Pay Better Attention to Attention: Head Selection in Multilingual and Multi-Domain Sequence Modeling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-head attention has each of the attention heads collect salient information from different parts of an input sequence, making it a powerful mechanism for sequence modeling. Multilingual and multi-domain learning are common scenarios for sequence modeling, where the key challenge is to maximize positive transfer and mitigate negative transfer across languages and domains. In this paper, we find that non-selective attention sharing is sub-optimal for achieving good generalization across all languages and domains. We further propose attention sharing strategies to facilitate parameter sharing and specialization in multilingual and multi-domain sequence modeling. Our approach automatically learns shared and specialized attention heads for different languages and domains to mitigate their interference. Evaluated in various tasks including speech recognition, text-to-text and speech-to-text translation, the proposed attention sharing strategies consistently bring gains to sequence models built upon multi-head attention. For speech-to-text translation, our approach yields an average of $+2.0$ BLEU over 13 language directions in multilingual setting and $+2.0$ BLEU over 3 domains in multi-domain setting.

1 Introduction

Recent progress on deep learning models, in particular multi-head attention, has brought significant gains to sequence modeling tasks including speech recognition (Moritz et al., 2020), text-to-text translation (Vaswani et al., 2017), and speech-to-text translation (Vila et al., 2018; Gangi et al., 2019). Attention mechanism allows a model to focus on informative parts of the inputs, and multi-head attention computes attention over inputs by multiple heads independently. With each head attending to different information, multi-head attention potentially captures more complicated data patterns and extracts sophisticated knowledge.

Sequence modeling has attracted a lot of research interest in multilingual and multi-domain settings, where a model is trained on data in multiple language directions and data from different domains respectively. Key advantages of these settings are better data efficiency and the support of knowledge transfer among languages or domains. This is critical for resource-limited scenarios. For example, multilingual translation enhances the performance of low-resource languages via knowledge transfer from high-resource languages (Gu et al., 2018; Inaguma et al., 2019b). Given the data scarcity in individual domains, a common practice is to combine the data from various domains to augment the training set (Wang et al., 2020d). Another appealing aspect of multilingual or multi-domain models is their low deployment and maintenance costs compared with numerous models trained for individual language pairs or domains.

Despite the positive knowledge transfer, negative interference has also been observed in multilingual (or multi-domain) training especially when languages (or domains) are dissimilar. Recent studies reveal from the optimization perspective that conflicting gradients in shared parameters is one cause of

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

interference between languages (or domains) (Yu et al., 2020). A promising direction for interference mitigation is to design better strategies of parameter sharing. In some previous works, sharing is based on the similarity between languages (or domains), which require expert knowledge or pre-computed relatedness (Wu et al., 2019). Recent studies also propose branches and components specific to languages (or domains) in addition to shared modules (Bapna and Firat, 2019; Guo et al., 2020).

In this work, we bring the mitigation of language and domain interference under a common umbrella, and tackle it by improving parameter sharing within multi-head attention. We propose strategies to select attention heads for different languages or domains. Instead of sharing everything across languages or domains, our model automatically learns to share heads among a subset of languages or domains. It encourages positive transfer within the subset and preserves their specificity without interference from outside the subset. The major contributions of this work are summarized below:

1. We propose attention head selection to mitigate interference in multilingual and multi-domain modeling;

2. The parameter sharing strategies are lightweight and preserve computational efficiency;

3. We extensively evaluate attention sharing strategies on various sequence modeling tasks including speech recognition, text-to-text and speech-to-text translation. Consistent gains are achieved across multiple benchmark datasets.

The paper is structured as follows. Section 2 discusses related works on sequence modeling in multilingual and multi-domain setting. In Section 3, we introduce the proposed strategies of head selection in multi-head attention. Section 4 describes the empirical evaluation, followed by a discussion in Section 5. We conclude this paper in Section 6.

2 Related Work

Multilingual learning. Multilingual modeling has the potential to improve low-resource language performance through knowledge transfer from high-resource languages, and it draws great interest from researchers in speech recognition and translation (Pratap et al., 2020; Heigold et al., 2013; Johnson et al., 2017; Dabre et al., 2020; Liu et al., 2020; Inaguma et al., 2019a; Li et al., 2020). Although impressive progress has been made for low-resource or zero-shot tasks, it is also found the multilingual model has inferior performance on high-resource tasks due to multilingual interference. In order to address this issue, some works focus on multilingual models with task-specific parameters. Different parameter sharing strategies are examined for the Transformer model (Sachan and Neubig, 2018). Attention dependent on target languages is proposed to enhance the multilingual translation performance (Blackwood et al., 2018). Treating multilingual modeling as an adaptation problem, Bapna and Firat (2019) first build a general multilingual model for all languages and then finetune newly added residual adapters for each language pair. Another thread of work is to increase the model capacity to compensate for the high-resource language loss (Pratap et al., 2020). Shazeer et al. (2017) propose mixture-of-experts and select RNN cells based on input tokens. Lepikhin et al. (2020) extend it to Transformer with FFN experts. Different from previous works, we propose strategies of attention sharing among languages in the level of attention heads for multilingual modeling.

Multi-domain learning. Similar to multilingual learning, multi-domain learning (MDL) can effectively utilize data from different domains but also suffers from interference due to inter-domain heterogeneity (Saunders, 2021; Pham et al., 2021). Previous works address this issue from two perspectives: optimization and model architecture. For the optimization aspect, attempts have been made to synchronize the learning speed of different tasks (Chen et al., 2018), adjust the gradients of individual tasks to alleviate gradient conflicts (Yu et al., 2020) and apply regularization to achieve better generalization in different domains (Dakwale and Monz, 2017; Khayrallah et al., 2018; Thompson et al., 2019). In terms of model architecture, domain-specific labels (Kobus et al., 2017), word embedding (Zeng et al., 2018a), sub-networks (Wang et al., 2020d) are adopted to address the issue of domain divergence. The architecture can be specified during the general training with the mixed data from multiple domains (Wang et al., 2020d) or during the finetuning in individual domains (Bapna and Firat, 2019). In this work, we deal with domain interference by leveraging domain-specific attention heads in multi-head attention.

Attention selection. Selective self-attention networks proposes to apply masking to the inputs and pay more attention to content words (Geng et al., 2020). Liu et al. (2021) selects text-related

image regions with attention in multi-modality translation. Compared to their methods, we conduct automatic attention head selection for different tasks and focus on mitigating task interference.

3 Model

In this section, we start with preliminaries of multi-head attention, and introduce our approach to attention interference mitigation. We put multilingual and multi-domain sequence modeling under the same umbrella in this study. For the simplicity of the following discussions, we refer to the two settings as multi-task modeling, where a task is one language or one domain. Different from the standard multi-head attention, our model provides more attention heads than those used in computation. Different subsets of heads are assigned to each task so that partial attention sharing enables knowledge transfer and meanwhile mitigates interference. We introduce latent variables to modulate head selection, and propose strategies to learn the head assignment to different tasks.

3.1 Preliminary

Multi-head attention. As a core module in Transformer, multi-head attention paramterizes each head with key, query and value transformation matrices (Vaswani et al., 2017). The token representation is transformed into key, query and value vectors via these transformations. Each head assigns the attention of this token over the input sequence based on the matching between its query vector and key vectors of other tokens. The value vectors are weighted by the attention as the contextualized token representation. It is passed through linear projection as the output of the attention head. Suppose that head h has output $\mathbf{x}^{(h)}$. Multi-head attention with H heads yields an output $\mathbf{x}$ for the given token, which is the concatenation of all head outputs.

$$\mathbf{x} = \mathbf{x}^{(1)} \oplus \cdots \oplus \mathbf{x}^{(h)} \oplus \cdots \oplus \mathbf{x}^{(H)}, \tag{1}$$

where $\oplus$ is vector concatenation.

Interference. Maximal parameter sharing aims to learn universal knowledge across languages (Wang et al., 2020e) and domains (Zeng et al., 2018b). To capture the task specificity, different languages or domains compete for model capacity, which is observed as the interference in previous studies. The interference results in degraded performance in jointly trained models. However, few works look into the improvement of parameter sharing within multi-head attention. This study explores head selection strategies to mitigate the inference in multilingual and multi-domain models.

3.2 Latent Variable for Head Selection

First, we outline our approach to learn a more general-purpose multi-head attention in Transformer from the Bayesian neural network perspective. Suppose that the input sequence is x and the output sequence is y. For conditional sequence modeling tasks such as machine translation, the posterior of $p(y \mid x)$ can be computed by marginalizing over the posterior of latent variable z, which modulates parameters Θ in the standard Transformer architecture:

$$p(y \mid x, \Theta) = \mathbf{E}_{p(z|\Theta)}[p(y \mid x, z)] = \int p(y \mid x, z) p(z|\Theta) \, \mathrm{d}z \tag{2}$$

Parameterization of z. In this work, we define z as modulating the selection of attention heads. We model $z_t^{(h)}$ as a discrete latent variable from a Bernoulli distribution with $z_t^{(h)} \sim \mathcal{B}(\pi)$, $\pi \in [0, 1]$ indicating whether task t selects attention head h. This modeling choice allows us to prune attention heads during computation, which preserves computation efficiency as well as regularizes training.

Marginalizing over z is intractable given numerous heads in neural models. Therefore, we use variational inference to derive an approximate solution. Specifically, we learn an inference network $q_\phi(z)$, which is paramterized with ϕ, to approximate the true distribution $p(z)$ and optimize the evidence lower bound (ELBO) of Eq. 2:

$$\log p(y \mid x) \geq \mathbf{E}_{q_\phi(z)}[\log p_\theta(y \mid x, z)] - \mathrm{KL}(q_\phi(z) \parallel p(z)), \tag{3}$$

where KL is the KL-divergence between two distributions. In our work, we assume identical probability of each head being selected. Therefore, we have $p(z = 1) = \frac{H}{H'}$, where H and H' are numbers of selected heads and all head candidates.

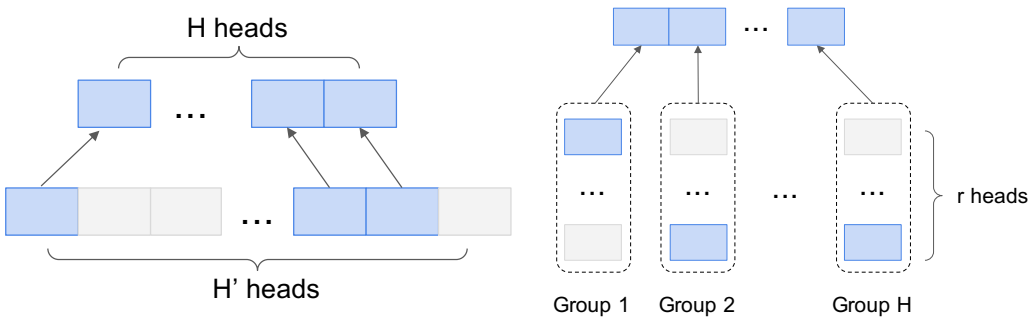
Training and interference. We use the Gumbel-Softmax reparameterization (Jang et al., 2017) to draw samples of z from the posterior $q_\phi(z)$. It makes the model end-to-end differentiable, while learning discrete policies of head selection without resorting to policy gradients. We adopt a lightweight estimator of $q_\phi(z)$ by directly learning the logit parameters $\{\phi_t^{(h)}\}$:

$$q_\phi(z_t^{(h)}) = \frac{\exp((\phi_t^{(h)}(1) + \epsilon(1))/\tau)}{\sum_{j\in\{0,1\}} \exp((\phi^{(h)}(j) + \epsilon(j))/\tau)}, \epsilon \sim \mathcal{G}(0,1) \tag{4}$$

where $\mathcal{G}(0,1)$ is the Gumbel distribution, and τ is a temperature hyperparameter which increases the discreteness of samples when $\tau \to 0$.

We will discuss different head selection strategies in Section 3.3, which make binary selection decisions based on real-valued posterior $q_\phi(z_t^{(h)})$.

3.3 Attention Selection Strategies



(a) Subset strategy.

(b) Group strategy.

Figure 1: Attention sharing strategies. The blue heads are selected while the grey heads are not.

Suppose that the output dimension of multi-head attention is d, and the dimension of each attention head is $\frac{d}{H}$. We provide a large pool of H' ($H' > H$) attention head candidates in every Transformer layer, and H' is a hyperparameter controlling the search space size of attention selection strategies. The model requires attention outputs to have a consistent dimension d, so each task needs to select H heads among H' candidates. Knowledge transfer is enabled among tasks accessing the same attention heads, and interference is disabled among tasks without attention sharing. We introduce two strategies for the attention head selection: subset strategy and group strategy.

Subset strategy. The strategy is straightforward, and we compare the posterior $\{q_\phi(z_t^{(h)}) : h \in [1, H']\}$ of all H' heads given a task t. A subset of H heads with the highest posterior are selected by the task, and there are $C_H^{H'}$ subset choices. The subset strategy is described in Fig. 1(a). The binary mask $s_t^{(h)}$ indicate whether an attention head h is assigned to task t.

$$s_t^{(h)} = \begin{cases} 1, & h \in \text{TopH}(\{q_\phi(z_t^{(h)})\}), \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where $\text{TopH}(\cdot)$ returns the top H heads with the highest values.

The outputs of the selected heads are concatenated as the attention output. Note that the subset strategy does not consider the order of the attention heads. For example, when head 2 and 3 are selected, head 2 contributes to the beginning part of attention output. With head 1 and 2 selected, the output of head 2 goes to the last part of the attention output.

Group strategy. We further propose group strategy to preserve the order of attention heads during head selection. Different from the subset strategy, the group strategy first divides H' heads into H groups. As is shown in Fig. 1(b), each group contains $r = \frac{H'}{H}$ candidates. Each task could choose one attention head from each group, and has access to H heads per layer. There are r^H

possible combinations of heads. The group strategy keeps the head order in that heads from group g only contribute to g's corresponding dimensions in the attention output. The head with the highest posterior in its group would be selected by a given task t. We use binary masks $\{s_t^{(h)}\}$ to indicate the selection of head h in group g.

$$s_t^{(h)} = \begin{cases} 1, & h = \text{argmax}(\{q_\phi(z_t^{(h')}) : h' \in g\}), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The output of group g is:

$$\mathbf{x}^{(g)} = \sum_{h \in g} s_t^{(h)} \cdot \mathbf{x}^{(h)}. \tag{7}$$

The outputs of H groups are concatenated as the output of the attention module for task t.

With either subset or group strategy, the sequence model is trained to assign attention heads to different tasks to maximize the lower bound in inequality (3). The number of additional parameters $\{\phi_t^{(h)}\}$ introduced by our attention selection is only $T \times H' \times L$, where T is the number of tasks, H' is the number of head candidates per layer, and L is the number of layers. It is small compared with the total parameter size of the model, and head selection is thus lightweight and memory efficient. Moreover, the head selection is inherently a pruning process. Regardless of the size of head candidates, only a fixed number of attention heads are involved in computation for a given task. Hence our approach is also computationally efficient in both training and inference.

4 Experiments

We evaluate sequence models in multilingual and multi-domain settings respectively. Various applications are considered including multilingual machine translation (MT), automatic speech recognition (ASR) and speech translation (ST) in both multilingual and multi-domain settings. We include widely used sequence models built on multi-head attention as strong baselines. All baselines have an encoder-decoder architecture, leveraging attention in each encoder and decoder layer.

1.Transformer (Vaswani et al., 2017). It is a state-of-the-art model in machine translation, which takes texts in source languages as inputs and generates texts in target languages.

2. S2T Transformer (Wang et al., 2020a). As a variant of Transformer for speech recognition and translation, S2T Transformer takes audio features and generates target texts. It is a stack of a convolutional subsampler and a Transformer model, where the subsampler processes input features (i.e., log mel-filter features in our experiments) and sends them to Transformer for text generation.

3. Adapter model (Bapna and Firat, 2019). Adapters have been shown as an effective approach to language and domain adaptation. Based on a well-trained Transformer or S2T Transformer, task-specific layers are added on top of each encoder and decoder layer. Parameters of adapter layers are trained with other model parameters frozen. A typical adapter layer consists of two feed-forward sub-layers. Adapter is applied to every task in our experiments.

We integrate attention selection strategies into the self-attention[1] module. Our implementation is based on the FAIRSEQ toolkit (Ott et al., 2019; Wang et al., 2020b).

4.1 Machine Translation

The task of machine translation is to translate a text from the source language into the target language. BLEU is an evaluation metric in translation, which measures the overlap between model translations and the ground truth (Papineni et al., 2002). Higher BLEU reflects better translation quality.

Dataset. We experiment with public multilingual machine translation datasets collected by WMT shared tasks as used by (Liu et al., 2020). The dataset consists of parallel sentences between English

[1]We also tried head selection in the encoder-decoder attention but did not observe big improvements when using it alone or in combination with self-attention head selection.

and other 14 languages[2]. Its data statistics are summarized in Appendix A.1. We evaluate models on both one-to-many (O2M) and many-to-one (M2O) translations, which are translation from English to 14 languages and from 14 languages to English respectively.

Model configurations. All models have 6 encoder layers and 6 decoder layers with 4 attention heads per layer (i.e., $H = 4$). The embedding dimension is 512 and the feed-forward dimension is 1024. They are trained with a batch size of 131k tokens and a learning rate of 0.0007. For O2M translation, attention selection models and Transformer are trained for 140k steps. As for M2O translation, they are trained for 100k steps. The attention selection is based on the source language in the encoder side for M2O translation, and is based on the target language in the decoder part for O2M translation.

The adapter model is initialized with parameters from the trained Transformer. It then trains the new parameters of the adapter layers for 40k steps with Transformer parameters frozen. The adapter layers are added to Transformer for each language direction, and they have an intermediate dimension of 256. The dimension is selected so that the number of parameters (460M) in the adapter model is close to the parameter size (420M) in attention selection models. Our attention selection sets the number of attention head candidates as 8 in each layer (i.e., $H' = 8$) for both subset and group strategies. We will discuss how the hyperparamter H' affects model performance in Section 5.

Table 1: BLEU (↑) of Machine Translation on WMT Datasets (AVG-A: average BLEU over 14 directions, High and Low are average BLEU over high- and low-resource languages respectively.)

	O2M			M2O		
	AVG-A	High	Low	AVG-A	High	Low
Transformer	20.1	25.7	16.0	22.8	27.9	19.0
Adapter	20.9	26.7	**16.6**	23.3	28.7	19.3
Group strategy	**21.0**	**27.1**	16.4	**23.5**	**28.8**	**19.6**
Subset strategy	20.9	27.0	16.4	23.3	28.7	19.4

Results. We group 14 language directions based on their amount of training data. We have 6 high-resource languages with more than 10M parallel sentences, 8 low-resource languages with fewer than 10M sentence pairs. Table 1 shows model performance on WMT datasets. Both attention selection and adapter models demonstrate gains over the multilingual Transformer. Group strategy achieves +0.9 and +0.7 BLEU on average of 14 language directions in O2M and M2O translations respectively. Adapter has comparable performance to both group and subset strategies. We note that adapter increases computational costs due to the additional 12 adapter layers added to Transformer, while the attention head selection approaches preserve the computation efficiency.

4.2 Speech Recognition

The task of Automatic Speech Recognition (ASR) is to transcribe source audios in the same language. Word error rate (WER) is ASR evaluation metric, which measures the difference of model outputs from the ground truth (Klakow and Peters, 2002). Lower WER indicates better recognition.

Model configuration. Models included in the experiments of speech recognition are S2T Transformer, S2T Transformer with adapter layers and S2T Transformer with attention selection. Following the setup of (Salesky et al., 2021), all models have 1024 channels in the input convolutional subsampler, 12 encoder layers and 6 decoder layers with 4 attention heads per layer. The embedding dimension is 256 and the feed-forward dimension is 2048. We set a batch size of 320k tokens and a learning rate of 0.0005 during training. Attention selection models and S2T Transformer are trained for 250 epochs. Adapter model is initialized with parameters of the trained S2T Transformer, and is then trained for another 200 epochs with only adapter layer parameters tuned. The intermediate dimension of adapter layers is again set as 256. To prevent over-fitting, we stop the model training when the model does not improve on the validation set for 10 epochs. To reduce the performance variance, we average checkpoints of the last 10 epochs, and use the averaged model for evaluation.

[2]The 14 languages are: Chinese, Czech, Estonian, Finnish, French, German, Gujarati, Kazakh, Latvian, Lithuanian, Romanian, Russian, Spanish, Turkish

4.2.1 Multilingual Speech Recognition

Dataset. We use the multilingual TEDx (mTEDx) dataset for speech recognition (Salesky et al., 2021). It collects audio recordings from TEDx talks. Eight languages are covered including Arabic (ar), German (de), Greek (el), Spanish (es), French (fr), Italian (it), Portuguese (pt) and Russian (ru).

Table 2: WER (↓) of Speech Recognition on mTEDx Dataset

	AVG	ar	de	el	es	fr	it	pt	ru
S2T Transformer	49.0	109.5	72.3	43.3	23.9	27.8	28.6	31.0	55.3
Adapter	41.1	93.4	57.2	33.0	21.4	25.3	24.3	27.2	46.7
Group strategy	**40.0**	94.2	59.8	33.5	18.2	22.0	21.9	24.6	45.5
Subset strategy	44.7	97.3	65.3	38.7	22.4	25.8	26.4	29.0	52.4

Results. S2T transformer share all parameters among languages. Attention selection models select attention heads based on the source and target languages in the multilingual setting. Adapter adds adapter layers based on the language directions. We report the ASR results in Table 2. Attention selection with either group or subset strategy is shown to reduce the word error rate in comparison with S2T Transformer. Adapter model also achieves lower WER than S2T Transformer. Group strategy yields the lowest WER, achieving an average drop of 18.4% compared with S2T Transformer. Adapter reduces the WER of S2T Transformer by 16.1%.

4.2.2 Multi-Domain Speech Recognition

Dataset. Besides mTEDx data, we include two other public datasets, CoVoST and EuroParl, which are commonly used for speech translation. Since source audios are accompanied by transcripts, we could use their source audio-text data for speech recognition tasks. We investigate multi-domain modeling with these three datasets.

1. CoVoST (Wang et al., 2020c). With Common Voice as the audio source, CoVoST covers speech-to-text translations from 22 languages to English and from English to 15 languages.
2. EuroParl (Iranzo-Sánchez et al., 2020). It provides paired audio-text instances from and into 6 European languages, which are compiled from the debates in European Parliament.

Table 3: WER (↓) of Speech Recognition on mTEDx, CoVoST and EuroParl Dataset

	mTEDx	CoVoST	EuroParl
S2T Transformer (separate)	49.0	41.9	115.0
S2T Transformer (joint)	42.7	38.3	25.6
Adapter	41.7	37.0	**24.0**
Group strategy	**41.0**	**36.4**	24.3
Subset strategy	41.8	37.0	25.0

Results. In the multi-domain setting, attention selection models assign different heads to each domain, and adapter model adds domain-specific adapter layers to S2T Transformer. We train models for 400 epochs in the multi-domain setting. Table 3 reports WER of models in three domains: mTEDx, CoVoST and EuroParl respectively. The S2T Transformer jointly trained on multi-domain data (in the row of "S2T Transformer (joint)") reduces WER by 12.9%, 8.6% and 77.7% in three domains respectively, when compared with the models separately trained in individual domains (in the row of "S2T Transformer (separate)"). This demonstrates the benefits of positive transfer between domains.

The performance of speech recognition could be further improved by the mitigation of the domain interference. Both attention selection and adapter model achieve lower WER than the joint S2T Transformer. Attention selection with group strategy has the lowest WER on both mTEDx and CoVoST datasets, decreasing WER by 4.0% and 5.0% respectively in comparison with joint S2T Transformer. The best system on EuroParl is adapter model, yielding a WER reduction by 6.3% than the joint S2T Transformer.

4.3 Speech Translation

Now with a focus on the task of speech translation, we again design experiments in multilingual and multi-domain settings. In the multilingual setup, we train translation models with samples in multiple languages to investigate language interference. As for the multi-domain setup, the models are trained with data from multiple domains so that we could look into the domain interference. BLEU serves as the evaluation metric of speech translation systems.

Baselines. We use the same baselines as in speech recognition. As recommended by (Salesky et al., 2021), we initialize the encoders in speech translation with the encoders trained in the task of speech recognition in Section 4.2 for the purpose of improving training efficiency and performance.

Model configurations. All models are trained for up to 400 epochs. Other model configurations in ST are the same as those in ASR.

4.3.1 Multilingual Speech Translation

To explore language interference, we perform experiments on multilingual speech translation.

Dataset. We again use mTEDx dataset for multilingual speech translation. Besides speech recognition data, mTEDx also collects speech translation data from TEDx talks. Its test set covers 13 language directions. The training data is provided in 10 of these directions, so there are 3 zero-shot directions.

Table 4: BLEU (↑) of Speech Translation on mTEDx (AVG-A: average over all directions, AVG-T: average of 10 training directions, and AVG-Z: average of 3 zero-shot directions)

	AVG-A	AVG-T	AVG-Z
S2T Transformer	13.2	14.6	8.5
Adapter	-	14.8	-
Group strategy	**15.2**	**16.7**	**10.4**
Subset strategy	13.3	14.7	8.5

Results. Table 4 summarizes the multilingual speech translation results on mTEDx. Since adapter model brings in language-specific layers, it cannot deal with zero-shot translations. Both attention selection models and adapter model bring improvements over S2T Transformer which is jointly trained in 13 language directions. It suggests that multiple languages interfere within S2T Transformer whose parameters are shared by all languages. Attention selection with group strategy achieves the best translation performance. In comparison with S2T Transformer, group strategy achieves an average of +2.1 and +1.9 BLEU in training and zero-shot directions respectively. It leads to +2.0 BLEU on average of all directions.

4.3.2 Multi-Domain Speech Translation

In this experiment, we investigate interference across domains in the task of speech translation, and evaluate the effectiveness of different models in multi-domain training. The attention selection now is based on the data domain instead of languages, i.e., samples in different domains would choose their own attention heads. Similarly for adapter model, its adapter layers are domain-specific in this setup.

We again use CoVoST and EuroParl as additional domains. We focus on the 13 language directions in mTEDx test set, and use the subset of CoVoST and EuroParl corpora in the same directions. CoVoST has 5 common directions[3] and EuroParl has 11 common directions[4] as mTEDx. Details about these datasets are included in Appendix A.1.

Results. Table 5 shows the average BLEU of speech translation in mTEDx, CoVoST and EuroParl. We report results in the row of "S2T Transformer (joint)" when S2T Transformer is trained with a mixture of three datasets. The results are included in the row "S2T Transformer (separate)" when S2T Transformer is trained on each dataset independently. Zero-shot translations in mTEDx benefit a lot from additional data of CoVoST and EuroParl, as the joint S2T Transformer shows an average of +5.4 BLEU over separate S2T Transformer. However, there is a drop of 1.0 BLEU in its training directions, brought by the interference from CoVoST and EuroParl domains.

Again we observe that both attention selection and adapter model bring gains to the joint model in individual domains. Compared with the joint S2T Transformer, adapter model improves mTEDx

[3] {es, fr, it, pt, ru}-en

[4] es-{en, fr, it, pt}, fr-{en, es, pt}, it-{en, es}, pt-{en, es}

Table 5: BLEU (↑) of Speech Translation on mTEDx, CoVoST and EuroParl Dataset

	TEDx			CoVoST	EuroParl
	AVG-A	AVG-T	AVG-Z	AVG	AVG
S2T Transformer (separate)	13.2	14.6	8.5	17.6	19.1
S2T Transformer (joint)	13.7	13.6	13.9	17.3	19.0
Adapter	14.0	14.3	13.2	17.9	20.0
Group strategy	**15.6**	**15.9**	**14.8**	**19.6**	**20.8**
Subset strategy	13.8	14.3	13.1	17.9	19.2

translation by 0.3 BLEU, CoVoST translation by 0.6 BLEU and EuroParl by 1.0 BLEU on average. The attention selection with group strategy outperforms all other models. Its average BLEU gain over adapter model is 1.6 BLEU in mTEDx, 1.7 BLEU in CoVoST and 0.8 in EuroParl.

5 Discussion

Hyperparameter H'. The attention selection models set a hyperparameter H' as the total number of attention head candidates in multi-head attention, which controls the search space of attention sharing strategies. We now explore how the performance varies with H' for group and subset strategies.

Evaluated on the task of multilingual speech recognition, models have the same hyperparameters as those in multilingual ASR experiments except for H'. Attention selection models are configured with $H' = 4, 8, 12, 16$ respectively, and Figure 2 shows the change of WER with H'.

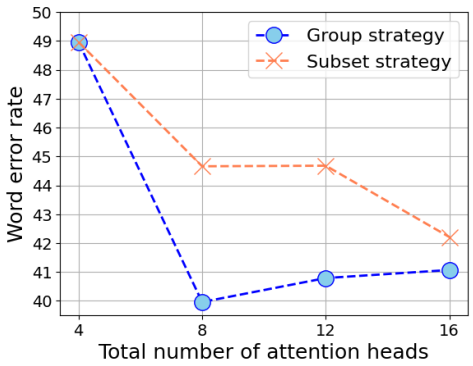


Figure 2: WER of speech recognition on mTEDx with different number of candidates.

When $H' = 4$, there is no attention selection and all attention heads are shared by different languages. We observe a large drop of error rate as H' increases from 4 to 8. For the subset strategy, WER keeps decreasing when the number of head candidates grows from 4 to 16. As for group strategy, $H' = 8$ is the optimal hyperparameter on the ASR task. As we continue increasing H' to 12 and 16, the error rate increases a bit. The performances of subset and group strategies are close when $H' = 16$.

The search space of group strategy is a strict subset of the space of subset strategy. However, we observe that group strategy shows comparable or better performance than subset strategy across tasks, including MT, ASR and ST. One possible explanation is that group strategy keeps the head order information while subset strategy does not. With a larger pool of head candidates, there is less sharing among tasks. The performance of the group strategy degrades a bit due to less positive transfer dependent on attention sharing. As for the subset strategy, better head assignments are learned with enlarged search space. Due to page limit, we include more discussions in Appendix A.3.

6 Conclusion

Research efforts in multilingual and multi-domain modeling have been driven by the increasing need to improve data efficiency and model performance. In this work, we propose head selection strategies to allow heads to be shared or specialized for different languages or domains. It effectively mitigates interference within multi-head attention which is a core part of strong sequence models, and demonstrates good empirical gains in various text generation tasks.

This work has several limitations left for future research. We did not explore head selection based on both language and data domain. We did not analyze model fairness and robustness. As a technology used for text generation, the model might have systemic bias or produce inappropriate outputs.

References

Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, 2019.

Graeme W. Blackwood, Miguel Ballesteros, and Todd Ward. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3112–3122. Association for Computational Linguistics, 2018.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 793–802. PMLR, 2018.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53:1 – 38, 2020. URL https://dl.acm.org/doi/pdf/10.1145/3406095.

Praveen Dakwale and Christof Monz. Finetuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, 117, 2017.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. Adapting transformer to end-to-end spoken language translation. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1133–1137. ISCA, 2019.

Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu. How does selective mechanism improve self-attention networks? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2986–2995. Association for Computational Linguistics, 2020.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, 2018.

Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, pages 3854–3863. PMLR, 2020.

Georg Heigold, Vincent Vanhoucke, Andrew W. Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. Multilingual acoustic models using distributed deep neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8619–8623, 2013.

H. Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577, 2019a.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 570–577. IEEE, 2019b.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

M. Johnson, M. Schuster, Quoc V. Le, M. Krikun, Y. Wu, Z. Chen, Nikhil Thorat, Fernanda B. Viégas, M. Wattenberg, G. Corrado, Macduff Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. Regularized training objective for continued training for domain adaptation in neural machine translation. In Alexandra Birch, Andrew M. Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 36–44. Association for Computational Linguistics, 2018.

Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 372–378. INCOMA Ltd., 2017.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *CoRR*, abs/2006.16668, 2020.

X. Li, Changhan Wang, Y. Tang, C. Tran, Yuqing Tang, J. Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv: Computation and Language*, 2020.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. Gumbel-attention for multi-modal machine translation. *CoRR*, abs/2103.08862, 2021. URL https://arxiv.org/abs/2103.08862.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Niko Moritz, Takaaki Hori, and Jonathan Le. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE, 2020.

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 Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

MinhQuang Pham, Josep Maria Crego, and François Yvon. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9:17–35, 2021.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Y. Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. In *INTERSPEECH*, 2020.

Devendra Singh Sachan and Graham Neubig. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 261–271. Association for Computational Linguistics, 2018.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757, 2021. URL https://arxiv.org/abs/2102.01757.

Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR*, abs/2104.06951, 2021. URL https://arxiv.org/abs/2104.06951.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2062–2068. Association for Computational Linguistics, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Laura Cross Vila, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. End-to-end speech translation with the transformer. In Jordi Luque, Antonio Bonafonte, Francesc Alías Pujol, and António J. S. Teixeira, editors, *Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings*, pages 60–63. ISCA, 2018.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, 2020a.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020b.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310, 2020c. URL https://arxiv.org/abs/2007.10310.

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9233–9241, 2020d.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, 2020e.

Sen Wu, Hongyang R Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 447–457. Association for Computational Linguistics, 2018a.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, 2018b.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We describe limitations in the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss the potential negative impact in the conclusion section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The data used in this work is publicly available, and we provide model configurations and training settings of our experiments to reproduce main results. We will also open source our code and model.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 4 describes training details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the mean results (WER or BLEU) averaged over multiple languages or domains in Section 4. The models for evaluation are the average version of their multiple checkpoints to reduce the result variance.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix A.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We do not create new data.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All datasets used in this work are publicly available for research purposes.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All datasets used in this work are checked and released by previous studies for research purposes.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]