Latent Semantic and Disentangled Attention

Jen-Tzung Chien and Yu-Han Huang

***Abstract*—Sequential learning using transformer has achieved state-of-the-art performance in natural language tasks and many others. The key to this success is the multi-head self attention which encodes and gathers the features from individual tokens of an input sequence. The mapping or decoding is performed to produce an output sequence via cross attention. There are threefold weaknesses by using such an attention framework. First, since the attention would mix up the features of different tokens in input and output sequences, it is likely that redundant information exists in sequence data representation. Second, the patterns of attention weights among different heads tend to be similar. The model capacity is bounded. Third, the robustness in an encoder-decoder network against the model uncertainty is disregarded. To handle these weaknesses, this paper presents a Bayesian semantic and disentangled mask attention to learn latent disentanglement in multi-head attention where the redundant features in transformer are compensated with the latent topic information. The attention weights are filtered by a mask which is optimized through semantic clustering. This attention mechanism is implemented according to Bayesian learning for clustered disentanglement. The experiments on machine translation and speech recognition show the merit of Bayesian clustered disentanglement for mask attention.**

***Index Terms*—Sequential learning, Bayesian learning, disentangled representation, mask attention, transformer.**

I. Introduction

SEQUENTIAL learning has been widely developed to build a variety of information systems in presence of audio signals, video streams, natural sentences or medical sequence data. It is crucial to enhance the learning representation by incorporating self attention within a sequence and cross attention between source and target sequences. Accordingly, the encoder-decoder attention network based on transformer [1] was proposed and has achieved state-of-the-art results in a wide range of sequence-to-sequence learning tasks such as speech recognition, machine translation, question answering, dialogue generation, to name a few. In spite of the success by using attention mechanism, the issues of computational complexity and representation redundancy still considerably constrain the efficiency in learning process [2]. In order to cope with theses issues, several extensions of transformer using mask attention have been proposed. These extensions consist of reformer [3], routing transformer [4], and sparse transformer [5] where the dot-product in attention was calculated by only considering a small portion of word tokens. The computational complexity was reduced by applying binary mask on attention weights. In contrast to binary mask, the real-valued mask was constructed in the adaptively sparse transformer [6] and the adversarial sparse transformer [7] where the α-entmax function [8] was presented to remove redundant features in calculation of attention weights. In [9], the distance between two tokens of a sequence was used to implement a dynamic attention mask in a mask attention network. In addition, the relation between a pair of tokens was measured to construct an attention mask, which was adopted to carry out the Gaussian-weighted self attention [10] in transformer [11]. On the other hand, there were a number of works pointing out that some of attention heads were redundant [6] and the semantic interpretation of attention weights was missing. In [12], the similarity of attention patterns between individual heads was high in standard transformer. Therefore, similar performance was obtained even though some attention heads were pruned. In [13], [14], the attention weights or mask coefficients were estimated through adversarial training. The resulting transformer obtained comparable outputs while the attention weights could be very different. This circumstance reveals that the estimated attention weights did not really convey sufficient semantic information.

To overcome these shortcomings in attention-based network [15], this paper presents a new sequential learning to enhance the semantic meaning as well as reduce the redundancy of attention weights within each individual head and across different heads. There are two stages in construction of the disentangled transformer with attention weights under a latent variable representation. Bayesian learning is implemented to enhance the robustness to random perturbation in model construction. At the first stage, the disentanglement of attention weights in individual heads is conducted. The semantics of these heads are characterized by latent clusters or topics in accordance with Bayesian sequential learning [16], [17] where the mixture of Gaussians as the prior density of topics is merged. Bayesian clustering is proposed to build a semantic mask which is operated over attention weights in latent space for those semantically-close tokens. The real-valued clusters of attention mask are learned to strengthen the attention mechanism based on the variational inference procedure. In addition, the second stage is devoted to reduce the redundancy of attention model and disentangle the multi-head attention across various heads. The disentanglement objective based on the mutual information of query vectors between two heads is constructed and minimized to enhance the compactness of attention patterns in attention-based

Manuscript received 1 October 2022; revised 20 April 2024; accepted 18 July 2024. Date of publication 23 July 2024; date of current version 5 November 2024. This work was supported by National Science and Technology Council, Taiwan under Grant NSTC 112-2634-F-A49-006. Recommended for acceptance by L. Sigal. *(Corresponding author: Jen-Tzung Chien.)*

The authors are with the Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: jtchien@nycu.edu.tw; yhhuang.ee08@nycu.edu.tw).

Digital Object Identifier 10.1109/TPAMI.2024.3432631

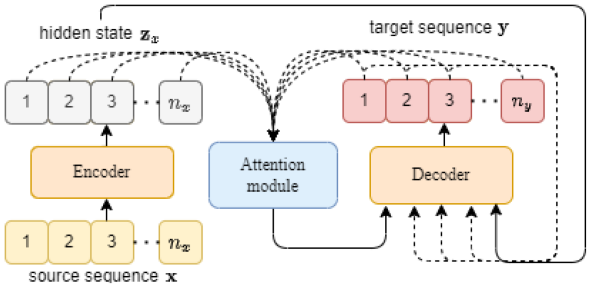

Fig. 1. Sequence-to-sequence learning with attention.

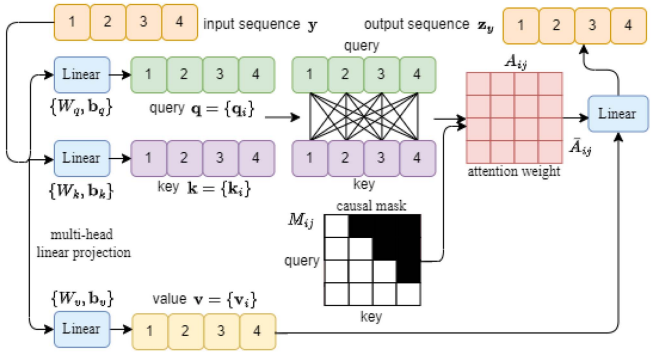

Fig. 2. Masked multi-head self attention for a target sequence $\mathbf{y}$.

sequential learning. There are threefold novelties presented in this study. First, the topic-based mask attention is developed to build a hierarchical latent representation which strengthens the attention weights for semantically related word tokens. Second, a Bayesian learning approach is proposed to carry out uncertainty modeling [18] for latent topic disentanglement over mask attention in sequence data representation. Third, a Bayesian sequence-to-sequence learning is implemented for a probabilistic transformer in accordance with an information-theoretic objective for disentanglement. The experiments on sequential learning in machine translation and speech recognition under various settings illustrate the merit of the proposed Bayesian semantic mask in an attention-based encoder-decoder network.

II. Unsupervised Sequential Learning

First of all, attention-based sequential learning and latent variable disentangled modeling are surveyed.

A. Mask Attention for Sequence Data

Sequence-to-sequence (S2S) model is learned for domain mapping from source sequence $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n_x}$ to target sequence $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^{n_y}$ with different lengths. Fig. 1 depicts a S2S learning model with attention consisting of an encoder to extract hidden features or states $\mathbf{z}_x = \{\mathbf{z}_{xi}\}_{i=1}^{n_x}$ from $\mathbf{x}$, a decoder to generate target samples $\mathbf{y}$ through a mapping, and an attention module to conduct self attention and cross attention over $\mathbf{x}$ and $\mathbf{y}$. Here, encoder and decoder calculates a kind of source and target embeddings, respectively. Attention scheme is performed to deal with the challenges of capturing and storing long-distance temporal information in long sequence modeling and mapping. Accordingly, transformer [1] is developed to fulfill an encoder-decoder framework for S2S learning where the attention module consists of three components including multi-head self attention in encoder, masked multi-head self attention in decoder, multi-head cross attention between encoder and decoder. The multi-head attention in three components shares the same calculation form. In particular, Fig. 2 shows the calculation of masked multi-head self attention by using an input or embedding sequence $\mathbf{y}$. In calculation of multi-head self attention, the same embedding $\mathbf{y}_i$ of a word token at time i is transformed to find query $\mathbf{q}_i^h = W_q^h \mathbf{y}_i + \mathbf{b}_q^h$, key $\mathbf{k}_i^h = W_k^h \mathbf{y}_i + \mathbf{b}_k^h$ and value $\mathbf{v}_i^h = W_v^h \mathbf{y}_i + \mathbf{b}_v^h$, respectively, where h is the head index The attention parameters consist of $\theta_a = \{W_q^h, \mathbf{b}_q^h, W_k^h, \mathbf{b}_k^h, W_v^h, \mathbf{b}_v^h\}$. A softmax function is calculated to find attention weight A_{ij}^h for a query $\mathbf{q}_i^h$ to attend a key $\mathbf{k}_j^h \in \mathbb{R}^{d_k}$ via

$$A_{ij}^h = \text{Softmax}\left(((\mathbf{k}_{1:J}^h)^\top \mathbf{q}_i^h)/\sqrt{d_k}\right)_j \tag{1}$$

from the set of J key tokens $\mathbf{k}_{1:J}^h$. Notably, a causal mask is employed to find the masked attention weight

$$\bar{A}_{ij}^h = \frac{M_{ij}^h A_{ij}^h}{\sum_{j'} M_{ij'}^h A_{ij'}^h} \tag{2}$$

where $M_{ij} = 1$ for $i \geq j$ and $M_{ij} = 0$ for $i < j$. Multi-head query, key and value are augmented as $\mathbf{q}_i = [\mathbf{q}_i^1; \ldots; \mathbf{q}_i^{n_h}]$, $\mathbf{k}_i = [\mathbf{k}_i^1; \ldots; \mathbf{k}_i^{n_h}]$ and $\mathbf{v}_i = [\mathbf{v}_i^1; \ldots; \mathbf{v}_i^{n_h}]$, respectively. The attended feature of an output sequence $\mathbf{z}_y = \{\mathbf{z}_{yi}\}$ is calculated at each time i by a linear transform $\mathbf{z}_{yi} = \bar{A}^\top \mathbf{v}_i$ where $\bar{A} = [\bar{A}^1; \ldots; \bar{A}^{n_h}]$ and $\bar{A}^h = \{\bar{A}_{ij}^h\}$. This study presents a new latent variable model for unsupervised learning of disentangled features and semantic clusters. In what follows, some basics of information-theoretic learning and Bayesian learning are introduced.

B. Latent Disentangled Representation

Disentangled representation is seen as a horizon of researches [19], [20], [21] which aim to factorize the latent representation into several independent low-dimensional features by optimizing an objective function for decomposition. These researches have been widely employed in various technical data such as images [22], sentences [23], [24] and voices [25]. This work infers the mutually independent latent variables from information-theoretic perspective by following the variation of information (VI). VI is defined as the distance between two clusterings of variables $\mathbf{z}_i$ and $\mathbf{z}_j$, which is closely related to the mutual information (MI) by

$$\text{VI}(\mathbf{z}_i, \mathbf{z}_j) = \mathbb{H}(\mathbf{z}_i) + \mathbb{H}(\mathbf{z}_j) - 2\mathbb{I}(\mathbf{z}_i, \mathbf{z}_j) \tag{3}$$

where $\mathbb{H}$ and $\mathbb{I}$ denotes the entropy and MI, respectively, and $\text{VI}(\mathbf{z}_i, \mathbf{z}_j) \geq 0$. The independent components can be optimized by minimizing MI or comparably maximizing VI. Importantly, VI is a distance metric and obeys the triangle inequality $\text{VI}(\mathbf{x}, \mathbf{z}_i) + \text{VI}(\mathbf{x}, \mathbf{z}_j) \geq \text{VI}(\mathbf{z}_i, \mathbf{z}_j)$ for any observed variable

$\mathbf{x}$ and its corresponding latent components $\mathbf{z}_i$ and $\mathbf{z}_j$. The equality happens when random vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ are parallel or statistically independent. The *disentanglement* or independence between latent variables $\mathbf{z}_i$ and $\mathbf{z}_j$ from the observed variable $\mathbf{x}$ is accordingly measured by the difference $\mathbb{D}$ between two sides of triangular inequality which can be manipulated and written by

$$\mathbb{D}(\mathbf{x};\mathbf{z}_i,\mathbf{z}_j) = 2(\mathbb{I}(\mathbf{z}_i,\mathbf{z}_j) - \mathbb{I}(\mathbf{x},\mathbf{z}_i) - \mathbb{I}(\mathbf{x},\mathbf{z}_j) + \mathbb{H}(\mathbf{x})) \tag{4}$$

where $\mathbb{D}(\mathbf{x};\mathbf{z}_i,\mathbf{z}_j) \geq 0$. To learn the disentangled latent representation, the term $\mathbb{H}(\mathbf{x})$ is a constant and the three MI terms in (4) are included to form a disentanglement objective. But, the direct calculation of MI does not exit. It is practical to implement information-theoretic disentanglement through estimating the upper and lower bounds of MI [26], [27]. The disentanglement or independence is therefore optimized by minimizing the difference $\mathbb{D}(\mathbf{x};\mathbf{z}_i,\mathbf{z}_j)$, or comparably minimizing the upper bound of MI $\mathbb{I}(\mathbf{z}_i,\mathbf{z}_j)$ and at the same time maximizing the lower bounds of both MIs $\mathbb{I}(\mathbf{x},\mathbf{z}_i)$ and $\mathbb{I}(\mathbf{x},\mathbf{z}_j)$. In this study, these bounds are empirically estimated and optimized to extract latent features in transformer.

C. Latent Semantic Clustering

Based on the latent disentanglement, this paper develops the semantic mask attention where the variational clustering using neural network is proposed to capture the latent semantic meaning in the process of mask attention. In [28], the latent embedding in neural network was assumed to be distributed by a Gaussian mixture model (GMM) which was feasible to implement the variational deep embedding. The GMM distribution of a latent variable $\mathbf{z} \in \mathbb{R}^d$ is formed by

$$p(\mathbf{z}) = \sum_{c=1}^{n_c} p(c)p(\mathbf{z}|c) = \sum_{c=1}^{n_c} \pi_c \mathcal{N}\left(\mathbf{z};\boldsymbol{\mu}_c,\text{diag}\{\boldsymbol{\sigma}_c^2\}\right) \tag{5}$$

where c denotes a latent cluster or semantic topic [29] corresponding to a latent sample $\mathbf{z}$ from an observed sample $\mathbf{x}$. GMM parameters consist of the mixture weights $\boldsymbol{\pi} = \{\pi_c\} \in \mathbb{R}^{n_c}$, the mean vectors $\boldsymbol{\mu} = \{\boldsymbol{\mu}_c\} \in \mathbb{R}^{n_c \times d}$ and the variance vectors of diagonal matrices $\boldsymbol{\sigma}^2 = \{\boldsymbol{\sigma}_c^2\} \in \mathbb{R}^{n_c \times d}$. A variational autoencoder (VAE) [30] is feasible to implement a variational clustering by maximizing the marginal likelihood over latent variables $\mathbf{z}$ and c where $\mathbf{z}$ is encoded from training data $\mathbf{x}$ and distributed by a GMM, and $\mathbf{x}$ is then decoded for reconstruction. The learning loss $\mathcal{L}_{\text{ELBO}}$ for semantic clustering is yielded by the negative evidence lower bound (ELBO) which is derived by

$$\begin{aligned} -\log p(\mathbf{x}) &= -\log \int \sum_c p(\mathbf{x},\mathbf{z},c)d\mathbf{z} \leq -\mathbb{E}_{\mathbf{z},c}[\log p(\mathbf{x}|\mathbf{z},c)] \\ &+ \mathcal{D}_{\text{KL}}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|c)) + \mathcal{D}_{\text{KL}}(q(c|\mathbf{x})\|p(c)) \triangleq \mathcal{L}_{\text{ELBO}}. \end{aligned} \tag{6}$$

In right-hand-side of (6), the first term is seen as a reconstruction error for training data $\mathbf{x}$ under a latent variable model and the other two terms reflect the Kullback-Leibler (KL) divergence $\mathcal{D}_{\text{KL}}$ due to two latent variables $\mathbf{z}$ and c characterized by a variational distribution $q(\mathbf{z},\mathbf{c}|\mathbf{x})$ which is decomposed into $q(\mathbf{z}|\mathbf{x})$ and $q(c|\mathbf{x})$ getting close to their priors $p(\mathbf{z}|c)$ and $p(c)$, respectively, by minimizing two KL terms. A Bayesian clustering is realized as a new variant of VAE in presence of GMM as the prior. Variational distribution $q(\mathbf{z},\mathbf{c}|\mathbf{x})$ is adopted for Bayesian learning. In this study, latent semantic clustering is implemented for deep embedding and clustering where similar samples with the same *latent semantic topics* are represented to carry out the semantic mask attention in transformer.

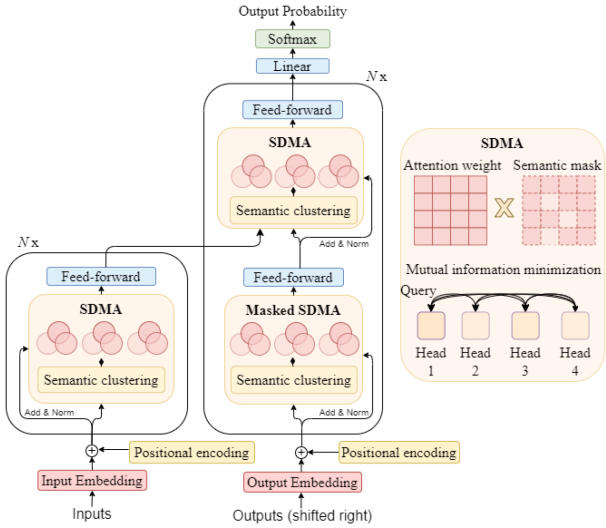

Fig. 3. Semantic and disentangled mask attention in transformer.

III. Semantic and Disentangled Attention

Fig. 3 shows the implementation of semantic and disentangled mask attention (SDMA) in an encoder-decoder framework, which is extended from a vanilla transformer, consisting of add & norm, multi-head attention and feed-forward network. The attention modules based on self attention and cross attention in the encoder and decoder of vanilla transformer are now replaced by SDMA. Different from self attended SDMA in encoder and cross attended SDMA in decoder, the masked SDMA runs the self attention to prevent attending to subsequent positions in the output prediction. There are two components which are combined with according to learning perspectives. First, the semantic clusters of latent features are inferred to carry out attention mechanism using the semantic mask. Variational Bayesian approach to a latent variable model is presented. Second, the disentanglement over attention heads is implemented to reduce the redundancy in learning representation using transformer. The information-theoretic learning is performed by maximizing the independence among n_h heads in the queries $\mathbf{q}$ of word sequences $\mathbf{x}$. A generic solution to semantic and disentangled mask attention is proposed.

A. Variational Sequence-to-Sequence Model

This study develops a latent variable model to carry out a Bayesian transformer for sequence-to-sequence classification where the conditional likelihood of target sequence $\mathbf{y}$ given

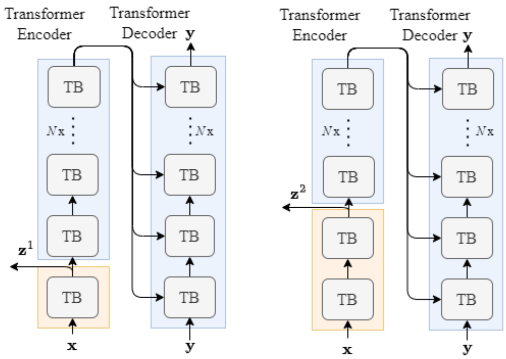

Fig. 4. Illustration of variational sequence-to-sequence model. $\mathbf{z}^n$ is the output of the nth transformer block (TB). The left and right transformers show the scenario when the first one and the first two TBs (shown by orange) are used to estimate the latent feature $\mathbf{z}$ of input $\mathbf{x}$, and the other TBs (shown by blue) are used to map a source sequence $\mathbf{x}$ to a target sequence $\mathbf{y}$ by given the latent variables $\mathbf{z}^1$ and $\mathbf{z}^2$, respectively.

by source sequence $\mathbf{x}$ is maximized. A hierarchical Bayesian model is configured by merging a GMM of clusters or topics $\mathbf{c}$ to characterize the prior density of latent variable $\mathbf{z}$. Latent semantic clustering in an unsupervised model in Section II-C is now extended to a supervised classification model which is feasible to machine translation and speech recognition. The S2S classification loss $\mathcal{L}_{\text{S2S}}$ is yielded as a negative ELBO consisting of cross entropy (CE) classification loss $\mathcal{L}_{\text{CE}}$ and KL loss $\mathcal{L}_{\text{KL}}^z$ as given by [31]

$$-\sum_{n,h}\Big\{\mathbb{E}_{(\mathbf{z}^{n,h},\mathbf{c}^{n,h})\sim q(\mathbf{z}^{n,h},\mathbf{c}^{n,h}|\mathbf{x})}\left[\log p(\mathbf{y}|\mathbf{z}^{n,h},\mathbf{c}^{n,h},\mathbf{x})\right] + \mathcal{D}_{\text{KL}}(q(\mathbf{z}^{n,h},\mathbf{c}^{n,h}|\mathbf{x})\|p(\mathbf{z}^{n,h},\mathbf{c}^{n,h}))\Big\} \triangleq \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}}^z. \tag{7}$$

Using transformer, $\mathbf{z}^{n,h}$ denotes the latent features from the outputs of feed-forward network of transformer block or layer $1 \le n \le N$ after calculating the multi-head self or cross attention using source sequence $\mathbf{x}$ where h denotes the head index. Also, $\mathbf{c}^{n,h}$ denotes the one-hot cluster vectors of block n and head h reflecting the latent semantic topics from word tokens in $\mathbf{x}$. Fig. 4 displays how this variational S2S model is performed according to the loss $\mathcal{L}_{\text{S2S}}$ by aggregating the negative ELBO from different transformer block or layer n of latent variable $\mathbf{z}^n$ and semantic topic $\mathbf{c}^n$.

A sophisticated attention is then proposed to implement Bayesian transformer by following (7). For example, the computational complexity for self attention is proportional to the square of sequence length n_x or n_y where the memory and computation requirements are extensive. In [32], the clustered attention applied the local sensitive hashing to calculate the attention via maximum inner product search so as to approximate the softmax attention. In [33], a linear transformer handled this issue by changing the order of matrix multiplication in attention mechanism. More attractively, mask attention is a line of researches which aim to reduce the redundancy or computational complexity of attention. A number of solutions were exploited to implement various types of sparse attention for transformer [34], [35], [36]. This paper proposes the *semantic mask attention* where the semantic relation among the embeddings of word tokens is characterized in a semantic mask which is used to adjust the corresponding attention weights.

B. Latent Semantic Mask Attention

A hierarchical latent variable model is constructed with two levels of variables $\mathbf{z}^{n,h}$ and $\mathbf{c}^{n,h}$. The prior density $p(\mathbf{z}^{n,h})$ is modeled via a GMM by referring to (5) with parameters θ_z. Generally, the inferred Gaussian component c naturally captures the distribution of latent semantic topics from the embedding sequence $\mathbf{z}$ or equivalently the word sequence $\mathbf{x}$. Latent semantic clustering is developed to carry out mask attention to enhance the attention-based representation using transformer. According to the stochastic gradient variational Bayes estimator [30], a Gaussian sampling process is run to draw latent sample $\mathbf{z}^{n,h}$ from the variational distribution $q(\mathbf{z}^{n,h}|\mathbf{x})$ by applying a reparameterization trick

$$\mathbf{z}^{n,h} = \text{Transformer}_{n,h}(\mathbf{x}) + \sigma^2\epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0,1). \tag{8}$$

$\text{Transformer}_{n,h}$ denotes the output of layer n and head h, and σ^2 is a sampling parameter. Then, the corresponding cluster variable $\mathbf{c}_z^h = c$ of an embedding $\mathbf{z}_i^h$ of head h of a word token $\mathbf{x}_i$ is determined by the clustering probability

$$p(\mathbf{c}_z^h = c|\mathbf{z}_i^h) = \frac{\pi_c^z \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_c^z, \text{diag}\{(\boldsymbol{\sigma}_c^z)^2\})}{\sum_{c'} \pi_{c'}^z \mathcal{N}(\mathbf{z}_i^h|\boldsymbol{\mu}_{c'}^z, \text{diag}\{(\boldsymbol{\sigma}_{c'}^z)^2\})}. \tag{9}$$

For ease of expression, the layer index n is ignored hereafter. In addition to the multi-head attention parameters in transformer layers $\theta_a = \{W_q^h, \mathbf{b}_q^h, W_k^h, \mathbf{b}_k^h, W_v^h, \mathbf{b}_v^h\}$, we incorporate the GMM parameters $\theta_z = \{\pi_c^z, \boldsymbol{\mu}_c^z, (\boldsymbol{\sigma}_c^z)^2\}$ which are jointly trained with θ_a by minimizing $\mathcal{L}_{\text{s2s}}$ in (7). This study estimates $\boldsymbol{\Theta} = \{\theta_a, \theta_z\}$ for S2S classification based on the latent semantic mask attention (SMA).

Given the estimated prior of semantic topic c using GMM, the semantic correlation between two latent features $\mathbf{z}_i^h$ and $\mathbf{z}_j^h$ corresponding to the word tokens $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed by constructing the semantic mask $M = \{M_{ij}^h\}$ based on aggregation of the clustering probabilities $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ and $p(\mathbf{c}_z^h = c|\mathbf{z}_j^h)$ of the tokens under the same cluster c via

$$M_{ij}^h = \frac{\sum_c p(\mathbf{c}_z^h = c|\mathbf{z}_i^h) p(\mathbf{c}_z^h = c|\mathbf{z}_j^h)}{\sum_j \sum_{c'} p(\mathbf{c}_z^h = c'|\mathbf{z}_i^h) p(\mathbf{c}_z^h = c'|\mathbf{z}_j^h)}. \tag{10}$$

Notably, this calculation reveals a probabilistic relation between $\mathbf{x}_i$ and $\mathbf{x}_j$ measured through different semantic clusters c. This semantic mask is real-valued by $0 \le M_{ij}^h \le 1$. This is different from the causal mask with the value of M_{ij}^h being 0 or 1 for self attention in decoder as mentioned in Section II-A. Typically, the mask calculated in (10) 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}^h. Here, original attention A_{ij}^h and masked attention $\bar{A}_{ij}^h$ have been shown in (1) and (2), respectively. In the implementation, the estimation of SMA model can be further improved by an interpolation scheme to find the smoothed attention for each word pair (i, j) as

$$\widehat{A}_{ij}^h = (1-\gamma)A_{ij}^h + \gamma \bar{A}_{ij}^h \tag{11}$$

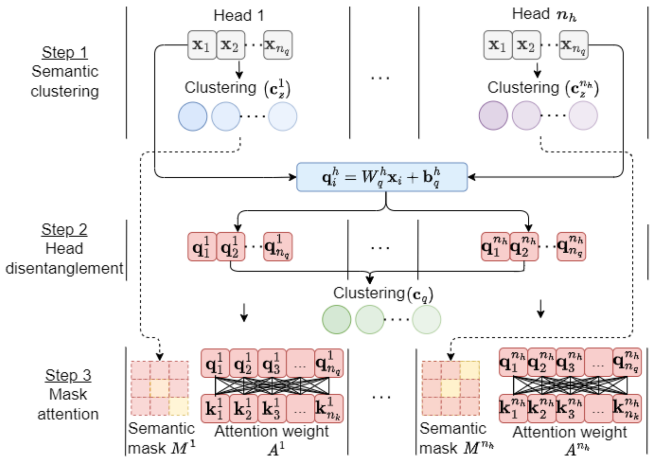

Fig. 5. Three steps for semantic and disentangled mask attention (SDMA). In the step of *semantic clustering*, the input token $\mathbf{x}_i$ at time i in head h is clustered by semantic clusters $\mathbf{c}_z^h$. The clustering probability is used to construct the semantic mask M in mask attention indicated by dotted line. In the step of *head disentanglement*, $\mathbf{x}_i$ 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. In the step of *mask attention*, 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 $\bar{A}^h$. The calculation of key $\mathbf{k}^h$ is omitted in the figure. n_h, n_q and n_k denotes the number of heads, queries and keys, respectively. SDMA is reduced to SMA without head disentanglement.

where an annealing mixing rate at each learning iteration t is selected as $\gamma = 1 - \max(1-\nu, e^{(5\times 10^{-4})t})$ with a hyperparameter of upper bound ν. New weights $\widehat{A} = \{\widehat{A}_{ij}^h\}$ are used to implement a semantic-aware transformer with the attended context vector $\mathbf{z}_i = \widehat{A}^\top \mathbf{v}_i$ in multi-head attention.

C. Latent Disentangled Mask Attention

Although the semantic mask $M = \{M_{ij}^h\}$ in (10) is calculated to reduce the redundancy in attention weights $A = \{A_{ij}^h\}$ driven by latent semantic topics c, the compactness of mask attention can be further strengthened by enhancing the disentanglement of multi-head representation. To optimize the mask attention for model capacity, this paper presents the semantic and disentangled mask attention (SDMA) where the disentangled representation based on $\mathbb{D}$ in (4) is learned. Fig. 5 depicts the overview of a new variant of attention mechanism based on SDMA, which is implemented by three steps including semantic clustering, head disentanglement and mask attention. In particular, the semantic mask attention is implemented by using the disentangled query vectors $\{\mathbf{q}^h\}_{h=1}^{n_h}$ in multi-head attention method. Following the disentanglement over two latent variables $\{\mathbf{z}_i, \mathbf{z}_j\}$ in (4), the disentanglement loss $\mathcal{L}_D \triangleq \mathbb{D}(\cdot)$ over multiple latent queries from n_h attention heads in SDMA is then extended and constructed by

$$\mathbb{D}(\mathbf{x}; \mathbf{q}^{h=1}, \ldots, \mathbf{q}^{h=n_h}) = \sum_{h=1}^{n_h} \sum_{h' \neq h}^{n_h} \mathbb{I}(\mathbf{q}^h, \mathbf{q}^{h'}) - \mathbb{I}(\mathbf{x}, \mathbf{q}^h) \tag{12}$$

where $\mathbf{q}^h = \{\mathbf{q}_i^h\}$ denotes n_q or n_x queries of input word tokens $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{n_q}$ in head h. Latent disentanglement is accordingly fulfilled by minimizing $\mathcal{L}_D$ which comparably promotes the *semantic disentanglement* for the queries $\{\mathbf{q}_i^h, \mathbf{q}_j^{h'}\}$ belonging to different heads $h' \neq h$. The semantic and disentangled mask attention is therefore developed via the semantic mask attention driven by latent semantic topics c of word tokens via GMM where the attention weights are calculated by using the disentangled queries which are learned by minimizing the disentanglement loss

$$\mathcal{L}_D = \mathcal{L}_{D_{qq}} - \mathcal{L}_{D_{xq}} \tag{13}$$

or equivalently minimizing the self MI loss $\mathcal{L}_{D_{qq}}$ and maximizing the cross MI loss $\mathcal{L}_{D_{xq}}$. The disentangled queries $\mathbf{q}^h = \{\mathbf{q}_i^h\}$ are used to enhance the compactness of the attention weights calculated across different heads.

However, the exact calculation of MI terms $\mathbb{I}(\mathbf{q}^h, \mathbf{q}^{h'})$ and $\mathbb{I}(\mathbf{q}^h, \mathbf{x})$ in (12) does not exist. Finding the estimator of these two terms is required for latent disentangled learning. In the implementation, the first MI term is minimized or equivalently the upper bound of this MI is minimized. This bound $\mathcal{L}_{D_{qq}}$ aims to pursue the disentanglement of queries across different heads h and h', i.e., $h' \neq h$, is given by [37]

$$\mathbb{I}(\mathbf{q}^h, \mathbf{q}^{h'}) \leq \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n_q}\left(\log p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) - \frac{1}{n_q - 1}\sum_{j\neq i}^{n_q}\log p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})\right)\right]. \tag{14}$$

Simultaneously, the second MI term is maximized, or equivalently the lower bound of this MI is maximized. The bound $\mathcal{L}_{D_{xq}}$ is designed to be maximized to assure the highest expressiveness of query sequence $\mathbf{q}^h$ or observation sequence $\mathbf{x}$ as yielded by [37]

$$\mathbb{I}(\mathbf{x}, \mathbf{q}^h) \geq \mathbb{E}\left[\frac{1}{n_q}\sum_{i=1}^{n_q}\left(\log \frac{\exp(f(\mathbf{q}_i^h, \mathbf{x}_i))}{\frac{1}{n_q}\sum_{j=1}^{n_q}\exp(f(\mathbf{q}_i^h, \mathbf{x}_j))}\right)\right] \tag{15}$$

where n_q is the number of samples in $\mathbf{q}^h$ or $\mathbf{q}^{h'}$.

To implement (14), a variational leave-one-out upper bound of MI between $\mathbf{q}^h$ and $\mathbf{q}^{h'}$ is calculated by using the conditional probabilities $p(\mathbf{q}_i^h|\mathbf{q}_i^{h'})$ and $p(\mathbf{q}_i^h|\mathbf{q}_j^{h'})$ within the same token $\mathbf{x}_i$ and across two tokens $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, as

$$p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) = \sum_c p(\mathbf{q}_i^h|\mathbf{c}_q = c)p(\mathbf{c}_q = c|\mathbf{q}_i^{h'}). \tag{16}$$

Here, the clustering probability of the query cluster $p(\mathbf{c}_q = c|\mathbf{q}_i^h)$ is computed similar to that of transformer block output $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ as shown in (9). The likelihood function of the query given a cluster is calculated by

$$p(\mathbf{q}_i^h|\mathbf{c}_q = c) = \frac{p(\mathbf{c}_q = c|\mathbf{q}_i^h)p(\mathbf{q}_i^h)}{\sum_{i'} p(\mathbf{c}_q = c|\mathbf{q}_{i'}^h)p(\mathbf{q}_{i'}^h)} \tag{17}$$

where a new GMM prior for query $\mathbf{q}_i^h$ of a token $\mathbf{x}_i$ given by query clusters $\mathbf{c}_q = c$ is introduced in a form of

$$p(\mathbf{q}_i^h) = \sum_c \pi_c^q \mathcal{N}(\boldsymbol{\mu}_c^q, \text{diag}\{(\boldsymbol{\sigma}_c^q)^2\}) \tag{18}$$

with parameters $\theta_q = \{\pi_c^q, \boldsymbol{\mu}_c^q, (\boldsymbol{\sigma}_c^q)^2\}$. In implementation of (15), an ELBO is measured by using a critic function $f(\mathbf{q}_i^h, \mathbf{x}_j)$ which reflects the semantic correlation between $\mathbf{q}_i^h$ and $\mathbf{x}_j$ by aggregating the correlation across query clusters or topics c by using the clustering probabilities

$$f(\mathbf{q}_i^h, \mathbf{x}_j) \triangleq \sum_c p(\mathbf{c}_q = c|\mathbf{q}_i^h) p(\mathbf{c}_q = c|\mathbf{q}_j^h). \tag{19}$$

The probabilistic correlation between $\mathbf{q}_i^h$ and $\mathbf{x}_j$ under the same cluster or query topic $\mathbf{c}_q = c$ is measured. This study considers two levels of semantic clustering. In addition to the first level of clustering over latent features $\mathbf{z}_i^h$, the second level of clustering over queries $\mathbf{q}_i^h$ is constructed to carry out the disentangled mask attention. Starting from the prior probability and clustering probability, the first bound is computed via $p(\mathbf{q}_i^h) \to p(\mathbf{c}_q = c|\mathbf{q}_i^h) \to p(\mathbf{q}_i^h|\mathbf{c}_q = c) \to p(\mathbf{q}_i^h|\mathbf{q}_i^{h'}) \to \mathcal{L}_{D_{qq}}$ while the second bound is computed via $p(\mathbf{q}_i^h) \to p(\mathbf{c}_q = c|\mathbf{q}_i^h) \to f(\mathbf{q}_i^h, \mathbf{x}_j) \to \mathcal{L}_{D_{xq}}$. The parameters of the proposed SDMA $\boldsymbol{\Theta} = \{\theta_a, \theta_z, \theta_q\}$ are arranged and estimated through a hierarchical Bayesian learning procedure.

D. Learning Objectives With Regularization

This study presents a S2S classifier based on transformer where the semantic and disentangled mask attention is implemented. A hierarchical latent variable model is constructed in presence of four latent variables consisting of transformer features and feature clusters $\{\mathbf{z}^{n,h}, \mathbf{c}_z^{n,h}\}$, head queries and query clusters $\{\mathbf{q}^{n,h}, \mathbf{c}_q^n\}$. Latent semantic clusters $\mathbf{c}_z^{n,h}$ and $\mathbf{c}_q^n$ are formed due to the word tokens in sequence $\mathbf{x} = \{\mathbf{x}_i\}$ or $\mathbf{y} = \{\mathbf{y}_i\}$ in the feature level and query level, respectively. This hierarchical model is learned via variational inference by minimizing the negative ELBO or the S2S classification loss $\mathcal{L}_{\text{S2S}}$ as shown in (7) where an additional set of latent variables $\{\mathbf{q}^{n,h}, \mathbf{c}_q^n\}$ is incorporated and an additional KL term $\mathcal{D}_{\text{KL}}(q(\mathbf{q}^{n,h}, \mathbf{c}_q^n|\mathbf{x}) \| p(\mathbf{q}^{n,h}, \mathbf{c}_q^n))$ is merged. SMA is upgraded to SDMA where the queries $\mathbf{q}^h = \{\mathbf{q}_i^h\}$ are disentangled across different attention heads h under the same layer n. Using SDMA, a KL term is imposed to regularize the variational distribution $q(\mathbf{q}_i^h|\mathbf{x})$ to get close to a shared GMM prior $p(\mathbf{q}_i^h)$. Thus, KL terms $\{\mathcal{L}_{\text{KL}}^z, \mathcal{L}_{\text{KL}}^q\}$ for latent variables $\{\mathbf{z}, \mathbf{q}\}$ and their clusters $\{\mathbf{c}_z, \mathbf{c}_q\}$ are formed. KL loss for semantic mask is given by

$$\mathcal{L}_{\text{KL}}^z = \frac{1}{N n_h n_q} \sum_{n,h,i} \Big\{ \mathcal{D}_{\text{KL}}(q(\mathbf{z}_i^{n,h}|\mathbf{x}) \| p(\mathbf{z}_i^{n,h}|\mathbf{c}_z^{n,h})) + \mathcal{D}_{\text{KL}}(q(\mathbf{c}_z^{n,h}|\mathbf{x}) \| p(\mathbf{c}_z^{n,h})) \Big\} \tag{20}$$

while that for disentangled head $\mathcal{L}_{\text{KL}}^q$ is similarly obtained.

Bayesian learning is fulfilled to build a semantic mask attention while the information-theoretic learning is performed to enhance the attention weight A_{ij} by enforcing the disentanglement over groups of queries $\mathbf{q}^h$ across various heads h. In implementation of the disentangled transformer, the estimates of MI bounds in $\mathcal{L}_D$ (12) are obtained by

$$\mathcal{L}_{D_{qq}} \approx -\frac{1}{n_q n_h^2} \sum_i \sum_h \sum_{h' \neq h} \log(1 - p(\mathbf{q}_i^h|\mathbf{q}_i^{h'})) \tag{21}$$

Algorithm 1: Inference Procedure for SDMA-T.

input : inputs $\mathbf{x} = \{\mathbf{x}_i\}$, n_h, d_k, scaling factor s
output : outputs $\mathbf{z} = \{\mathbf{z}_i^h\}$
loss : losses $\mathcal{L}_{\text{KL}}^z$, $\mathcal{L}_{\text{KL}}^q$, $\mathcal{L}_{D_{qq}}$, $\mathcal{L}_{D_{xq}}$, $\mathcal{L}_{\text{DV}}^z$, $\mathcal{L}_{\text{DV}}^q$
head separation
  $\{\mathbf{x}_i^h | \forall h \in [1, n_h]\} \leftarrow \frac{1}{s}\mathbf{x}_i$
semantic clustering of z or x
  accumulate losses $\mathcal{L}_{\text{KL}}^z, \mathcal{L}_{\text{DV}}^z$
  calculate semantic mask $M = \{M_{ij}^h\}$
  $\mathbf{x}_i^h \leftarrow s(\mathbf{x}_i^h + \sigma^2 \epsilon)$ where $\epsilon \sim \mathcal{N}(0, 1)$
linear projection
  $\mathbf{q}_i = \mathbf{x}_i W_q + \mathbf{b}_q$, $\mathbf{k}_i = \mathbf{x}_i W_k + \mathbf{b}_k$, $\mathbf{v}_i = \mathbf{x}_i W_v + \mathbf{b}_v$
head separation
  $\{\mathbf{q}_i^h | \forall h \in [1, n_h]\} \leftarrow \frac{1}{s}\mathbf{q}_i$, $\{\mathbf{k}_i^h | \forall h \in [1, n_h]\} \leftarrow \mathbf{k}_i$
  $\{\mathbf{v}_i^h | \forall h \in [1, n_h]\} \leftarrow \mathbf{v}_i$
semantic clustering of q
  accumulate losses $\mathcal{L}_{D_{qq}}, \mathcal{L}_{D_{xq}}, \mathcal{L}_{\text{KL}}^q, \mathcal{L}_{\text{DV}}^q$
  $\mathbf{q}_i^h \leftarrow s(\mathbf{q}_i^h + \sigma^2 \epsilon)$ where $\epsilon \sim \mathcal{N}(0, 1)$
multi-head attention
  $A_{i,j}^h = \text{Softmax}((\mathbf{k}_{1:J}^h)^\top \mathbf{q}_i^h / \sqrt{d_k})_j$
  $\bar{A}_{i,j}^h = \frac{(M \odot A^h)_{i,j}}{\sum_{j'} (M \odot A^h)_{i,j'}}$
  $\widehat{A}^h = (1 - \gamma) A^h + \gamma \bar{A}^h$, $\mathbf{z}_i^h = \sum_j \widehat{A}_{i,j}^h \mathbf{v}_i^h$
return z

$$\mathcal{L}_{D_{xq}} = \frac{1}{n_q n_h} \sum_i \sum_h \log \frac{\exp(f(\mathbf{q}_i^h, \mathbf{x}_i))}{\sum_j \exp(f(\mathbf{q}_i^h, \mathbf{x}_j))}. \tag{22}$$

Notably, the estimate of $\mathcal{L}_{D_{qq}}$ in (21) is approximated for a rapid calculation by only minimizing the first term of (14) and applying the property $\mathbb{I}(\mathbf{q}^h, \mathbf{q}^{h'}) \geq 0$ where mutual information cannot be negative.

In this study, the diversity of semantic clustering is further enhanced and regularized towards increasing the probabilistic correlation within each variable $\mathbf{z}_i^{n,h}$ via maximizing $\sum_c p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h}) p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h})$ and simultaneously decreasing the correlation between variables $\mathbf{z}_i^{n,h}$ and $\mathbf{z}_j^{n,h}$ via minimizing $\sum_c p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h}) p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_j^{n,h})$, accordingly different samples $\mathbf{z}_i^{n,h}$ and $\mathbf{z}_j^{n,h}$ likely go to different clusters c. An objective to enhance the cluster diversity (DV) is calculated by summing up all entries (i, j) of the squared values of a masked matrix [38]

$$\mathcal{L}_{\text{DV}} = \frac{1}{N n_h n_q^2} \sum_{n,h} \sum_{i=1}^{n_q} \sum_{j=1}^{n_q} \Big[\widehat{M} \odot (CC^\top - I_{n_q}) \Big]_{i,j}^2 \tag{23}$$

where I_{n_q} is an identity matrix, $\widehat{M} = \alpha I_{n_q} + \beta(\mathbf{1} - I_{n_q})$ is a mask with hyperparameters $\{\alpha, \beta\}$, $C = [C_{ic}]$, $C_{ic}^z \triangleq p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h})$ for GMM of $\mathbf{z}_i^{n,h}$, and $C_{ic}^q \triangleq p(\mathbf{c}_q^n = c|\mathbf{q}_i^h)$ for GMM of $\mathbf{q}_i^h$. $\mathbf{1}$ is a matrix of 1 in all entries. $\odot$ is the element-wise product. The mask $\widehat{M}$ is used to control the ratio between the diagonal terms for *negative* within-correlation of $\mathbf{z}_i^{n,h}$ via α and the non-diagonal terms for *positive* between-correlation of $\mathbf{z}_i^{n,h}$ and $\mathbf{z}_j^{n,h}$ via β. Diagonal term in $CC^\top - I_{n_q}$ is always negative since

$$0 < [CC^\top]_{i,i} = \sum_c p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h}) p(\mathbf{c}_z^{n,h} = c|\mathbf{z}_i^{n,h}) < 1.$$

The measures of cluster diversity in two levels $\{\mathcal{L}_{\text{DV}}^z, \mathcal{L}_{\text{DV}}^q\}$ are enhanced. The total loss $\mathcal{L}$ for the transformer based on SDMA is formed by

$$\mathcal{L}_{\text{CE}}+\lambda_k(\mathcal{L}_{\text{KL}}^z+\mathcal{L}_{\text{KL}}^q)+\lambda_q\mathcal{L}_{D_{qq}}-\lambda_x\mathcal{L}_{D_{xq}}+\lambda_d(\mathcal{L}_{\text{DV}}^z+\mathcal{L}_{\text{DV}}^q) \tag{24}$$

where the regularization parameters $\{\lambda_k, \lambda_q, \lambda_x, \lambda_d\}$ are adopted in various objective terms. The parameters Θ of transformation matrices of query, key and value $\theta_a = \{W_q^h, \mathbf{b}_q^h, W_k^h, \mathbf{b}_k^h, W_v^h, \mathbf{b}_v^h\}$ in encoder and decoder for self attention and cross attention, and GMMs $\{\theta_z, \theta_q\} = \{\pi_c^z, \boldsymbol{\mu}_c^z, (\boldsymbol{\sigma}_c^z)^2, \pi_c^q, \boldsymbol{\mu}_c^q, (\boldsymbol{\sigma}_c^q)^2\}$ for semantic clustering and disentanglement are estimated by finding the gradients of $\mathcal{L}$ over individual parameters in $\Theta = \{\theta_a, \theta_z, \theta_q\}$. Algorithm 1 shows the inference procedure for SDMA transformer (SDMA-T) where the losses $\mathcal{L}_{\text{KL}}^z$, $\mathcal{L}_{\text{KL}}^q$, $\mathcal{L}_{D_{qq}}$, $\mathcal{L}_{D_{xq}}$, $\mathcal{L}_{\text{DV}}^z$, $\mathcal{L}_{\text{DV}}^q$ are accumulated in a process from input embeddings $\mathbf{x}$ to output features $\mathbf{z}$. The classification loss $\mathcal{L}_{\text{CE}}$ is calculated from the transformer outputs given by the targets $\mathbf{y}$.

IV. Experiments

This paper conducted the experiments on machine translation tasks by using IWSLT'14 De-En, WMT'14 En-De and WMT'17 Zh-En as well as Chinese speech recognition task by using Aishell-1 [39]. The models were preprocessed and trained by using Fairseq [40] and ESPnet [41] toolkits based on a PC with a GPU via GeForce RTX 3090 Ti 24 GB, a CPU via Intel Core i9-10900K, and a memory with 128 G RAM.

A. Experimental Settings

1) Dataset Descriptions: German-to-English (De-En), English-to-German (En-De) and Chinese-to-English (Zh-En) translation tasks were evaluated. IWSLT 2014 De-En task contained 160K, 7K and 7K of German and English sentence pairs as training, validation and test data, respectively. WMT 2014 En-De task contained 4M, 6K and 7K of English and German sentence pairs as training, validation and test data, respectively. For the WMT 2017 Zh-En task, the results were reported by only using 227K Chinese and English sentence pairs in 'news-commentary-v12' set where training, validation and test data contained 208.1K, 2.1K and 2.0K sentence pairs. respectively. The byte-pair-encoding (BPE) dictionary of the models in WMT'14 En-De were shared for English and German in encoder and decoder. For the other two tasks, encoder and decoder had individual BPE dictionaries for different languages. Jieba https://github.com/fxsjy/jieba was used as the tokenization tool for Chinese sentences. In addition, Aishell-1 [39], [42] contained Chinese speech over 400 speakers with various accents in 11 domains including finance, sports, news, etc. which were used to evaluate the results on S2S learning for speech recognition. The training, validation and test sets consisted of speech utterances with lengths of 150, 20, 10 hours, respectively.

2) Evaluation Metrics: The proposed method was evaluated for machine translation tasks IWSLT'14 De-En and WMT'17 Zh-En where BLEU score was calculated by using the script provided by Fairseq [40]. The BLEU score of the models using WMT'14 En-De task was evaluated after applying the compound splitting similar to the setting in [43]. In addition to BLEU score, there were two metrics introduced to measure the redundancy of attention weights $\widehat{A} = \{\widehat{A}_{i,j}^{n,h}\}$ based on the Jensen-Shannon (JS) distance [6], [12]

$$\text{LR} \triangleq \frac{1}{N}\sum_{n=1}^{N}\left(\log_2 n_h - \frac{1}{n_q}\sum_i \text{JS}(\widehat{A}_{i,:}^{n,h=1}, \ldots, \widehat{A}_{i,:}^{n,h=n_h})\right)$$

$$\text{HR} \triangleq \frac{1}{N^2 n_h^2}\sum_{n_1,n_2}^{N}\sum_{h_1,h_2}^{n_h}\frac{1}{n_q}\sum_i\left(1 - \text{JS}(\widehat{A}_{i,:}^{n_1,h_1}, \widehat{A}_{i,:}^{n_2,h_2})\right)$$

which are the layer redundancy (LR) and head redundancy (HR), respectively, where $\widehat{A}_{i,:}^{n,h}$ denotes the ith row of attention matrix $\widehat{A}^{n,h}$. Basically, LR measures the similarity of attention weights among n_h attention heads under the same layer n while HR measures the similarity of attention weights between individual pairs of attention heads h_1 and h_2 at layers n_1 and n_2 in the whole model, respectively. The lower the values of redundancies LR and HR, the larger the difference of attention weights between individual layers and individual heads, respectively. Moreover, the character error rate (CER) was measured to evaluate the performance of Chinese speech recognition. To analyze the computational complexity, the number of floating-point operations (FLOPs) was evaluated under different models and settings.

3) Model Configurations: This study viewed vanilla transformer (denoted as V-T) with standard attention as the baseline model for comparison with the proposed transformer using SDMA (denoted as SDMA-T). For ablation study, SDMA-T without head disentanglement, denoted by SMA-T, was implemented. State-of-the-art model based on DeLighT [44] was included for comparison. The settings in different transformers were consistent for fair comparison. This paper adopted a trick of magnifying the gradient for updating the cluster centroids $\boldsymbol{\mu}_c^z$ and $\boldsymbol{\mu}_c^q$ of GMMs via a learning rate η and a scalar hyperparameter c_g in $\boldsymbol{\mu}_c \leftarrow \boldsymbol{\mu}_c - \eta(c_g \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_c})$. This trick was helpful to speed up the convergence of training GMMs. Table I shows the experimental settings for four tasks including dimensions of embedding (Embed) and feed-forward network (FFN) layers, number of layers or TBs N in encoder and decoder, number of heads n_h, number of clusters n_c in GMM parameters θ_z and θ_q, regularization hyperparameters λ_k, λ_q, λ_x, λ_d, c_g, BPE dictionary size, number of mini-batch iterations and number of model parameters. Warmup iterations were changed from 4K to 8K. The other hyperparameters α=1.25, β=0.75, σ^2=0.01 and $\nu = 0.9$ were fixed. The scaling factor s of the scaled-dot production attention in (1) as shown in Algorithm 1 was reduced from 10 to 1. Adam optimizer was used with maximum learning rate 5×10^{-4}, dropout rate 0.3 and weight decay parameter 10^{-4}. Before evaluation, the weight parameters were averaged over those of the last 5 epochs. During evaluation, the beam search was applied to predict output sequence with beam size 5. In the implementation, the variational distributions $q(\mathbf{z}_i^h|\mathbf{x})$ and $q(\mathbf{q}_i^h|\mathbf{x})$ of an input token $\mathbf{x}_i$ were simply modeled by a single Gaussian with a constant variance $\widetilde{\sigma}^2 = 0.1$ and $\mathbf{z}_i^h$ and $\mathbf{q}_i^h$ as means, i.e., $q(\mathbf{z}_i^h|\mathbf{x}) = \mathcal{N}(\mathbf{z}_i^h, \text{diag}\{\widetilde{\sigma}^2\})$

TABLE I
MODEL CONFIGURATIONS OF SDMA TRANSFORMER IN DIFFERENT TASKS

Task	Config	Embed/FFN	N	n_h	n_c	λ_k	λ_q	λ_x	λ_d	c_g	BPE	Iters	Params
IWSLT'14 De-En	*base*	512/1024	6/6	4	8/8	0.01	100	10	1	10	6K/8K	55K	39.6M(39.5M)
	small	384/1024	6/6	4	4/4	0.01	100	10	1	10	6K/8K	55K	26.1M(26.0M)
	tiny	256/512	6/6	4	4/4	0.01	50	5	1	10	6K/8K	99K	11.9M(11.8M)
WMT'14 En-De	*base*	512/2048	6/6	8	4/4	0.01	100	10	1	10	44K/44K	196K	66.6M(66.5M)
	small	384/2048	6/6	8	4/4	0.01	100	10	1	10	44K/44K	313K	46.4M(46.3M)
	tiny	256/1024	6/6	4	4/4	0.01	50	5	1	10	44K/44K	330K	20.2M(20.1M)
WMT'17 Zh-En	*base*	512/1024	6/6	4	4/4	0.01	50	5	1	5	25K/20K	79K	55.1M(55.0M)
Aishell-1	–	256/1024	6/6	4	8/4	0.001	30	4	1	5	25K	20K	16.3M(16.2M)

Embed and FFN denote the dimensions in input embedding and feed-forward network layers, respectively. Vanilla transformer follows the same configurations. Iters denote the number of parameter updates in SDMA transformer. Numbers of clusters n_c for θ_z/θ_q are shown. Params denote the number of parameters in SDMA and vanilla (shown in brackets) transformers.

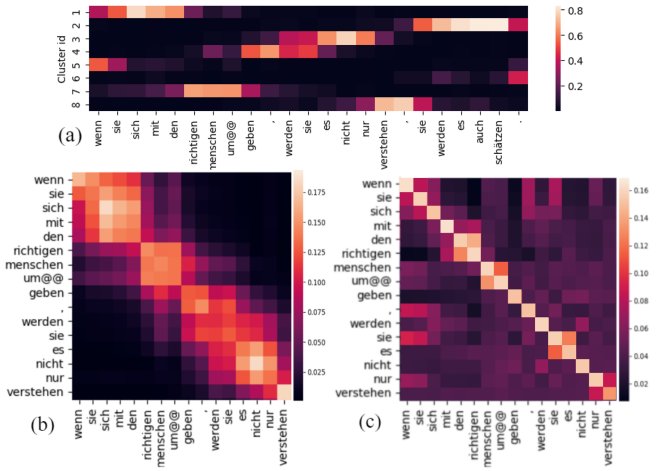

Fig. 6. (a) Clustering probability ($p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ in (9)) and (b) semantic mask (M_{ij}^h in (10)) of SDMA transformer, and (c) self attention weight (A_{ij}^h in (1)) of transformer in the first head of the sixth encoder layer where the BPE tokens in two axes of (b) and (c) are from the tokens of a German sentence in (a). IWSLT'14 De-En test set is evaluated.

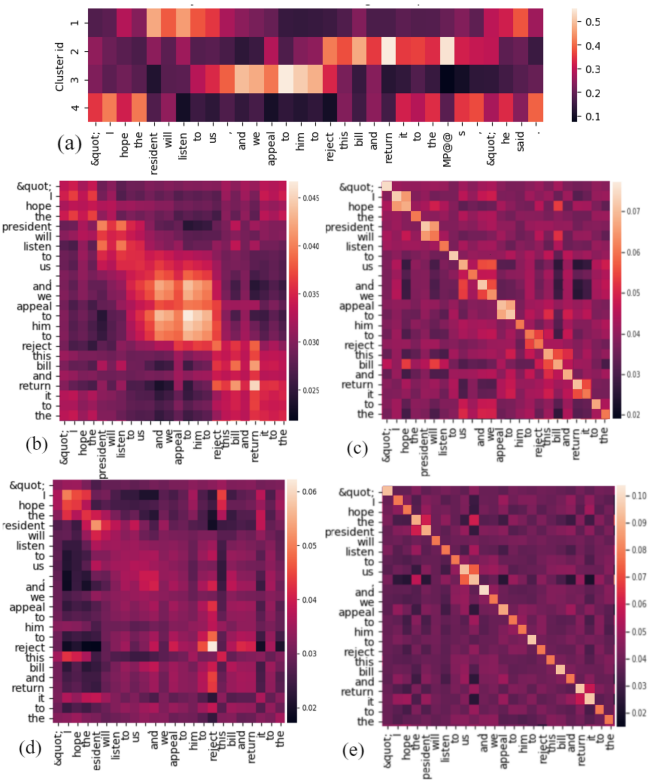

Fig. 7. (a) Clustering probability and (b) semantic mask of SDMA transformer, and (c) self attention weight of transformer in the first head of the second encoder layer. (d) Semantic mask of SDMA transformer and (e) self attention weight of transformer in the first head of the sixth encoder layer. Tokens of a test English sentence in WMT'14 En-De are evaluated.

and $q(\mathbf{q}_i^h|\mathbf{x}) = \mathcal{N}(\mathbf{q}_i^h, \text{diag}\{\widetilde{\sigma}^2\})$, respectively. The semantic mask in the transformer block of the masked SDMA in decoder was disregarded in test phase. To evaluate the proposed model under different model sizes, SDMA transformer was built with three configuration types, which were *base*, *small* and *tiny* where the dimensions of embedding and FFN layers were changed.

B. Experimental Analyses

1) Analysis on Semantic Clustering: Fig. 6 depicts the clustering probability $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ (9) of cluster c given token i and the corresponding semantic mask M_{ij}^h (10) of another token j in SDMA transformer. A number of German BPE tokens from IWSLT'14 De-En dataset are evaluated with eight GMM clusters for parameters θ_z. The measures of $p(\mathbf{c}_z^h = c|\mathbf{z}_i^h)$ and M_{ij}^h corresponding to the first head in the sixth encoder layer are shown. For comparison, the self attention weights A_{ij}^h (1) in vanilla transformer are shown. Typically, each token is likely related to different clusters without collapsing into the same cluster or semantic topic. This is affected due to the regularization term $\mathcal{L}_{\text{DV}}^z$. It is also found that the semantic mask is composed of several rectangles which express local features during semantic clustering. Nevertheless, each token is still closely related with an individual cluster where the last encoder layer is examined. The semantic masks are estimated to express the semantic relation between tokens with long distance. The weights A_{ij}^h in transformer are highly self attended on tokens themselves without clustering effect. In addition, Fig. 7 displays the clustering probability and semantic mask of SDMA transformer, and self attention weight of transformer where WMT'14 En-De task is evaluated. The first head in the second encoder layer is examined with a number of English tokens. The results corresponding

Fig. 8. Query samples $\mathbf{q}_i^h$ of four heads h (shown by colors) from 104 sentences (13,312 tokens) in the last decoder layer of (a) transformer and (b) SDMA transformer. IWSLT'14 De-En test set is used.

Fig. 9. Query samples of eight heads using (a) transformer and (b) SDMA transformer. 88 test sentences from WMT'14 En-De are used.

to the first head in the sixth encoder layer are also shown. Comparing with the semantic masks in the second and sixth layers, the second layer (Fig. 7(b)) captures local dependencies in BPE tokens while the sixth layer (Fig. 7(d)) characterizes long-distance semantic dependencies. Self attention weights (Fig. 7(c)(e)) simply capture the token-level connection without clear semantic topic features. In this task, we find that the topic words, namely the tokens with high clustering probabilities, corresponding to a specific cluster contain 'es', 'ing', 'ed', 'ies', etc which are semantically similar. Also, the semantically-close topic tokens 'year', 'time', 'Thursday', and 'times' are found to be associated with the same cluster. The semantic clustering and masking via GMM in a joint estimation could capture the semantic meaning.

2) Analysis on Latent Representations: Figs. 8 and 9 depict two-dimensional query samples $\mathbf{q}_i^h$ of tokens $\mathbf{x}_i$ calculated by using the attention weights $\{W_q^h, \mathbf{b}_q^h\}$ in the last decoder layer of vanilla and SDMA transformers where IWSLT'14 De-En and WMT'14 En-De translation tasks are investigated, respectively. Different heads are shown by different colors. Dimension reduction $\mathbb{R}^{d_k} \rightarrow \mathbb{R}^2$ using the t-distributed stochastic neighbor embedding [45] is applied. Typically, using V-T and SDMA-T, the query samples from different heads h properly represent high-level semantic meanings in different regions of latent space. Since the query vectors in SDMA-T are disentangled via minimizing $\mathcal{L}_{D_{qq}}$ and maximizing $\mathcal{L}_{D_{xq}}$, the resulting latent queries within the same head are more diverse and the gaps of

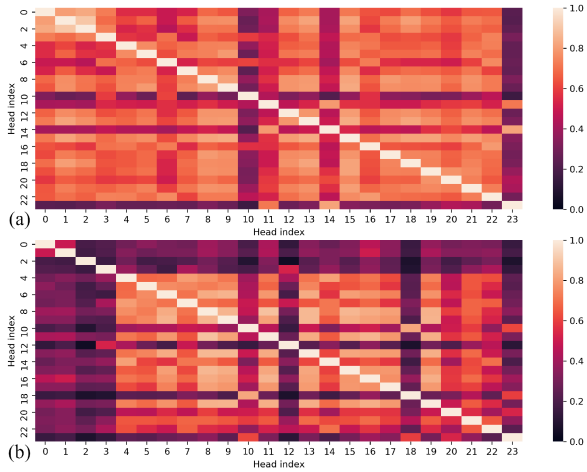

Fig. 10. Head redundancy in encoder of (a) transformer and (b) SDMA transformer is evaluated by using IWSLT'14 De-En test set.

queries between different heads are more separate than those of V-T where latent disentanglement is disregarded. The semantic diversity of query vectors using SDMA-T is truly increased in both tasks.

3) Analysis on Attention Redundancy: Next, the effect of query samples on attention weights is evaluated. Figs. 10 and 11 depict the head redundancy (HR) between individual heads in IWSLT'14 De-En and WMT'14 De-En translation tasks where the self attention weights over all 24 heads (6 layers $\times$ 4 heads)

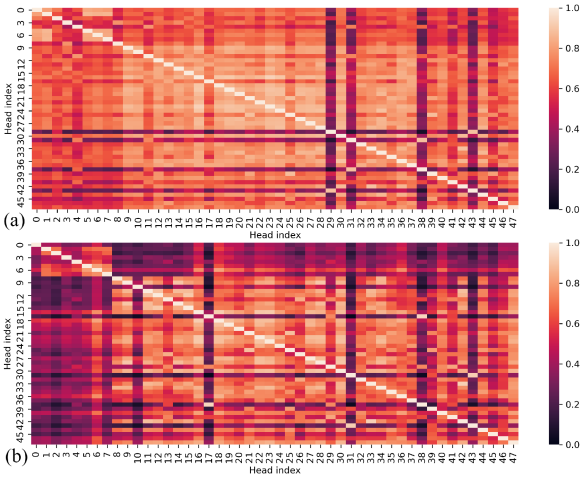

Fig. 11. Head redundancy in encoder of (a) transformer and (b) SDMA transformer is evaluated by using WMT'14 En-De test set.

and all 48 heads (6 layers $\times$ 8 heads) in the encoder are shown, respectively. The results of using vanilla and SDMA transformers are compared. It is found that the similarity of attention weights between individual heads in SDMA-T is generally smaller than that in V-T. This phenomenon likely happens since the semantic diversities of SDMA query samples in the same heads as well as different heads are both increased as shown in Figs. 8 and 9. As a result, the attention weights of SDMA-T are more diverse than those of V-T. HR using SDMA-T is consistently smaller than HR using V-T for both tasks with either 4 or 8 attention heads per layer. The proposed SDMA encourages different attention heads to capture various semantic relations between individual tokens. This is an evidence to illustrate the efficiency in learning representation by reducing the redundancy of attention weights in different layer-wise heads using SDMA.

C. Experimental Results

1) Comparison on Machine Translation: The experimental results on three machine translation tasks are investigated. The results of LR, HR and BLEU on vanilla transformer (V-T) [1], [44], DeLighT [44] and the proposed SDMA transformer (SDMA-T) with different number of parameters (or configurations *base*, *small* and *tiny*) and number of BPE tokens in encoder/decoder. SDMA-T without disentanglement (SMA-T) is included for comparison. Number of clusters n_c in GMMs of θ_z and θ_q is both fixed to be four. Tables II, III and IV report the results on using IWSLT'14 De-En, WMT'14 En-De and WMT'17 Zh-En tasks, respectively. For all three translation tasks, SDMA-T consistently achieves higher BLEU score, lower layer redundancy (LR) and lower head redundancy (HR) than V-T. For the results on IWSLT'14 De-En and WMT'14 En-De tasks, SDMA-T is consistently better than SMA-T in terms of LR, HR and BLEU. The performance is improved due to head disentanglement. Among the results in IWSLT'14 De-En task, the highest BLEU is obtained by SDMA-T (*base*) which absolutely improved 0.8 BLEU score over V-T. It is interesting that the BLEU 34.96 using SDMA-T (*tiny*) with model size as small as 11.9M is even higher than 33.80 using the competitive work DeLighT [44] with model size 14.0M. This BLEU 34.96 is higher than those of V-T where 3.32 times of parameters are required. In addition, the pre-trained BERT [46] is incorporated to implement the BERT-fused V-T (BV-T) and SDMA-T (BSDMA-T) with *base* model where the positional encoding and input/output embeddings are replaced by BERT contextual embeddings [47]. It is found that BLEU is increased by using BV-T and BSDMA-T relative to V-T and SDMA-T, respectively. In this comparison, BSDMA-T (*base*) considerably reduces LR and HR and consistently obtains the highest BLEU both in IWSLT'14 De-En and WMT'14 En-De tasks as reported in Tables II and III, respectively. In IWSLT'14 De-En task, the lowest LR 0.59 and HR 0.46 are obtained by using BSDMA-T (*base*). The absolute reduction of LR and HR relative to those of V-T (*base*) is 0.15 and 0.19, respectively. In WMT'14 En-De task, BSDMA-T (*base*) obtains 1.98 absolute improvement of BLEU relative to V-T (*base*). Again, SDMA-T (*small*) with model size 46.4M obtains even higher BLEU than DeLighT [44] with model size 54.0M. Under the same task, the improvement of SDMA-T (*tiny*) over DeLighT [44] (*tiny*) with smaller model size and higher BLEU is similarly obtained. In

TABLE II
COMPARISON OVER DIFFERENT METHODS IN IWSLT'14 DE-EN TASK

Model	Params	BPE	LR	HR	BLEU	FLOPs
V-T (*base*, [1], [44])	42.0M	10K	–	–	34.30	–
V-T (*base*, ours)	39.5M	6/8K	0.74	0.65	34.50	2.2G
BV-T (*base*)	39.5M	6/8K	0.69	0.60	34.78	2.3G
SMA-T (*base*)	39.6M	6/8K	0.65	0.58	34.93	2.3G
SDMA-T (*base*)	39.6M	6/8K	0.63	0.52	35.22	2.3G
BSDMA-T (*base*)	39.6M	6/8K	0.59	0.46	**35.99**	2.5G
SMA-T (*small*)	26.0M	6/8K	0.65	0.58	34.88	1.4G
SDMA-T (*small*)	26.1M	6/8K	0.64	0.57	35.00	1.4G
DeLighT [44]	14.0M	10K	–	–	33.80	–
SDMA-T (*tiny*)	11.9M	6/8K	0.63	0.58	34.96	0.6G

BPE denotes the number of tokens in the dictionary of encoder/decoder. 'Ours' denotes the result using the model trained by ourselves.

TABLE III
COMPARISON OVER DIFFERENT METHODS IN WMT'14 EN-DE TASK

Model	Params	BPE	LR	HR	BLEU	FLOPs
V-T (*big*, [1])	213.0M	37K	–	–	28.40	–
V-T (*base*, [1], [44])	67.0M	44K	–	–	27.70	–
V-T (*base*, ours)	66.5M	44K	0.79	0.69	27.75	5.5G
BV-T (*base*)	66.5M	44K	0.74	0.62	28.05	5.6G
SMA-T (*base*)	66.6M	44K	0.76	0.63	27.98	5.6G
SDMA-T (*base*)	66.6M	44K	0.73	0.58	28.95	5.7G
BSDMA-T (*base*)	66.6M	44K	0.67	0.52	**30.73**	5.9G
DeLighT [44]	54.0M	44K	–	–	28.00	–
SDMA-T (*small*)	46.4M	44K	0.70	0.57	28.16	4.1G
DeLighT [44]	23.0M	44K	–	–	26.70	–
SDMA-T (*tiny*)	20.2M	44K	0.68	0.56	27.73	2.0G

TABLE IV
COMPARISON OVER DIFFERENT METHODS IN WMT'17 ZH-EN TASK

Model	Params	BPE	LR	HR	BLEU	FLOPs
V-T (*base*, ours)	55.0M	25/20K	0.71	0.60	12.76	3.8G
SDMA-T (*base*)	55.1M	25/20K	0.62	0.55	**13.63**	3.9G

TABLE V
ABLATION STUDY ON CLUSTER NUMBER n_c UNDER IWSLT'14 DE-EN TASK

Model	Params	BPE	n_c	LR	HR	BLEU
SDMA-T (*base*)	39.56M	6/8K	4	0.63	0.52	35.22
SDMA-T (*base*)	39.65M	6/8K	8	0.62	0.53	**35.31**
SDMA-T (*base*)	39.74M	6/8K	12	0.62	0.53	35.29

case of very different languages in WMT'17 Zh-En task, it is found that the absolute increase of BLEU, decrease of LR and decrease of HR using SDMA-T (*base*) relative to V-T (*base*) are 0.87, 0.09 and 0.05, respectively. These results illustrate that SDMA transformer obtains higher BLEU than the other models for translation with grammatically similar and different languages. LR and HR are consistently reduced. Efficiency in learning representation is revealed.

2) Analyses on Model Size And Computation Cost: In the experiments, only two additional GMMs θ_z and θ_q are introduced for latent semantic and disentangled attention. The additional model parameters using SDMA-T compared to V-T are very limited in model size. In IWSLT'14 De-En task (Table II), SDMA-T (*tiny*) obtains 0.46 BLEU score higher than V-T (*base*) by using only 30% of model parameters. Compared to DeLighT [44], SDMA-T (*tiny*) obtains 1.16 BLEU score higher by using 5% smaller number of parameters. In WMT14'En-De task (Table III), SDMA-T (*small*) attains 0.41 BLEU score higher than V-T (*base*) by reducing 30% of model parameters. On the other hand, the performance of SDMA-T (*base*) is even higher than V-T (*big*) [1] by 0.55 BLEU score, but using only one third of model parameters. Introducing the objectives of variational clustering and disentangled attention heads into transformer does strengthen the latent representation of sequence data. Similar or even higher BLEU is obtained by using much smaller model. On the other hand, the effect on FLOPs by using different methods is comparable with that on parameter size [48]. Relative to V-T, the computational overhead using SDMA-T is very limited, but the improvement on BLEU is clear.

3) Ablation Study on Cluster Number: In what follow, the ablation study is conducted by varying the number of clusters n_c in IWSLT'14 De-En task as shown in Table V. The BLEU score of SDMA-T is absolutely improved by 0.09 relative to V-T when increasing the cluster number to eight. However, BLEU is decreased after increasing the cluster number to twelve. Choosing a proper cluster number is an issue of model selection for semantic representation. Basically, the size of GMMs does not change too much the overall model size. LR and HR are comparable for SDMA-T with various cluster numbers.

4) Ablation Studies on Dictionary Size and BERT Fusion: The ablation study is further implemented by varying BPE dictionary size in WMT'14 En-De task as reported in Table VI where the sizes of 37K and 44K are compared by fixing $n_c = 4$. Typically, the models with larger dictionary tend to have higher BLEU scores. SDMA-T with BPE dictionary size of 37K even obtains 0.19 BLEU score higher than V-T with BPE dictionary size of 44K while using 5% smaller number of parameters. This is due to the merit of semantic information of SDMA-T through semantic clustering so that similar performance as V-T can be achieved by using SDMA-T even when the number of parameters is reduced.

TABLE VI
ABLATION STUDY ON DICTIONARY SIZE UNDER WMT'14 EN-DE TASK

Model	Params	BPE	LR	HR	BLEU
V-T (*base*, [1])	65.0M	37K	–	–	27.30
V-T (*base*, ours)	63.08M	37K	0.79	0.70	27.44
SDMA-T (*base*)	63.16M	37K	0.73	0.61	27.94
V-T (*base*, ours)	66.52M	44K	0.79	0.69	27.75
SDMA-T (*base*)	66.61M	44K	0.73	0.58	**28.95**

TABLE VII
COMPARISON OVER DIFFERENT METHODS IN AISHELL-1 TASK FOR CHINESE SPEECH RECOGNITION

Model	LR	HR	Dev	Test	ERR	Params
LAS [49], [50]	–	–	–	10.6	–	–
V-T (ours)	0.55	0.43	7.8	8.6	–	16.2M
SMA-T	0.52	0.40	7.1	7.8	9.3	16.3M
SDMA-T	0.48	0.38	6.8	**7.4**	14.0	16.3M

Character error rates (%) on development/test data and error rate reduction (ERR) (%) on test data are shown.

5) Comparison on Speech Recognition: In evaluation of Chinese speech recognition on Aishell-1 with recipe [41], [42], a speech frame was expressed by 80-dimensional filter-bank features and then transformed into a 256-dimensional embedding vector where a convolution module consisting of two layers of time-domain 2D-convolution with 256 channels and kernel size 3 was used. In implementation of end-to-end speech recognition, the joint decoding for the connectionist temporal classification (CTC) (output of encoder with an additional linear layer) and the attention output (output of decoder) was performed. Table VII shows the results of LR, HR and CERs over development and test data by using listen-attend-spell (LAS) [49], [50], V-T [1], [50], SMA-T and SDMA-T. In this comparison, SDMA-T achieved the lowest LR, HR and CERs over development and test data. Without head disentanglement, the performance of semantic mask attention is degraded but still significantly better than V-T. The error rate reduction using SDMA-T relative to V-T reaches 14% with only 0.6% overhead of model parameters. Source codes are given at https://github.com/NYCU-MLLab.

V. CONCLUSION

This paper has presented a variant of transformer with mask attention in encoder and decoder where the semantic representation and head disentanglement were imposed. A Bayesian semantic mask attention was implemented by introducing the Gaussian semantic clusters of features and queries with clustering probabilities where the classification and disentanglement were jointly optimized with regularization. A semantic and disentangled transformer was constructed. The experiments on machine translation and speech recognition illustrated the merit of this work in reducing the redundancies in attention weights and improving the accuracies in sequence prediction where a number of different variants of transformer were compared. The

semantic diversities of queries within the same head and between different heads were both enhanced. The proposed semantic and disentangled attention was exploited for encoder-decoder in transformer and will be extended to implement Bayesian attention in other machines.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[2] J.-T. Chien and Y.-H. Chen, "Learning continuous-time dynamics with attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1906–1918, Feb. 2023.

[3] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Representations*, 2020.

[4] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 53–68, 2021.

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

[6] G. M. Correia, V. Niculae, and A. F. T. Martins, "Adaptively sparse transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 2174–2184.

[7] S. Wu, X. Xiao, Q. Ding, P. Zhao, Y. Wei, and J. Huang, "Adversarial sparse transformer for time series forecasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17105–17115.

[8] B. Peters, V. Niculae, and A. F. T. Martins, "Sparse sequence-to-sequence models," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1504–1519.

[9] Z. Fan et al., "Mask attention networks: Rethinking and strengthen transformer," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 1692–1701.

[10] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6649–6653.

[11] M. Guo, Y. Zhang, and T. Liu, "Gaussian transformer: A lightweight approach for natural language inference," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6489–6496.

[12] Y. Bian, J. Huang, X. Cai, J. Yuan, and K. Church, "On attention redundancy: A comprehensive study," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 930–945.

[13] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 3543–3556.

[14] H. Lio, S.-E. Li, and J.-T. Chien, "Adversarial mask transformer for sequential learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 4178–4182.

[15] J.-T. Chien and Y.-H. Huang, "Bayesian transformer using disentangled mask attention," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2022, pp. 1761–1765.

[16] H. Bahuleyan, L. Mou, O. Vechtomova, and P. Poupart, "Variational attention for sequence-to-sequence models," in *Proc. Int. Conf. Comput. Linguistics*, 2018, pp. 1672–1682.

[17] S. Watanabe and J.-T. Chien, *Bayesian Speech and Language Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[18] W. Chen and Y. Li, "Calibrating transformers via sparse Gaussian processes," in *Proc. Int. Conf. Learn. Representations*, 2023.

[19] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4114–4124.

[20] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5670–5679.

[21] J.-T. Chien and S.-E. Li, "Contrastive disentangled learning for memory-augmented transformer," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2023, pp. 2958–2962.

[22] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in VAEs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2615–2625.

[23] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, "Disentangled representation learning for non-parallel text style transfer," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 424–434.

[24] P. Cheng et al., "Improving disentangled text representation learning with information-theoretic guidance," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7530–7541.

[25] 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. Int. Conf. Learn. Representations*, 2021.

[26] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[27] J.-T. Chien and S.-J. Huang, "Learning flow-based disentanglement," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2390–2402, Feb. 2024.

[28] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1965–1972.

[29] J.-T. Chien and C.-H. Lee, "Deep unfolding for topic models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 318–331, Feb. 2018.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[31] J.-T. Chien and C.-W. Wang, "Hierarchical and self-attended sequence autoencoder," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4975–4986, Sep. 2022.

[32] A. Vyas, A. Katharopoulos, and F. Fleuret, "Fast transformers with clustered attention," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21665–21674.

[33] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5156–5165.

[34] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5243–5253.

[35] M. Zaheer et al., "Big bird: Transformers for longer sequences," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17283–17297.

[36] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv: 2004.05150*.

[37] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5171–5180.

[38] Z. Lin et al., "A structured self-attentive sentence embedding," in *Proc. Int. Conf. Learn. Representations*, 2017.

[39] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, 2017, pp. 1–5.

[40] M. Ott et al., "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 48–53.

[41] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2207–2211.

[42] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online compressive transformer for end-to-end speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2021, pp. 2082–2086.

[43] F. Wu, A. Fan, A. Baevski, Y. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," in *Proc. Int. Conf. Learn. Representations*, 2019.

[44] S. Mehta, M. Ghazvininejad, S. Iyer, L. Zettlemoyer, and H. Hajishirzi, "DeLighT: Deep and light-weight transformer," in *Proc. Int. Conf. Learn. Representations*, 2021.

[45] L. Van der Maaten and G. Hinton, "Visualizing data using *t*-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[46] J. D. Kenton, M.-W. Chang, and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[47] J. Zhu et al., "Incorporating BERT into neural machine translation," in *Proc. Int. Conf. Learn. Representations*, 2020.

[48] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 15908–15919.

[49] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 4960–4964.

[50] Z. Tian, J. Yi, Y. Bai, J. Tao, S. Zhang, and Z. Wen, "Synchronous transformers for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 7884–7888.

Jen-Tzung Chien is currently the Lifetime Chair Professor with National Yang Ming Chiao Tung University, Hsinchu, Taiwan. He has authored more than 250 peer-reviewed articles in machine learning, deep learning, and Bayesian learning with applications on natural language processing and computer vision, and three books including *Bayesian Speech and Language Processing*, Cambridge University Press, 2015, *Source Separation and Machine Learning*, Academic Press, 2018, and *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020. He was a tutorial speaker of AAAI, IJCAI, ACL, MM, KDD, ICASSP, CIKM, WSDM, COLING and Interspeech. He received the Best Paper Award in IEEE Workshop on Automatic Speech Recognition and Understanding in 2011, and IEEE International Workshop on Machine Learning for Signal Processing in 2023.

Yu-Han Huang received the BS degree in electrical engineering from National Chung Hsing University, Taichung, in 2019, and the MS degree in electrical and computer engineering from the National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2021. His interest focuses on sequential learning, generative modeling, and Bayesian machine learning.