exploiting adapters for cross-lingual low-resource speech recognition

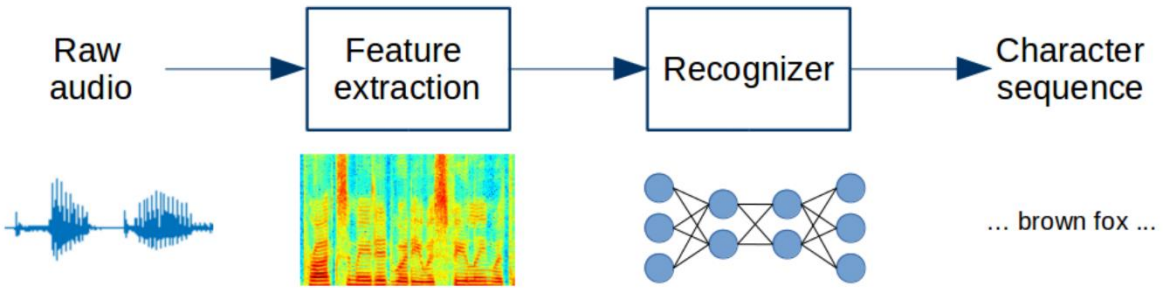
Presenter: Aein Koupaei

Hou, W., Zhu, H., Wang, Y., Wang, J., Qin, T., Xu, R., & Shinozaki, T. (2021). Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 317-329.

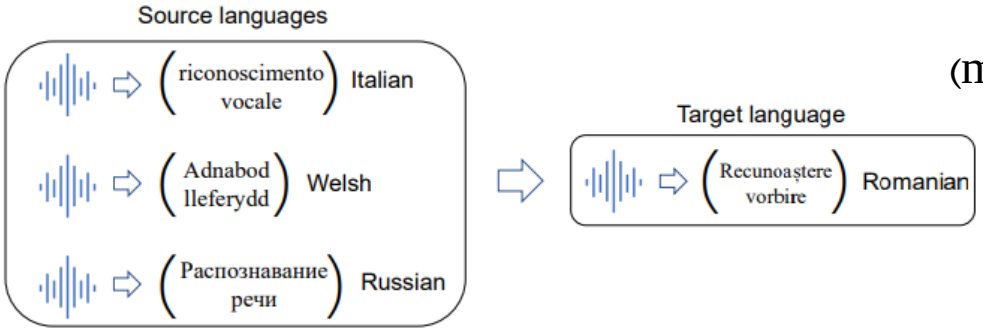
سیستم بازشناسی گفتار E2E (End-to-end ASR)

- دنباله‌ای از ویژگی‌های صوتی را به دنباله‌ای از حروف یا کلمات نگاشت می‌کند.
- برای رسیدن به یک عملکرد خوب به داده‌های آموزشی زیادی نیاز دارد:

  دارای عملکرد ضعیف برای زبان‌های کم منبع



- توسعه سیستم‌های ASR چندزبانه(multilingual) و استفاده از یادگیری انتقالی برای کمک به سیستم‌های ASR در برخورد با زبان‌های کم منبع



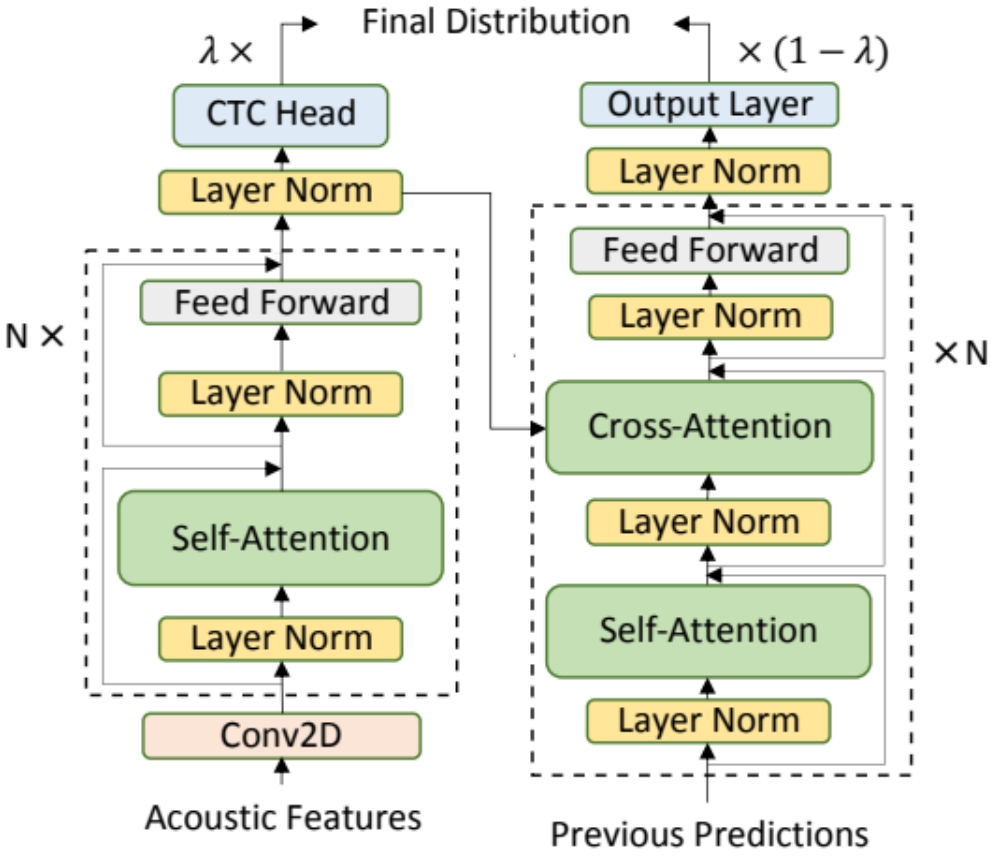
مشکلات تطبیق مدل چندزبانه با زبان هدف:

- Fine-tune کردن کل پارامترهای یک مدل چندزبانه با مقیاس بزرگ، از نظر محاسباتی پرهزینه و پیچیده است.
- عملکرد نامناسب مدل و وقوع overfitting در شرایطی که زبان هدف بسیار کم منبع باشد.
- تفسیرناپذیری مدل: عدم آگاهی از اینکه چه زبان‌هایی دارای اشتراکات بیشتری با زبان هدف هستند.

راه حل پیشنهادی:

- افزودن ماژول adapter به مدل و fine-tune کردن پارامترهای adapter به جای پارامترهای کل مدل
- استفاده از رویکرد meta-learning در ترکیب با adapter برای غلبه بر مشکل overfiting در زبان‌های بسیار کم منبع
- استفاده از مکانیزم توجه برای یادگیری دانش مشترک بین زبان‌ها و افزایش تفسیرپذیری مدل

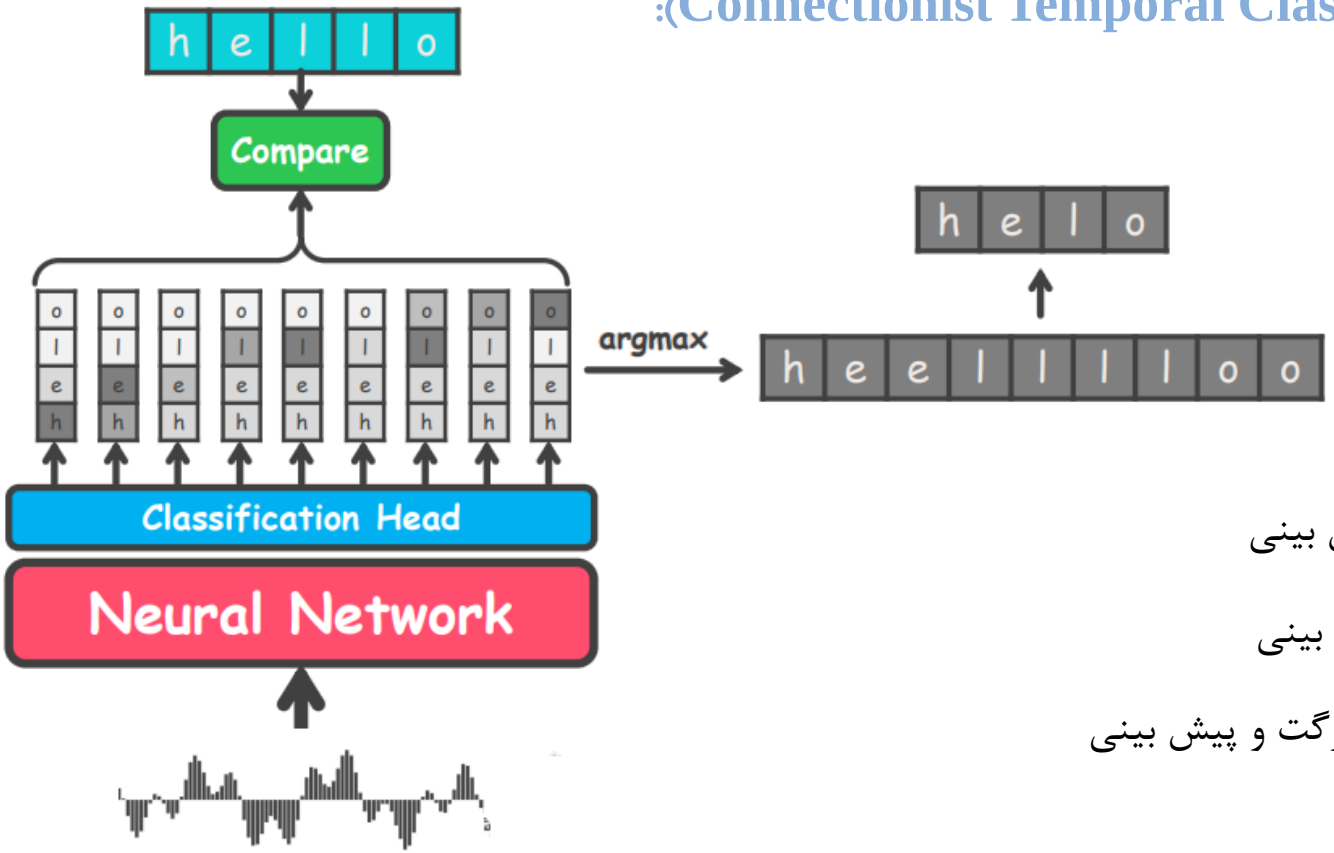
Super Multilingual Transformer ASR Model



- Backbone روش پیشنهادی یک مدل ASR ۴۲ زبانه مستقل از زبان (language-independant) مبتنی بر transformer است. ← LID-42
- ورودی مدل: ویژگی‌های صوتی ۸۳ بعدی (فیلتر بانک‌های ۸۰ بعدی و ویژگی‌های pitch سه بعدی) محاسبه شده با فریم شیفت ۱۰ میلی‌ثانیه و طول فریم ۲۵ میلی‌ثانیه
- استفاده از ساختار ترکیبی CTC-attention

$$\mathcal{L}_{\text{ASR}} = (1-\lambda)\mathcal{L}_{\text{ATT}} + \lambda\mathcal{L}_{\text{CTC}}$$

Connectionist Temporal Classification) CTC Loss:

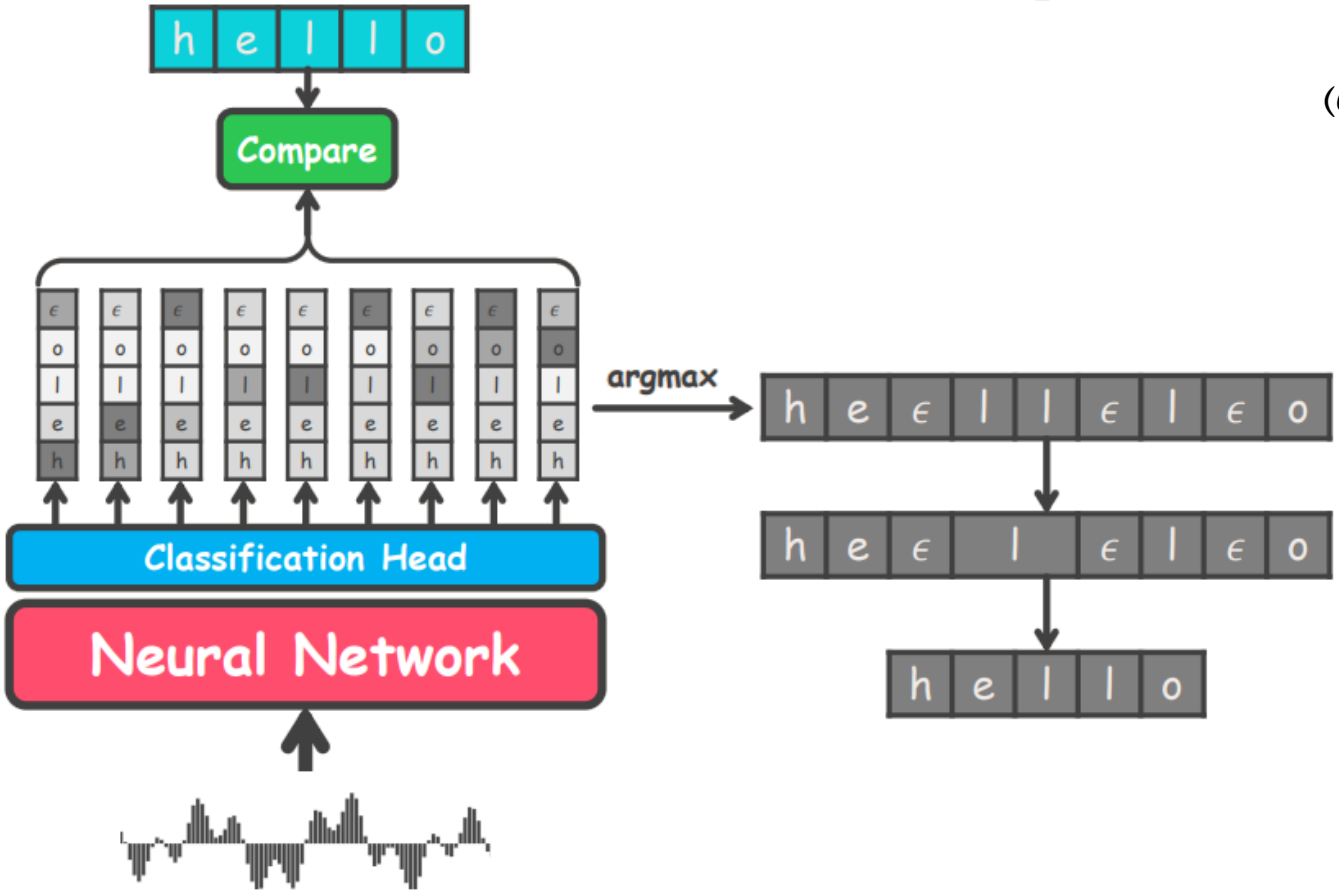


چالش‌ها:

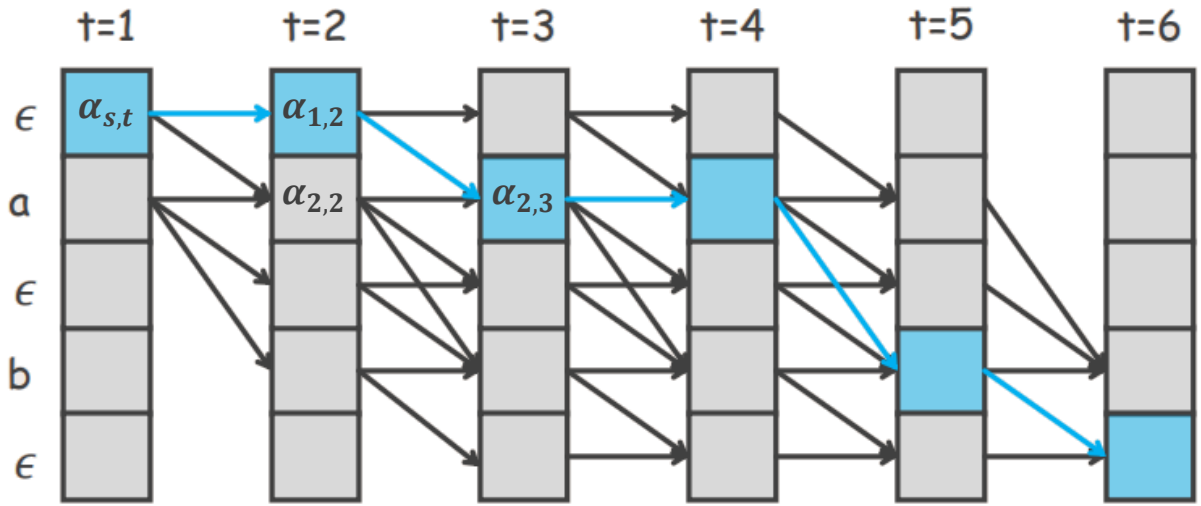
- طول متفاوت تارگت و پیش بینی
- نسبت متغیر تارگت و پیش بینی
- عدم امکان تطبیق دقیق تارگت و پیش بینی

CTC Loss (Connectionist Temporal Classification):

- افزودن توکن خالی (ϵ)



CTC Loss (Connectionist Temporal Classification):

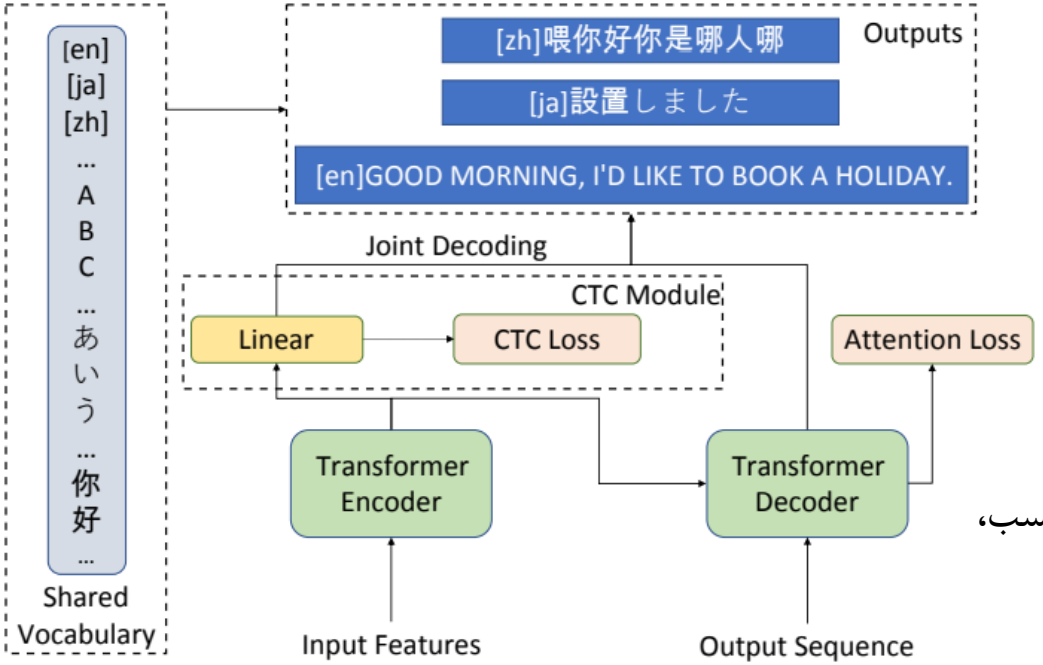


- نحوه محاسبه اتلاف

$$\alpha_{2,3} = (\alpha_{1,2} + \alpha_{2,2})y_{2,3}$$

$$P(Y|X) = \sum_{A} \prod_{t=1}^{T} P_t(y_t|X) \qquad L = -\sum logP(Y|X)$$

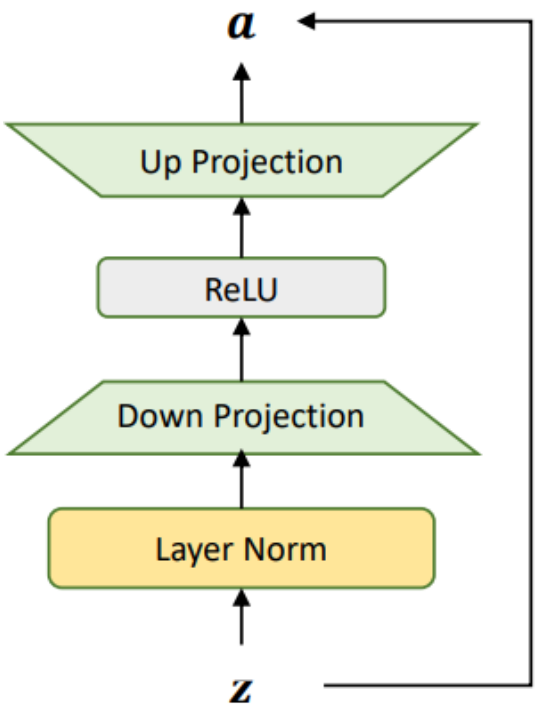
All alignments



- تحقق آموزش و تشخیص مستقل از زبان:

  استفاده از یک shared vocabulary شامل کاراکترها/زیرکلمات و توکن‌های نشان دهنده زبان (<en>, <fr>) مربوط به ۴۲ زبان

- افزودن توکن زبان در ابتدای هر برچسب
- افزودن توکن نشان‌دهنده زبان در ابتدای هر برچسب، تشخیص زبان پیش از تشخیص محتویات گفتار
- آموزش LID-42 بر روی حدود ۵۰۰۰ ساعت داده گفتاری برچسب‌دار (ترکیب ۱۱ پیکره شامل ۴۲ زبان)



ساختار متداول در Adapter شامل:

- یک layer normalization
- یک لایه down-projection
- یک تابع فعالساز غیرخطی
- یک لایه up-projection

$$\mathbf{a}^l = \text{Adapter}(\mathbf{z}^l) = \mathbf{z}^l + \mathbf{W}_u^l \text{ReLU}\left(\mathbf{W}_d^l \left(\text{LN}\left(\mathbf{z}^l\right)\right)\right)$$

layer normalization

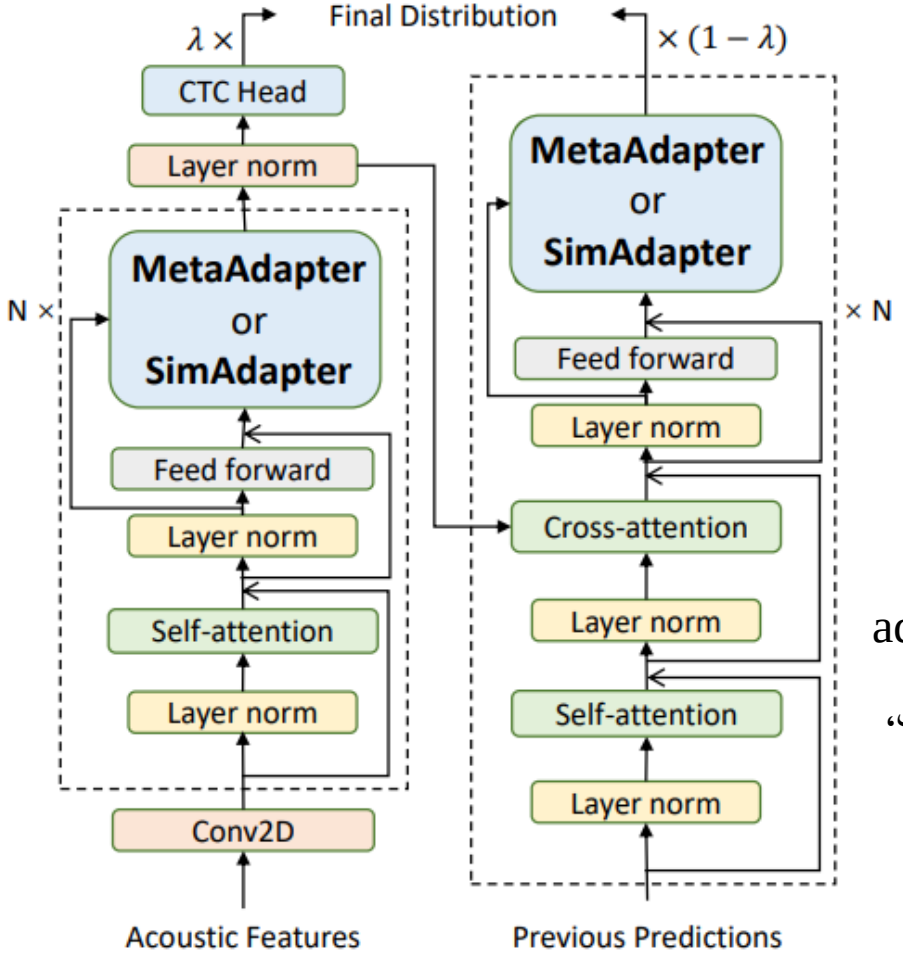
خروجی لایه L

وزن‌های لایه up-projection

وزن‌های لایه down-projection

ماژول MetaAdapter

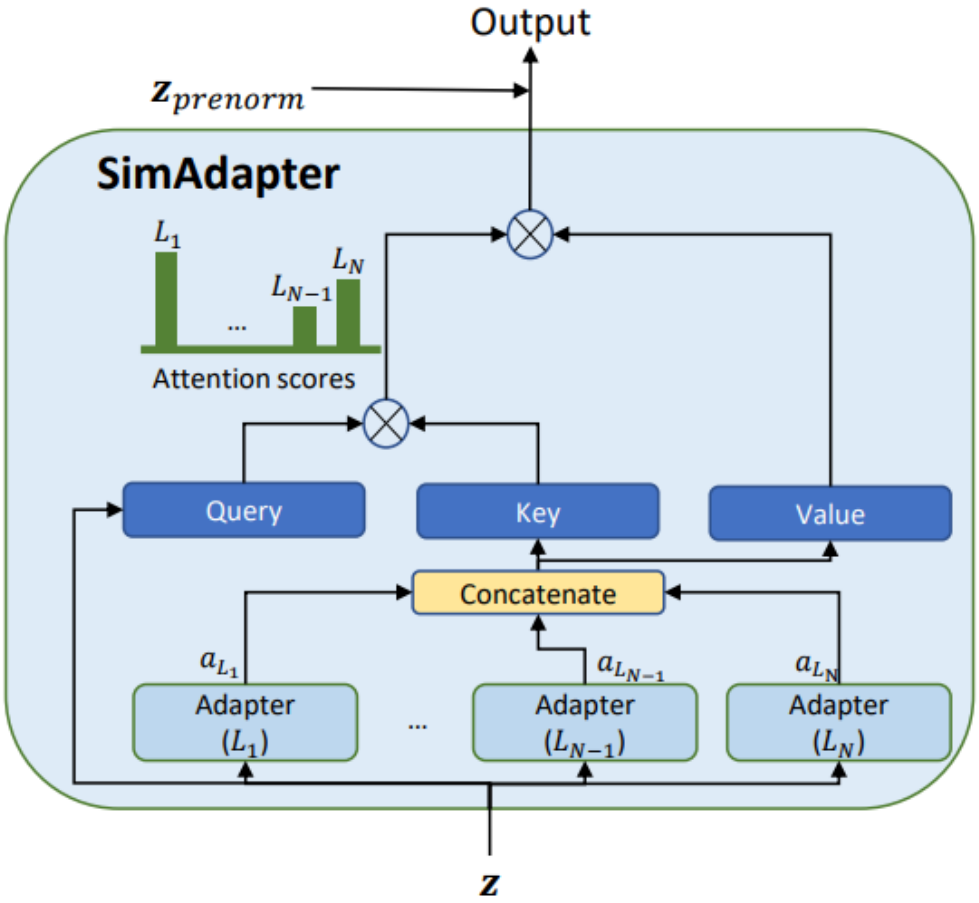
- ترکیب ماژول adapter با رویکرد meta-learning
- بهبود عملکرد مدل برای زبان‌های بسیار کم منبع
- استفاده از پارامترهای کمتر و در نتیجه تسریع تطبیق
- آموزش MetaAdapter با استفاده از الگوریتم MAML (Model-Agnostic Meta-Learning) به دلیل مقاومت نسبت به overfiting
- افزودن ماژول MetaAdapter به مدل speech-transformer ASR آموزش دیده
- دارای دو فاز : ۱. pre-training : adapter ها روی تعدادی زبان source، meta-train می‌شوند
  ۲. fine-tuning رو زبان هدف



- دارای دو گروه پارامتر :
  ۱. پارامترهای backbone (θ_b)
  ۲. پارامترهای adapter (θ_a)
- پارامترهای backbone در طول آموزش و fine-tune ماژول MetaAdapter فریز می‌شوند.
- برای N زبان source، پارامترهای adapter آموزش می‌بینند و سپس بر روی زبان هدف، fine-tune می‌شوند.

Meta Adapter

- زبان‌های مختلف بر اساس ویژگی‌های geological و تحولات فرهنگی شباهت‌هایی دارند.
- ماژول SimAdapter با یکسری زبان منبع آموزش می‌بیند.
- به تعداد زبان‌های مبداء + زبان هدف، adapter دارد.
- با استفاده از مکانیزم توجه، شباهت بین زبان‌های منبع با زبان هدف یاد گرفته می‌شود.



$$\text{SimAdapter}(\mathbf{z}, \mathbf{a}_{\{S_1,S_2,\ldots,S_N\}})$$

$$= \sum_{i=1}^{N} \text{Attn}(\mathbf{z}, \mathbf{a}_{S_i}) \cdot (\mathbf{a}_{S_i}\mathbf{W}_V)$$

Language-agnostic features

Language-specific features

$$\text{Attn}(\mathbf{z}, \mathbf{a}) = \text{Softmax}\left(\frac{(\mathbf{z}\mathbf{W}_Q)(\mathbf{a}\mathbf{W}_K)^\top}{\tau}\right)$$

- W_Q و W_k به صورت رندم مقداردهی اولیه می‌شوند اما برای W_v عناصر روی قطر با مقدار ۱ و بقیه ماتریس با مقدار ۰.۰۰۰۰۰۱ ($1e-6$) مقداردهی اولیه می‌شود (با هدف حفظ بازنمایی‌های adapter).
- برای جلوگیری از تغییرات شدید ویژگی، یک عبارت regularization معرفی شده است:

$$\mathcal{L}_{\text{reg}} = \sum_{i,j} \left((\mathbf{I}_V)_{i,j} - (\mathbf{W}_V)_{i,j} \right)^2$$

$\mathbf{I}_V$: identity matrix with the same size as W_v

- Fusion guide loss برای هر لایه fusion: با هدف توجه مناسب ماژول به adapter زبان هدف

لایه fusion

$$\mathcal{L}^f_{\text{guide}} = -\frac{1}{K \times S} \sum_{s=1}^{S} \sum_{k=1}^{K} \log \alpha^s_{f,k}$$

$\alpha^s_{f,k}$: Attention score of target language's adapter

در encoder: تعداد فریم
در decoder: تعداد توکن

S: تعداد sample

$$\mathcal{L}_{\text{guide}} = \sum_f \mathcal{L}^f_{\text{guide}}$$

- Loss کل:

$$\mathcal{L} = \mathcal{L}_{\text{ASR}} + \eta \mathcal{L}_{\text{reg}} + \gamma \mathcal{L}_{\text{guide}}$$

SimAdapter+

- Adapter های منبع را با adapter هدفی که با MetaAdapter آموزش داده‌ایم، با استفاده از SimAdapter ترکیب می‌کنیم.
- می‌تواند به عنوان فرآیند انتقال دانش دو مرحله‌ای در نظر گرفته شود.

دیتاست:

- Common Voice 5.1
- انتخاب ۵ زبان rich-resource به عنوان منبع و ۵ زبان low-resource به عنوان هدف

TRAINING / VALIDATION / TESTING HOURS

	Language	Train	Valid	Test
Source	Russian (ru)	80.61	11.78	12.61
	Welsh (cy)	74.84	4.35	4.25
	Italian (it)	88.74	19.74	20.85
	Basque (eu)	73.26	7.46	7.85
	Portuguese (pt)	37.40	5.40	5.85
Target	Romanian (ro)	3.04	0.42	1.66
	Czech (cs)	20.66	2.84	3.13
	Breton (br)	2.84	1.51	1.75
	Arabic (ar)	7.87	2.01	2.09
	Ukrainian (uk)	17.35	2.30	2.36

COMPARISON OF NUMBER OF TRAINABLE PARAMETERS.

Method	# Trainable Parameters
Hybrid DNN/HMM	14,247K
Full Model	27,235K
Head	77K
Head+(Meta-)Adapter	676K
Head+(Meta-)Adapter+SimAdapter	4,224K

WORD ERROR RATES (WER) ON THE CROSS-LINGUAL ASR TASKS

Target	DNN/HMM	Trans.(B)	Trans.(S)	Head	Full-FT	Full-FT+L2	Part-FT	Adapter	SimAdapter	MetaAdapter	SimAdapter+
Romanian (ro)	70.14	97.25	94.72	63.98	53.90	52.74	52.92	48.34	47.37	**44.59**	47.29
Czech (cs)	63.15	48.87	51.68	75.12	34.75	35.80	54.66	37.93	35.86	37.13	**34.72**
Breton (br)	-	97.88	92.05	82.80	61.71	61.75	66.24	58.77	**58.19**	58.47	59.14
Arabic (ar)	69.31	75.32	74.88	81.70	47.63	50.09	58.49	47.31	47.23	46.82	**46.39**
Ukrainian (uk)	77.76	64.09	67.89	82.71	**45.62**	46.45	66.12	50.84	48.73	49.36	47.41
AVG	-	76.68	76.24	77.26	48.72	49.37	59.69	48.64	47.48	47.27	**46.99**
Weighted AVG	-	72.28	72.50	77.54	46.72	47.50	59.43	47.38	46.08	46.12	**45.45**

Refrences

- Hou, W., Zhu, H., Wang, Y., Wang, J., Qin, T., Xu, R., & Shinozaki, T. (2021). Exploiting adapters for cross-lingual low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 317-329.

- Kim, S., Hori, T., & Watanabe, S. (2017, March). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4835-4839). IEEE.

- Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J., & Shinozaki, T. (2020). Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. Babel, 37(4k), 10k.

- Hou, W., Wang, Y., Gao, S., & Shinozaki, T. (2021, June). Meta-adapter: Efficient cross-lingual adaptation with meta-learning. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7028-7032). IEEE.