ACL 2020 paper recap

P. R. Sullivan[1]

[1]School of Information
UBC

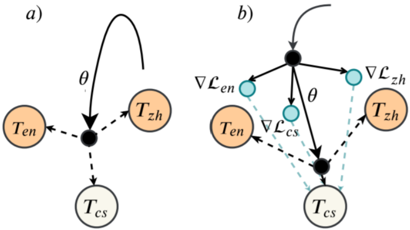
UBC DL-NLP, July 2020

Table of Contents

MAML for Code-switched speech recognition

Model Agnostic Meta-Learning is the process of training on different tasks (or corpora) to allow for fast adaptation to any specific training task.

Joint-training (a) vs. MAML (b)



MAML for Code-switched speech recognition

Use Meta-learning to harness large monolingual dataset through updating parameters on how it does at a task and then calculating the final loss based on how that update would do on the target dataset.

Algorithm 1 Meta-Transfer Learning

Require: $\mathscr{D}_{src}, \mathscr{D}_{tgt}$
Require: α, β: step size hyperparameters

Randomly initialize θ
while not done **do**
  Sample batch data $\mathcal{D}^{tra} \sim (\mathscr{D}_{src}, \mathscr{D}_{tgt})$, $\mathcal{D}^{val} \sim \mathscr{D}_{tgt}$
  for all $\mathcal{D}^{tra}_{\mathcal{T}_i} \in \mathcal{D}^{tra}$ **do**
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{D}^{tra}_{\mathcal{T}_i}}(f_\theta)$ using $\mathcal{D}^{tra}_{\mathcal{T}_i}$
    Compute adapted parameters with gradient descent: $\theta'_{\mathcal{T}_i} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{D}^{tra}_{\mathcal{T}_i}}(f_\theta)$
  end for
  $\theta \leftarrow \theta - \beta \sum_i \nabla_\theta \mathcal{L}_{\mathcal{D}^{val}} \left(f_{\theta'_{\mathcal{T}_i}} \right)$
end while

[1] From [1]

MAML for Code-switched speech recognition

This approach gives a modest improvement over a jointly-pre-train/fine-tune procedure, at the expense of increased memory cost.

Model	**CER**
Winata et al. (2019)	32.76%
+ Pointer-Gen LM	31.07%
Only *CS*	34.51%
Joint Training (*EN* + *ZH*)	98.29%
+ Fine-tuning	31.22%
Joint Training (*EN* + *CS*)	34.77%
Joint Training (*ZH* + *CS*)	33.93%
Joint Training (*EN* + *ZH* + *CS*)	32.87%
+ Fine-tuning	31.90%
+ Pointer-Gen LM	31.74%
Meta-Transfer Learning (*EN* + *CS*)	32.35%
Meta-Transfer Learning (*ZH* + *CS*)	31.57%
Meta-Transfer Learning (*EN* + *ZH* + *CS*)	30.30%
+ Fine-tuning	29.99%
+ Pointer-Gen LM	**29.30%**

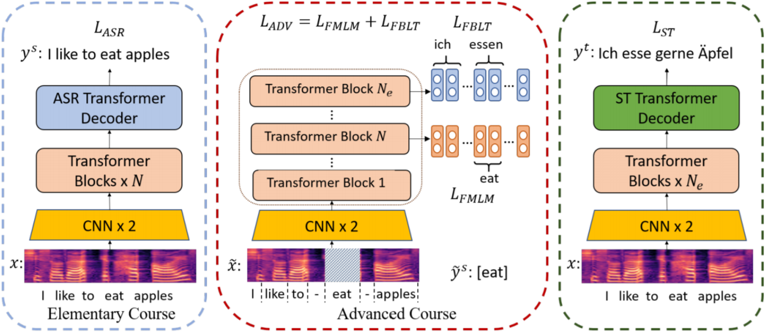
Table of Contents

Curriculum Pre-training for End-to-end speech translation

Curriculum training is the process of scaffolding the difficulty of training examples, so that models are more likelier to converge. In this paper, they apply this to pre-training on different intermediate tasks (ASR, then bilingual lexicon prediction, finally translation).

Curriculum Pre-training for End-to-end speech translation

- Start using ASR
- Transition to predicting segments of audio based on layer.
- Add a decoder in final stage.



[1] [2]

Curriculum Pre-training for End-to-end speech translation

- Lots of work for around .6 BLEU
- However, on expanded setting near MT limit.

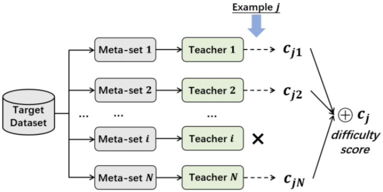
Method	Enc pre-train	Dec pre-train	BLEU
MT(Berard et al., 2018)*	-	-	19.3
MT(Inaguma et al., 2019)	-	-	18.3
base setting			
LSTM ST (Berard et al., 2018)*			12.9
+pre-train+multitask (Berard et al., 2018)*	✓	✓	13.4
LSTM ST+pre-train (ESPnet)	✓	✓	16.68
Transformer+pre-train (Liu et al., 2019)	✓	✓	14.30
+knowledge distillation(Liu et al., 2019)			17.02
TCEN-LSTM (Wang et al., 2019b)	✓	✓	17.05
Transformer+ASR pre-train	✓		15.97
Transformer+curriculum pre-train	✓		**17.66**
expanded setting			
LSTM+pre-train+SpecAugment(Bahar et al., 2019)	✓(236h)	✓	17.0
Multilingual ST+pre-train (Inaguma et al., 2019)	✓(472h)		17.6
Transformer+ASR pre-train	✓(960h)		16.90
Transformer+curriculum pre-train	✓(960h)		**18.01**

[1] [2]

Table of Contents

Curriculum Learning for Natural Language Understanding

- Split up examples in training set into N meta-sets.
- Train a teacher model based on each of these sets.
- Score each training item (via Teachers)
- Sort training items into buckets
- Train by sampling from buckets, moving to harder buckets as training continues.



[1] [3]

Curriculum Learning for Natural Language Understanding

	MNLI-m	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	Avg
results on dev									
BERT Large	**86.6**	92.3	91.3	70.4	93.2	88.0	60.6	90.0	84.1
BERT Large*	**86.6**	92.5	91.5	74.4	93.8	91.7	63.5	90.2	85.5
BERT Large+CL	**86.6**	**92.8**	**91.8**	**76.2**	**94.2**	**91.9**	**66.8**	**90.6**	**86.4**
results on test									
BERT Large	**86.7**	91.1	89.3	70.1	**94.9**	89.3	60.5	87.6	83.7
BERT Large*	86.3	92.2	**89.5**	70.2	94.4	89.3	60.5	87.3	83.7
BERT Large+CL	**86.7**	**92.5**	**89.5**	**70.7**	94.6	**89.6**	**61.5**	**87.8**	**84.1**

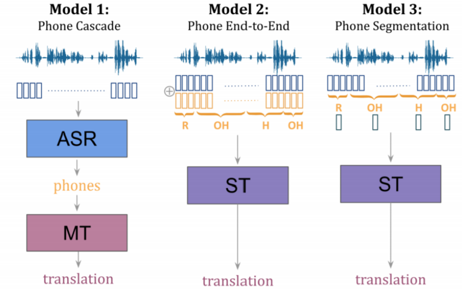
Method	**SQuAD 2.0**		**Δ**
	EM	**F1**	
No Curriculum	-	76.30	-
No Curriculum*	73.66	76.78	-
Rarity+**Annealing**	73.75	76.90	+0.12
Answer+**Annealing**	74.02	77.15	+0.37
Question+**Annealing**	74.35	77.37	+0.59
Paragraph+**Annealing**	74.45	77.54	+0.76
Cross-Review+*Naive order*	74.31	77.29	+0.51
Cross-Review+Annealing	**74.96**	**77.93**	**+1.15**

[1] [3]

Table of Contents

Phone Features Improve Speech Translation

- Use seq2seq model to generate per-frame phone features (e.g. /R/)
- concat with audio and feed to ST model
- Results show 10 point BLEU increase with High resource setting (160hr) and 22 point increase with low setting (20hr)



[1] [4]

References I

[1] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-transfer learning for code-switched speech recognition," *arXiv preprint arXiv:2004.14228*, 2020.

[2] C. Wang, Y. Wu, S. Liu, M. Zhou, and Z. Yang, "Curriculum pre-training for end-to-end speech translation," *arXiv preprint arXiv:2004.10093*, 2020.

[3] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, and Y. Zhang, "Curriculum learning for natural language understanding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6095–6104.

[4] E. Salesky and A. W. Black, "Phone features improve speech translation," *arXiv preprint arXiv:2005.13681*, 2020.