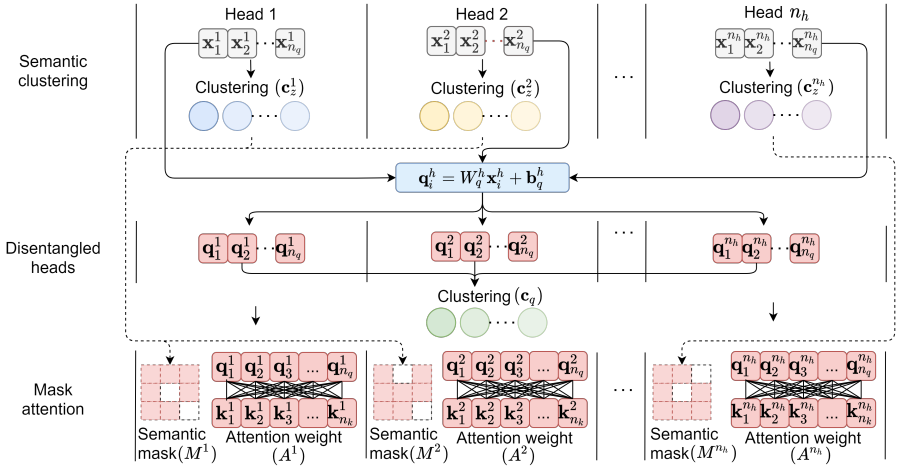

Figure 2: Implementation procedure of the disentangled mask attention, where the procedure details of key $\mathbf{k}$ is omitted in figure, n_q and n_k denotes the number of query and key tokens, respectively. First, the input $\mathbf{x}_i^h$ of token i in head h is clustered by semantic clusters $\mathbf{c}_z^h$. The clustering probability is used for constructing semantic mask M in mask attention indicated by dotted line. After, $\mathbf{x}_i^h$ is transformed to query $\mathbf{q}_i^h$ by parameters $\{W_q^h, \mathbf{b}_q^h\}$, and is clustered by a set of additional clusters $\mathbf{c}_q$ for the disentanglement of $\mathbf{q}_i^h$ in different head. Finally, the semantic mask M is applied on the attention weight A^h calculated by $\mathbf{q}^h$ and $\mathbf{k}^h$ to construct a new attention weight.

ment objective in Eq. (8) is implemented by minimizing the upper bound of the first MI term

$$\mathrm{I}(\mathbf{q}^h, \mathbf{q}^{h'}) \leq \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n_q}\left(\log p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) - \frac{1}{n_q-1}\sum_{j\neq i}^{n_q}\log p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})\right)\right] \triangleq \mathcal{L}_{D_{qq}} \tag{9}$$

and simultaneously maximizing the lower bound of the second MI term (Poole et al., 2019)

$$\mathrm{I}(\mathbf{x}, \mathbf{q}^h) \geq \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n_q}\left(\log \frac{\exp(f(\mathbf{q}_i^h, \mathbf{x}_i))}{\frac{1}{n_q}\sum_{j=1}^{n_q}\exp(f(\mathbf{q}_i^h, \mathbf{x}_j))}\right)\right] \triangleq \mathcal{L}_{D_{xq}} \tag{10}$$

where n_q is the number of samples in $\mathbf{q}^h$ or $\mathbf{q}^{h'}$. In Eq. (9), 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 $p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})$ can be estimated by a trainable neural network. In Eq. (10), 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 $\mathbf{x}_j$. In this study, the critic function is defined by $f(\mathbf{q}_i^h, \mathbf{x}_j) \triangleq \sum_c p(\mathbf{c}_q = c|\mathbf{q}_i^h)p(\mathbf{c}_q = c|\mathbf{q}_j^h)$ which 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 block output $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ as shown in Eq. (5).

3.3 Learning Criteria

This study presents the semantic mask attention based on the disentangled attention heads. A disentangled mask attention (DMA) is proposed for sequence-to-sequence learning via transformer. There are four latent variables in such a 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 the same layer n. The latent

variable model is trained by minimizing the negative ELBO or sequence-to-sequence classifier loss

$$\begin{aligned}\mathcal{L}_{\text{s2s}} = -\sum_{n} \mathbb{E}_{(\mathbf{z}^{n,h},\mathbf{q}^{n,h})\sim q(\mathbf{z}^{n,h},\mathbf{q}^{n,h}|\mathbf{x})}\left[\log p(\mathbf{y}|\mathbf{z}^{n,h},\mathbf{q}^{n,h},\mathbf{x})\right] \\ + \text{KL}(q(\mathbf{z}^{n,h},\mathbf{c}_z^{n,h}|\mathbf{x})\|p(\mathbf{z}^{n,h},\mathbf{c}_z^{n,h})) + \text{KL}(q(\mathbf{q}^{n,h},\mathbf{c}_q^{n}|\mathbf{x})\|p(\mathbf{q}^{n,h},\mathbf{c}_q^{n}))\end{aligned} \tag{11}$$

which is extended from Eq. (4). 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 $p(\mathbf{q}_i^h)$. Importantly, 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 using the estimators of MI or the upper bound $\mathcal{L}_{D_{qq}}$ and lower bound $\mathcal{L}_{D_{xq}}$ of MI. Disentanglement objective is 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

$$p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) = \sum_c p(\mathbf{q}_i^h|\mathbf{c}_q = c)p(\mathbf{c}_q = c|\mathbf{q}_i^{h'}). \tag{12}$$

In Eq. (12), $p(\mathbf{c}_q = c|\mathbf{q}_i^{h'})$ is computed by referring Eq. (5) and the first term is computed via GMM

$$p(\mathbf{q}_i^h|\mathbf{c}_q = c) = \frac{p(\mathbf{c}_q = c|\mathbf{q}_i^h)p(\mathbf{q}_i^h)}{\sum_{i'} p(\mathbf{c}_q = c|\mathbf{q}_{i'}^h)p(\mathbf{q}_{i'}^h)}. \tag{13}$$

In addition, the loss functions $\mathcal{L}_{D_{qq}}$ and $\mathcal{L}_{D_{xq}}$ in Eqs. (9) and (10) are calculated over all queries over n_h heads including n_q samples in each head. In this study, the diversity of semantic clustering is further enhanced and regularized towards increasing the probabilistic correlation within each variable $\mathbf{z}_i^h$ via $\sum_c p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ and simultaneously decreasing the correlation between variables $\mathbf{z}_i^h$ and $\mathbf{z}_j^h$ via $\sum_c p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)p(\mathbf{c}_z^h = c|\mathbf{z}_j^h)$, accordingly different samples $\mathbf{z}_i^h$ and $\mathbf{z}_j^h$ likely go to different clusters c. An objective to enhance the cluster diversity is calculated by summing up all entries (i, j) of the squared values of a matrix (Lin et al., 2017) shown below

$$\mathcal{L}_{\text{dv}} = \frac{1}{n_q^2}\sum_i\sum_j (CC^T - I)_{i,j}^2 \tag{14}$$

where I is a $n_q \times n_q$ identity matrix, $C = [C_{ic}]$ and $C_{ic} \triangleq p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ for GMM of $\mathbf{z}_i^h$, and $C_{ic} \triangleq p(\mathbf{c}_q = c|\mathbf{q}_i^h)$ for GMM of $\mathbf{q}_i^h$. The overall loss $\mathcal{L}$ for the transformer with disentangled mask attention is composed of classification loss, disentanglement loss and diversity loss as $\mathcal{L} = \mathcal{L}_{\text{s2s}} + \mathcal{L}_D + \mathcal{L}_{\text{dv}}$ where the regularization parameters are adopted in three 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 to evaluate the performance of the proposed model. All models in the following experiments were trained and evaluated by using Fairseq (Ott et al., 2019) toolkit.

4.1 DATASET DESCRIPTIONS

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 roughly 212K of Chinese and English sentence pairs in WMT'17 Zh-En. 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. In addition, Jieba[1] was used as the tokenization tool for Chinese sentences.

[1]https://github.com/fxsjy/jieba

4.2 EVALUATION METRICS

The performance of translation models were evaluated in terms of BLEU (Papineni et al., 2002) score. The BLEU score of models in IWSLT'14 De-En and WMT'17 Zh-En was calculated by using the evaluation script[2] provided by Fairseq. The BELU score of models in WMT'14 En-De was evaluated after applying compound splitting[3], similar to the setting in (Wu et al., 2019).

In addition, two metrics were introduced to measure the redundancy of attention weights based on the Jensen-Shannon (JS) distance (Correia et al., 2019; Bian et al., 2021). One is the layer redundancy (LR), LR $\triangleq \frac{1}{N}(\sum_{n=1}^{N}(\log_2 n_h - \frac{1}{n_q}\sum_i \text{JS}(\widehat{A}_i^{n,h=1}, \ldots, \widehat{A}_i^{n,h=n_h})))$, and the other is the head redundancy (HR), HR $\triangleq \frac{1}{N^2 {n_h}^2}\sum_{n_1,n_2}^{N}\sum_{h_1,h_2}^{n_h}\frac{1}{n_q}\sum_i(1-\text{JS}(\widehat{A}_i^{n_1,h_1}, \widehat{A}_i^{n_2,h_2}))$ where N is the number of transformer layers, and A_i denotes the ith row of attention matrix A. 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.

4.3 MODEL CONFIGURATIONS

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 $N = 6$ layers of encoder and decoder. To reduce the computational cost, the term $p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})$ in $\mathcal{L}_{D_{qq}}$ was neglected, and $q(\mathbf{z}_i^h|\mathbf{x})$ and $q(\mathbf{q}_i^h|\mathbf{x})$ were estimated by a single Gaussian with a constant variance $\sigma^2 = 0.1$ and $\mathbf{z}_i^h$ and $\mathbf{q}_i^h$ as means, i.e. $q(\mathbf{z}_i^h|\mathbf{x}) = \mathcal{N}(\mathbf{z}_i^h, \text{diag}\{\sigma^2\})$ and $q(\mathbf{q}_i^h|\mathbf{x}) = \mathcal{N}(\mathbf{q}_i^h, \text{diag}\{\sigma^2\})$. In addition, the semantic mask was disregarded in the transformer block of masked DMA in test phase. DMA transformer worked well under this setting. In order to rapid the convergence of GMMs, the gradient of cluster centroid was multiplied by a constant c_{grad} during training. To evaluate the proposed model under different model sizes, DMA transformer was built with three configuration types, which were *base*, *small*, and *tiny*. The only difference between these configurations was the size of attention and feed-forward layers. 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 $\mathcal{L}_{s2s}$. The details of settings are listed in Appendix A.2.

4.4 EXPERIMENTAL RESULTS

Comparison on model performance: The experiment results on three tasks are shown and compared. Generally speaking, the proposed DMA transformer achieved higher BLEU score than baseline transformer 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 illustrate that DMA transformer performs better than other models for translation with grammatically similar and different languages.

Model	Params	BPE	n_c	LR	HR	BLEU
Transformer (*base*, (Mehta et al., 2021))	42.0M (1.06x)	10K	-	-	-	34.30
Transformer (*base*, ours)	39.5M (1.00x)	6K/8K	-	0.74	0.65	34.50
DMA transformer (*base*)	39.7M (1.00x)	6K/8K	8	0.62	0.53	**35.31**
DeLighT (Mehta et al., 2021)	14.0M (0.35x)	10K	-	-	-	33.80
DMA transformer (*small*)	26.1M (0.66x)	6K/8K	4	0.64	0.57	35.00
DMA transformer (*tiny*)	11.9M (0.30x)	6K/8K	4	0.63	0.58	34.96

Table 1: Experimental results on IWSLT'14 De-En translation task. 'BPE' denotes the number of tokens in the dictionary of encoder and decoder (encoder/decoder). 'Ours' denotes the model is trained by ourselves. 'Params' denotes the number of parameters in model.

[2]https://github.com/pytorch/fairseq/blob/main/fairseq_cli/generate.py

[3]https://github.com/pytorch/fairseq/blob/main/scripts/compound_split_bleu.sh

Model	Params	BPE	LR	HR	BLEU
Transformer (*base*, (Mehta et al., 2021))	67.0M (1.01x)	44K	-	-	27.70
Transformer (*base*, ours)	66.5M (1.00x)	44K	0.79	0.69	27.75
DMA transformer (*base*)	66.6M (1.00x)	44K	0.73	0.58	**28.35**
DeLighT (Mehta et al., 2021)	54.0M (0.81x)	44K	-	-	28.00
DMA transformer (*small*)	46.4M (0.70x)	44K	0.70	0.57	28.16
Transformer (*big*, (Vaswani et al., 2017))	213.0M (3.20x)	37K	-	-	28.40

Table 2: Experimental results on WMT'14 En-De translation task. DMA with $n_c = 4$ is used.

Model	Params	BPE	n_c	LR	HR	BLEU
Transformer (ours)	55.0M (1.00x)	25K/20K	-	0.71	0.60	12.76
DMA transformer	55.1M (1.00x)	25K/20K	4	0.62	0.55	**13.13**

Table 3: Experimental results on WMT'17 Zh-En translation task with very different languages.

Analysis on model size: The number of additional parameters introduced by 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 (Mehta et al., 2021), with smallest model size, DMA transformer (*tiny*) obtained about 1.16 BLEU score higher with using 5% less of parameters.

In WMT14'En-De (Table 2), DMA transformer (*small*) achieved 0.41 BLEU score higher than transformer with using 30% less of parameters. On the other hand, 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. In conclusion, introducing the objectives of variational 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.

Analysis on attention redundancy: Figure 3 depicts the head redundancy between individual attention heads in encoder. The similarity of DMA transformer attention weights between each head is generally smaller than that of transformer. Diversity is improved. In addition, LR and HR metrics of DMA transformer in three translation tasks 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.

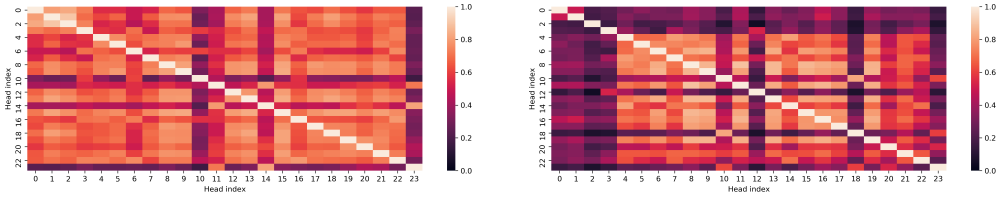

Figure 3: Head redundancy in encoder of transformer (left) and DMA transformer (right), where IWSLT'14 De-En test set is used for evaluation. The head index in x-axis and y-axis are accumulated from first head in first layer to last head in last layer.

Analysis on semantic clustering: Figure 4 depicts top 15 tokens captured by one of clusters in DMA transformer, where the punctuations and common stopwords are removed. This cluster captures the semantic meaning of topic words, which can be observed from the followings tokens, 'es', 'ing', 'ed', 'ies'. In addition, Figure 5 depicts the clustering probability distribution and the corresponding semantic mask of another cluster in DMA transformer. With the help of applying $\mathcal{L}_{dv}$, each token was clustered into different clusters without collapsing into same cluster. In the figure of corresponding semantic mask, the mask was composed of several rectangles. This indicates this set of semantic clusters were learned to find out and enrich the local features. In addition, we also

found there were some semantic masks formed to identifying the semantic relation between token in longer distance.

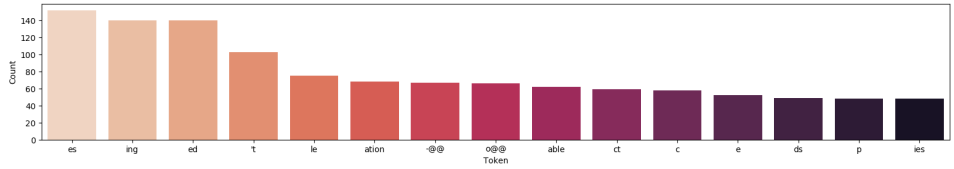

Figure 4: Top 15 tokens captured by the second cluster in the first head of the last decoder layer by using DMA transformer, where the text in x-axis shows the BPE tokens, and IWSLT'14 De-En test set is used for evaluation.

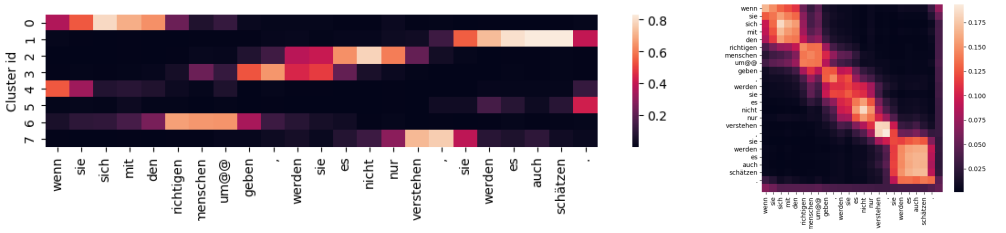

Figure 5: Clustering probability distribution (left) and the corresponding semantic mask (right) in the 1st head of 6th DMA transformer encoder layer, where the BPE tokens in two axes of (right) are same as the tokens in x-axis of (left), and IWSLT'14 De-En test set is used for evaluation.

Analysis on latent representations: Figure 6 depicts the distribution of query on latent space by using t-SNE for dimension reduction. The gaps between queries of different heads in DMA transformer are larger than transformer, and the distribution is more diverse within the same head. The head redundancy is reduced because the proposed methods increase the semantic diversity of query vectors in same and different heads, and this also improves the performance of model.

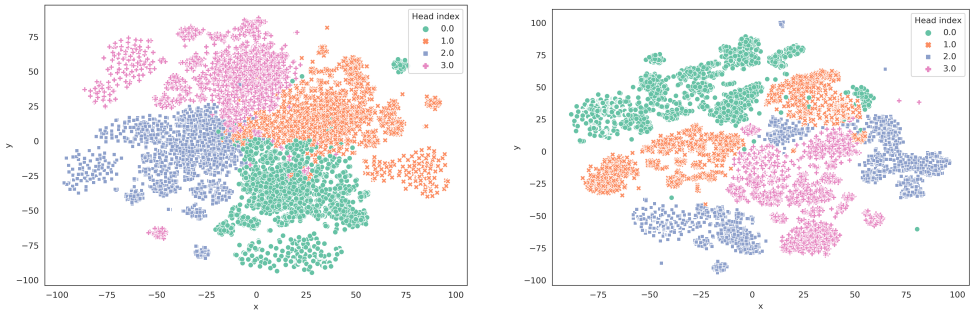

Figure 6: Query distribution of 104 sentences (13,312 tokens) in the last decoder layer of transformer (left) and DMA transformer (right), where the point with different colors indicates the query of different heads, and IWSLT'14 De-En test set is used for evaluation.

5 Conclusions

This paper presented a variant of transformer, DMA transformer, which replaced the vanilla attention with the disentangled mask attention. 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 that the redundancy of attention weight was reduced, and the semantic diversity of query within same head and between different heads was increased. By applying the proposed methods, the compact DMA transformer outperformed other transformers in different translation tasks either with identical or smaller model size.

REFERENCES

Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational attention for sequence-to-sequence models. In *Proc. of International Conference on Computational Linguistics*, pp. 1672–1682, 2018.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Church. On attention redundancy: A comprehensive study. In *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, pp. 930–945, 2021.

Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Proc. of International Conference on Neural Information Processing Systems*, pp. 2615–2625, 2018.

Pengyu Cheng, Martin Renqiang Min, Dinghan Shen, Christopher Malon, Yizhe Zhang, Yitong Li, and Lawrence Carin. Improving disentangled text representation learning with information-theoretic guidance. In *Proc. of Annual Meeting of Association for Computational Linguistics*, pp. 7530–7541, 2020.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, 2019.

Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Proc. of International Conference on Neural Information Processing Systems*, pp. 4417–4426, 2017.

Zhihao Fan, Yeyun Gong, Dayiheng Liu, Zhongyu Wei, Siyuan Wang, Jian Jiao, Nan Duan, Ruofei Zhang, and Xuanjing Huang. Mask attention networks: Rethinking and strengthen transformer. In *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, pp. 1692–1701, 2021.

Maosheng Guo, Yu Zhang, and Ting Liu. Gaussian transformer: A lightweight approach for natural language inference. *Proc. of AAAI Conference on Artificial Intelligence*, 33:6489–6496, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proc. of International Conference on Learning Representations*, 2017.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, 2019.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *Proc. of International Joint Conference on Artificial Intelligence*, pp. 1965–1972, 2017.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 424–434, 2019.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement. In *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 6649–6653, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proc. of International Conference on Learning Representations*, 2014.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *Proc. of International Conference on Learning Representations*, 2020.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *Proc. of International Conference on Learning Representations*, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. of International Conference on Machine Learning*, pp. 4114–4124, 2019.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. DeLighT: Deep and light-weight transformer. In *Proc. of International Conference on Learning Representations*, 2021.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Demonstrations*, pp. 48–53, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proc. of Annual Meeting of Association for Computational Linguistics*, pp. 1504–1519, 2019.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proc. of the International Conference on Machine Learning*, pp. 5171–5180, 2019.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, pp. 53–68, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *Proc. of International Conference on Learning Representations*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *Proc. of International Conference on Learning Representations*, 2019.

Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, and Junzhou Huang. Adversarial sparse transformer for time series forecasting. In *Proc. of International Conference on Neural Information Processing Systems*, pp. 17105–17115, 2020.

Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *Proc. of International Conference on Machine Learning*, pp. 5670–5679, 2018.

Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, and Lawrence Carin. Improving zero-shot voice style transfer via disentangled representation learning. In *Proc. of International Conference on Learning Representations*, 2021.

A EXPERIMENTAL DETAILS

A.1 STATISTICS OF DATASET

The number of sentence pairs after preprocessing in WMT'14 En-De, IWSLT'14 De-En and WMT'17 Zh-En are listed in Table 4.

Task	Train	Validation	Test
IWSLT'14 De-En	160K	7K	7K
WMT'14 En-De	3.9M	39K	3K
WMT'17 Zh-En	208K	2K	2K

Table 4: Number of sentence pairs in different translation tasks.

A.2 CONFIGURATION OF MODELS

The details of model configuration in three translation tasks are listed in Table 5, where $c_{\text{kl}} = 0.01$ in all configurations. Before evaluation, the weights of model were averaged over last 5 epochs. During evaluation, beam search was applied on the predicted output with beam size 5 for all tasks.

Task	Configuration	Embed/FFN	n_h	c_{qq}	c_{xq}	c_{grad}	Iterations
WMT'14 En-De	*base*	512/2048	8	100	10	10	196K
	small	384/2048	8	100	10	10	313K
WMT'17 Zh-En	-	512/1024	4	50	5	5	79K
IWSLT'14 De-En	*base*	512/1024	4	100	10	10	55K
	small	384/1024	4	100	10	10	55K
	tiny	256/512	4	50	5	10	99K

Table 5: Model configurations of DMA transformer in different tasks. Embed and FFN denotes the dimension of hidden layer in attention and feed-forward layers, respectively. Iterations denotes the number of parameters updates.

A.3 ABLATION ON DICTIONARY SIZE

The ablation study was established by varying the size of dictionary in WMT'14 En-De. The model with larger dictionary tends to have better performance. From results in Table 6, DMA transformer with smaller dictionary obtained 0.19 BLEU score higher than transformer with larger dictionary, while using 5% less of parameters. This is due to the semantic information of DMA transformer is enhanced by semantic clustering, so that similar performance as transformer can be achieved by DMA transformer with using smaller number of parameters.

Model	Params	BPE	n_c	LR	HR	BLEU
Transformer (*base*, ours)	63.08M (1.00x)	37K	-	0.79	0.70	27.44
DMA transformer (*base*)	63.16M (1.00x)	37K	4	0.73	0.61	27.94
Transformer (*base*, ours)	66.52M (1.05x)	44K	-	0.79	0.69	27.75
DMA transformer (*base*)	66.61M (1.06x)	44K	4	0.73	0.58	**28.35**

Table 6: Ablation study of dictionary size on WMT'14 En-De translation task.

A.4 ABLATION ON CLUSTER NUMBERS

The ablation study was established by varying the number of clusters in IWSLT'14 De-En. The BLEU score of DMA transformer is slightly impacted by different number of clusters within the

range of 0.1. From the trend in Table 7, suitable number of clusters is required to achieve best performance. Fewer or more number of clusters may make the clusters fail to capture the semantic meaning or some clusters become redundant.

Model	Params	BPE	n_c	LR	HR	BLEU
DMA transformer (*base*)	39.56M (1.00x)	6K/8K	4	0.63	0.52	35.22
DMA transformer (*base*)	39.65M (1.00x)	6K/8K	8	0.62	0.53	**35.31**
DMA transformer (*base*)	39.74M (1.01x)	6K/8K	12	0.62	0.53	35.29

Table 7: Ablation of cluster numbers n_c on IWSLT'14 De-En translations task.

B Experimental Analysis

B.1 Additional analysis on latent representations

Figure 7 and Figure 8 depict the query distribution of transformer or DMA transformer in WMT'14 En-De and WMT'17 Zh-En translation tasks, where t-SNE was used for dimension reduction. Similar to query distribution in IWSLT'14 De-En (Figure 6), the query vectors in DMA transformer within same head or between different heads were more diverse than transformer. And the gaps between different heads were larger. In addition, by using either 4 heads in WMT'17 Zh-En or 8 heads in WMT'14 En-De, the semantic diversity of query were all increased by proposed methods.

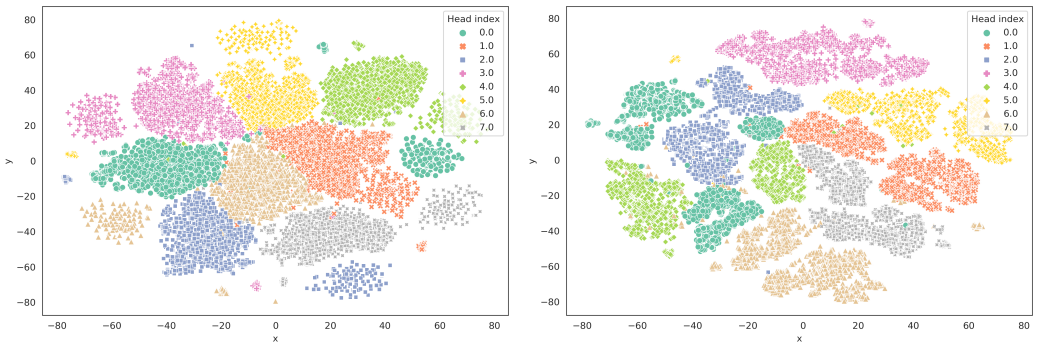

Figure 7: Query distribution of 88 sentences (30,272 tokens) in last decoder layer of transformer (left) and DMA transformer (right), where WMT'14 En-De test set is used for evaluation.

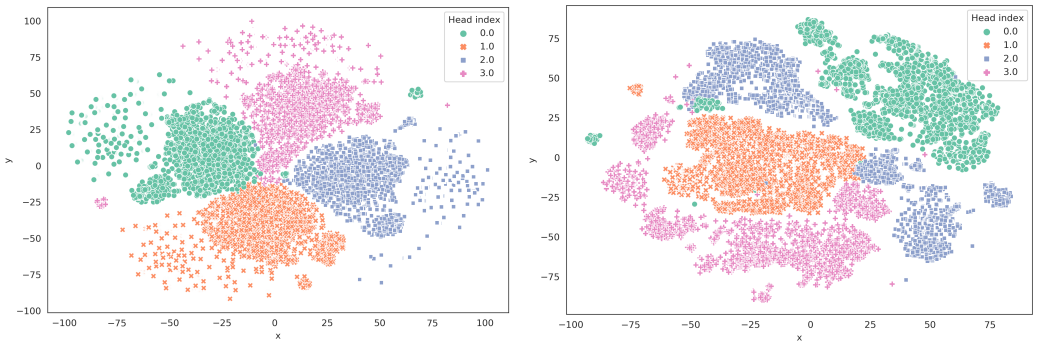

Figure 8: Query distribution of 72 sentences (15,552 tokens) in last decoder layer of transformer (left) and DMA transformer (right), where WMT'17 Zh-En test set is used for evaluation.

B.2 ADDITIONAL ANALYSIS ON ATTENTION REDUNDANCY

Figure 9 and Figure 10 depict the head redundancy of transformer or DMA transformer encoder in WMT'14 De-En and WMT'17 Zh-En translation tasks. Similar to the results in IWSLT'14 De-En (Figure 3), head redundancy of DMA transformer with either 4 or 8 attention heads were all smaller than transformer. These results illustrates the proposed methods successfully reduce the attention redundancy between each head and work for any number of attention heads.

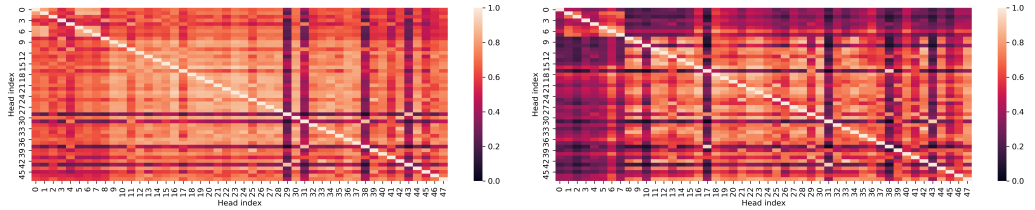

Figure 9: Head redundancy in encoder of transformer (left) and DMA transformer (right), where WMT'14 En-De test set was used for evaluation.

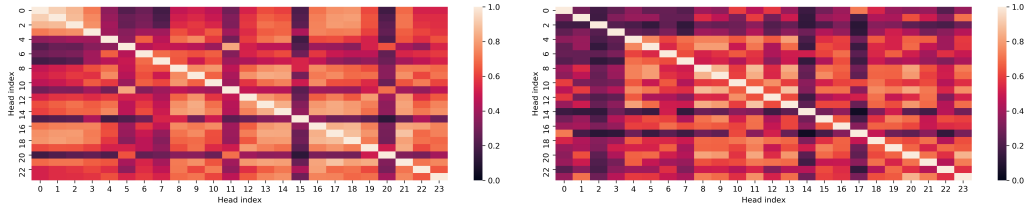

Figure 10: Head redundancy in encoder of transformer (left) and DMA transformer (right), where WMT'17 En-De test set is used for evaluation.

B.3 ADDITIONAL ANALYSIS ON SEMANTIC CLUSTERING

Figure 11 depicts the top 15 tokens captured by one cluster of DMA transformer in WMT'14 En-De. The semantic meaning of topic words can be observed from following tokens, 'year', 'time', 'Thursday', and 'times'.

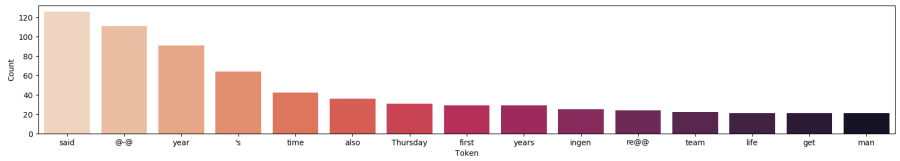

Figure 11: Top 15 tokens captured by the second cluster in the first head of the 5th encoder layer by using DMA transformer, where WMT'14 En-De test set is used for evaluation.

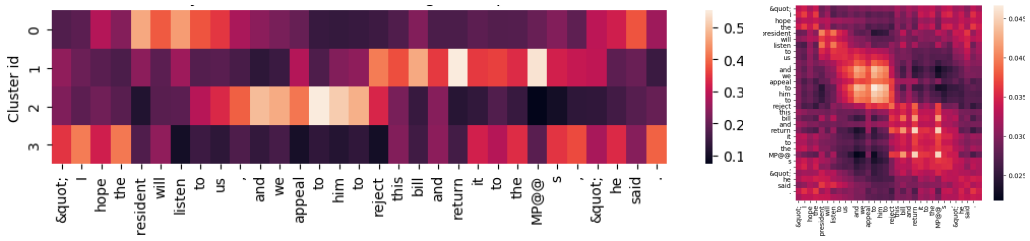

Figure 12: Clustering probability distribution (left) and the semantic mask (right) in the 1st head of 2nd DMA transformer encoder layer, where WMT'14 En-De test set is used for evaluation.

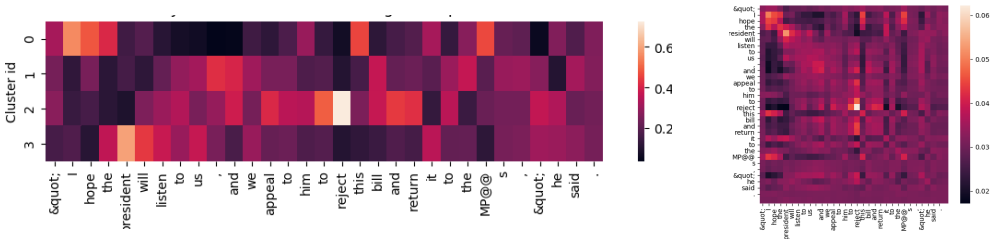

Figure 13: Clustering probability distribution (left) and the semantic mask (right) in the 1st head of 6th DMA transformer encoder layer, where WMT'14 En-De test set is used for evaluation.

Figure 12 and Figure 13 depict the probability distribution of semantic clustering and corresponding semantic mask of different DMA transformer encoder layers in WMT'14 En-De. The semantic mask in Figure 12 captures the local dependencies, while the semantic mask in Figure 13 captures the semantic dependency in longer distance.