Low Resource Speech Recognition

SiRui Li

2022/09/09

Improving Low-Resource Speech Recognition with Pretrained Speech Models: Continued Pretraining vs. Semi-Supervised Training

- **Task**

CoPT vs. SST on the XLSR-53 pretrained model in several low-resource languages

- **Motivation**

finetuning dependent on the amount of in-language or similar-to-in-language data included in the pretraining dataset

- **Datasets**

Build sets： Georgian, Farsi, Somali（36-65h） and Tagalog（150h） （CS NB TB）

BABEL sets： Georgian Tagalog （CS）

Youtube sets： For each language （ 1000 h ）

- **Experiments**

For each language, we consider four different CoPT setups:

- None: The baseline setup, no CoPT
- BUILD: Using the **build** data for CoPT
- YouTube: Using the 1000 hour unlabeled in-language YouTube audio for CoPT
- BUILD + YouTube: Using both the **build** data and corresponding 1000 hour unlabeled in-language YouTube audio for CoPT

In all four setups we start from the publicly available XLSR-53 pretrained model[3].

Table 1: *WER results comparing CoPT with different pretraining data with and without SST*

CoPT	Finetuning	Georgian	Farsi	Somali	Tagalog
None	BUILD	18.7	30.7	51.1	34.6
None	BUILD + SST	18.0	27.9	50.4	26.6
BUILD	BUILD	18.5	32.1	50.4	33.1
BUILD	BUILD + SST	17.7	27.5	50.2	26.1
YouTube	BUILD	17.5	27.6	49.4	26.6
YouTube	BUILD + SST	17.7	27.3	49.9	25.8
BUILD + YouTube	BUILD	**17.4**	27.2	**48.9**	26.5
BUILD + YouTube	BUILD + SST	17.6	27.2	49.9	25.8
BUILD + YouTube	BUILD + SST Best	17.6	**26.5**	49.0	**24.4**

Table 2: *MATERIAL **analysis** WER by genre with finetuning using **build** only.*

Language	CoPT	CS	NB	TB
Georgian	None	33.8	11.9	19.0
	BUILD + YT	32.7	10.7	17.4
Farsi	None	41.5	26.6	30.8
	BUILD + YT	38.2	23.1	27.2
Somali	None	59.0	47.1	52.1
	BUILD + YT	56.4	44.6	49.5
Tagalog	None	45.2	25.3	35.6
	BUILD + YT	43.5	19.0	27.6

Table 3: *BABEL dev set WER with CoPT with finetuning using corresponding train set.*

Language	CoPT	WER
Georgian	None	31.9
	YT	30.7
Tagalog	None	36.3
	YT	35.2

- **Summarize**

CoPT provides similar to, or better, results than SST
CoPT and SST are complementary
CoPT performs best with in-domain data

Adaptive Activation Network for Low Resource Multilingual Speech Recognition

- **Task**

Adaptive activation network：cross-lingual learning、 multilingual learning

- **Motivation**

The existing models mostly established a bottleneck (BN) layer by pre-training on a large source language, and transferring to the low resource target language.

- **Datasets**

IARPA Babel： source (Guarani, Igbo, Lithuanian) target (Amharic, Cantonese)

- **Methods**

Adaptive Activation Network：

$$O_n^l = F_n^l(W_n O_{n-1}^l + b_n) \tag{8}$$

different language l at n-th layer. The defination of F_n^l is a set of basis functions as follows:

$$F_n^l(\cdot) = \sum_{i=1}^{M} \lambda_n^l(i)\sigma_i(\cdot) \tag{9}$$

where $\{\sigma_i\}_{i=1}^M$ represents M different basis activation function, and $\lambda_n^l = [\lambda_n^l(1), ..., \lambda_n^l(M)] \in \mathbb{R}^M$ are the coordinates of bases $\{\sigma_i\}_{i=1}^M$. In this work, we also chose the adaptive piecewise linear (APL) activation units [20], to parameterize $F_n^l(\cdot)$. The definition of $F_n^l(\cdot)$ is as following:

$$F_n^l(x) = max(0, x) + \sum_{i=1}^{M} \lambda_n^l(i) max(0, -x + b_i) \tag{10}$$

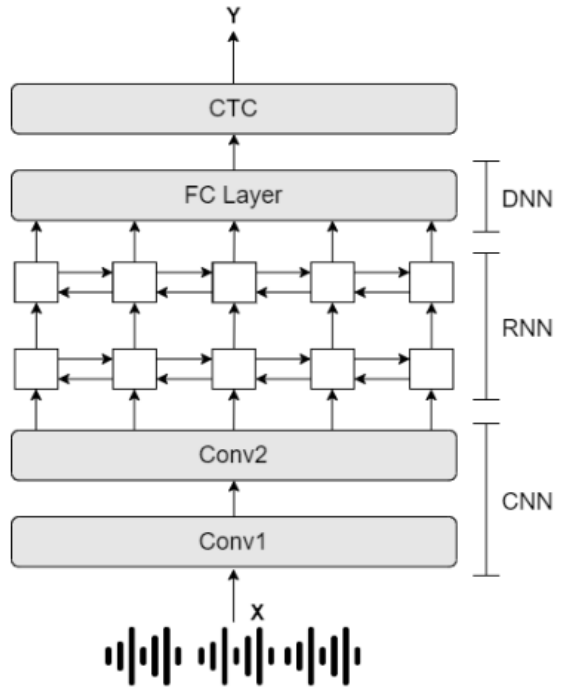
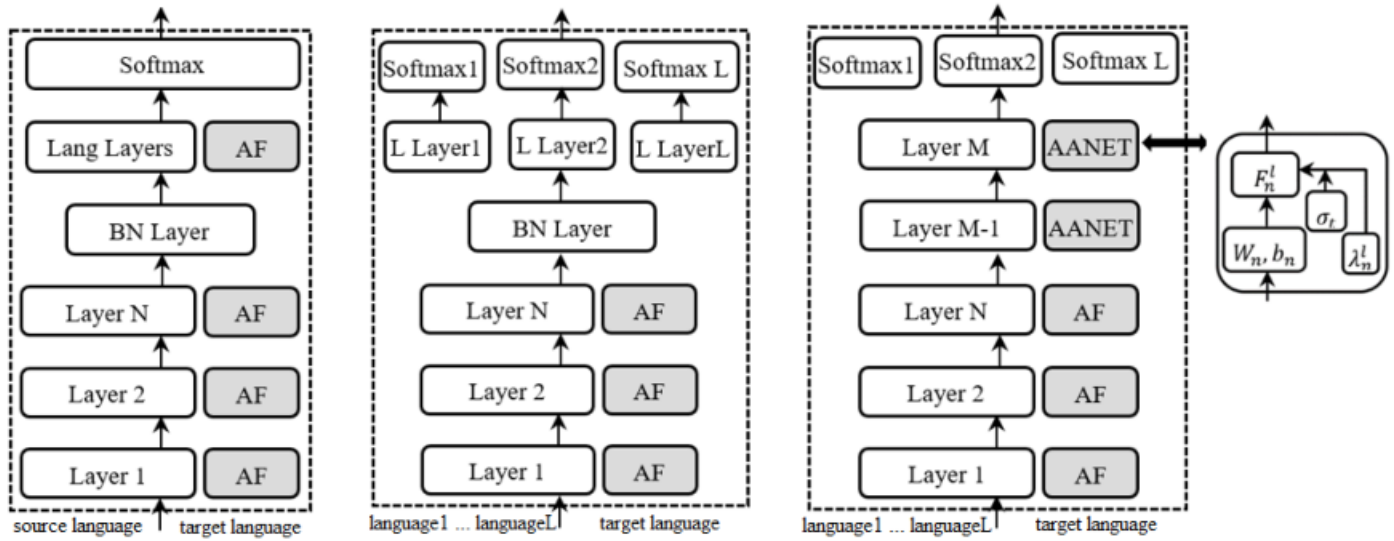


Fig. 1. The architecture of CNN-RNN-DNN network for low resource speech recognition task. The bottom convolutional layers (CNN) extract the local features of the audio. The middle recurrent layers (RNN) model the long-time dependency of the feature sequence. The upper deep neural layers (DNN) map the features to the vocabulary of target language. The whole network is trained by the CTC loss between the output hidden states and gold text sequence.



(a) Traditional Transfer Learning (b) Traditional Multilingual Structure (c) The Proposed AANET Architecture

Fig. 2. The traditional transfer learning (a) and multilingual (b) structures are training the models in source large or multilingual corpus, to produce the bottleneck (BN) layer features. And then, the upper layers are finetuned in the target corpus for target ASR task. By contrast, the proposed Adaptive Activation Network (AANET) architecture (c) only replaces the Activation Function (AF) of the upper layers, and uses these AANET to model the relevance and difference among different languages.

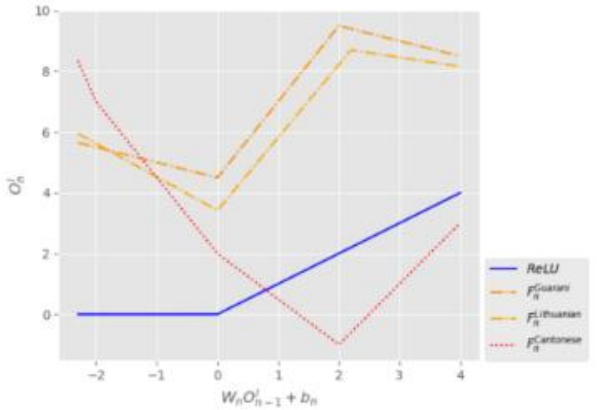


Fig. 3. The tradition ReLU activation function (blue) and the adaptive activation networks of Guarani, Lithuanian and Cantonese (orange, dark orange and red) in layer n. The adaptive activation networks of Guarani and Lithuanian are more similar, and differ from that of Cantonese a lot. Adaptive activation network introduces more non-linearity into the neural network and increases the learning capability of ASR system.

Cross-lingual Learning

$$\mathcal{L}_{pre} = \mathcal{L}_{ctc}^{l_0} \tag{11}$$

More specifically, our cross-lingual learning method is: (1) pre-training the model, which contains adaptive activation network $F_n^{l_0}$, in a large source corpus l_0 by the source CTC loss $\mathcal{L}_{ctc}^{l_0}$, (2) applying a new adaptive activation network $F_n^{l_1}$ to the upper layers, and maintaining the weight and bias paramters, (3) fine-tuning these new adaptive activations for the target language l_1 by the target CTC loss $\mathcal{L}_{ctc}^{l_1}$. It should be noted that only upper layers' activation functions are replaced. Because bottom layers are leveraged to extract speech feature, and could not distinguish different unique features among different languages.

$$\mathcal{L}_{fine} = \mathcal{L}_{ctc}^{l_1} \tag{12}$$

Multilingual Learning

$$\mathcal{L}_{multi} = \sum_{i=1}^{L} \mathcal{L}_{ctc}^{l_i} + \alpha \mathcal{L}_{mtl} \tag{13}$$

We introduced the trace-norm function to reflect the relevance of different languages. The definition of multi-task languages loss $\mathcal{L}_{mtl}$ is as following:

$$\mathcal{L}_{mtl} = trace(\sqrt{\lambda_n \lambda_n^T}) \tag{14}$$

- **Experiments：**

TABLE IV
RESULTS OF DIFFERENT TRAINING STRATEGIES WITH ADAPTIVE ACTIVATION NETWORK, WER (%)

Model	Pre-training Data	Fine-tuning Data	Amharic	Cantonese
CRD-Small + FS	-	Amharic, Cantonese	71.2	58.3
CRD-Small + BN	Guarani, Igbo, Lithuanian	Amharic, Cantonese	69.1	55.1
CRD-Small + CL	Guarani, Igbo, Lithuanian	Amharic, Cantonese	68.2	53.2
CRD-Small + ML	-	Guarani, Igbo, Lithuanian, Amharic, Cantonese	68.9	56.2
CRD-Small + CL & ML	Guarani, Igbo, Lithuanian	Guarani, Igbo, Lithuanian, Amharic, Cantonese	**67.3**	**52.9**
CRD-Large + FS	-	Amharic, Cantonese	68.9	57.7
CRD-Large + BN	Guarani, Igbo, Lithuanian	Amharic, Cantonese	66.3	54.6
CRD-Large + CL	Guarani, Igbo, Lithuanian	Amharic, Cantonese	66.2	51.3
CRD-Large + ML	-	Guarani, Igbo, Lithuanian, Amharic, Cantonese	67.8	54.1
CRD-Large + CL & ML	Guarani, Igbo, Lithuanian	Guarani, Igbo, Lithuanian, Amharic, Cantonese	**66.3**	**51.1**

FS: From-Scratch Training
BN: Bottleneck Features
CL: Cross-lingual Learning
ML: Multilingual Learning

- **Summarize**

  introduced adaptive activation network to the low resource multilingual speech recognition
  Propose a cross-lingual learning approach
  Propose a multilingual learning approach, jointly the CTC loss + trace-norm function
  Combine the cross-lingual learning and multilingual learning together

Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation

- **Task**

learnable and interpretable framework to combine SF and SSL representations

- **Motivation**

the quality of SSL representations depends highly on the relatedness between the SSL training domain(s) and the target data domain

- **Datasets**

Totonac（10h） Arabic from Common voice (20h) Mboshi-French(4h)

- **Methods**

Feature extraction

$$f_{\mathrm{i}}(S) = (f_{\mathrm{i}}^{t}(S) \in \mathbb{R}^{D_{\mathrm{i}}} | t = 1, \cdots, T_{\mathrm{i}}),\ i \in \{\mathrm{SF}, \mathrm{SSL}\} \quad (1)$$

Learnable combinations

$$f_{\mathrm{FUSE}}(S) = \textsc{Linear}(\textsc{Transform}(f_{\mathrm{SF}}(S), f_{\mathrm{SSL}}(S)) \quad (2)$$

$$Q_{\mathrm{i}} = f_{\mathrm{i}}(S)W_{\mathrm{i}}^{\mathrm{Q}}, \quad K_{\mathrm{i}} = f_{\mathrm{i}}(S)W_{\mathrm{i}}^{\mathrm{K}}, \quad V_{\mathrm{i}} = f_{\mathrm{i}}(S)W_{\mathrm{i}}^{\mathrm{V}} \quad (3)$$

$$h_{\mathrm{SF}} = \textsc{SoftMax}(\frac{Q_{\mathrm{SF}} \cdot K_{\mathrm{SSL}}}{\sqrt{D}})V_{\mathrm{SSL}} + f_{\mathrm{SF}}(S) \quad (4)$$

$$h_{\mathrm{SSL}} = \textsc{SoftMax}(\frac{Q_{\mathrm{SSL}} \cdot K_{\mathrm{SF}}}{\sqrt{D}})V_{\mathrm{SF}} + f_{\mathrm{SSL}}(S) \quad (5)$$

$$f_{\mathrm{FUSE}}(S) = \textsc{Linear}(h_{\mathrm{SF}} \parallel h_{\mathrm{SSL}}) \quad (6)$$

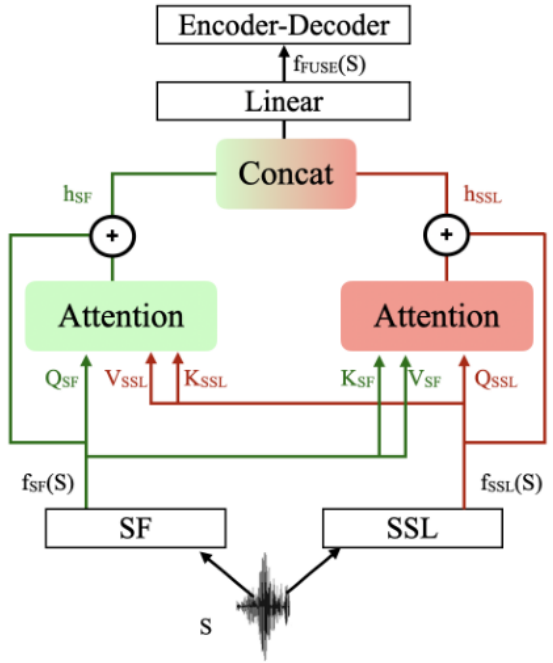


Figure 1: *Architecture of our proposed co-attention based fusion. Raw signal S is passed through SF and SSL feature extractors. The extracted features, $f_{SF}(S)$ and $f_{SSL}(S)$, attend to each other through two distinct attention mechanisms. Output features are then concatenated, projected and passed to the speech model.*

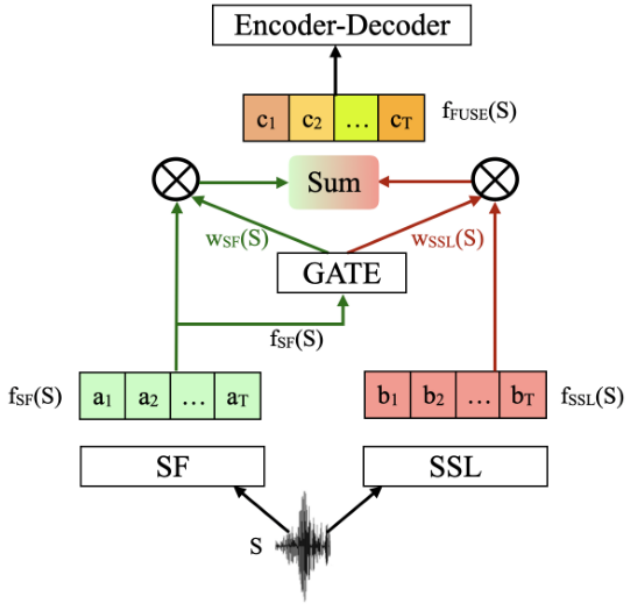


Figure 2: *Architecture of the model combining SF and SSL through a gating mechanism. For a given utterance, the features are extracted by the two models (a_i for SF and b_i for SSL, $i \in \{1, ..., T\}$). Each model gets confidence scores and features are then summed. The c_i variables indicates the weighted sum. Colors of c_i frames are used to show how each frame gets a specific combination of SF (green) and SSL (red) features.*

Eq. (7), where $w(S) \in \mathbb{R}^{T \times 2}$ is the obtained weight matrix.

$$w(S) = \Theta(f_{\text{SF}}(S)W_{\text{MoE}}), \tag{7}$$

with $W_{\text{MoE}} \in \mathbb{R}^{D \times 2}$ a learnable matrix, and $\Theta(\cdot)$ a gating-type function such as SOFTMAX. For clarity, we introduce $w_{\text{SF}}(S)$, $w_{\text{SSL}}(S) \in \mathbb{R}^T$, the column vectors of $w(S)$. The final combined feature is computed following Eq. (8), where $[x]^{\text{tr}}$ denotes the transpose vector of x.

$$f_{\text{FUSE}}(S) = \sum_{i \in \{\text{SF,SSL}\}} [w_{\text{i}}(S)]^{\text{tr}} f_{\text{i}}(S) \tag{8}$$

- **Experiments:**

Table 1: *ASR and ST results over models described in Sec. 4.2. The two first experiments are our FBANK and SSL baselines. The following lines are the proposed Linear, Convolutional, co-Attention, and Mixture of Experts models.*

	CER ↓		BLEU ↑
Exp	Totonac	Arabic	Mboshi-French
Base	17.2	15.4	10.9
SSL	14.2	8.1	10.6
Linear	14.0	6.6	**11.6**
Conv.	13.9	7.2	11.3
co-Att.	**13.4**	**5.4**	10.9
MoE	13.7	6.2	11.2

Table 2: *Two views on HuBERT representations quality over Totonac and Arabic data. The first column presents $\overline{w_{SSL}(S)}$, the mean **MoE** weights for HuBERT front-end. The second column is the character error reduction rate reduction (CERR[8]) between the FBANK baseline and the HuBERT baseline.*

Language	$\overline{w_{\text{SSL}}(S)}$	CERR(**Base** → **SSL**)
Totonac	0.17	17%
Arabic	0.51	47%