Data Efficient Direct Speech-to-Text Translation with Modality Agnostic Meta-Learning

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu,
Beomseok Lee, Insoo Chung, Sangha Kim, Chanwoo Kim
Samsung Research, Seoul, South Korea
{s.indurthi, h.j.han, n07.kumar, bsgunn.lee, sangha01.kim, insooo.chung, chanw.com}@samsung.com

Abstract

End-to-end Speech Translation (ST) models have several advantages such as lower latency, smaller model size, and less error compounding over conventional pipelines that combine Automatic Speech Recognition (ASR) and text Machine Translation (MT) models. However, collecting large amounts of parallel data for ST task is more difficult compared to the ASR and MT tasks. Previous studies have proposed the use of transfer learning approaches to overcome the above difficulty. These approaches benefit from weakly supervised training data, such as ASR speech-to-transcript or MT text-to-text translation pairs. However, the parameters in these models are updated independently of each task, which may lead to sub-optimal solutions. In this work, we adopt a meta-learning algorithm to train a modality agnostic multi-task model that transfers knowledge from source tasks=ASR+MT to target task=ST where ST task severely lacks data. In the meta-learning phase, the parameters of the model are exposed to vast amounts of speech transcripts (e.g., English ASR) and text translations (e.g., English-German MT). During this phase, parameters are updated in such a way to understand speech, text representations, the relation between them, as well as act as a good initialization point for the target ST task. We evaluate the proposed meta-learning approach for ST tasks on English-German (En-De) and English-French (En-Fr) language pairs from the Multilingual Speech Translation Corpus (MuST-C). Our method outperforms the previous transfer learning approaches and sets new state-of-the-art results for En-De and En-Fr ST tasks by obtaining 9.18, and 11.76 BLEU point improvements, respectively.

1 Introduction

The Speech Translation (ST) task takes audio as input and generates text translation as output. Traditionally it is achieved by cascading Automatic Speech Recognition (ASR) and Machine Translation (MT) models (Ney 1999). However, the cascaded model suffers from compounding errors between ASR and MT models, higher latency due to sequential inference from the two models, and higher memory and computational resource requirements.

End-to-end neural models for Automatic Speech Recognition (ASR) (Graves, Mohamed, and Hinton 2013) and Machine Translation (MT) (Bahdanau, Cho, and Bengio 2015) are evolving into end-to-end neural model for Speech Translation (ST)(Bérard et al. 2016). Such models overcome the above limitations of cascaded systems. However, training such end-to-end ST models requires huge amounts of speech-to-text parallel data. Huge amounts of parallel data along with the advancements in sequence-to-sequence models led to successful ASR and MT neural systems. However, collecting such amounts of parallel data for training ST system is very challenging.

To alleviate the issue of collecting vast amounts of parallel data for ST task, Bérard et al.; Jia et al. (2018; 2019) proposed pre-training based approaches such as transfer learning. Although these approaches have been shown to improve the performance of the ST task, they have some limitations. In the transfer learning strategy, the pre-trained model parameter updates are based on the current task at hand and are not optimized towards adapting to new tasks. In multi-task learning (Weiss et al. 2017; Anastasopoulos and Chiang 2018), a variant of transfer learning, the parameters are shared across multiple tasks, and thus limits the performance of individual tasks. Moreover, the parameters of the model are updated independently based on individual task performance, and this may lead to sub-optimal solutions in these approaches.

To overcome the limitations of transfer-learning and its variants, we propose a multi-task learning approach based on a meta-learning algorithm, for the ST task. We adopt the model-agnostic meta-learning algorithm (MAML) (Finn, Abbeel, and Levine 2017) to train on tasks with different input modalities. We use ASR and MT as source tasks during the meta-learning phase. These two tasks have different input modalities; ASR with speech input and MT with text input. The learned parameters from the meta-learning phase are used to initialize the parameters of our target ST model in the fine-tuning phase. There are two advantages of this approach over transfer learning: 1) The parameter updates of the model are not only based on the source ASR and MT tasks but also how good these works as initialization parameters for the target ST task. 2) We can utilize both the ASR and MT data at the same time without sharing parameters between auxiliary ASR, MT tasks, and the target ST task.

We conducted several experiments on English-German

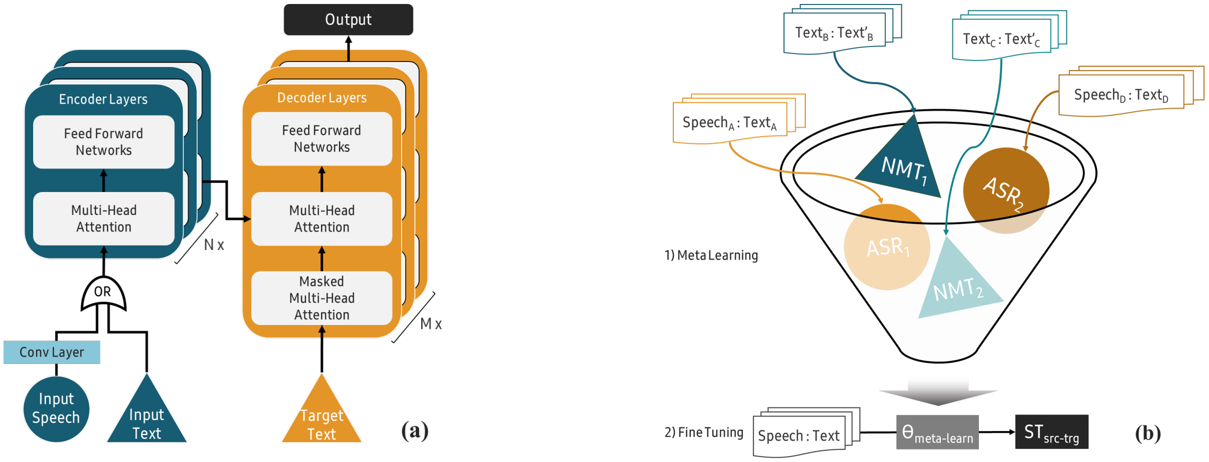

Figure 1: (a) Overview of the base seq2seq architecture. (b) Overview of the modality agnostic meta learning model for ST.

(En-De) and English-French (En-Fr) speech translation tasks from the MuST-C corpus (Di Gangi et al. 2019) to evaluate the effectiveness of the proposed meta-learning approach. Our experiments reveal that the proposed approach achieves state-of-the-art performance on En-De, En-Fr ST tasks by obtaining 22.11 and 34.05 BLEU points, respectively.

2 Speech Translation with Meta-Learning

2.1 Problem Formalization

A typical Sequence-Sequence (seq2seq) architecture (Sutskever, Vinyals, and Le 2014) generates a target sequence $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ given a source sequence $\boldsymbol{x} = \{x_1, \cdots, x_m\}$ by modeling the conditional probability, $p(\boldsymbol{y}|\boldsymbol{x}, \theta)$. In general, the seq2seq architecture consists of two components: (1) an encoder which computes a set of representations $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_{\boldsymbol{1}}, \cdots, \tilde{\boldsymbol{x}}_{\boldsymbol{m}}\} \in \mathbb{R}^{m \times d}$ corresponding to $\boldsymbol{x}$, and a decoder coupled with attention mechanism (Bahdanau, Cho, and Bengio 2015) dynamically reads encoder's output and predicts the distribution of each token in the target language. It is trained on a dataset of D parallel sequences to maximize the log likelihood:

$$\ell(D; \theta) = -\frac{1}{|D|} \sum_{i=1}^{N} \log p\left(\boldsymbol{y}^i | \boldsymbol{x}^i; \theta\right), \quad (1)$$

where θ are parameteres of the model.

The ASR, MT, and ST tasks in our work share the same seq2seq architecture. The ASR and ST tasks take speech signal as input, and the input to the MT task is a sequence of characters or wordpiece tokens. The output of all the models is a sequence of tokens consisting of either characters or wordpiece tokens.

2.2 Base Seq2Seq Model

In recent years, the non-recurrent Transformer network Vaswani et al. (2017) has achieved best translation quality for MT task. The encoder and decoder blocks of the Transformer are composed of a stack N and M identical layers, respectively. Each layer in the encoder contains two sublayers, a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Each decoder layer consists of three sublayers; the first and third sub-layers are similar to the encoder sub-layers, and the additional second sub-layer is used to compute the encoder-decoder attention (context) vector based on the soft-attention based approaches (Bahdanau, Cho, and Bengio 2015). Please refer to Vaswani et al. (2017) for detailed architecture.

Here we adopt the above transformer network as our base seq2seq model for our ASR, MT, and ST tasks. Specifically, we adapt it to ASR and ST tasks by appending a compression layer. The speech sequence, represented using Mel bank features, is commonly a few times longer than the text sequence. Therefore, we stack 3×3 CNN layers with stride 2 for both time and frequency dimensions to compress the length and exploit the structure locality of the speech signal. This compressed signal is later sent to the self-attention layers of the encoder. The overview of the base seq2se model is shown in Figure 1(a).

2.3 Modality Agnostic Meta-Leaning for ST

The base seq2seq model described above is known to easily overfit and result in an inferior performance when the training data is limited. We mitigate this issue by sharing the knowledge between low and high resource tasks using the MAML algorithm. The approach of MAML is to use a set of high resource tasks as source tasks $\{\tau^1, \cdots, \tau^s\}$ to find a good parameter initialization point θ^0 for the low resource target task τ^0.

Meta-Learning Phase: In this paper, we extend the idea of MAML to meta-learn on tasks with different input modalities. The source tasks in our model are ASR and MT with speech-text and text-text modalities, respectively. Later, we fine-tune the target ST task from the parameters of the meta-learned model (θ^m). The overview of the proposed approach

Algorithm 1: Meta-Learning Algorithm for ST task

```
Input: Training examples from source tasks,
  $T = \{ASR, MT\}$ and target ST task.
Input: Hyperparameters such as learning rates, $\alpha$ and $\beta$
Randomly initialize model parameters $\theta^m$.
while not done do
    Sample task, $\tau$ from $T$
    Assign $\theta^a = \theta^m$
    Sample K data points, $D_\tau = \{x_{(i)}, y_{(i)}\}_{i=1}^{k}$ from $\tau$
    compute $\nabla_{\theta^m} \ell(D_\tau; \theta^m)$ using $D_\tau$ and $\theta^m$
    Meta-Train: update $\theta^a$ using Eq. 4
    sample l data points, $D_\tau^{'} = \{x_{(i)}^{'}, y_{(i)}^{'}\}_{i=1}^{l}$ from $\tau$
    compute $\nabla_{\theta^a} \ell(D_\tau^{'}; \theta^a)$ using $D_\tau^{'}$ and $\theta^m$
    Meta-Test: update $\theta^m$ using Eq. 5
end
Assign $\theta = \theta^m$
while not done do
    sample m data points, $D_{st} = \{x_{(i)}, y_{(i)}\}_{i=1}^{m} \in$ ST
      task
    compute $\nabla_\theta \ell(D_{st}; \theta)$ using $D_{st}$ and $\theta$
    Finetune: Update $\theta$ with gradient descent:
      $\theta = \theta - \gamma \nabla_\theta \ell(D_{st}; \theta)$
end
Return: $\theta$
```

is shown in Figure 1(b). The process can be understood as

$$\theta^* = \mathrm{Learn(ST; Meta - Learn(ASR, MT))}. \quad (2)$$

We find the initialization θ^0 for ST task by simulating low resource scenarios using source ASR and MT tasks. We define the meta objective function $\ell(\theta^m)$ to get $\theta^0 = \theta^m$ following Finn, Abbeel, and Levine (2017):

$$\ell(\theta^m) = E_\tau E_{D_k, D_k^{'}} \left[\ell \left(D_\tau; \ell \left(D_\tau^{'}; \theta^m \right) \right) \right], \quad (3)$$

where τ refers to the randomly sampled task to carry-out one meta-learning step. The set of samples D_τ and $D_\tau^{'}$ follow the uniform distribution over τ's dataset.

We maximize the meta-objective function in eq. 3 using gradient descent. For each meta-learning step, we uniformly sample one source task (τ) at random from the set, $\{\mathrm{ASR}, \mathrm{MT}\}$. We then sample two batches of training examples, D_τ and $D_\tau^{'}$, independently from the chosen source task, τ. We use D_τ to simulate task-specific learning and the $D_\tau^{'}$ to evaluate its outcome. We call the gradient step to simulate task-specific learning (the *auxiliary-gradient* step). The auxiliary parameters (θ^a) are updated using the auxiliary-gradient step with the learning parameter α, which is given as:

$$\theta_\tau^a = \theta^m - \alpha \nabla_{\theta^m} \ell(D_\tau; \theta^m). \quad (4)$$

Once the task-specific learning is done, we evaluate the auxiliary parameters θ^a against the previously sampled batch of training examples, $D_\tau^{'}$. The gradient computed on ($\ell(D_\tau^{'}; \theta^a)$) during this evaluation is called the *meta-gradient*. The meta parameters (θ^m) are updated using this meta-gradient and is computed as follows:

$$\theta_\tau^m = \theta^m - \beta \nabla_{\theta^a} \ell(D_\tau^{'}; \theta^a), \quad (5)$$

where β is the learning rate. Use of second derivates when estimating the meta-gradient through the auxiliary gradient in eq. 3 requires expensive Hessian matrix computation. Therefore, by following the vanilla MAML algorithm, we also use first-order approximation while computing the meta-gradients.

The meta-learned parameters θ^m, updated through eq. 5, can adapt to a new learning task using only a small number of training examples.

Dealing with Different Modalities: The vanilla MAML algorithm does not handle tasks with different input-output modalities. Moreover, we use additional compression layer on the input speech signal and it is not required for input text sequence. To deal with these limitations: (1) We create a universal vocabulary from all the tasks by following Gu et al. (2018a). (2) We dynamically disable the compression layer whenever we sample from the MT task during the meta-learning phase. That is, the MT examples do not affect the parameters of the compression layer.

Fine-tuning Phase: During the meta-learning phase, the parameters of the model (θ^m) are exposed to vast amounts of speech-to-transcripts and text-to-text translation datasets via ASR and MT tasks. This allows the parameters of all the sublayers in the model such as compression, encoder, decoder, encoder-decoder attention, and output layers to learn individual language representations and translation relations between them. Hence, the meta-learned parameters (θ^m) may not be suitable for the ST task on its own but can act as a good starting point to learn the target ST task. The model parameters are initialized from θ^m and further updated based on the target ST task evaluations. During fine-tuning phase, model training proceeds like in usual neural network training without involving auxiliary updates. An overview of the proposed modality agnostic meta-learning approach is given in Algorithm 1.

3 Experiments

3.1 Datasets and Metrics

Target Tasks: We used the MuST-C corpus (Di Gangi et al. 2019) to test the effectiveness of our proposed approach. MuST-C is a corpus for ST from English to 8 different target languages (German, Spanish, French, Italian, Dutch, Portuguese, Romanian and Russian). This corpus is created from English TED talks, which are automatically aligned at the sentence level with corresponding English transcriptions and translations to the target languages mentioned above. This dataset is larger than any other publicly available ST corpus.

In our experiments, we focus on two target languages, German and French. The En-De corpus consists of around 408 hours of English speech, which corresponds to 234k sentences of paired speech-transcript-translation data,

Task	Dataset	Domain	Train	Dev	Test
ST	MuST-C English-German	TED Talks	229K	1.4K	2.6K
	MuST-C English-French		275K	1.4K	2.6K
ASR	Spoken Wikipedia Corpus-English	Wikipedia Articles	347K	2.7K	2.0K
MT	WMT16 English-German	News articles & European Parliment proceedings	4.5M	3K	3K
	WMT16 English-French		40.8M	3K	3K

Table 1: Statistics of the datasets used in our experiments.

whereas the En-Fr corpus has 492 hours of speech, corresponding to 280k sentences, see Table 1.

Source Tasks: We use the Spoken Wikipedia Corpus (SWC, Baumann, Köhn, and Hennig (2019)) for training the English ASR tasks. The SWC is a collection of time-aligned spoken Wikipedia articles for Dutch, English, and German using a fully automated pipeline to download, normalize and align the data. We use the English speech and English transcripts of the SWC Corpus to train our ASR models in transfer learning and meta-learning phase. This corpus contains 352k sentences in 395 hours of speech read by a diverse set of 413 speakers.

For training MT tasks during meta-learning, we use the WMT16 En-De and En-Fr language pairs. The datasets are created by extracting language pairs from news articles and proceedings of the European Parliament.

The datasets used in ST, ASR, and MT tasks come from different domains. The datasets of ST are from TED talks, ASR from Wikipedia and MT from news articles. The primary reason to use datasets from different domains is that it is difficult to gather all the task-specific datasets from the same domain. Therefore, here we test the generalization performance of ST task trained with the help of ASR and MT tasks collected from different domains. The statistics of all the datasets used in our experiments are shown in Table 1.

Data Processing and Evaluation Metrics: The speech signal in ASR and ST is represented by log Mel 80-dimensional features. The text sequence in all the tasks is split into characters preserving word boundaries. We report the case sensitive BLEU scores on test sets for ST and MT tasks and are obtained using 4-gram NIST BLEU score (Papineni et al. 2002). ASR performance is measured in terms of word error rate (WER). We choose the best models based on the dev set performance and report the results on the testset.

3.2 Implementation Details

The proposed model is implemented based on Tensor2Tensor framework (Vaswani et al. 2018). The number of convolutional layers in the compression layer is set to 2. We use eight encoder and decoder layers in our experiments. We apply dropout rate of 0.2 to the output of each sublayer before it is added to the sublayer input and normalized. We use a batch size of 1.5M frames for ASR and ST tasks and a batch size of 4096 tokens for MT task. All other hyperparameters such as optimization algorithm, learning rate schedule are set similar to Vaswani et al. (2017). All the models are trained on 4*NVIDIA V100 GPUs.

3.3 Baselines

We present the reported results from Di Gangi et al. (2019) as Baseline 1 to compare with our models. The architecture of Baseline 1 is based on (Bérard et al. 2018) and it is an attention-based end-to-end ST model. The encoder of the model is based on feedforward, convolutional layers, and three stacked LSTMs. The decoder consists of a two-layered deep transition LSTM (Pascanu et al. 2013). The system is trained using transfer learning strategy by first pre-training the ASR model followed by the ST model. The ASR model is trained on speech-to-transcripts available from the MuST-C corpus.

In order to show the effectiveness of the proposed Meta-Learning (ML) approach compared to Transfer Learning (TL) and Multi-Task Learning (MTL) approaches, we design two baselines, called as Baseline 2 and Baseline 3. The architecture of these baselines is precisely similar to the seq2seq model used in our meta-learning experiments. These baselines are more powerful compared to Baseline 1 and act as better baselines to compare the effectiveness of the proposed meta-learning approach. The Baseline 2 is pre-trained on the ASR task, and Baseline 3 is pre-trained on all the three tasks (ASR, MT, and ST) simultaneously. The two baselines are further fine-tuned on the ST task. The datasets used during the pretraining phase in these approaches are same as the meta-learning phase.

3.4 Main Results

Baseline 1 vs. Baseline 2: We compare our non-recurrent based Baseline 2 model (in Table 2, model no. 2) against the Baseline 1. Even though the ASR dataset used during transfer learning of our *Baseline 2* is out of domain with ST task, it achieved significant BLEU score improvement for both En-De ($\uparrow 2.67$) and En-Fr ($\uparrow 4.65$) ST tasks compared to the *Baseline 1*. Therefore, we use *Baseline 2* to compare our proposed meta-learning approach.

Meta-Learning vs. Transfer Learning: Here, we compare the performance of Baseline 2, Baseline 3, and the proposed modality agnostic meta-learning model for ST task. During the meta-learning phase, we use the SWC English dataset for the ASR task and WMT dataset of English to the corresponding ST task target language for the MT task. The parameters of the model are updated using the proposed meta-learning approach, as described in Section 2.3. From Table 2, we can see that the ST model trained using the proposed meta-learning approach outperforms Baseline 2 and Baseline 3 by achieving 17.20 and 29.19 BLEU score on En-De and En-Fr lan-

No.	Model	Char / Wordpiece	Synthetic Data Augmentation	ST (BLEU)	
				En-De	En-Fr
1	Transfer Learning (Di Gangi et al. 2019) (`Baseline 1`)	char	No	12.93	22.29
	This Work				
2	Transfer Learning (`Baseline 2`)	char	No	15.60	26.94
3	Multi-Task Learning (`Baseline 3`)	char	No	16.00	26.20
4	**Meta-Learning**	**char**	**No**	**17.20**	**29.19**
5	Cascade (`Baselien 4`)	wordpiece	Yes	20.86	33.7
6	**Meta-Learning**	**wordpiece**	**Yes**	**22.11**	**34.05**

Table 2: Performance of various models on En-De and En-Fr speech translation tasks.

guage pairs, respectively. The results show that the meta-learning phase helps to learn the individual language representations and relations between them. Moreover, we can see that the meta-learning algorithm helps the target task despite being trained on the source tasks coming from different domains.

ST Task	TL	ML	
	ASR (wer ↓)	ASR (wer ↓)	MT (bleu ↑)
En-De	37.43	42.95	17.16
En-Fr	37.43	39.56	24.70

Table 3: Performace of various source tasks used in transfer learning (TL) and meta-learning (ML) approaches.

We also report the performance of ASR, MT tasks used in transer and meta-learning phase in Table 3. The performance of the ASR model used in transfer learning approach is significantly better than the ASR model in meta-learning approach. However, we achieved significantly better results for target ST task with the meta-learning approach. This is expected given that in the meta-learning phase, we update the parameters with a focus to adapt to the target ST task instead of focusing heavily on learning the particular source task. This also applies to the MT model, whose performance is lower than the standard MT model.

3.5 Impact of Initialization

To study the effectiveness of the meta-learned parameters (θ^m) as an initialization point and check quick adaptability to the target ST task, we analyze the BLEU scores and training losses for the first few steps. We compare the models obtained by fine-tuning from the meta-learned parameters (θ^m) against the transfer learning parameters (θ^t). We measure the BLEU score for every 1K steps for the first 10K steps for both the models on the test set. We can see from Figures 2(a) and 2(b) that the meta learned parameters act as a much better initialization point for the target task. We also plot the training loss on the ST task for the first 10k steps from the models obtained by fine-tuning from (θ^m) vs. (θ^t). From Figures 3(a) and 3(b), we can observe that the ST model initialized from the meta-learned parameters has significantly lesser loss than that from the transfer learned parameters. Here, we reported only the first 10K steps of BLEU score and training loss to see the effect near the initialization step. However, these trends continued in the later training steps as well.

3.6 Sample Translations

We present a few sample translations from the testset in Table 4 for the transfer and meta-learning apporaches. These are generated using model numbers 2 and 4 in Table 2. Analyzing these samples gives an insight into the proposed meta-learning approach for the ST task. We can see that the translations from the meta-learning approach preserve the context better than the ones from the transfer learning approach. We also see that it mitigates the speech pronunciation issues by leveraging language representations learnt via MT task during the meta-learn phase (For example, Confucianism versus Confisherism in the last example in Table 4).

3.7 Further Improvements

We further improved the results of our approach by (1) Augmenting the MuST-C corpus ST data using synthetic data (Jia et al. 2019). We first train the MT model using the Transformer network on the WMT16 MT dataset and later generate new translations using MuST-C transcripts. We combine the generated translation and original speech signal to create the synthetic training point. (2) Training wordpiece (Sennrich, Haddow, and Birch 2016) vocabulary based models, instead of character-based models. We adopt (Gu et al. 2018a) to create a universal vocabulary based on all the tasks present in our meta-learning approach.

Our meta-learning approach with these additional improvements achieves new state-of-the-art results on En-De and En-Fr ST tasks by obtaining a BLEU score of 22.11 and 34.05, respectively, and surpasses cascaded system (`Baseline 4`). The ASR model in the cascaded system is trained on the SWC corpus and fine-tuned on MuST-C transcripts, and the MT model is pre-trained on WMT16 and fine-tuned on Must-C transcript-translation data.

4 Related Work

End-to-End Speech Translation: Traditionally, speech translation is implemented as a cascade of ASR and MT (Ney 1999; Post et al. 2013). However, it has its own limitations. Starting with the attempt to align source speech and target translation text without transcription (Duong et

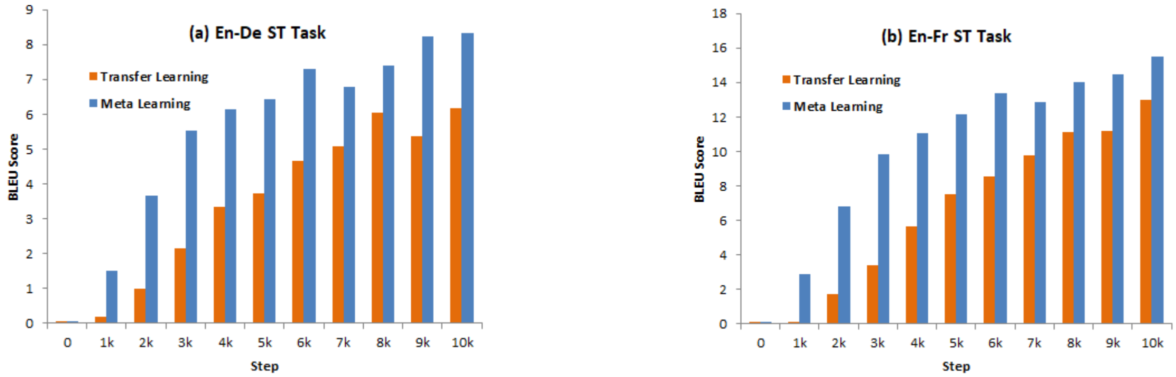

Figure 2: ST model performance on testset obtained from the checkpoints 0 to 10k.

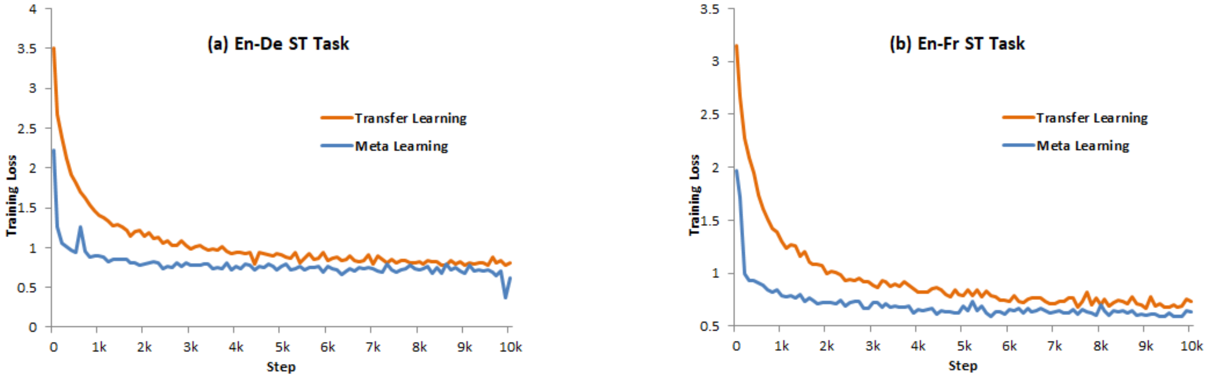

Figure 3: ST model loss from the training steps 0 to 10k.

al. 2016; Anastasopoulos, Chiang, and Duong 2016), several works have been proposed to realize the end-to-end speech translation system (Bérard et al. 2016; Weiss et al. 2017; Bérard et al. 2018). However, collecting such huge speech-to-translation corpus is relatively more challenging compared to the collection of MT and ASR corpora. This challenge leads to attempting various methods to moderate its paucity, including data augmentation with MT or TTS (Text-to-Speech) models, and utilization of data of other related tasks by employing transfer learning. Augmenting training data with synthesized audio using TTS is also adopted (Bérard et al. 2016; Kano, Sakti, and Nakamura 2018; Jia et al. 2019). Several variants of transfer learning approach such as multi-task learning have been explored by simultaneously training either ASR+ST or MT+ST pairs (Weiss et al. 2017; Liu et al. 2019). However, the performance gains were similar to the transfer learning approach (Bérard et al. 2018). The above approaches result in sub-optimal solutions for the target ST task due to the reasons discussed in Section 1.

Meta-Learning: In general, meta-learning, or learning-to-learn, aims to solve the problem of adapting to new tasks with few examples. The meta-learning focuses more on learning aspects instead of wholly focusing on a particular task at hand. Several approaches have been proposed for meta-learning to acquire an ability of fast adaptation. Bengio et al.; Andrychowicz et al.; Ha, Dai, and Le (1992; 2016; 2016) approach the meta-learning by learning a meta-policy, while Finn, Abbeel, and Levine; Vinyals et al. (2017; 2016) learn to find a good initialization point for a new task. Our work is based on the later approaches, specifically, it is based on the recent model agnostic meta-learning (MAML, Finn, Abbeel, and Levine (2017)) that can be readily applied to any gradient descent based neural network. Our work is similar in spirit to the work of low resource neural machine translation (Gu et al. 2018b). However, we focus on adapting meta-learning to tasks with different input modalities and solve the more challenging ST task.

5 Conclusion

In this work, we introduce a modality agnostic meta-learning to solve the low resource end-to-end speech translation task. The proposed approach adopts from MAML and extends it to work on tasks with different modalities during the meta-learning phase. Our approach has several benefits. It makes use of vast amounts of data available from MT and ASR tasks and does not share parameters across the source and target tasks. It finds a good initialization point during the meta-learning using the source tasks=ASR+MT and adapts quickly to the target ST task during the fine-tuning phase. To test the effectiveness of the proposed approach, we conducted several experiments on En-De and En-Fr ST tasks. Our approach significantly outperforms the existing

TS	**(En)** So the same as we saw before.
OT	**(De)** Also genauso, wie wir es vorher gesehen haben. **(BT)** Just as we have seen before.
TL	**(De)** Das Gleiche gilt für uns. **(BT)** The same applies to us.
ML	**(De)** Das Gleiche haben wir also schon früher gesehen. **(BT)** So we saw the same thing earlier.
TS	**(En)** This is what, in engineering terms, you would call a real time control system.
OT	**(De)** Dies würden Sie, in Ingenieurteams, eine Echtzeit-Kontrollsystem nennen. **(BT)** You would call this, in engineering teams, a real-time control system.
TL	**(De)** Das ist es, was man im Ingenieurssystem nennen könnte. **(BT)** That's what you could call in the engineering system.
ML	**(De)** Das ist es, was man in Ingenieurwissenschaften als Echtzeit-Kontrollsystem bezeichnen könnte. **(BT)** That's what you could call a real-time control system in engineering.
TS	**(En)** They even can bring with them some financing.
OT	**(Fr)** Ils peuvent même apporter avec eux des financements. **(BT)** They can even bring with them funding.
TL	**(Fr)** Ils peuvent même les amener avec eux et financer. **(BT)** They can even bring them with them and finance.
ML	**(Fr)** Ils peuvent même apporter avec eux des financements. **(BT)** They can even bring with them funding.
TS	**(En)** She grew up at a time when Confucianism was the social norm and the local mandarin was the person who mattered.
OT	**(Fr)** Elle a grandi à une poque où le confucianisme était la norme sociale et le mandarin local était la personne qui importait. **(BT)** She grew up at a time when Confucianism was the social norm and local Mandarin was the person who mattered.
TL	**(Fr)** Elle a grandi à un moment où le confisherisme était le norme social, et la mandarine locale était la personne qu'il avait importée. **(BT)** She grew up at a time when Confisherism was the social norm, and the local mandarin was the person he had imported.
ML	**(Fr)** Elle a grandi à une époque où le confucianisme était la norme sociale, et la mandarine locale était la personne qui comptait. **(BT)** She grew up in a time when Confucianism was the social norm, and the local mandarin was the person who counted.

Table 4: Sample translation from transfer and meta-learning (TL, ML) approches for En-De and En-Fr ST tasks. We provided transcripts (TS), original translations (OT), and back translations (BT) from De/Fr→En to help the the readers.

approaches of transfer learning on both the ST tasks. We further improved the performance of the proposed method by augmented synthetic data and using wordpiece vocabularies.

The proposed approach brings new opportunities to build efficient end-to-end ST systems with a limited amount of training data. First, the approach incorporates ASR and MT tasks in a principled way to leverage additional sources of data. Second, it is a generic framework that can comfortably accommodate existing and future end-to-end ST models.

6 Acknowledgments

We thank Yoonsuck Choe for valuable discussions and comments.

References

[Anastasopoulos and Chiang 2018] Anastasopoulos, A., and Chiang, D. 2018. Tied multitask learning for neural speech translation. *arXiv preprint arXiv:1802.06655*.

[Anastasopoulos, Chiang, and Duong 2016] Anastasopoulos, A.; Chiang, D.; and Duong, L. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. *arXiv preprint arXiv:1609.08139*.

[Andrychowicz et al. 2016] Andrychowicz, M.; Denil, M.; Gomez, S.; Hoffman, M. W.; Pfau, D.; Schaul, T.; Shillingford, B.; and De Freitas, N. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, 3981–3989.

[Bahdanau, Cho, and Bengio 2015] Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate.

[Baumann, Köhn, and Hennig 2019] Baumann, T.; Köhn, A.; and Hennig, F. 2019. The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening. *Lang. Resour. Eval.* 53(2):303–329.

[Bengio et al. 1992] Bengio, S.; Bengio, Y.; Cloutier, J.; and Gecsei, J. 1992. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, 6–8. Univ. of Texas.

[Bérard et al. 2016] Bérard, A.; Pietquin, O.; Servan, C.; and Besacier, L. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

[Bérard et al. 2018] Bérard, A.; Besacier, L.; Kocabiyikoglu, A. C.; and Pietquin, O. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6224–6228. IEEE.

[Di Gangi et al. 2019] Di Gangi, M. A.; Cattoni, R.; Bentivogli, L.; Negri, M.; and Turchi, M. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the As-*

sociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2012–2017. Minneapolis, Minnesota: Association for Computational Linguistics.

[Duong et al. 2016] Duong, L.; Anastasopoulos, A.; Chiang, D.; Bird, S.; and Cohn, T. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 949–959.

[Finn, Abbeel, and Levine 2017] Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.

[Graves, Mohamed, and Hinton 2013] Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 38.

[Gu et al. 2018a] Gu, J.; Hassan, H.; Devlin, J.; and Li, V. O. 2018a. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 344–354. New Orleans, Louisiana: Association for Computational Linguistics.

[Gu et al. 2018b] Gu, J.; Wang, Y.; Chen, Y.; Cho, K.; and Li, V. O. 2018b. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

[Ha, Dai, and Le 2016] Ha, D.; Dai, A.; and Le, Q. V. 2016. Hypernetworks. *arXiv preprint arXiv:1609.09106*.

[Jia et al. 2019] Jia, Y.; Johnson, M.; Macherey, W.; Weiss, R. J.; Cao, Y.; Chiu, C.-C.; Ari, N.; Laurenzo, S.; and Wu, Y. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7180–7184. IEEE.

[Kano, Sakti, and Nakamura 2018] Kano, T.; Sakti, S.; and Nakamura, S. 2018. Structured-based curriculum learning for end-to-end english-japanese speech translation. *arXiv preprint arXiv:1802.06003*.

[Liu et al. 2019] Liu, Y.; Xiong, H.; He, Z.; Zhang, J.; Wu, H.; Wang, H.; and Zong, C. 2019. End-to-end speech translation with knowledge distillation. *CoRR* abs/1904.08075.

[Ney 1999] Ney, H. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, 517–520. IEEE.

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Pascanu et al. 2013] Pascanu, R.; aglar Gülehre; Cho, K.; and Bengio, Y. 2013. How to construct deep recurrent neural networks. *CoRR* abs/1312.6026.

[Post et al. 2013] Post, M.; Kumar, G.; Lopez, A.; Karakos, D.; Callison-Burch, C.; and Khudanpur, S. 2013. Improved speech-to-text translation with the fisher and callhome spanishenglish speech translation corpus. In *International Workshop on Spoken Language Translation (IWSLT 2013)*.

[Sennrich, Haddow, and Birch 2016] Sennrich, R.; Haddow, B.; and Birch, A. 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)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 3104–3112. Cambridge, MA, USA: MIT Press.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

[Vaswani et al. 2018] Vaswani, A.; Bengio, S.; Brevdo, E.; Chollet, F.; Gomez, A. N.; Gouws, S.; Jones, L.; Kaiser, L.; Kalchbrenner, N.; Parmar, N.; Sepassi, R.; Shazeer, N.; and Uszkoreit, J. 2018. Tensor2tensor for neural machine translation. *CoRR* abs/1803.07416.

[Vinyals et al. 2016] Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, 3630–3638.

[Weiss et al. 2017] Weiss, R. J.; Chorowski, J.; Jaitly, N.; Wu, Y.; and Chen, Z. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.