Multilingual Mix: Example Interpolation Improves Multilingual Neural Machine Translation

Yong Cheng, Ankur Bapna, Orhan Firat, Yuan Cao,
Pidong Wang, and **Wolfgang Macherey**
Google Research
{chengyong,ankurbpn,orhanf,yuancao,pidong,wmach}@google.com

Abstract

Multilingual neural machine translation models are trained to maximize the likelihood of a mix of examples drawn from multiple language pairs. The dominant inductive bias applied to these models is a shared vocabulary and a shared set of parameters across languages; the inputs and labels corresponding to examples drawn from different language pairs might still reside in distinct subspaces. In this paper, we introduce multilingual crossover encoder-decoder (*mXEncDec*) to fuse language pairs at an instance level. Our approach interpolates instances from different language pairs into joint 'crossover examples' in order to encourage sharing input and output spaces across languages. To ensure better fusion of examples in multilingual settings, we propose several techniques to improve example interpolation across dissimilar languages under heavy data imbalance. Experiments on a large-scale WMT multilingual dataset demonstrate that our approach significantly improves quality on English-to-Many, Many-to-English and zero-shot translation tasks (from $+0.5$ BLEU up to $+5.5$ BLEU points). Results on code-switching sets demonstrate the capability of our approach to improve model generalization to out-of-distribution multilingual examples. We also conduct qualitative and quantitative representation comparisons to analyze the advantages of our approach at the representation level.

1 Introduction

Multilingual modeling has been receiving increasing research attention over the past few years, arising from successful demonstrations of improved quality across a variety of tasks, languages and modalities (Lample and Conneau, 2019; Arivazhagan et al., 2019b; Conneau et al., 2021). The success of these models is typically ascribed to vocabulary sharing, parameter tying and implicit pivoting through dominant languages like English (Conneau et al., 2020). These conventional techniques are effective, but might not be exploiting the full potential of multilingual models to learn the underlying inductive bias: *the learning signal from one language should benefit the quality of other languages* (Caruana, 1997; Arivazhagan et al., 2019b).

Here we study two related issues that exist in the context of multilingual Neural Machine Translation (NMT) training (Dong et al., 2015; Firat et al., 2016a; Johnson et al., 2017). First, NMT models (Bahdanau et al., 2015; Vaswani et al., 2017) are trained with maximum likelihood estimation which has a strong tendency to overfit and even memorize observed training examples, particularly posing challenges for low resource languages (Zhang et al., 2018). Second, training examples from distinct language pairs are separately fed into multilingual NMT models without any explicit instance-level sharing (with the exception of multi-source NMT (Zoph and Knight, 2016; Firat et al., 2016b)); as a consequence, given large enough capacity, the models have the liberty to map representations of different languages into distinct subspaces, limiting the extent of cross-lingual transfer.

In this work, we introduce multilingual crossover encoder-decoder (*mXEncDec*) to address these issues following the recent work on *XEncDec* (Cheng et al., 2021) and *mixup* (Zhang et al., 2018; Cheng et al., 2020; Guo et al., 2020). Inspired by chromosomal crossovers (Rieger et al., 2012), *mXEncDec* fuses two multilingual training examples to generate crossover examples inheriting the combinations of traits of different language pairs, which is capable of explicitly capturing cross-lingual signals compared to the standard training which mechanically combines multiple language pairs. *mXEncDec* has the following advantages:

1. *Enhancing the cross-lingual generalization.* Thanks to crossover examples generated by fusing different language pairs, the multilingual NMT is encouraged to learn to transfer

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4092 - 4102
May 22-27, 2022 ©2022 Association for Computational Linguistics

explicitly via more languages rather than implicitly via the predominant languages.

2. *Improving the model generalization and robustness.* As vicinity examples around each example in the multilingual corpus (akin to Vicinal Risk Minimization (Chapelle et al., 2001)), crossover examples produced by *mXEncDec* can enrich the support of the training distribution and lead to better generalization and robustness respectively on general and noisy inputs (Zhang et al., 2018).
3. *Alleviating overfitting to low-resource languages. mXEncDec* can increase the diversity of low-resource languages by fusing low-resource examples with others, instead of the simple duplication in the standard training.

In *mXEncDec*, we randomly pick up two training examples drawn from the multilingual training corpus and first interpolate their source sentences where we have to prudently deal with language tags. Then we leverage a mixture decoder to produce a virtual target sentence. To account for heavy data imbalance of each language pair, we propose a pairwise sampling strategy to adjust interpolation ratios between language pairs. We also propose to simplify the target interpolation to cope with noisy attention and fusions of dissimilar language pairs. Different from *XEncDec* fusing two heterogeneous tasks (Cheng et al., 2021), we attempt to adapt it to deeply fuse different language pairs.

Experimental results on a large-scale WMT multilingual dataset show that *mXEncDec* yields improvements of $+1.13$ and $+0.47$ BLEU points averagely on xx-en and en-xx test sets over a vanilla multilingual Transformer model. We also evaluate our approaches on zero-shot translations and obtain up to $+5.53$ BLEU points over the baseline method, which corroborates the better transferabilty of multilingual models with our approaches. The more stable performance on noisy input text demonstrates the capability of our approach to improve the model robustness. To further explain the model behaviors at the representation level, qualitative and quantitative comparisons on representations manifest that our approach learns better multilingual representations, which indirectly explicates the BLEU improvements.

2 Background

Multilingual Neural Machine Translation. NMT (Bahdanau et al., 2015; Vaswani et al., 2017) optimizes the conditional probability $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ of translating a source-language sentence $\mathbf{x}$ into a target-language sentence $\mathbf{y}$. The encoder reads the source sentence $\mathbf{x} = x_1, ..., x_I$ as a sequence of word embeddings $e(\mathbf{x})$. The decoder acts as a conditional language model over embeddings $e(\mathbf{y})$ and the encoder outputs with a cross-attention mechanism (Bahdanau et al., 2015). For clarity, we denote the input and output in the decoder as $\mathbf{z}$ and $\mathbf{y}$, *i.e.*, $\mathbf{z} = \langle s \rangle, y_1, \cdots, y_{J-1}$ as a shifted copy of $\mathbf{y}$, where $\langle s \rangle$ is a sentence start token. Then the decoder generates $\mathbf{y}$ as $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} P(y_j|\mathbf{z}_{\leq j}, \mathbf{x};\boldsymbol{\theta})$. The cross-attention matrix is denoted as $\mathbf{A} \in \mathbb{R}^{J\times I}$. NMT optimizes the parameters $\boldsymbol{\theta}$ by maximizing the likelihood of a parallel training set $\mathcal{D}$:

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}[\ell(f(\mathbf{x},\mathbf{y};\boldsymbol{\theta}), v(\mathbf{y}))], \tag{1}$$

where ℓ is the cross entropy loss between the model prediction $f(\mathbf{x},\mathbf{y};\boldsymbol{\theta})$ and label vectors $v(\mathbf{y})$ for $\mathbf{y}$. $v(\mathbf{y})$ could be a sequence of one-hot vectors with smoothing in Transformer (Vaswani et al., 2017).

Multilingual NMT extends NMT from the bilingual to the multilingual setting, in which it learns a one-to-many, many-to-one or many-to-many mapping from a set of languages to another set of languages (Firat et al., 2016a; Johnson et al., 2017). More specifically, the multilingual NMT model is learned over parallel corpora $\mathcal{M} = \{\mathcal{D}^{l_i}\}_{i=1}^{L}$ where L is the number of language pairs:

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\mathcal{D}^{l_i}\in\mathcal{M}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}^{l_i}}[\ell(f(\mathbf{x},\mathbf{y};\boldsymbol{\theta}), v(\mathbf{y}))], \tag{2}$$

where all the parallel training sets are fed into the NMT model.

XEncdec: Crossover Encoder-Decoder. *XEncDec* aims to fuse two parallel examples (called parents) in the encoder-decoder model (Cheng et al., 2021). The parents' source sentences are shuffled into a sentence (the offspring's source) on the encoder side, and a mixture decoder model predicts a virtual target sentence (the offspring's target). Given a pair of examples $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ where their lengths are different in most cases, padding tokens are appended to the shorter one to align their lengths. The crossover example $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ (offspring) is generated by carrying out *XEncDec* over $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ (parents).

The crossover encoder combines embeddings of the two source sequences into a new sequence of embeddings:

$$e(\tilde{x}_i) = e(x_i)m_i + e(x'_i)(1 - m_i), \qquad (3)$$

where $\mathbf{m} = m_1, \cdots, m_{|\tilde{\mathbf{x}}|} \in \{0, 1\}^{|\tilde{\mathbf{x}}|}$ is sampled from a distribution or constructed according to a hyperparameter ratio p; e.g., $p = 0.15$ means that 15% of elements in $\mathbf{m}$ are 0. $|\tilde{\mathbf{x}}|$ is the length of $\tilde{\mathbf{x}}$, which is equal to $\max(|\mathbf{x}|, |\mathbf{x}'|)$.

On the crossover decoder side, a mixture conditional language model is employed for the generation of the virtual target sentence. The input embedding $e(\tilde{z}_j)$ and output label $v(\tilde{y}_j)$ for the decoder at the j-th position are calculated as:

$$e(\tilde{z}_j) = e(y_{j-1})t_{j-1} + e(y'_{j-1})(1 - t_{j-1}), \qquad (4)$$

$$v(\tilde{y}_j) = v(y_j)t_j + v(y'_j)(1 - t_j), \qquad (5)$$

where $\mathbf{t} = t_1, ..., t_{|\tilde{\mathbf{y}}|} \in [0, 1]^{|\tilde{\mathbf{y}}|} \subset \mathbb{R}^{|\tilde{\mathbf{y}}|}$. In contrast to a common language model fed with a single word y_{j-1} for predicting y_j at the j-th position, the crossover decoder aims to generate an interpolated vector $v(\tilde{y}_j)$ by averaging $v(y_j)$ and $v(y'_j)$ with t_j, on condition that the current input embedding is also weighted on embeddings $e(y_{j-1})$ and $e(y'_{j-1})$ with t_{j-1}. The weight vector $\mathbf{t}$ used for interpolating target inputs and labels is computed as:

$$t_j = \frac{\sum_{i=1}^{I} A_{ji} m_i}{\sum_{i=1}^{I} A_{ji} m_i + \sum_{i=1}^{I'} A'_{ji}(1 - m_i)}, \qquad (6)$$

where $\mathbf{A}$ and $\mathbf{A}'$ are the alignment matrices for $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$. In practice the cross-attention scores in the NMT model are utilized as an alternative noisy alignment matrix (Garg et al., 2019).

The cross-entropy is utilized to compute the loss for *XEncDec* when feeding $e(\tilde{\mathbf{x}})$, $e(\tilde{\mathbf{z}})$ and $v(\tilde{\mathbf{y}})$ into the encoder-decoder model, denoted as:

$$\ell(f(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \boldsymbol{\theta}), v(\tilde{\mathbf{y}})) = \sum_j KL(v(\tilde{y}_j) \| P(y | \tilde{\mathbf{z}}_{\leq j}, \tilde{\mathbf{x}}; \boldsymbol{\theta})). \qquad (7)$$

3 mXEncDec

In this work, we aim to leverage *XEncDec* to encourage multilingual NMT models to better exploit cross-lingual signals with crossover examples created by explicitly fusing different language pairs. We introduce its variant, called *mXEncDec* as shown in Figure 1, in which the parent examples could belong to either the same or different language pairs. The subsequent subsections discuss how to address new challenges of *mXEncDec* for multilingual NMT.

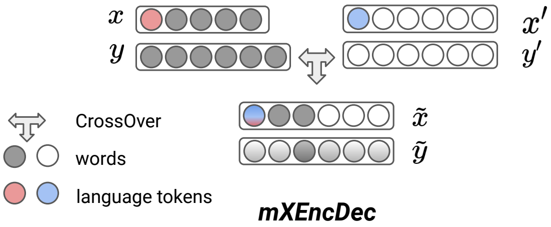


Figure 1: An illustration of multilingual crossover encoder-decoder (*mXEncDec*). The language tokens in the source sentences are softly interpolated based on the proportion of their words in $\tilde{\mathbf{x}}$.

Language Interpolation. As multilingual NMT involves a large number of language pairs, several techniques have been adopted to distinguish translation directions among them, such as prepending a language tag to source inputs (Johnson et al., 2017) or both source and target sentences (Wang et al., 2018), training language-specific embeddings for different languages (Lample and Conneau, 2019), and so on (Dabre et al., 2020). When following Lample and Conneau (2019), it is natural to interpolate language-specific embeddings as we do for token embeddings. However, if we want to adopt a language tag in the first word of a source sentence to indicate the target language (Johnson et al., 2017), we need to address how to interpolate them. As Figure 1 shows, to make the sentence $\tilde{\mathbf{x}}$ still carry language-specific information from $\mathbf{x}$ and $\mathbf{x}'$, we conduct a soft combination over their language tags, that is:

$$e(\tilde{x}_1) = e(x_1)\frac{\sum_{i=2}^{|\mathbf{m}|} m_i}{|\mathbf{m}| - 1} + e(x'_1)\frac{\sum_{i=2}^{|\mathbf{m}|}(1 - m_i)}{|\mathbf{m}| - 1}, \qquad (8)$$

where $|\mathbf{m}|$ is the length of $\mathbf{m}$. $e(\tilde{x}_1)$ captures the proportion of words in $\tilde{x}$ coming from the translation pairs $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$.

Simplified Target Interpolation. In comparison to bilingual NMT, attention matrices learned in multilingual NMT models are excessively noisy, which results in an inappropriate design of using the attention-based target interpolation in Eq. (6) for *mXEncDec*. Instead, we can employ a simple linear interpolation by setting $\mathbf{t}$ as a constant vector, here exemplified by the case of using language tags:

$$t_j = \frac{\sum_{i=2}^{|\mathbf{m}|} m_i}{|\mathbf{m}| - 1}, \quad \forall j \in \{1, ..., |\tilde{\mathbf{y}}|\}, \qquad (9)$$

A similar equation can be obtained for using language embeddings. In addition, dispensing with attention can improve the parallel efficiency with 10% speed-up gain.

Hard Target Input Interpolation. For multilingual NMT with multiple languages on the target side, i.e., one-to-many and many-to-many models, we need to carefully design combinations of target input word embeddings. As representations from the same language are usually close to each other, it can still augment the representation space by linearly interpolating target embeddings in Eq. (4). But for dissimilar languages, in particular distantly related languages, the interpolation points between them are comparatively unreliable. To tackle this issue, we simply quantize t_j to 1 if $t_j > 0.5$, otherwise $t_j = 0$ when interpolating target input embeddings for two different target languages in Eq. (4). A better solution should consider varying the interpolation ratio based on the language similarity or encourage interpolations of similar languages. We leave this for future exploration.

Pairwise Sampling. The multilingual corpus is usually heavily imbalanced: most of its data distribution concentrates on high-resource language pairs (Arivazhagan et al., 2019b). When interpolating high-resource and low-resource sentence pairs, we assume the fusion should be encouraged to be in favor of high-resource language pairs because the representation space supported by high-resource sentences is relatively reliable and stable (Kudugunta et al., 2019). This indicates a more frequent small p (e.g. $p < 0.5$) to weigh high-resource sentences over low-resource sentences if $(\mathbf{x}, \mathbf{y}) \in D^{l_i}$ is a high-resource sentence and $(\mathbf{x}', \mathbf{y}') \in D^{l_j}$ is a low-resource sentence. To this end, we propose a pairwise sampling method to sample the source shuffle ratio p_{l_i,l_j} for interpolating language pair l_i and l_j:

$$g \sim Bernoulli(1/(1 + exp(-\tau d(l_i, l_j))), \quad (10)$$

$$p_{l_i,l_j} = gp + (1 - g)(1 - p), \quad (11)$$

where τ is a temperature hyperparameter to control the tendency of g towards 0 or 1 for the Bernoulli distribution. $d(l_i, l_j)$ can be an arbitrary metric to measure the relationship between language l_i and l_j. Here we use $d(l_i, l_j) = |D^{l_i}|/|D^{l_j}|$ where $|D^{l_i}|$ denotes the data size of the language pair l_i.

Algorithm 1: Computing *mXEncDec* Loss.

```
Input: Corpus $\mathcal{M}$, temperature $\tau$, ratio $p$.
Output: Batch Loss $\mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta})$.
Function mXEncDec ($\mathcal{M}$, $\tau$, $p$):
    $(X', Y') \leftarrow$ shuffle $(X, Y) \in \mathcal{M}$ along batch;
    foreach $(\mathbf{x}, \mathbf{y}, \mathbf{x}', \mathbf{y}') \in (X, Y, X', Y')$ do
        $p_{l_i,l_j} \leftarrow$ sample a shuffle ratio in Eq. (10) and (11) with $\tau$ and $p$;
        $(e(\tilde{\mathbf{x}}), e(\tilde{\mathbf{z}}), v(\tilde{\mathbf{y}})) \leftarrow$ compute them using Eq. (3)-(5), (8), (6) or (9), and $p_{l_i,l_j}$;
        $\mathcal{L}_{\mathcal{X}} \leftarrow$ Eq. (7) with $(e(\tilde{\mathbf{x}}), e(\tilde{\mathbf{z}}), v(\tilde{\mathbf{y}}))$;
    end
    return $\mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta})$
```

Computing Loss. We calculate the training loss over *mXEncDec* as:

$$\mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\mathcal{D}^{l_i}\in\mathcal{M}} \mathop{\mathbb{E}}_{\mathcal{D}^{l_j}\in\mathcal{M}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}^{l_i}} \mathop{\mathbb{E}}_{(\mathbf{x}',\mathbf{y}')\in\mathcal{D}^{l_j}} [\ell(f(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \boldsymbol{\theta}), v(\tilde{\mathbf{y}}))], \quad (12)$$

where the generation of $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ depends on $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$. Algorithm 1 shows how to compute Eq. (12) effectively. We shuffle the min-batch consisting of all the language pairs. Then the shuffled batch and original batch can be used to generate $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ to compute the *mXEncDec* loss. Instead of using one-hot labels $v(y_j)$ in Eq. (5), we adopt label co-refinement (Li et al., 2019) by linearly combing the ground-truth one-hot label with the model prediction, that is $v(y_j)\beta + f_j(\mathbf{x}, \mathbf{y}; \hat{\boldsymbol{\theta}})(1 - \beta)$. Finally, our approach optimizes the model loss involving two training losses, Eq. (2) and Eq. (12):

$$\theta^* = \text{argmin}\{\mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta}) + \mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta})\}. \quad (13)$$

4 Experiments

Data and Evaluation. We conduct experiments on the English-centric WMT multilingual dataset composed of 16 languages (including English) and 30 translation directions from past WMT evaluation campaigns before and on WMT'19 (Barrault et al., 2019). The data distribution is highly skewed, varying from roughly 10k examples in En-Gu to roughly 60M examples in En-Cs. Two non-English test sets, Fr-De and De-Cs, are used to verify zero-shot translations. In addition, we also use multi-

$\tau =$	-2	-0.8	-0.4	0	0.4	0.8	2
xx-en	27.22,	27.42,	27.21,	27.41,	27.46,	**27.60,**	27.41
en-xx	21.76,	21.83,	21.74,	21.87,	21.89,	**22.01,**	21.87

Table 1: Effect of the temperature τ in the pairwise sampling. We tune this hyperparameter on *mXEncDec*-A for many-to-many models. *mXEncDec*-A: the target interpolation is computed based on attention.

Method	Many-to-One					One-to-Many				
			xx-en					en-xx		
	Low	Med.	High	Avg	WR	Low	Med.	High	Avg	WR
MLE	21.28	29.96	31.85	26.53	-	14.92	22.52	29.42	21.27	-
mixup	+0.95	+0.28	+0.05	+0.52	**93.33**	+0.49	-0.46	-0.26	+0.05	46.66
mXEncDec-A	+0.50	+0.44	+0.30	+0.42	86.67	+0.51	+0.06	+0.17	+0.31	80.00
+Hard	-	-	-	-	-	+0.47	**+0.08**	+0.31	+0.34	**86.66**
mXEncDec-S	**+1.76**	**+0.62**	**+0.36**	**+1.06**	**93.33**	+0.45	-0.25	-0.04	+0.15	73.33
+Hard	-	-	-	-	-	**+0.78**	-0.05	**+0.35**	**+0.47**	**86.66**

Table 2: Baseline comparisons for many-to-one and one-to-many models on the WMT multilingual translation. *mXEncDec*-A: the target interpolation is computed based on attention. *mXEncDec*-S: the target interpolation is simplified as a constant vector. WR: winning ratio. xx-en: other languages to English. en-xx: English to other languages. Hard: hard target input interpolation when interpolating different languages.

way test sets in FLORES-101 (Goyal et al., 2021) to analyze the trained multilingual models.

To mitigate the data imbalance in the WMT multilingual corpus, we follow Arivazhagan et al. (2019b) and adopt a temperature-based data sampling strategy to over-sample the low-resource languages where the temperature is set to 5. We apply SentencePiece (Kudo and Richardson, 2018) to learn a vocabulary of $64k$ sub-words. We perform experiments in three settings: many-to-one, one-to-many and many-to-many translations. The 15 test language pairs are cast into three groups according to their data size: High ($> 10M$, 5 languages), Low ($< 1M$, 7) and Medium ($> 1M\& < 10M$, 3). We report not only the average detokenized BLEU scores for each group as calculated by the SacreBLEU script (Post, 2018) but also winning ratio (WR) indicating the ratio of all the test sets on which our approach beats the baseline method.

Models and Hyperparamters. Following Chen et al. (2018), we select the Transformer Big (6 layer, 1024 model dimension, 8192 hidden dimension) as the backbone model and implement them with the open-source *Lingvo* (Shen et al., 2019). Adafactor (Shazeer and Stern, 2018) is adapted as our training optimizer, in which the learning rate is set to 3.0 and adjusted with $40k$ warm-up steps. We use a beam size of 4 and a length penalty of 0.6 for all the test sets. We apply language-specific embeddings to both many-to-one and one-to-many models while languages in many-to-many models are specified with language tags. Many-to-one and one-to-many models are optimized for $150k$ steps while many-to-many models run for $300k$ steps. All Transformer models utilize a large batch of around 5600×64 tokens over 64 TPUv4/TPUv3 chips. We average the last 8 checkpoints to report model performance. We tune p over the set: $\{0.10, 0.15, 0.25, 0.50\}$ and set it to 0.15 except for many-to-one using 0.25. The temperature τ used in Eq. (10) to sample the shuffle ratio is selected over the set $\{0, \pm 0.4, \pm 0.8, \pm 2.0\}$. $\tau = 0.8$ is selected for many-to-many models while $\tau = 0$ is for others as Table 1 suggests. The parameter β in label co-refinement is annealed from 0 to 0.7 in the first $40K$ steps. We find that a non-zero and non-one β can not only better capture informative label but also substantially improve the training stability.

Training Efficiency. If we adopt the simplified target interpolation, the loss computations for $\mathcal{L}_{\mathcal{M}}(\boldsymbol{\theta})$ and $\mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta})$ in Eq. (13) are totally independent. But we have to halve the batch size to load interpolation examples ($\mathcal{L}_{\mathcal{X}}(\boldsymbol{\theta})$) into memory. To make the

Method	Many-to-Many									
	xx-en					en-xx				
	Low	Med.	High	Avg	WR	Low	Med.	High	Avg	WR
MLE	23.2	29.02	31.19	27.03	-	15.86	22.34	29.49	21.70	-
mixup	+0.79	-0.11	-0.12	+0.31	60.00	+0.32	-0.28	-0.48	-0.06	33.33
mXEncDec-A	+0.88	+0.28	+0.31	+0.57	93.33	+0.64	**-0.01**	+0.04	+0.31	**73.33**
$\tau = 0$	+0.88	+0.20	-0.22	+0.38	73.33	+0.58	-0.14	-0.22	+0.17	66.66
+Hard	+0.92	+0.30	+0.16	+0.54	**100**	+0.52	-0.20	-0.14	+0.15	66.66
mXEncDec-S	+0.62	+0.34	+0.27	+0.45	86.66	+0.45	-0.10	+0.18	+0.25	60.00
$\tau = 0$	+0.87	+0.06	-0.10	+0.38	66.66	+0.43	-0.40	-0.29	+0.02	37.50
+Hard	**+1.78**	**+0.35**	**+0.71**	**+1.13**	**100**	**+0.66**	-0.14	**+0.53**	**+0.46**	60.00

Table 3: Baseline comparisons for many-to-many models on the WMT multilingual translation.

Method	Many-to-Many								
	WMT				FLORES				
	de→fr	fr→de	de→cs	cs→de	de→fr	fr→de	de→cs	cs→de	Avg
MLE	16.84	16.50	6.52	10.65	15.30	9.94	5.18	10.94	11.48
mixup	+2.66	+1.02	-3.35	+1.01	+2.16	+0.18	-2.61	+0.95	+0.25
mXEncDec-A	+3.70	+1.45	+2.33	+4.07	+2.54	+0.83	+1.82	+4.14	+2.61
+Hard	**+4.98**	+3.66	**+5.53**	+4.36	+5.02	+2.99	**+5.11**	+4.28	**+4.49**
mXEncDec-S	+4.94	+3.50	+0.18	**+5.31**	**+5.26**	**+3.30**	-0.26	**+4.56**	+3.34
+Hard	+3.45	**+3.82**	+3.50	+3.52	+2.46	+2.98	+3.44	+3.76	+3.37

Table 4: Results of WMT many-to-many models on zero-shot translations from WMT and FLORES.

baseline models and our models observe the same amount of parallel examples per step, we double the number of TPUs to compensate for it.

4.1 Main Results

We validate two variants of *mXEncDec* on many-to-one, one-to-many and many-to-many settings:

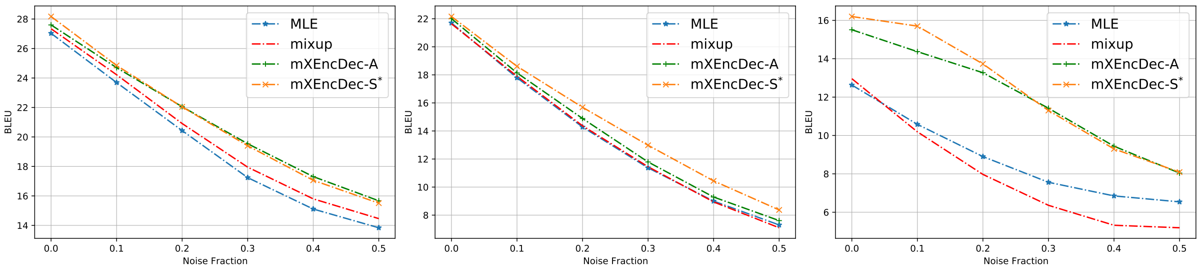
- *mXEncDec*-A: the target interpolation $\mathbf{t}$ is computed by normalizing attention in Eq. (6).
- *mXEncDec*-S: the target interpolation $\mathbf{t}$ is simplified to a constant vector in Eq. (9).

We compare *mXEncDec* to the baseline methods:

- MLE: the vanilla Multilingual NMT is trained with maximum likelihood estimation.
- *mixup*: we adapt *mixup* (Zhang et al., 2018) to multilingual NMT by mixing source and target sequences following the methods proposed in Cheng et al. (2020) and Guo et al. (2020). For a fair comparison, we also mix co-refined labels rather than one-hot labels.

Table 2 shows results on the WMT multilingual dataset for many-to-one and one-to-many models. The comparisons between the baseline MLE and our approach suggest that *mXEncDec* can improve the translation performance on both xx-en and en-xx translation settings (up to +1.06 BLEU & 93.33 WR on xx-en and +0.47 BLEU & 86.66 WR on en-xx). In particular, using simplified target interpolation to substitute the noisy attention-based interpolation (*mXEncDec*-S vs. *mXEncDec*-A) can achieve better results on xx-en translations (+0.64 BLEU) while slightly performing worse on en-xx translations (-0.16 BLEU). After incorporating quantized target interpolation, it yields an additional improvement for *mXEncDec*-S on en-xx translations (+0.32 BLEU). The improvement differences between xx-en and en-xx (+1.06 BLEU vs. +0.47 BLEU) to some extent imply that interpolations on the target side are more favourable to similar languages, and interpolations on the encoder side are not sensitive to language types.

Table 3 shows results for many-to-many models. Among all the training methods, our approaches still obtain the best results for both xx-en and en-



(a) xx-en (b) en-xx (c) zero-shot

Figure 2: Results on artificial code-switching noisy data. We plot the BLEU changes of many-to-many models when varying the noise fraction on xx-en, en-xx and zero-shot test sets.

xx translations (up to +1.13 BLEU & 100 WR on xx-en and +0.46 BLEU & 73.33 WR). We consistently find that *mXEncDec*-S benefits much more from the quantized target interpolation with +0.68 BLEU on xx-en and +0.21 BLEU on en-xx. Although this technique slightly impairs the performance of *mXEncDec*-A on both xx-en and en-xx translations, it significantly boosts its zero-shot translations as shown in Table 4. We also observe that removing the pairwise sampling with $\tau = 0$ has big negative effects on high-resource language pairs for many-to-many models. Pairwise sampling can not only stabilize the performance on low-resource language pairs and significantly improve high-resource language pairs.

Compared to *mixup*, our approaches still attain better performance except that *mXEncDec*-A on xx-en performs slightly worse. *mixup* trains models on linear interpolations of examples and their labels. By contrast, *mXEncDec* combines training examples in a non-linear way on the source side, and encourages the decoder to decouple the non-linear interpolation with a ratio related to the source end.

4.2 Zero-shot Translation

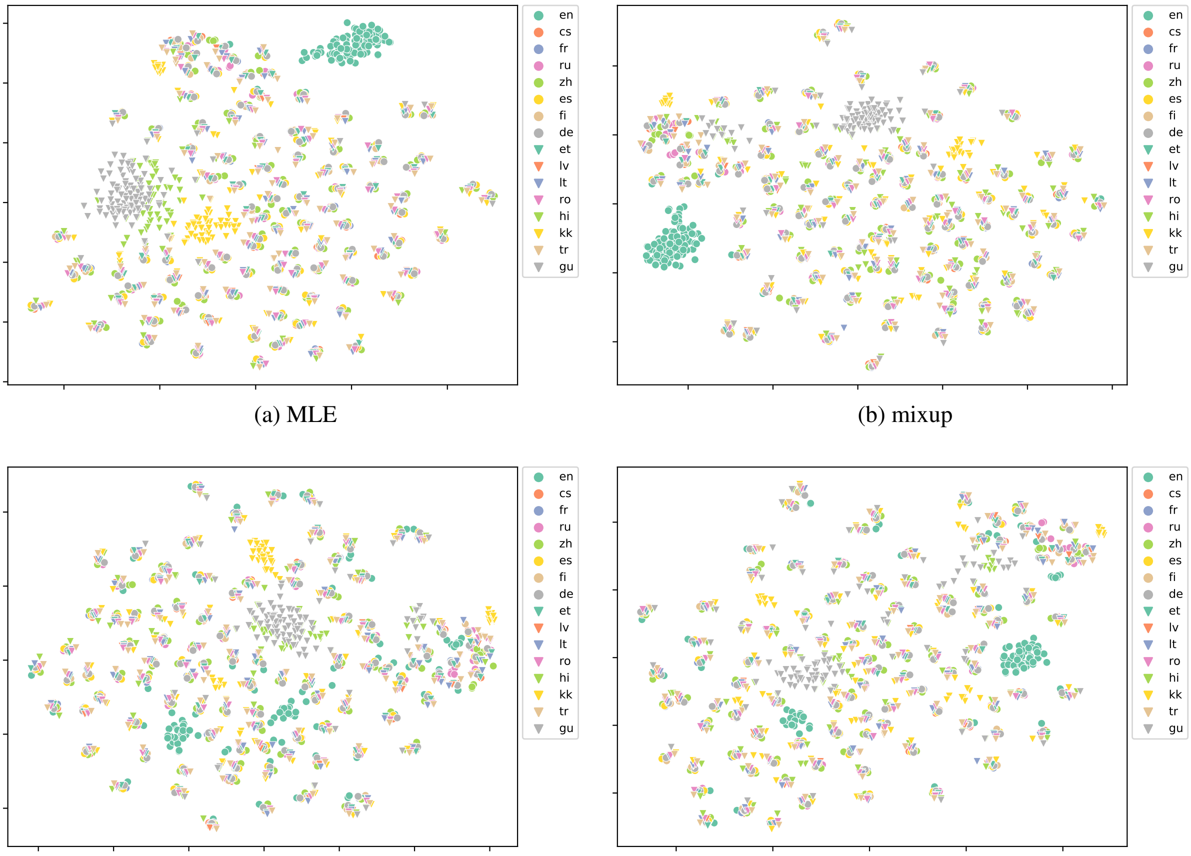
To further verify cross-lingual transfer of our approaches, we utilize many-to-many models to decode language pairs not pesent in the training data, i.e., zero-shot sets from WMT and FLORES. In Table 4, our approaches achieve notable improvements across all the test sets compared to baseline methods. On average, our best approach (*mXEncDec*-A + Hard) can gain up to +4.49 BLEU over MLE. Interestingly, this model is not the best on general translations but delivers the best results on zero-shot translations. These substantial improvements demonstrate the strong transferability of our approaches.

4.3 Multilingual Robustness

We construct a noisy test set comprising code-switching noise to test the robustness of multilingual NMT models (Belinkov and Bisk, 2018; Cheng et al., 2019). Following the method proposed in Cheng et al. (2021), we randomly replace a certain ratio of English/non-English source words with non-English/English target words by resorting to an English-centric dictionary. From results in Figure 2, we find our approaches to exhibit higher robustness with larger improvements as the noise fraction increases. *mXEncDec*-A shows similar robustness to *mXEncDec*-S* on zero-shot translations and even higher robustness on xx-en translations although its performance on clean test sets falls behind *mXEncDec*-S*. *mXEncDec*-S* performs significantly better on en-xx translations compared to other approaches. Moreover, it is noteworthy that our approaches have better stability on xx-en translations where we replace non-English words with English counterparts, which is in complete agreement with the finding in section 4.4 that English representations tend to be fused into non-English representations by virtue of our approaches.

4.4 Representation Analyses

To better interpret the advantages of our approaches over baselines, we attempt to delve deep into the representations incurred by models. A common method is to study the encoder representations of multilingual NMT models (Kudugunta et al., 2019), which we follow. We aggregate the sentence representations by averaging the encoder outputs. The data computing representations come from FLORES (Goyal et al., 2021) as it provides a high quality of multi-way translations implying that sentences from each language are semantically equivalent to each other. We use the first 100 sentences in



(a) MLE

(b) mixup

(c) *mXEncDec*-A

(d) *mXEncDec*-S*

Figure 3: t-SNE visualizations of encoder representations on xx-en translations for comparing many-to-many models trained with MLE, *mixup*, *mXEncDec*-A and *mXEncDec*-S*.[1] *mXEncDec*-S*: *mXEncDec*-S + Hard.

each language to visualize representations.

We argue that *the encoder in a good multilingual NMT model prefers to distribute sentence representations based on their semantic similarities rather than language families.* Figure 3 depicts visualisations of representations plotted by t-SNE (Van der Maaten and Hinton, 2008) on xx-en translations. We make the following observations:

1. In each figure, sentences with the same semantics incline to form a single cluster.
2. For MLE in Figure (a), most sentences are dispersed into each cluster based on semantics while extremely low-resource languages (Hi, Gu, Kk) and English possess their own distinct clusters.
3. For *mixup*, *mXEncDec*-A and *mXEncDec*-S* in Figure (b)-(d), sentences from extremely low-resource languages start to be assimilated into their own semantic clusters.
4. For *mXEncDec*-A and *mXEncDec*-S* in Figure (c)-(d), English sentences attempt to fuse into representations of other languages.

English sentences prefer to become an individual cluster. Because when using the language tag "<2en>" to compute English encoder representations, it is treated as a copy task instead of translation tasks for computing representations of other languages. However, our approach promotes English sentences to be closer to their semantic equivalents in other languages. This leads to enhanced robustness toward code-switching noise when translating sentences in languages that are mixed with English codes. The evident representation amelioration for extremely low-resource languages corroborates significant BLEU improvements on low-resource translations in Table 2 and Table 3. The encoder learned by our approach performs the best and complies with our argument. We also conduct quantitative analyses to evaluate the clustering effect of each method in Figure 3. In Table 5, we adopt three clustering metrics, SC (Silhouette Coefficient), CH (Calinski-Harabaz Index), and DB

[1] We also have similar findings from visualizations for en-xx translations.

Method	SC ↑	CH ↑	DB ↓
MLE	0.1625	15.02	1.896
mixup	0.1821	16.56	1.796
mXEncDec-A	0.1796	16.52	1.806
mXEncDec-S*	**0.1924**	**18.38**	**1.739**

Table 5: Quantitative analysis of clusters produced by methods in Figure 3. Three popular metrics to evaluate the quality of clustering are used: SC (Silhouette Coefficient), CH (Calinski-Harabaz Index), DB (Davies-Bouldin Index). *mXEncDec*-S*: *mXEncDec*-S + Hard.

(Davies-Bouldin Index). Although these metrics cannot adequately assess multilingual representations as they advocate distinct separations between different clusters and tight closeness within the same cluster, we believe they can still measure the within-cluster variance in part. Among them, *mXEncDec*-S* performs the best while *mixup* and *mXEncDec*-A yield similar performance.

5 Related Work

Multilingual NMT has made tremendous progress in recent years (Dong et al., 2015; Firat et al., 2016a; Johnson et al., 2017; Arivazhagan et al., 2019b; Fan et al., 2021). Recent research efforts to improve the generalization of multilingual models concentrate on enlarging the model capacity (Huang et al., 2019; Zhang et al., 2020; Lepikhin et al., 2020), incorporating hundreds of languages (Fan et al., 2021), pretraining multilingual models (Liu et al., 2020), and introducing additional regularization constraints (Arivazhagan et al., 2019a; Al-Shedivat and Parikh, 2019; Yang et al., 2021). Our work is related to the last three ones in that they try to enable models to better transfer across languages by introducing an alignment loss to learn an interlingua (Arivazhagan et al., 2019a) or imposing an agreement loss on translation equivalents (Al-Shedivat and Parikh, 2019; Yang et al., 2021). However, we propose to utilize *mXEncDec* to directly combine language pairs for better exploitation of cross-lingual signals.

Another related research line is data mixing. Since *mixup* (Zhang et al., 2018; Yun et al., 2019) was proposed in computer vision, we have observed great success in NLP (Guo et al., 2019; Cheng et al., 2020; Guo et al., 2020; Cheng et al., 2021). *mXEncDec* shares the commonality of combining example pairs as inspired by *XEncDec* (Cheng et al., 2021). To the best of our knowledge, we are the first to fuse different language pairs to improve cross-lingual generalization and robustness for multilingual NMT.

6 Conclusion

We have presented *mXEncDec* to fuse different language pairs at instance level for multilingual NMT, which enables the model to better exploit cross-lingual signals. Experimental results on general, zero-shot and noisy test sets demonstrate that our approach can significantly improve the cross-lingual generalization, zero-shot transfer and robustness of multilingual NMT models. Representation analyses further confirms that our approach is capable of learning better multilingual representations, which coincides with improvements in BLEU. We plan to investigate whether this approach can improve the model generalization in a broader scope like domain generalization. We find that *mXEncDec* can easily achieve notable improvements for xx-en translations because they share an identical target language. However, there still exits huge headroom for en-xx translations. We plan to explore how to interpolate target languages more effectively, for example, possibly considering language similarity.

References

Maruan Al-Shedivat and Ankur P Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *North American Association for Computational Linguistics (NAACL)*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In

Proceedings of the Fourth Conference on Machine Translation.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Rich Caruana. 1997. Multitask learning. *Machine learning*.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. 2001. Vicinal risk minimization. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Yong Cheng, Wei Wang, Lu Jiang, and Wolfgang Macherey. 2021. Self-supervised and supervised joint training for resource-rich machine translation. In *International Conference on Machine Learning (ICML)*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. *ArXiv*, abs/2006.13979.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research (JMLR)*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *North American Association for Computational Linguistics (NAACL)*.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman-Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. *arXiv preprint arXiv:1909.02074*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Demi Guo, Yoon Kim, and Alexander M Rush. 2020. Sequence-level mixed sample data augmentation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*.

Junnan Li, Richard Socher, and Steven CH Hoi. 2019. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

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

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT*.

Rigomar Rieger, Arnd Michaelis, and Melvin M Green. 2012. *Glossary of genetics and cytogenetics: classical and molecular*. Springer Science & Business Media.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning (ICML)*.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, et al. 2019. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research (JMLR)*.

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 *Conference on Neural Information Processing Systems (NeurIPS)*.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual neural machine translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *North American Association for Computational Linguistics (NAACL)*, pages 30–34.