MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training

Microsoft Research Asia

정영석

Introduction

▪ Music Understanding

- Music Understanding
  - ✓ including tasks like genre classification, emotion classification, music pieces matching
  - ✓ A better understanding of melody, rhythm, and music structure

    → not only beneficial for music information retrieval but also helpful for music generation
  - ✓ Similar to natural language, music is usually represented in symbolic data format (e.g., MIDI).

    → previous works leverage unlabeled music data to learn music token embeddings, similar to word embeddings in natural language tasks

➢ Unfortunately, due to their shallow structures and limited unlabeled data, such embedding-based approaches have limited capability to learn powerful music representations.

Introduction

▪ Difference between music and language

- Music songs and language is structurally different!!
  - ✓ First, since music songs are more structural (e.g., bar, position) and diverse (e.g., tempo, instrument, and pitch) encoding symbolic music is more complicated than natural language
  - ✓ Song are too long to be processed by pre-trained models
- Requiring other pretext tasks for training music embedding.
  - ✓ The pre-training mechanism (e.g., the masking strategy like the masked language model in BERT) should be carefully designed to avoid information leakage in pre-training
- Scarce dataset for learning music embedding

Introduction

▪ Contribution

1. We pre-train MusicBERT on a large-scale symbolic music corpus that contains more than 1 million music songs and fine-tune MusicBERT on some music understanding tasks, achieving state-of-the-art results

2. We propose OctupleMIDI, an efficient and universal music encoding for music understanding, which leads to much shorter encoding sequences and is universal for various kinds of music.

3. We design a bar-level masking strategy as the pre-training mechanism for MusicBERT, which significantly outperforms the naive token-level masking strategy used in natural language pretraining.

Methodology

- ## Model architecture

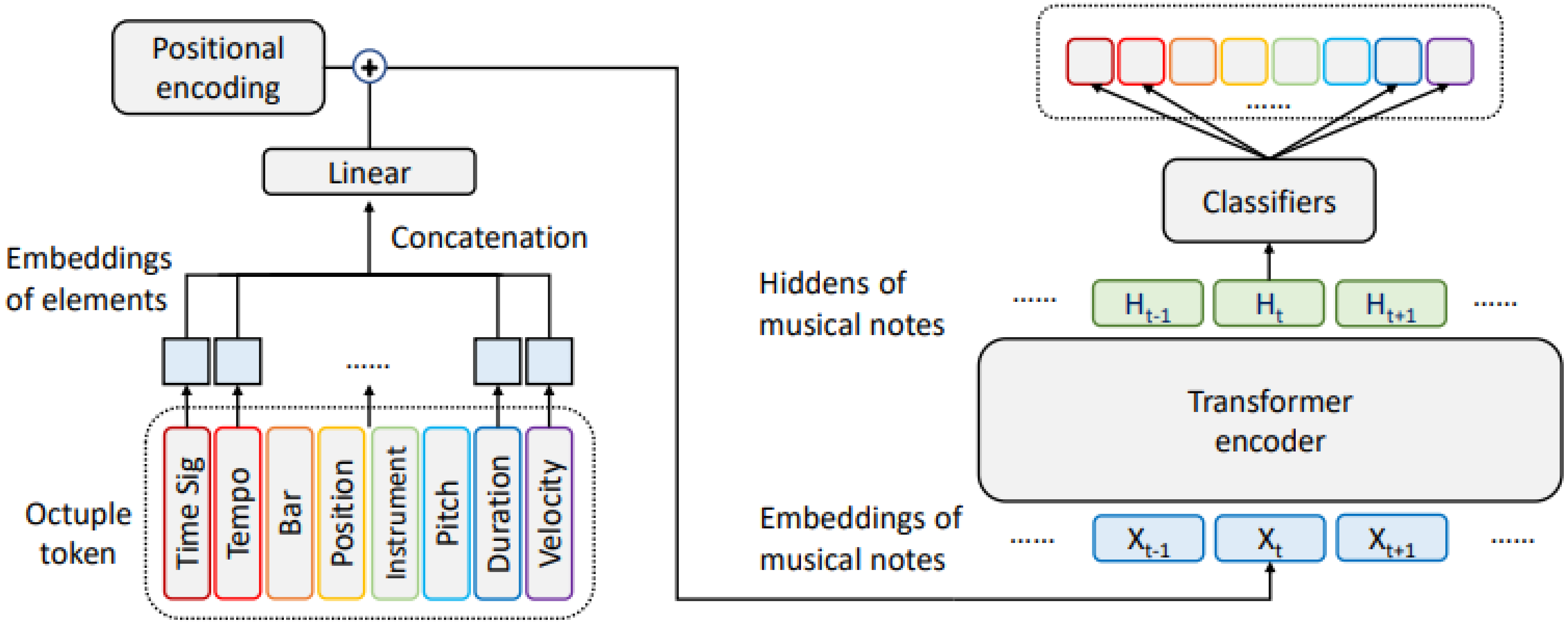


Figure 1: Model structure of MusicBERT.

Methodology

OctupleMIDI Encoding

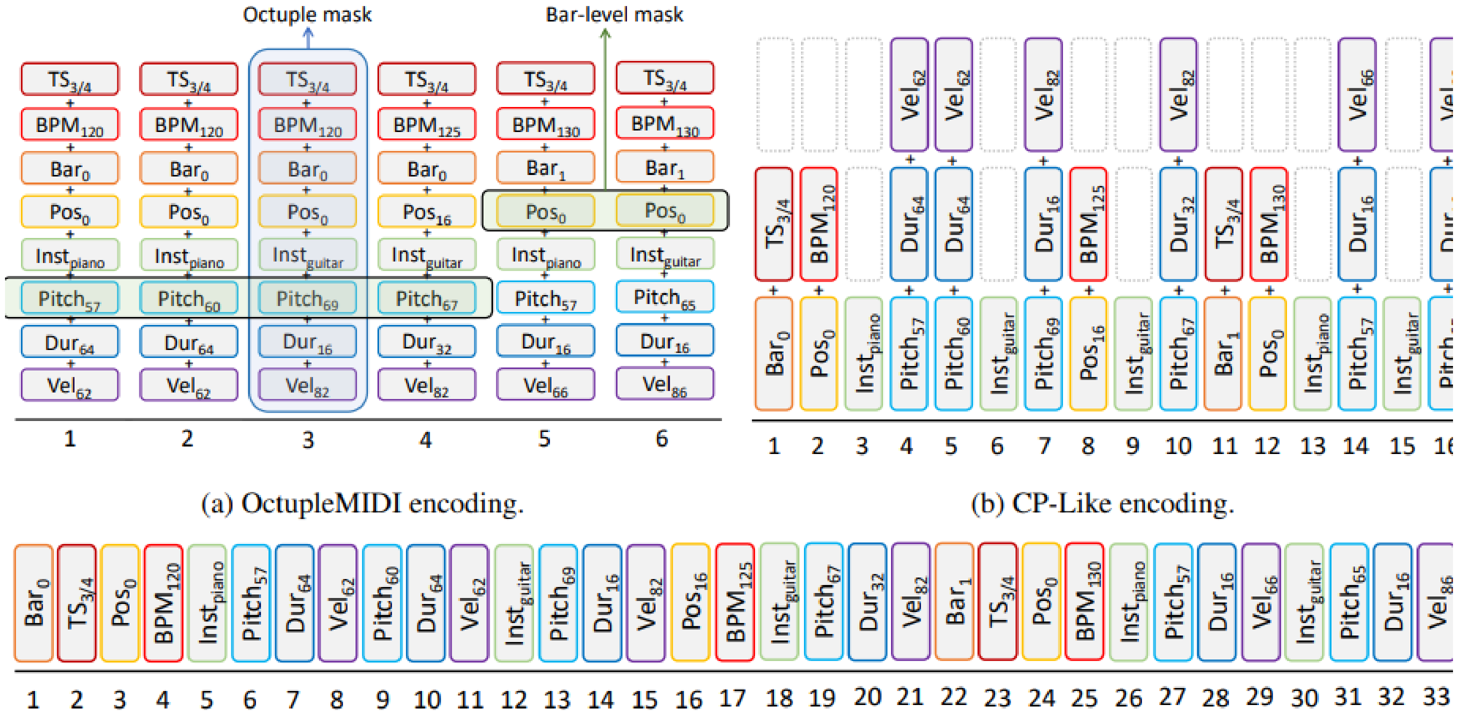
- Time Signature
- Bar
- Tempo

- Pitch
- Position
- Duration
- Velocity

- Instrument

Methodology

▪ OctupleMIDI Encoding & Masking Strategy



(a) OctupleMIDI encoding.

(b) CP-Like encoding.

(c) REMI-Like encoding.

Figure 2: Different encoding methods for symbolic music.

Experiments

- Melody Completion & Accompaniment Suggestion & Classification

Model	Melody Completion					Accompaniment Suggestion					Classification	
	MAP	HITS @1	HITS @5	HITS @10	HITS @25	MAP	HITS @1	HITS @5	HITS @20	HITS @25	Genre F1	Style F1
melody2vec$_F$	0.646	0.578	0.717	0.774	0.867	-	-	-	-	-	0.649	0.299
melody2vec$_B$	0.641	0.571	0.712	0.772	0.866	-	-	-	-	-	0.647	0.293
tonnetz	0.683	0.545	0.865	0.946	0.993	0.423	0.101	0.407	0.628	0.897	0.627	0.253
pianoroll	0.762	0.645	0.916	0.967	0.995	0.567	0.166	0.541	0.720	0.921	0.640	0.365
PiRhDy$_{GH}$	0.858	0.775	0.966	0.988	0.999	0.651	0.211	0.625	0.812	0.965	0.663	0.448
PiRhDy$_{GM}$	0.971	0.950	0.995	0.998	0.999	0.567	0.184	0.540	0.718	0.919	0.668	0.471
MusicBERT$_{small}$	0.982	0.971	0.996	0.999	1.000	0.930	0.329	0.843	0.993	0.997	0.761	0.626
MusicBERT$_{base}$	**0.985**	**0.975**	**0.997**	**0.999**	**1.000**	**0.946**	**0.333**	**0.857**	**0.996**	**0.998**	**0.784**	**0.645**

Experiments

▪ Ablation Study

Encoding	Melody	Accom.	Genre	Style
CP-like	95.7	87.2	0.719	0.510
REMI-like	92.0	86.5	0.689	0.487
OctupleMIDI	**96.7**	**87.9**	**0.730**	**0.534**

Table 5: Results of different encoding methods. "Accom." represents accompaniment suggestion task.

Mask	Melody	Accom.	Genre	Style
Random	96.3	87.8	0.708	0.533
Octuple	96.0	87.3	0.722	0.530
Bar	**96.7**	**87.9**	**0.730**	**0.534**

Table 6: Results of different masking strategies.

Model	Melody	Accom.	Genre	Style
No pre-train	92.4	76.9	0.662	0.395
MusicBERT	**96.7**	**87.9**	**0.730**	**0.534**

Table 7: Results with and without pre-training.

Contrastive Learning of general-purpose audio representations

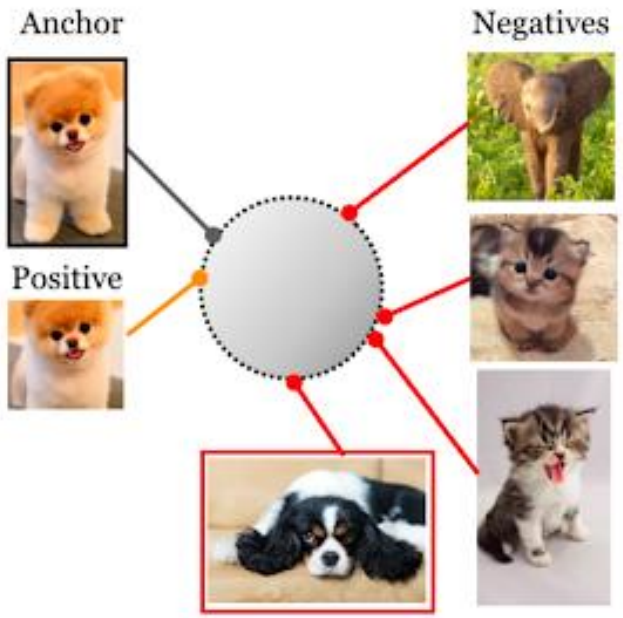
Google Research

정영석

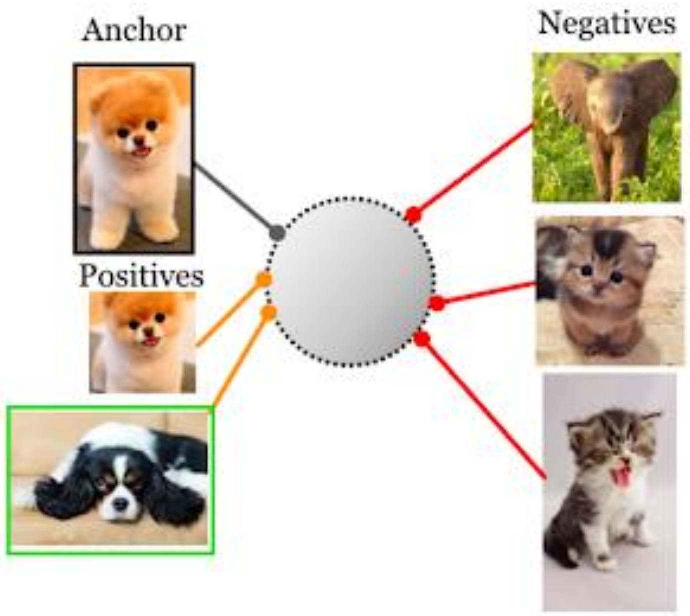
▪ Introduction

- Contrastive Learning concept

  ✓ Learning a representation which assigns high similarity to audio segments extracted from the same recording while assigning lower similarity to segments from different recordings.



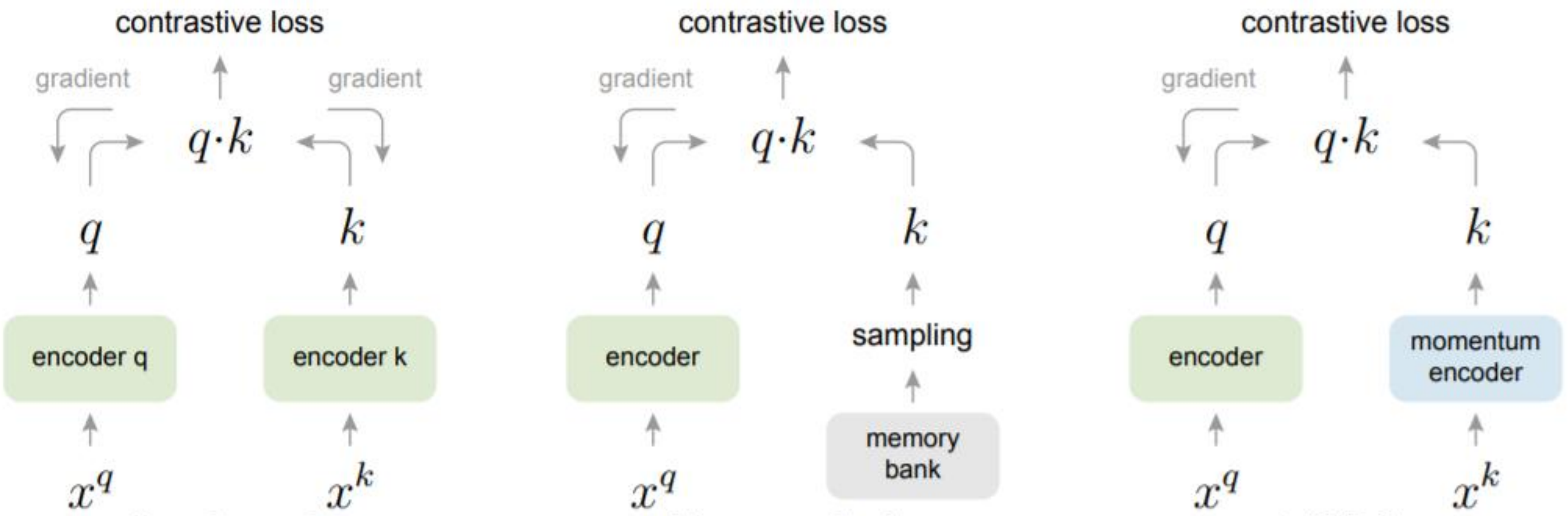
Self Supervised Contrastive



Supervised Contrastive

▪ Introduction

- Contrastive Learning model
  - ✓ 3 kinds of learning model(end-to-end, memory bank, MoCo)



(a) end-to-end (b) memory bank (c) MoCo

▪ Introduction

- The Proposed Methods

  ✓ The model learn general purpose-representations of sounds beyond speech

  ✓ The simple methods for sampling positive & negative example.

  ✓ Using bilinear similarity which shows better performance than cosine similarity.

Methodology

▪ Methodology

- Proposed Model

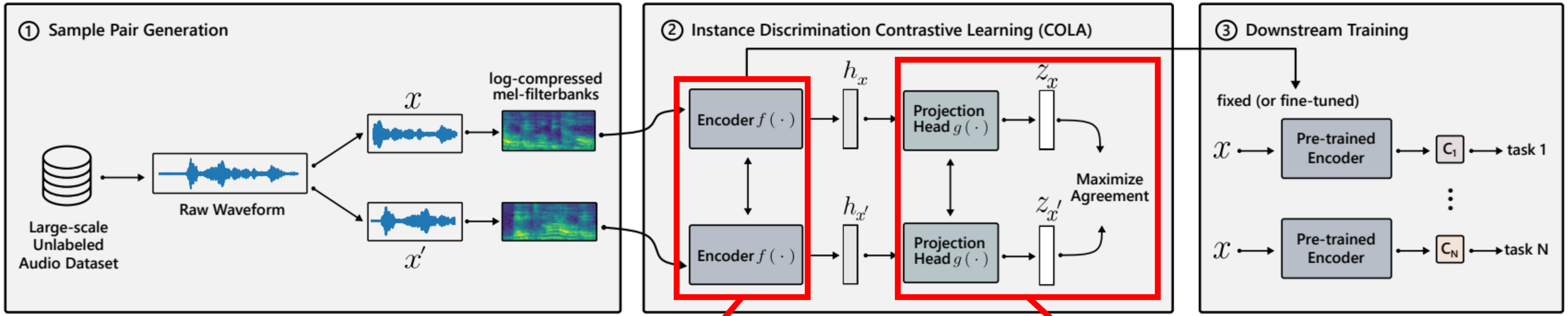


Fig. 1. Overview of the contrastive self-supervised learning for audio.

$$h = \underline{f(x)} \in \mathbb{R}^d \qquad s(x, x') = g(f(x))^\top W \, g(f(x')).$$

Efficientnet-b0

▪ Methodology

- Contrastive Loss

$$\mathcal{L} = -\log \frac{\exp\left(\mathrm{s}(x, x^{+})\right)}{\sum\limits_{x^{-} \in \mathcal{X}^{-}(x) \cup \{x^{+}\}} \exp\left(\boxed{\mathrm{s}(x, x^{-})}\right)}$$

$$\mathrm{s}(x, x') = g(f(x))^{\top}\, W\, g(f(x')).$$

Experiment

- Experiment

Table 1. Test accuracy (%) on downstream tasks.

Task	Random Init.	Supervised	COLA Frozen	COLA Fine-tuned
Speaker Id. (LBS)	0.4	**100.0**	**100.0**	**100.0**
Speech commands (V1)	62.9	97.2	71.7	**98.1**
Speech commands (V2)	4.0	94.3	62.4	**95.5**
Acoustic scenes	8.6	98.2	94.1	**99.2**
Speaker Id. (Voxceleb)	0.0	31.7	29.9	**37.7**
Birdsong detection	49.6	79.4	77.0	**80.2**
Music, Speech and Noise	56.8	99.3	99.1	**99.4**
Language Id.	59.1	**85.0**	71.3	82.9
Music instrument	20.8	70.7	63.4	**73.0**
Average	29.1	83.9	74.3	**85.1**

Table 3. Test accuracy (%) with different similarity functions

	Cosine Similarity	Bilinear Similarity
Speaker Id. (LBS)	99.9	**100.0**
Speech commands (V1)	64.5	**71.7**
Speech commands (V2)	42.4	**62.4**
Acoustic scenes	87.5	**94.1**
Speaker Id. (Voxceleb)	15.2	**29.9**
Birdsong detection	76.5	**77.0**
Music, Speech and Noise	99.0	**99.1**
Language Id.	62.3	**71.3**
Music instrument	58.3	**63.4**
Average	67.2	**74.3**

Experiment

- ## Experiment

Table 2. Test accuracy (%) of a linear classifier trained on top of COLA embeddings or baseline pre-trained representations.

	CBoW [16, 25]	SG [16, 25]	TemporalGap [16, 25]	Triplet Loss [16, 25]	TRILL [13]	COLA
Speaker Id. (LBS)	99.0	**100.0**	97.0	**100.0**	-	**100.0**
Speech commands (V2)	30.0	28.0	23.0	18.0	-	**62.4**
Acoustic scenes	66.0	67.0	63.0	73.0	-	**94.0**
Birdsong detection	71.0	69.0	71.0	73.0	-	**77.0**
Music, Speech and Noise	98.0	98.0	97.0	97.0	-	**99.1**
Music instrument	33.5	34.4	35.1	25.7	-	**63.4**
Speech commands (V1)	-	-	-	-	**74.0**	71.7
Speaker Id. (Voxceleb)	-	-	-	-	17.7	**29.9**
Language Id.	-	-	-	-	**88.1**	71.3
Average (TRILL tasks)	-	-	-	-	**59.9**	57.6
Average (non-TRILL)	66.25	66.0	64.3	64.4	-	**82.5**