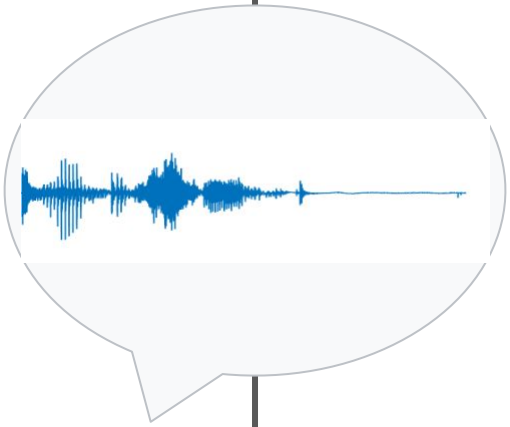
Universal Paralinguistic Speech Representations using Self-Supervised Conformers

Joel Shor, Aren Jansen, Wei Han, Daniel Park, Yu Zhang

1 Problem statement & technique

Problem: Create representation for paralinguistic speech tasks



Lexical

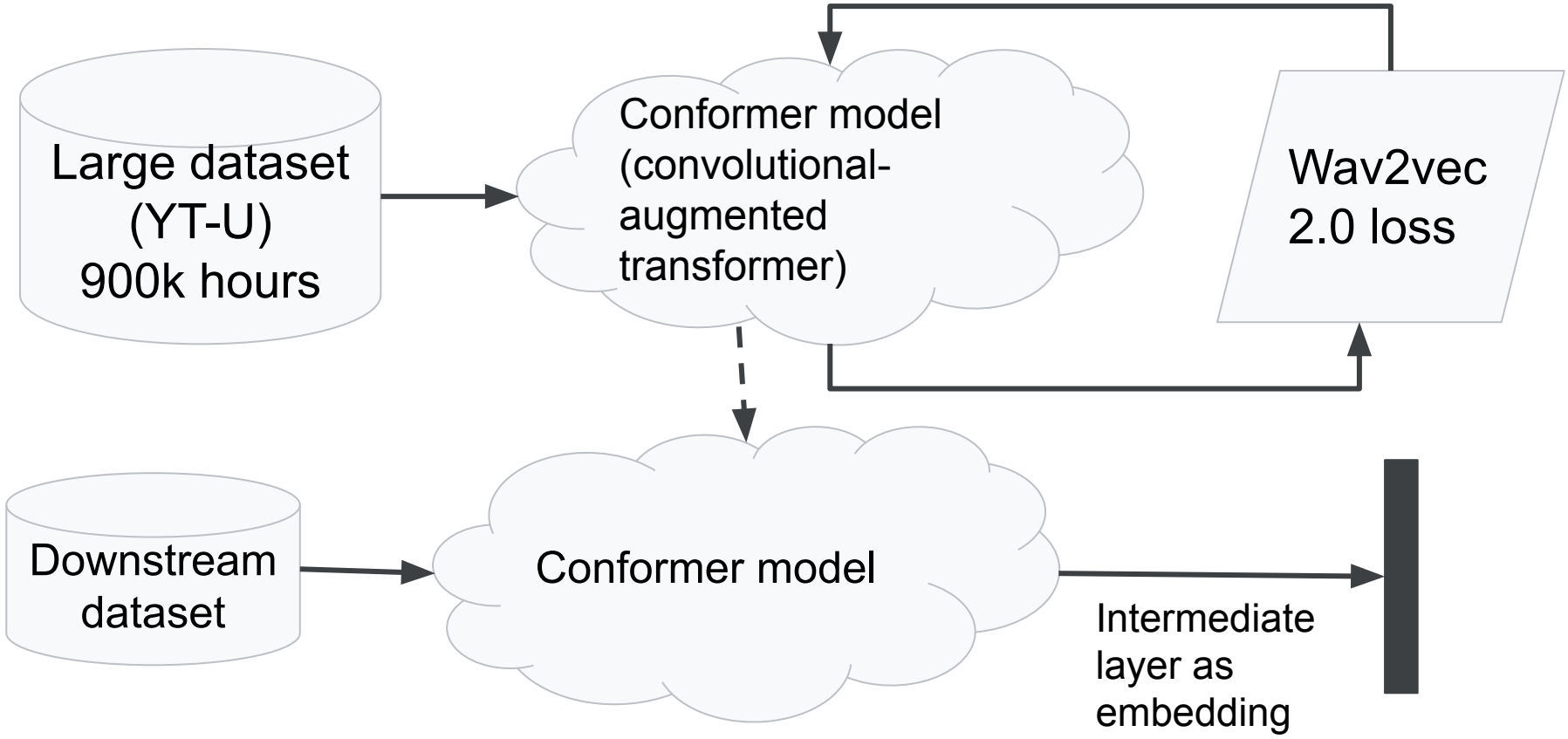
1. Automatic Speech Recognition ex: "Hi, my name is Joel. Nice to meet you!

Paralinguistic

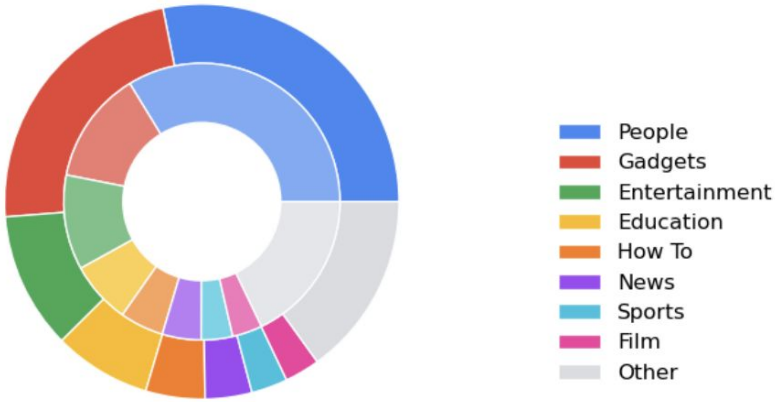
1. Emotion ex."Excitement"
2. Speaker identification
3. Language identification
4. Wearing a mask?
5. Real or fake?
6. Accent
7. Dysarthria

EN/US

Main flow


Large dataset
(YT-U)
900k hours
Conformer model
(convolutional-
augmented
transformer)
Wav2vec
2.0 loss
Downstream
dataset
Conformer model
Intermediate
layer as
embedding

Dataset: YT-U [1]



Video categories by length (outer) and number (inner)

Dataset creation process:

1. Randomly collect 3 million hours of audio from "speech-heavy" YouTube videos, including lectures, news and interviews, filtered by language.
2. Remove non-speech segments to yield approximately a million hours of unlabeled audio data.
3. Uniformly sampled to 16 KHz quality—any audio with a different native sampling rate is either up-sampled or downsampled.

Conformer [1]

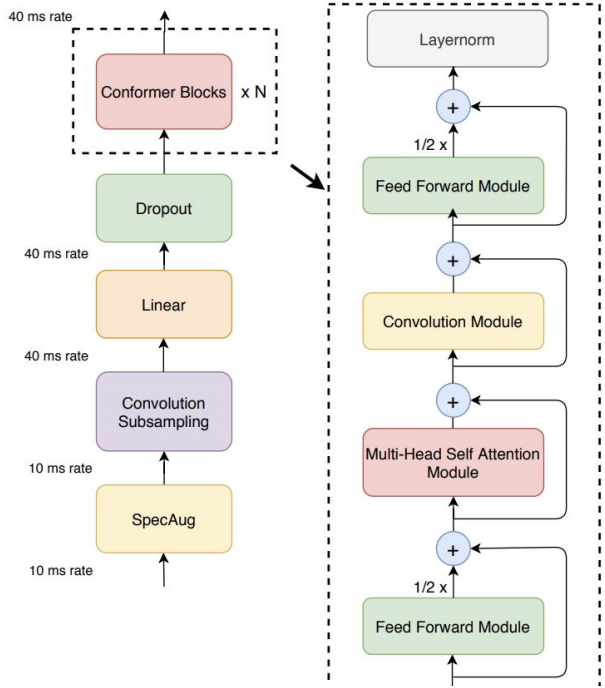
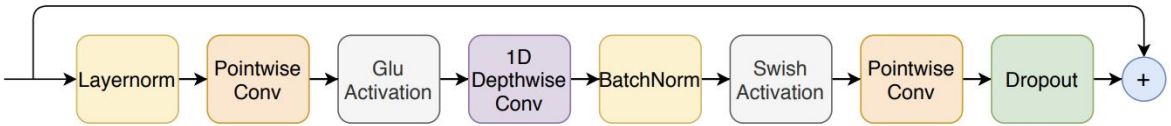
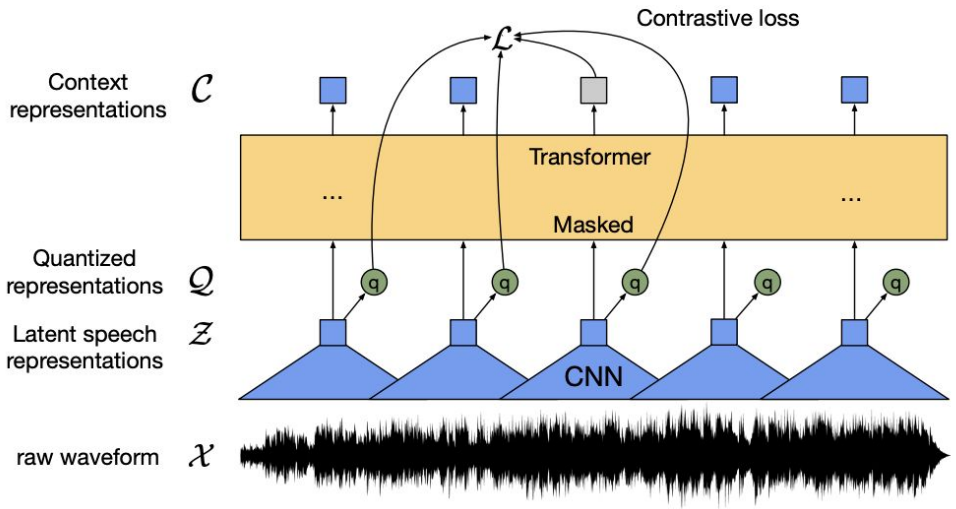


Table 1: Comparison of models. Resnetish50 [10]. MobileNetv3 [27]. RNN-T [28]. EfficientNet [29]. Conformer [14]. AudioSet [30]. YT-U [16], LL is Libri-Light [31]. *"RA" stands for "relative attention."

Name	Architecture	Params	Training data	Labels required
YAMNet [1]	MobileNetv1	3.7M	Audioset	Y
TRILL [1]	Resnetish50	24.5M	Audioset	N
FRILL [18]	MobileNetv3	10.1M	Audioset	N
COLA [2]	EfficientNetB0	4.0M	Audioset	N
ASR Emb [11]	RNN-T	122M	-	Y
Conformer XL (No) RA* YT (LL)	Conformer	608M	YT-U (LL)	N
Conformer XXL YT (LL)	Conformer	1.0B	YT-U (LL)	N
Conformer G	Conformer	8.0B	YT-U	N



Wav2Vec 2.0 training [1]



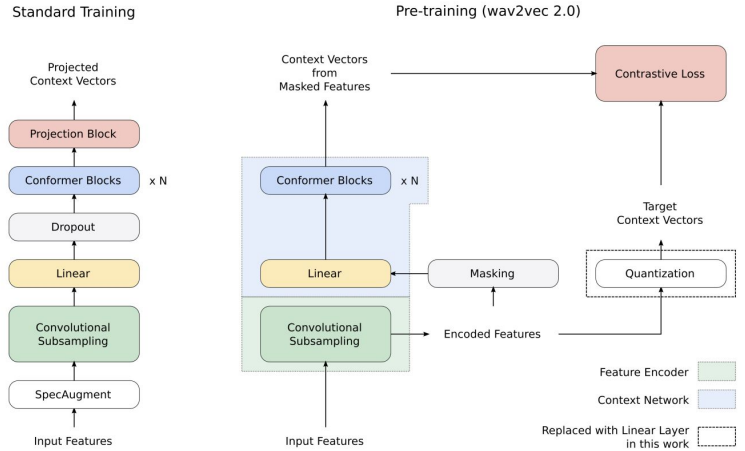
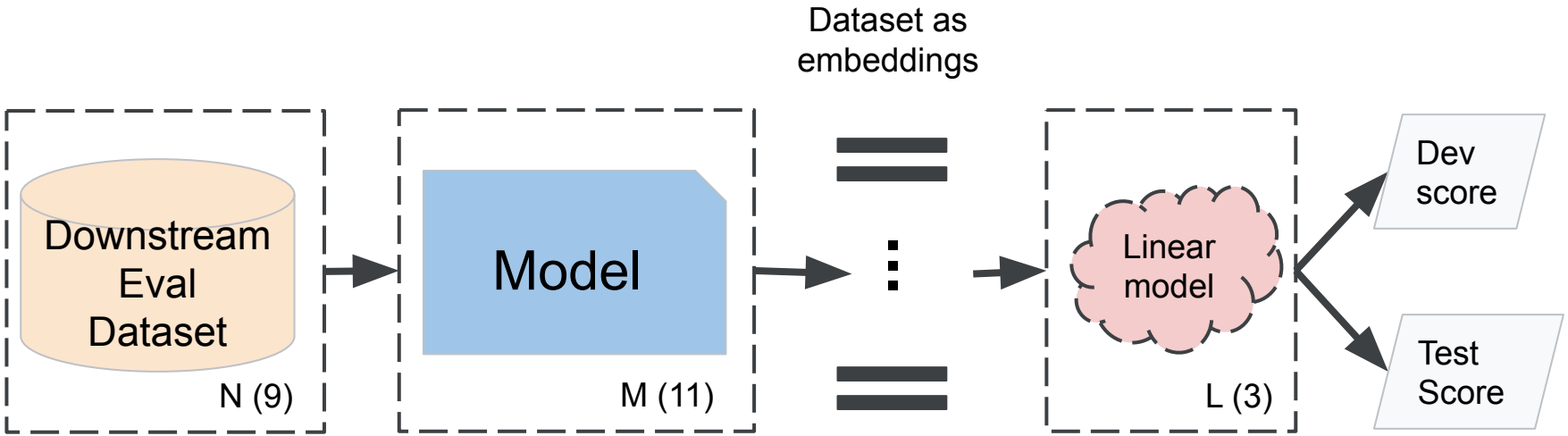


Fig. 3: The Conformer encoder and wav2vec 2.0 pre-training.

Evaluation [1]



1. N downstream eval datasets
2. M candidate embedding models
3. Train linear models (logistic regression, balanced logistic regression, LDA)
4. Best model (by dev set score) performance on test set is the score for that (dataset, model) pair

Evaluation datasets / tasks

Table 2: Downstream evaluation datasets. *Results in our study used a subset of Voxceleb filtered according to YouTube's privacy guidelines.

Dataset	Target	Classes	Samples	Avg length (s)
VoxCeleb* [32]	Speaker ID	1,251	12,052	8.4
VoxForge [33]	Language ID	6	176,438	5.8
Speech Commands[34]	Command	12	100,503	1.0
Masked Speech [19]	Mask wearing	2	36,554	1.0
ASVSpoof [20]	Synthetic or not	2	121,461	3.2
Euphonia [22]	Dysarthria	5	15,224	6.4
CREMA-D [35]	Emotion	6	7,438	2.5
IEMOCAP [21]	Emotion	4	5,531	4.5
SAVEE [36]	Emotion	7	480	3.8

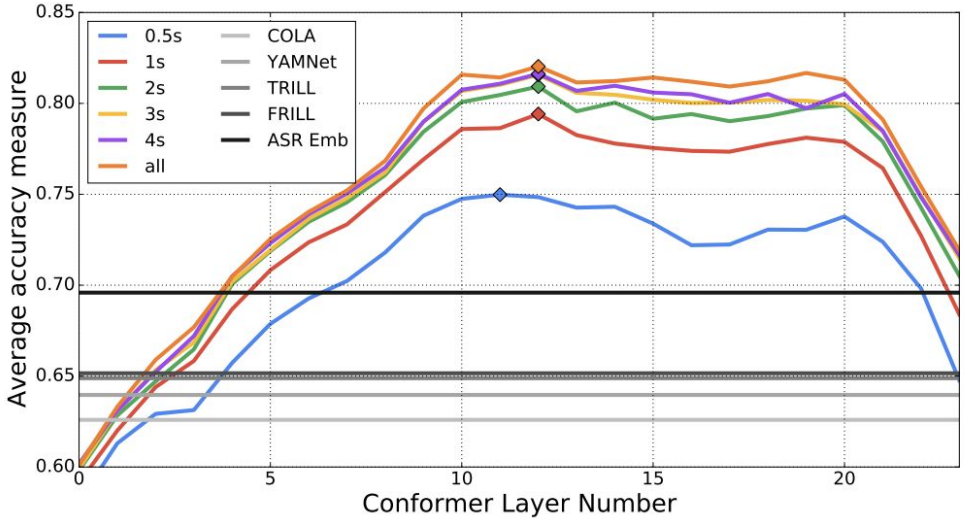
2 Results

Main result [1]

Model	Voxceleb1†	Voxforge	Speech Commands	Masked Speech‡	ASVSpoof 2019**	Euphonia#	CREMA-D	IEMOCAP	SAVEE††
Prev SoTA	-	95.4 [37]	**97.9** [38]	73.0 [39]	5.11 [17]	45.9 [11]	74.0* [40]	67.6+ [17]	84.0* [36]
Baselines									
YAMNet++ [1]	10.9	79.8	78.5	59.7	9.23	43.0	66.4	57.5	69.2
TRILL [1]	12.6	84.5	77.6	65.2	7.46	48.1	65.7	54.3	65.0
FRILL [18]	13.8	78.8	74.4	67.2	7.45	46.6	71.3	57.6	63.3
COLA [2]	11.7	71.0	60.6	65.0	4.58	47.6	69.3	63.9	59.2
ASR Emb [11]	5.2	98.9	96.1	54.4	11.2	**54.5**	71.8	65.4	85.0
Conformers									
Best per-task§ (model, layer #)	**53.5** (XXL-YT, 25)	**99.8** (G-YT, 19)	97.5 (CAP, 16)	**74.2** (XL-LL RA, 5)	**2.5** (CAP, 12)	53.6 (CAP, 13)	87.2 (G, 26)	**79.2** (CAP, 15)	**92.5** (CAP, 15)
Best CAP per task (layer #)	50.3 (11)	99.7 (14)	97.5 (16)	73.4 (10)	**2.5** (12)	53.6 (13)	**88.2**§ (12)	**79.2** (15)	92.5 (15)
Best single layer (CAP12)	51.0§	99.7	97.0	68.9	**2.5**	51.5	**88.2**§	75.0	81.7

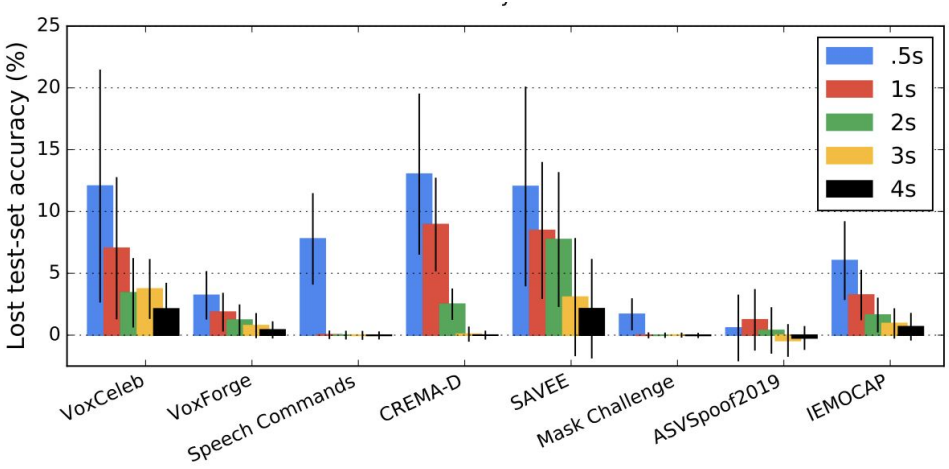
- Top row (**Prev SoTA**) has task-specific models that are arbitrarily complex. The others come from linear models on time-averaged candidate embeddings
- "**Conformers**" are the category of model that we explored in this study
- "**CAP**" is the name of our model with the best overall paralinguistic representations
- "**CAP12**" is the name of the overall best performing representation (layer 12 of CAP)

Analysis 1: How important is the Conformer's large context window?



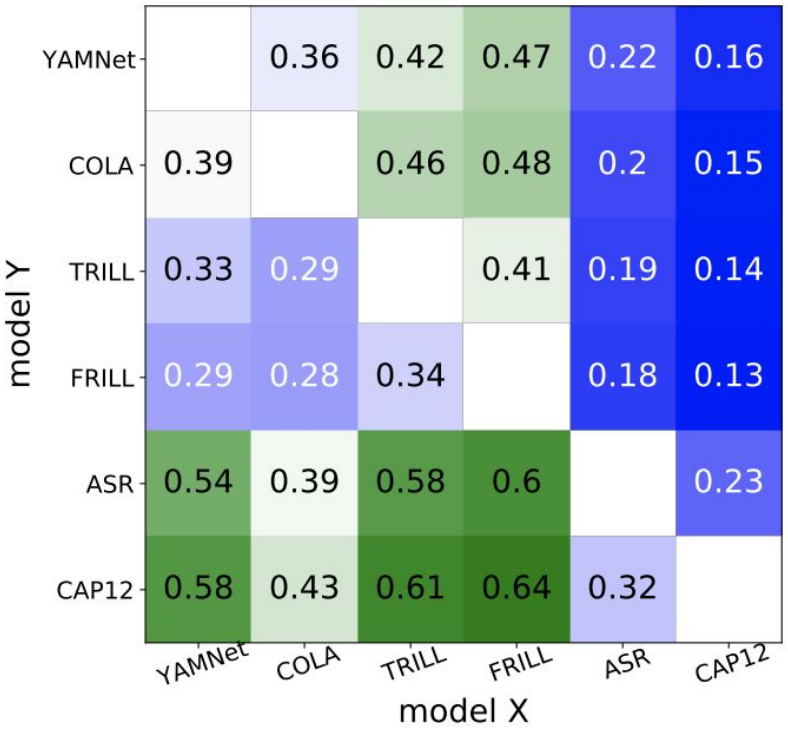
- **3 second** context windows is **99% as performant**
- **2 second** context window is **98% as performant**
- Paralinguistic is **captured** by intermediate layers but **lost** in the final layers

Analysis 2: Which tasks require larger context?



- Speech emotion recognition and speaker ID tasks require larger context windows
- Lang ID, mask challenge, fake speech are fine with 1 sec

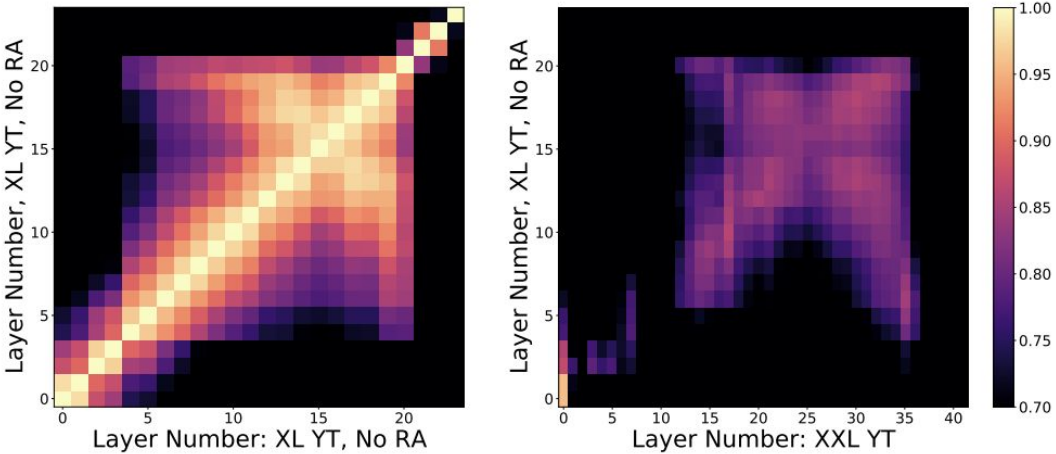
Analysis 3: Are better representations complementary or strictly better?



Each square is the probability that Model Y correctly predicts an example given that Model X and Model Y disagree on the prediction. The result is averaged over task.

- CAP12 strictly outperforms embeddings (except possibly ASR embedding)

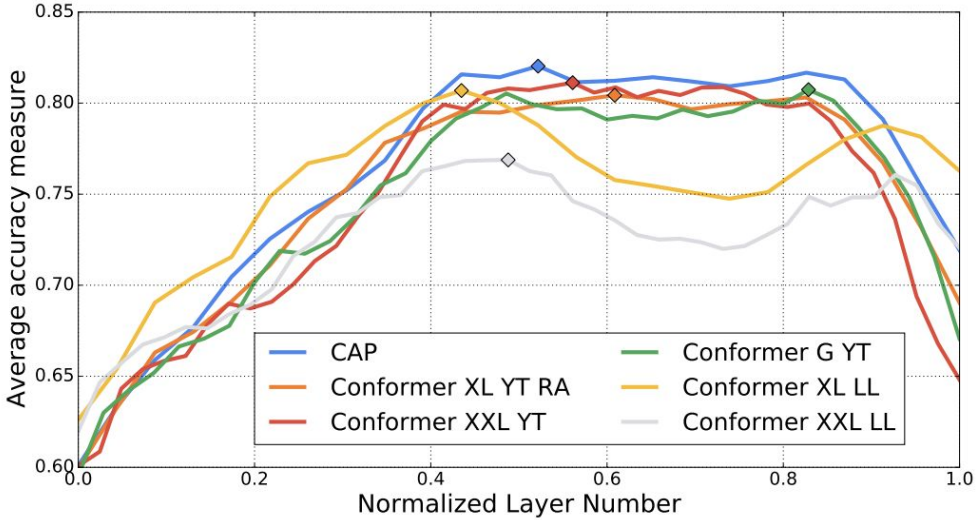
Analysis 1: How similar are representations of different layers, and different networks?



- Peaks occur at similar fractional layer
- Overall shape is similar

Linear CKA scores between all pairs of layers: (left) within the Conformer XL YT network and (right) across the top performing Conformer XL YT and XXL YT networks. The colormap is truncated at 0.7 as is common to both images

Observation 1: The representations are similar between different networks as a fraction of network depth



- Peaks occur at similar fractional layer
- Overall shape is similar

Aggregate NOSS score for 6 different Conformer models as a function of layer index normalized to [0.0, 1.0] using (layer #) / (# of layers), where # of layers is different for different models