

Multi-Unit Transformers for Neural Machine Translation

Jianhao Yan Fandong Meng Jie Zhou

Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China

{elliottyan,fandongmeng,withtomzhou}@tencent.com

Abstract

Transformer models (Vaswani et al., 2017) achieve remarkable success in Neural Machine Translation. Many efforts have been devoted to deepening the Transformer by stacking several units (i.e., a combination of Multi-head Attentions and FFN) in a cascade, while the investigation over multiple parallel units draws little attention. In this paper, we propose the **Multi-Unit TransformErs (MUTE)**, which aim to promote the expressiveness of the Transformer by introducing diverse and complementary units. Specifically, we use several parallel units and show that modeling with multiple units improves model performance and introduces diversity. Further, to better leverage the advantage of the multi-unit setting, we design biased module and sequential dependency that guide and encourage complementarity among different units. Experimental results on three machine translation tasks, the NIST Chinese-to-English, WMT’14 English-to-German and WMT’18 Chinese-to-English, show that the MUTE models significantly outperform the Transformer-Base, by up to +1.52, +1.90 and +1.10 BLEU points, with only a mild drop in inference speed (about 3.1%). In addition, our methods also surpass the Transformer-Big model, with only 54% of its parameters. These results demonstrate the effectiveness of the MUTE, as well as its efficiency in both the inference process and parameter usage.¹

1 Introduction

Transformer based models (Vaswani et al., 2017) have been proven to be very effective in building the state-of-the-art Neural Machine Translation (NMT) systems via neural networks and attention mechanism (Sutskever et al., 2014; Bahdanau

et al., 2014). Following the standard *Sequence-to-Sequence* architecture, Transformer models consist of two essential components, namely the encoder and decoder, which rely on stacking several identical layers, i.e., multihead attentions and position-wise feed-forward network.

Multihead attentions and position-wise feed-forward network, together as a basic unit, plays an essential role in the success of Transformer models. Some researchers (Bapna et al., 2018; Wang et al., 2019a) propose to improve the model capacity by stacking this basic unit many times, i.e., deep Transformers, and achieve promising results. Nevertheless, as an orthogonal direction, investigation over multiple parallel units draws little attention.

Compared with single unit models, multiple parallel unit layout is more expressive to capture complex information flow (Tao et al.; Meng et al., 2019; Li et al., 2018, 2019) in two aspects. First, this multiple-unit layout boosts the model by its varied feature space composition and different attentions over inputs. With this diversity, multi-unit models advance in expressiveness. Second, for the multi-unit setting, one unit could mitigate the deficiency of other units and compose a more expressive network, in a complementary way.

In this paper, we propose the **Multi-Unit TransformErs (MUTE)**, which aim to promote the expressiveness of transformer models by introducing diverse and complementary parallel units. Merely combining multiple identical units in parallel improves model capability and diversity by its varied feature compositions. Furthermore, inspired by the well-studied bagging (Breiman, 1996) and gradient boosting algorithms (Friedman, 2001) in the machine learning field, we design biased units with a sequential dependency to further boost model performance. Specifically, with the help of a module named bias module, we apply different kinds of noises to form biased inputs for corre-

¹Code is available at https://github.com/ElliottYan/Multi_Unit_Transformer

sponding units. By doing so, we explicitly establish information gaps among units and guide them to learn from each other. Moreover, to better leverage the power of *complementariness*, we introduce sequential ordering into the multi-unit setting, and force each unit to learn the residual of its preceding accumulation.

We evaluate our methods on three widely used Neural Machine Translation datasets, NIST Chinese-English, WMT’14 English-German and WMT’18 Chinese-English. Experimental results show that our multi-unit model yields an improvement of +1.52, +1.90 and +1.10 BLEU points, over the baseline model (Transformer-Base) for three tasks with different sizes, respectively. Our model even outperforms the Transformer-Big on the WMT’14 English-German by 0.7 BLEU points with only 54% of parameters. Moreover, as an interesting side effect, our model only introduces mild inference speed decrease (about 3.1%) compared with the Transformer-Base model, and is faster than the Transformer-Big model.

The contributions of this paper are threefold:

- We propose the **Multi-Unit Transformers (MUTE)**, to promote the expressiveness of Transformer models by introducing diverse and complementary parallel units.
- Aside from learning with identical units, we extend the **MUTE** by introducing bias module and sequential ordering to further model the diversity and complementariness among different units.
- Experimental results show that our models substantially surpass baseline models in three NMT datasets, NIST Chinese-English, WMT’14 English-German and WMT’18 Chinese-English. In addition, our models also show high efficiency in both the inference speed and parameter usage, compared with Transformer baselines.

2 Transformer Architecture

The Transformer Architecture (Vaswani et al., 2017) for Neural Machine Translation (NMT) generally adopts the standard encoder-decoder paradigm. In contrast to RNN architectures, the Transformer stacks several identical self-attention based layers instead of recurrent units for better parallelization. Specifically, given an input sequence $X = \{x_1, x_2, \dots, x_n\}$ in source language (e.g., English), the model is asked to predict its corre-

sponding translation $Y = \{y_1, y_2, \dots, y_m\}$ in target language (e.g., German).

Encoder. Digging into the details of the model, a Transformer encoder consists of N_e stacked layers, where each layer consists of two sub-layers, a multihead self-attention sub-layer and a position-wise feed-forward network (FFN) sub-layer.

$$s^k = \text{SelfAttn}(X^k) + X^k, \quad (1)$$

$$F^e(X^k) = s^k + \text{FFN}(s^k), \quad (2)$$

where $X^k \in \mathbb{R}^{n \times d}$ and $F^e(X^k) \in \mathbb{R}^{n \times d}$ denote the inputs and outputs of the k -th encoder layer, respectively, and d is the hidden dimension.

Decoder. The decoder follows a similar architecture, with an additional multihead cross-attention sub-layer for each of N_d decoder layers.

$$s^k = \text{SelfAttn}(Y^k) + Y^k, \quad (3)$$

$$c^k = \text{CrossAttn}(s^k, F^e(X^{N_e})) + s^k, \quad (4)$$

$$F^d(Y^k) = c^k + \text{FFN}(c^k), \quad (5)$$

where $Y^k \in \mathbb{R}^{m \times d}$ and $F^d(Y^k) \in \mathbb{R}^{m \times d}$ represent the inputs and outputs of k -th decoder layer.

Here, we omit layer norms among sub-layers for simplicity. We take the bundle of cascading sub-components (i.e., attention modules and FFN) as a unit, and refer to the original Transformer and its variants with such cascade units as *Single-Unit Transformer*. For ease of reading, we refer to a single unit of encoder and decoder as F^e and F^d in the following sections.

Then, the probability $P(Y|X)$ is produced with another Softmax layer on top of decoder outputs,

$$P(Y|X) = \text{Softmax}(W_s \cdot F^d(Y^{N_d}) + b), \quad (6)$$

where $W_s \in \mathbb{R}^{d \times |V|}$ and $b \in \mathbb{R}^{|V|}$ are learnable parameters, and $|V|$ denotes the size of target vocabulary. Then, a cross-entropy objective is computed by,

$$\mathcal{L}_{CE} = \sum_{t \in (1, m)} Y_t \log P(Y_t|X), \quad (7)$$

where t represents the t -th step for decoding phase.

3 Model Layout

3.1 MUTE

In this section, we briefly describe the Multi-Unit Transformer (MUTE), shown in Figure 1(a). As mentioned before, we take the bundle of cascading sub-components (i.e., attention modules and FFN) as a single unit. By combining units in parallel, we have a basic Multi-Unit Transformer (MUTE) layer.

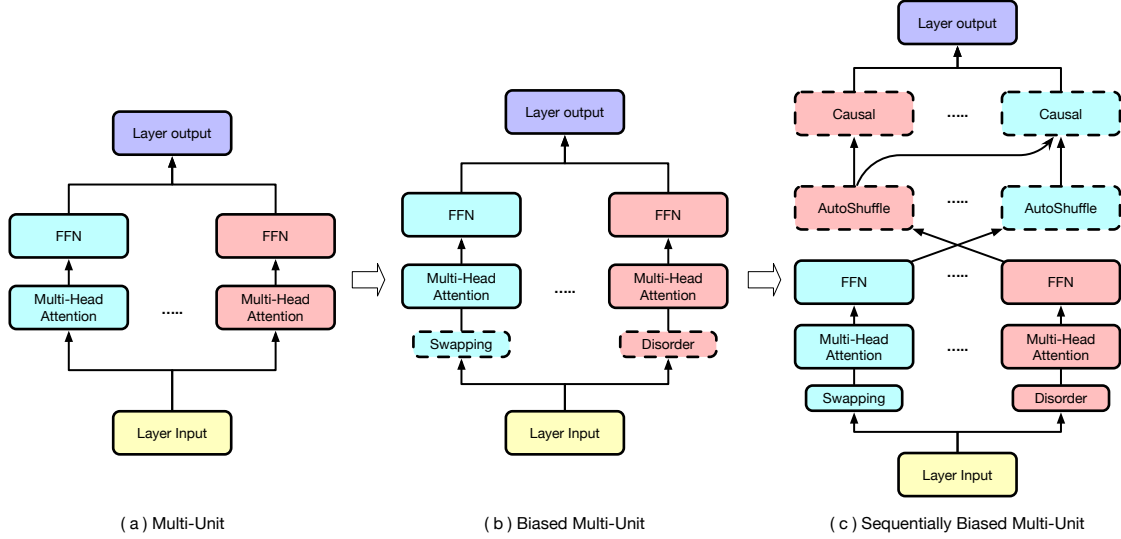


Figure 1: Layer architecture for Multi-Unit Transformer. White arrow indicates model change from Multi-Unit to Biased Multi-Unit, to Sequentially Biased Multi-Unit, with dashed lines representing the newly added modules.

Then, we follow the standard usage by stacking several MUTE layers to constitute our encoder and decoder.

In general, the encoder and decoder of the Transformer network share a similar architecture and can be improved with the same techniques. Without losing generality, we take the encoder as an example to further illustrate the MUTE. Given input X^k of k -th layer, we feed it into I identical units $\{F_1, \dots, F_i, \dots, F_I\}$ with different learnable parameters.

$$s_i^k = \text{SelfAttn}_i(X^k) + X^k, \quad (8)$$

$$F_i^e(X^k) = s_i^k + \text{FFN}_i(s_i^k), \quad (9)$$

where i denotes the i -th unit. After collecting outputs for all I units, we combine them by a weighted sum,

$$F^e(X^k) = \sum_{i \in (1, I)} \alpha_i \cdot F_i^e(X^k), \quad (10)$$

where $\alpha_i \in \mathbb{R}^1$ represents the learnable weight for the i -th unit (Section 5.8) and $F^e(X^k) \in \mathbb{R}^{n \times d}$ is the final output for the k -th layer.

3.2 Biased MUTE

The multi-unit setting for Transformer resembles the well-known ensemble techniques in machine learning fields, in that it also combines several different modules into one and aims for better performance. In that perspective, borrowed from the idea of bagging (Breiman, 1996), we propose to use biased units instead of identical units, which results in creating information gaps among units and makes them learn from each other.

More specifically, in training, we introduce a *Bias-Module* to create biased units, as shown in Figure 1(b). For each layer, instead of giving the same inputs $X^k \in \mathbb{R}^{n \times d}$ to all units, we transform each input with corresponding type of noises (e.g., swap, reorder, mask), in order to force the model to focus on different parts of inputs:

$$X_i^k = \text{Bias}_i(X^k), \quad (11)$$

$$F^e(X^k) = \sum_{i \in (1, I)} \alpha_i \cdot F_i^e(X_i^k), \quad (12)$$

where Bias_i denotes the noise function for i -th unit. The noise operations² we investigated include,

- **Swapping**, randomly swap two input embeddings up to a certain range (i.e., 3).
- **Disorder**, randomly permute a subsequence within a certain length (i.e., 3).
- **Masking**, randomly replace one input embedding with a learnable mask embedding.

Note that, the identity mapping (i.e., no noise) can be seen as a special case of bias module and is included in our model design.

Additionally, to get deterministic outputs, we disable the noises in the testing phase, which brings in the inconsistency between training and testing. Hence, we propose a switch mechanism with a sample rate p_β that determines whether to enable the bias module in training. This mechanism forces the model to adapt to golden inputs and mitigate the aforementioned inconsistency.

²To avoid a distortion of input sequence, each operation is performed only once for corresponding bias module.

3.3 Sequentially Biased MUTE

Although the bias module guides units of learning from each other by formulating such information gaps, it still lacks explicit complementarity modeling, i.e., mitigating these gaps. Here, based on the Biased MUTE, we propose a novel method to explicitly introducing a deep connection among units by utilizing the power of order (Figure 1(c)).

Sequential Dependency. Given the outputs from biased units $F_i^e(X_i^k)$, we permute these outputs by a certain ordering function $p(i)$ (e.g., $i \in \{1, 2, 3, 4\}$ to $p(i) \in \{4, 2, 3, 1\}$),

$$\{G_i^e = F_{p(i)}^e(X_{p(i)}^k) | i \in (1, I)\}, \quad (13)$$

where G_i^e is the i -th permuted output. The implementation of $p(i)$ will be illustrated later.

Then, we explicitly model the complementarity among units by introducing sequential dependency. Specifically, we compute an accumulated sequence $\{\hat{G}_i^e | i \in (1, I)\}$ over the permuted outputs G_i^e ,

$$\hat{G}_i^e = \hat{G}_{i-1}^e + G_i^e, \quad (14)$$

where $\hat{G}_i^e \in \mathbb{R}^{n \times d}$ is the i -th *accumulated* output, and $\hat{G}_0^e = 0$. Through this sequential dependency, each permuted output G_i^e learns the residual of previous *accumulated* outputs \hat{G}_{i-1}^e (He et al., 2016) and serves as a complement to previous accumulated outputs.

Finally, we normalize this accumulated sequence to keep the output norm stable and fuse all *accumulated* outputs.

$$F^e(X^k) = \sum_{i \in (1, I)} \alpha_i \cdot \frac{\hat{G}_i^e}{i}. \quad (15)$$

Autoshuffle. Until now, we have modeled the sequential dependency between each of the units. The only problem left is how to gather a proper ordering of units. We propose to use the AutoShuffle Network (Lyu et al., 2019). Mathematically, shuffling with specific order equals to a multiplication by a permutation matrix (i.e., every row and column contains precisely a single 1 with 0s elsewhere). Nevertheless, a permutation matrix only contains discrete values and can not be optimized by gradient descent. Therefore, we use a continuous matrix $M \in \mathbb{R}^{I \times I}$ with non-negative values $M_{i,j} \in (0, 1)$, $i, j \in (1, I)$ to approximate the discrete permutation matrix.

Particularly, M is regarded as a learnable matrix

and is used to multiply the outputs of units,

$$[\cdots; F_{p(i)}^e(X_{p(i)}^k); \cdots] = M^T \times [\cdots; F_i^e(X_i^k); \cdots], \quad (16)$$

where $[\cdot; \cdot]$ means the concatenation operation.

To ensure M remains an approximation for the permutation matrix during training, we normalize M after each optimization step.

$$M_{i,j} = \max(M_{i,j}, 0), \quad (17)$$

$$M_{i,j} = \frac{M_{i,j}}{\sum_{\hat{i}} M_{\hat{i},j}}, M_{i,j} = \frac{M_{i,j}}{\sum_{\hat{j}} M_{i,\hat{j}}}. \quad (18)$$

Then, we introduce a Lipschitz continuous non-convex penalty, as proposed in Lyu et al. (2019) to guarantee M converge to a permutation matrix.

$$\begin{aligned} \mathcal{L}_p = & \sum_{i=1}^I [\sum_{j=1}^I |M_{i,j}| - (\sum_{j=1}^I M_{i,j}^2)^{\frac{1}{2}}] \\ & + \sum_{j=1}^I [\sum_{i=1}^I |M_{i,j}| - (\sum_{i=1}^I M_{i,j}^2)^{\frac{1}{2}}], \end{aligned} \quad (19)$$

Please refer to Appendix D for proof and other details.

Finally, the penalty is added to cross-entropy loss defined in equation (7) as our final objective,

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \sum_k \mathcal{L}_p^k, \quad (20)$$

where λ is a hyperparameter to balance two objectives and \mathcal{L}_p^k is the penalty for the k -th layer.

4 Experimental Settings

In this section, we elaborate our experimental setup on three widely-studied Neural Machine Translation tasks, NIST Chinese-English (Zh-En), WMT' 14 English-German (En-De) and WMT' 18 Chinese-English.

Datasets. For NIST Zh-En task, we use 1.25M sentences extracted from LDC corpora³. To validate the performance of our model, we use the NIST 2006 (MT06) test set with 1664 sentences as our validation set. Then, the NIST 2002 (MT02), 2003 (MT03), 2004 (MT04), 2008 (MT08) test sets are used as our test sets, which contain 878, 919, 1788 and 1357 sentences, respectively.

For the WMT' 14 En-De task, following the same setting in Vaswani et al. (2017), we use 4.5M pre-processed data, which has been tokenized and split using byte pair encoded (BPE) (Sennrich et al.,

³The corpora include LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

System	#Param.	Valid	MT02	MT03	MT04	MT08	Δ
<i>Existing NMT systems</i>							
Wang et al. (2018) (Base)	95M	45.47	46.31	45.30	46.45	35.66	–
Cheng et al. (2018) (Base)	95M	45.78	45.96	45.51	46.49	36.08	–
Cheng et al. (2019) (Base)	95M	46.95	47.06	46.48	47.39	37.38	–
<i>Baseline NMT systems</i>							
Transformer (Base)	95M	45.79	46.34	45.32	47.92	36.16	<i>ref</i>
Transformer + Relative (Base)	95M	46.57	46.64	45.67	47.26	38.03	+0.53
Transformer + Relative (Big)	277M	46.52	47.23	46.43	48.35	37.31	+0.86
<i>Our NMT systems</i>							
MUTE (Base, 4 Units)	152M	47.23 [†]	47.38 [†]	46.24 [†]	47.81 [†]	38.48	+1.12
+ Bias (Base, 4 Units)	152M	47.45 ^{††}	47.40 ^{††}	47.11^{††}	48.44^{††}	38.31	+1.44
+ Bias + Seq. (Base, 4 Units)	152M	47.80^{††}	47.72^{††}	46.60 ^{††}	48.30 ^{††}	38.70	+1.52

Table 1: Case-insensitive BLEU scores (%) of NIST Chinese-English (Zh-En) task. For all models with **MUTE**, we use four units. **#Params.** means the number of learnable parameters in the model. Δ denotes the average BLEU improvement over dev set and test sets, compared with the “Transformer (Base)”. **Bold** represents the best performance. “[†]”: significantly better than “Transformer + Relative (Base)” ($p < 0.05$); “^{††}”: significantly better than “Transformer + Relative (Base)” ($p < 0.01$).

2016) with 32k merge operations and a shared vocabulary for English and German. We use *newstest2013* as our validation set and *newstest2014* as our test set, which contain 3000 and 3003 sentences, respectively.

For the WMT’18 Zh-En task, we use 18.4M preprocessed data, which is also tokenized and split using byte pair encoded (BPE) (Sennrich et al., 2016). We use *newstest2017* as our validation set and *newstest2018* as our test set, which contains 2001 and 3981 sentences, respectively.

Evaluation. For evaluation, we train all the models with maximum 150k/300k/300k steps for NIST Zh-En, WMT En-De and WMT Zh-En, respectively, and we select the model which performs the best on the validation set and report its performance on the test sets. We measure the case-insensitive/case-sensitive BLEU scores using *multi-bleu.perl*⁴ with the statistical significance test (Koehn, 2004)⁵ for NIST Zh-En and WMT’14 En-De, respectively. For WMT’18 Zh-En, we use case sensitive BLEU scores calculated by Moses *mteval-v13a.pl* script⁶.

⁴<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

⁵<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/analysis/bootsrap-hypothesis-difference-significance.pl>

⁶<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl>

Model and Hyper-parameters. For all our experiments, we basically follow two model settings illustrated in (Vaswani et al., 2017), namely *Transformer-Base* and *Transformer-Big*. In *Transformer-Base*, we use 512 as hidden size, 2048 as filter size and 8 heads in multihead attention. In *Transformer-Big*, we use 1024 as hidden size, 4096 as filter size, and 16 heads in multihead attention.

Besides, since noise types like swapping and reordering are of no effect on Transformer models with absolute position information, the MUTE models are implemented using relative position information (Shaw et al., 2018). In addition, we only apply multi-unit methods to encoders, provided that the encoder is more crucial to model performance (Wang et al., 2019a). All experiments on MUTE models are conducted with *Transformer-Base* setting. For the basic MUTE model, we use four identity units. As for the Biased MUTE and Sequentially Biased MUTE, we use four units including one identity unit, one swapping unit, one disorder unit and one masking unit. The sample rate p_β is set to 0.85. For more implementation details and experiments on sample rate, please refer to Appendix A and B.

5 Results

Through experiments, we first evaluate our model performance (Section 5.1 and 5.2). Then, we analyze how each part of our model works (Section 5.4 and 5.6). Finally, we conduct experiments to further understand the behavior of our models (Sec-

System	#Param.	BLEU	Δ
<i>Existing NMT systems</i>			
Vaswani et al. (2017) (Base)	81M	27.3	ref.
Bapna et al. (2018) (Base, 16L)	137M	28.0	-
Wang et al. (2019a) (Base, 30L)	137M	29.3	-
Vaswani et al. (2017) (Big)	251M	28.4	-
Chen et al. (2018) (Big)	379M	28.5	-
He et al. (2018) (Big)	*251M	29.0	-
Shaw et al. (2018) (Big)	*251M	29.2	-
Ott et al. (2018) (Big)	251M	29.3	-
<i>Existing Multi-Unit Style NMT Systems</i>			
Ahmed et al. (2017) (Base)	*81M	28.4	-
Li et al. (2019) (Base)	118M	28.5	-
Li et al. (2018) (Base)	*81M	28.5	-
<i>Baseline NMT systems</i>			
Transformer (Base)	81M	27.4	ref.
Transformer + Relative (Base)	81M	28.2	+0.8
Transformer + Relative (Big)	251M	28.6	+1.2
<i>Our NMT systems</i>			
MUTE (Base, 4 Units)	130M	28.8 ^{††}	+1.4
+ Bias (Base, 4 Units)	130M	29.1 ^{††}	+1.7
+ Bias + Seq. (Base, 4 Units)	130M	29.3^{††}	+1.9

Table 2: Case-sensitive BLEU scores (%) of WMT’14 English-German (En-De) task. Δ denotes the improvement over *newstest2014*, compared with Transformer (Base). * denotes an estimated value. “††”: significantly better than “Transformer + Relative (Base)” ($p < 0.01$).

tion 5.7 and 5.8).

5.1 Results on NIST Chinses-English

As shown in Table 1, we list the performance of our re-implemented Transformer baselines and our approaches. We also list several existing strong NMT systems reported in previous work to validate the effectiveness of our models. By investigating results in Table 1, we have the following observations.

First, compared with existing NMT systems, our re-implemented Transformers are strong baselines.

Second, all of our approaches substantially outperform our baselines, with improvement ranging from 1.12 to 1.52 BLEU points. Comparing our methods to “Transformer + Relative (Base)”, our best approach (i.e., “MUTE + Bias + Seq. (Base, 4 Units)”) still achieves a significant improvement of 1.0 BLEU points on multiple test sets.

Third, among our approaches, we find that, even though our basic MUTE model has already surpassed existing strong NMT systems and our baselines, the bias module and sequential dependency

can further boost the performance (i.e., from +1.12 to +1.52), which demonstrates that introducing complementariness does help the Multi-Unit Transformers.

Fourth, we find it interesting that compared with the “Transformer + Relative (Big)”, the basic “MUTE (Base, 4 Units)” can achieve better BLEU performance with only 54% of parameters, which indicates that our multi-unit approaches can leverage parameters more effectively and efficiently.

5.2 Results on WMT’14 English-German

The results on WMT’14 En-De are shown in Table 2. We list several competitive NMT systems for comparison, which are divided into models based on *Transformer-Base* and models based on *Transformer-Big*.

First of all, our models show significant BLEU improvements over two baselines in the *Transformer-Base* setting, ranging from +1.4 to +1.9 for “Transformer (Base)” and from +0.6 to +1.1 for “Transformer+Relative (Base)”. That proves our methods perform consistently across languages and are still useful in large scale datasets.

Next, among our own NMT methods, the sequentially biased model further improves the BLEU performance over our strong Multi-Unit model (from 28.8 to 29.3), which is consistent with our findings in the Zh-En⁷ task and further proves the power of complementariness.

Finally, compared with the existing NMT systems, we find that our models achieve comparable / better performance with much fewer parameters. The only exception is (Wang et al., 2019a), which learns a very deep (30 layers) Transformer. We regard these deep Transformer methods as orthogonal methods to ours, and it can be integrated with our MUTE models in future work. Additionally, we list several systems related to our multi-unit setting (“*Existing Multi-Unit Style NMT Systems*”), with diversity modeling or features space composition. As shown, MUTE models also outperform these methods, demonstrating the superiority of our methods in diverse and complementary modeling.

5.3 Results on WMT’18 Chinese-English

In this section, we represent our results on WMT18 Chinese-English. The results are shown in Table

⁷We find that improvements on the larger scale WMT’14 En-De dataset is bigger than that on NIST Zh-En. We attribute this phenomenon to the overfitting problem caused by the small Zh-En dataset.

System	#Param.	BLEU	Δ
Transformer (Base)	81M	23.4	ref.
Transformer + Relative (Base)	81M	23.8	+0.4
MUTE + Bias + Seq.	130M	24.5	+1.1

Table 3: Case-sensitive BLEU scores (%) of WMT’18 Chinese-English (Zh-En) task.

Methods	BLEU Score
MUTE + Bias + Seq.	47.80
- identity unit.	47.01 (-0.79)
- swapping unit.	47.13 (-0.67)
- disorder unit.	47.15 (-0.65)
- mask unit.	47.36 (-0.44)
- bias module	47.28 (-0.52)
- sequential dependency	47.45 (-0.35)

Table 4: Ablation study for BLEU scores (%) over the NIST Zh-En validation set.

3. As we can see, our MUTE model still strongly outperforms the baseline “Transformer (Base)” and “Transformer+Relative (Base)” with +1.1 and +0.7 BLEU points. Noting that WMT’18 Zh-En has a much larger dataset (18.4M), and these findings proves that our model perform consistently well with different size of datasets.

5.4 Ablation Study

In this section, we conduct the ablation study to verify each part of our proposed model. The results are shown in Table 4. From our strongest Sequentially Biased model, we remove each of the four different units to validate which unit contributes the most to the performance. Then, we remove the bias module and sequential dependency independently to investigate each module’s behavior.

We come to the following conclusions:

(1) All units make substantial contributions to “MUTE + Bias + Seq.”, ranging from 0.44 to 0.79, proving the effectiveness of our design.

(2) Among all units, the identity unit contributes most to our performance, which is consistent with our intuition that the identity unit should be responsible most for complementing other biased units.

(3) The bias module and sequential dependency both contribute much to our Multi-Unit Transformers. We find it intriguing that, without the bias module, sequential dependency only provides marginal improvements. We conjecture that the complementary effect becomes minimal with no information gap among units.

Methods	BLEU Score
Transformer + Relative	44.83
+ average last 5 saves	44.97
+ average 5 seeds	<i>Fails</i>
MUTE	45.43
MUTE + Bias + Seq.	45.82

Table 5: Comparison with averaging checkpoints in terms of BLEU scores (%) over the NIST Zh-En datasets. The reported BLEU scores is the average of all test sets.

5.5 Comparison with Averaging Checkpoints

As we mentioned before, our MUTE models are inspired by ensembling methods. Therefore, it is necessary to compare our model with representative ensemble methods, e.g., averaging model checkpoints. Here, the comparison results of our MUTE model and averaging checkpoints in NIST Zh-En are shown in Table 5. Since averaging checkpoints often leads to better generalization, we report the average BLEU scores over all test sets.

We adopt two settings, namely averaging the last several (i.e., 5) checkpoints and averaging over models initialized with different seeds. The experiment with different initialization seeds fails. We conjecture the reason is that different seeds make models fall in different sub-optimal, and brutally combining them together makes the model perform badly. Then we average checkpoints over the last 5 saves, which gives us 44.97 BLEU points, which only outperforms the best checkpoint marginally (+0.14 in average), and MUTE performs much better (45.43 and 45.82 BLEU points). Specifically, our naive MUTE model surpasses the averaging checkpoint method, and the sequential ordering and bias module enable a better interaction over different units.

5.6 Quantitative Analysis of Model Diversity

Here, we empirically investigate which granularity should be used for better diversity among units. To verify the impact of multiple units compared with the single unit, we evaluate three different models:

- **4 Self. + 4 FFN**, the model with four different self-attention modules and four different FFNs.
- **4 Self. + 1 FFN**, the model with four self-attention modules and one shared FFN.
- **1 Self. + 4 FFN**, the model with one shared self-attention module and four different FFNs.

Methods	Att Pos.	Att Sub.	FFN Sub.
MUTE + Bias + Seq.	0.475	0.602	0.460
4 Self. + 4 FFN	0.447	0.576	0.449
4 Self. + 1 FFN	0.420	0.520	0.404
1 Self. + 4 FFN	-	-	0.381

Table 6: Quantitative analysis on diversity scores. **Att Pos.**, **Att Sub.** and **FFN Sub.** denote the diversity scores for self-attention weights, self-attention outputs and FFN outputs, respectively. A larger score means better diversity.

To control variables, these models include neither bias module nor sequential dependency.

For each model, we evaluate the diversity among units for three outputs: (1) the outputs of self-attention modules, (2) the attention weights of self-attention modules, (3) the outputs of FFN modules. The diversity scores are computed by the exponential of the negative cosine distance among units, the same as proposed in (Li et al., 2018).

$$\text{DIV} = \exp\left(-\frac{o_i^\top o_j}{|o_i| \cdot |o_j|}\right), \quad (21)$$

where $\text{DIV} \in (0, 1)$ represents the diversity score for module outputs $o_i \in \mathbb{R}^d$ and $o_j \in \mathbb{R}^d$. The results are shown in Table 6.

Above all, we find that multiple FFN layers can introduce diversity. As seen, “1 Self. + 4 FFN” produces a 0.381 diversity score on “FFN Sub.”. Since we use a shared self-attention layer, which brings no diversity in the input-side of FFN layers, the difference is only brought by different FFN modules. We think the reason is that the RELU activation inside the FFN module serves as selective attention to filter out input information.

Next, “4 Self. + 1 FFN” achieves 0.420 and 0.520 diversity scores for self-attention modules, which indicates that multiple self-attention modules focus on different parts of the input sequence and lead to diverse outputs.

Then, “4 Self. + 4 FFN” has higher diversity scores than “4 Self. + 1 FFN” and “1 Self. + 4 FFN”, which verifies our choice of using a combination of self-attention module and FFN as a basic unit.

Finally, concerning the diversity scores among all models, we find that our full model with biased inputs and sequential dependency achieves the best diversity scores for all three outputs, which also achieves the best BLEU scores in previous experi-

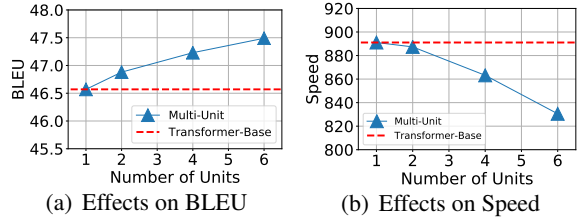


Figure 2: The effect on the number of units: (a) BLEU scores (%) against number of units used in the MUTE model. (b) Inference speed (tokens per second) on GPU against number of units used in the MUTE model. Experiments are conducted on the valid set of NIST Zh-En with decoding batch size = 30 and beam size = 4.

5.7 Effects on the Number of Units

Another concern is how the MUTE models perform when increasing the number of units. Thus, in this section, we empirically investigate the effects on the number of units, in terms of model performance and inference speed. Here we use the basic Multi-Unit Transformer⁸.

As shown in Figure 2(a) and 2(b), increasing the number of units from 1 to 6 yields consistent BLEU improvement (from 46.5 to 47.5) with only mild inference speed decrease (from 890 tokens/sec to 830 tokens/sec). Besides, our model with four unit used in other experiments is faster than Transformer-Big (863 tokens/sec vs 838 tokens/sec). These results prove the computational efficiency of our MUTE model. We attribute this mild speed decrease (about 3.1% for four units and 6.7% for six units) for two reasons. First, the multi-unit model is naturally easy for parallelization. Each unit can be computed independently without waiting for other functions to finish. Second, we only widen the encoder, which is only computed once for each sentence translation.

5.8 Visualization

We also present a visualization example of the learnable weights α for units, shown in Figure 3.

As we can see, learnable weights α show similar trends in “MUTE” and “MUTE + Bias”. The weights for each unit within the same layer fall in a similar range (0.2 to 0.35), dispelling the worries that the biased units may be omitted or skipped. As for the “MUTE + Bias + Seq.”, the weight distribution is very different. The weights nearly increase progressively when the unit index increases.

⁸The number of combinations of different noise types is exponentially large. Moreover, introducing sequential dependency and bias module hardly affects the inference speed.

Because the model weights are learned with back-propagation, the larger the weight, the more the model favors the corresponding unit. Thus, to some extent, this phenomenon demonstrates that the latter accumulated outputs are more potent than the preceding ones, and therefore, the latter single unit output complements the previous accumulation.

6 Related Work

Recently Transformer-based models (Vaswani et al., 2017; Ott et al., 2018; Wang et al., 2019a) become the de facto methods in Neural Machine Translation, owing to high parallelism and large model capacity.

Some researchers devise new modules to improve the Transformer model, including combining the transformer unit with convolution networks (Wu et al., 2019; Zhao et al., 2019; Lioutas and Guo, 2020), improving the self-attention architecture (Fonollosa et al., 2019; Wang et al., 2019b; Hao et al., 2019), and deepening the Transformer architecture by dense connections (Wang et al., 2019a). Since our multi-unit framework makes no limitation about its unit, these models can be easily integrated into our multi-unit framework.

There are also some works utilizing the power of multiple modules to capture complex feature representations in NMT. Shazeer et al. (2017) use a vast network and a sparse gated function to select from multiple experts (i.e., MLPs). Ahmed et al. (2017) train a weighted Transformer by replacing the multi-head attention by self-attention branches. Nevertheless, these models ignore the modeling of relations among different modules. Then, some multihead attention variants (Li et al., 2018, 2019) introduce modeling of diversity or interaction among heads. However, complementariness is not taken into account in their approaches. Our MUTE models differ from their methods in two aspects. First, we use a powerful unit with a strong performance in diversity (Section 5.6). Second, we explicitly model the complementariness with bias module and sequential dependency.

7 Conclusion

In this paper, we propose Multi-Unit Transformers for NMT to improve the expressiveness by introducing diverse and complementary units. In addition, we propose two novel techniques, namely bias module and sequential dependency to further improve the diversity and complementariness among

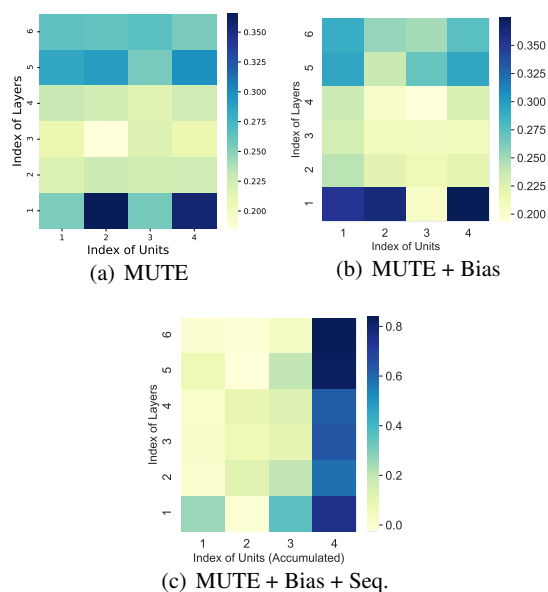


Figure 3: Visualization of our models’ learnable weights for units in each layer.

units. Experimental results show that our methods can significantly outperform the baseline methods and achieve comparable / better performance compared with existing strong NMT systems. In the meantime, our methods use much fewer parameters and only introduce mild inference speed degradation, which proves the efficiency of our models.

References

- Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. 2017. Weighted transformer network for machine translation. *arXiv preprint arXiv:1711.02132*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Ankur Bapna, Mia Xu Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033.
- Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.
- Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

- Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.
- Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766.
- José AR Fonollosa, Noe Casas, and Marta R Costa-jussà. 2019. Joint source-target self attention with locality constraints. *arXiv preprint arXiv:1905.06596*.
- Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. Multi-granularity self-attention for neural machine translation. *arXiv preprint arXiv:1909.02222*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 7944–7954.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. **OpenNMT: Open-source toolkit for neural machine translation**. In *Proc. ACL*.
- Philipp Koehn. 2004. **Statistical significance tests for machine translation evaluation**. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2897–2903.
- Jian Li, Xing Wang, Baosong Yang, Shuming Shi, Michael R Lyu, and Zhaopeng Tu. 2019. Neuron interaction based representation composition for neural machine translation. *arXiv preprint arXiv:1911.09877*.
- Vasileios Lioutas and Yuhong Guo. 2020. Time-aware large kernel convolutions. *arXiv preprint arXiv:2002.03184*.
- Jiancheng Lyu, Shuai Zhang, Yingyong Qi, and Jack Xin. 2019. Autosufflenet: Learning permutation matrices via an exact lipschitz continuous penalty in deep convolutional neural networks. *arXiv preprint arXiv:1901.08624*.
- Fandong Meng, Jinchao Zhang, Yang Liu, and Jie Zhou. 2019. Multi-zone unit for recurrent neural networks. *arXiv preprint arXiv:1911.07184*.
- Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. **Neural machine translation of rare words with subword units**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019a. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*.
- Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2019b. Self-attention with structural position representations. *arXiv preprint arXiv:1909.00383*.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo. 2019. Muse: Parallel multi-scale attention for sequence to sequence learning. *arXiv preprint arXiv:1911.09483*.

A Implementation Details

All models are trained using *Opennmt-py* framework (Klein et al., 2017). The batch size for each GPU is set to 4096 tokens. The beam size is set to 4, and the length penalty is 0.6 among all experiments. For the WMT’14 En-De and WMT’18 Zh-En task, all experiments are conducted using 4 NVIDIA Tesla V100 GPUs, while we use 2 GPUs for the NIST Zh-En task. That gives us about 8k/16k/16k tokens per update for NIST Zh-En/WMT’14 En-De/WMT’18 Zh-En. All models are optimized using Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, and learning rate is set to 1.0 for all experiments. Label smoothing is set to 0.1 for all three tasks. We use dropout of 0.4/0.2/0.1 for NIST Zh-en/WMT En-de/WMT Zh-en, respectively. All our Transformer models contain 6 encoder layers and 6 decoder layers, following the standard setting in (Vaswani et al., 2017).

All the hyperparameters are empirically set by our previous experiences or manually tuned. The criterion for selecting hyperparameters is the BLEU score on validation sets for both tasks. The average runtimes are one GPU day for NIST Zh-En and 4 GPU days for WMT’14 En-De and WMT’18 Zh-EN.

B Effect of Sample Rate

As we mentioned above, we use sample rate to mitigate the inconsistency between training and testing, and force the model to adapt to golden inputs. Here, we conduct hyper-parameter search for sample rate with 4 empirically set values, 0.5, 0.75, 0.85, 1.0. The results are shown in Figure 4.

As seen, we make the following observations: 1. The trends of BLEU scores for sequentially biased model and biased model are very similar, when we increases the sample rate p_β . The best performance for both models appear when the sample rate is 0.85. Therefore, we set sample rate to 0.85 among all of our experiments. 2. Introducing sample rate is essential for both biased model and sequentially biased model. Compared with always injecting noises, i.e., the sample rate $p_\beta = 1.0$, consistent improvements is observed for both models. BLEU scores increase from 47.27 to 47.48 (+0.21) for biased model, and from 47.56 to 47.80 (+0.24) for sequentially biased model.

The aforementioned observations prove that introducing sample rate when using bias module is an essential way to mitigate the inconsistency between

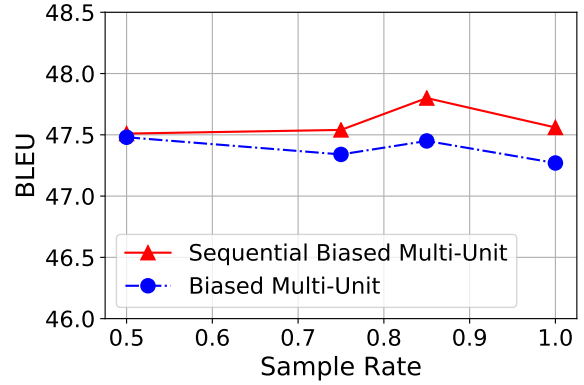


Figure 4: BLEU scores (%) changes on NIST Zh-En valid set when increasing the sample rate p_β . Bigger value means a better chance to inject noises. 1.0 for sample rate means always injecting noises.

training and testing.

C Case Study

We also provide a visualization example to show how our proposed methods improve complementarity among units. As shown in Figure 5, the top one is the attention weights of Basic Multi-Unit model. The attention weights for each of the units mainly focus on same parts of the inputs. In contrast to this phenomenon, the biased model and sequentially biased model attend with more diverse weights. Moreover, we find that the sequentially biased model pays more attention to what the previous unit miss, which proves that our proposed method can actually encourage complementarity.

D Simple Proof for Lipschitz Penalty

Here we give a intuitive proof why the Lipschitz penalty (Lyu et al., 2019) would lead our learnable matrix M towards a permutation matrix, i.e., every row and column contains precisely a single 1 with 0s elsewhere.

Recall the penalty for matrix M ,

$$\mathcal{L}_p = \sum_{i=1}^I \left[\sum_{j=1}^I |M_{i,j}| - \left(\sum_{j=1}^I M_{i,j}^2 \right)^{\frac{1}{2}} \right] + \sum_{j=1}^I \left[\sum_{i=1}^I |M_{i,j}| - \left(\sum_{i=1}^I M_{i,j}^2 \right)^{\frac{1}{2}} \right]. \quad (22)$$

By the Cauchy-Schwarz inequality, we have

$$\left(\sum_{j=1}^I |M_{i,j}| \right) - \left(\sum_{j=1}^I M_{i,j}^2 \right)^{\frac{1}{2}} \geq 0, \forall i, \quad (23)$$

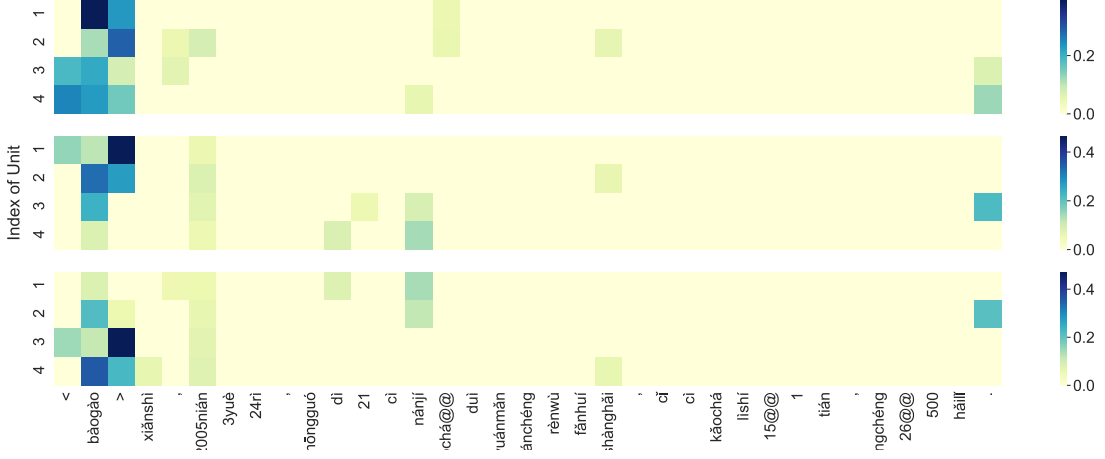


Figure 5: A visualization of example attention weights of the *Multi-Unit*, *Biased Multi-Unit* and *Sequentially Biased Multi-Unit* models, top to bottom. We filter out the weights smaller than 0.05 for ease of understanding.

with the equality holds if and only if there exists maximum one 1 for each row,

$$|\{j : M_{i,j} \neq 0\}| \leq 1, \forall i. \quad (24)$$

Then, in conjunction with our normalization that preserves,

$$M_{i,j} \geq 0, \forall i, j; \quad (25)$$

$$\sum_{j=1}^I M_{i,j} = 1, \forall i; \sum_{i=1}^I M_{i,j} = 1, \forall j, \quad (26)$$

the equality only holds if and only if there exists only one 1 for each row,

$$|\{j : M_{i,j} \neq 0\}| = 1, \forall i. \quad (27)$$

Likewise,

$$\left(\sum_{i=1}^I |M_{i,j}|\right) - \left(\sum_{i=1}^I M_{i,j}^2\right)^{\frac{1}{2}} \geq 0, \forall j, \quad (28)$$

with equality if and only if $|\{j : M_{i,j} \neq 0\}| = 1, \forall i$. Therefore, when the penalty (22) becomes 0, M converges to a permutation matrix. For more detailed mathematical proof about the Lipschitz penalty, we refer readers to [Lyu et al. \(2019\)](#).