Temporal Convolution for Real-time Keyword Spotting on Mobile Devices

Seungwoo Choi*, Seokjun Seo*, Beomjun Shin*, Hyeongmin Byun,
Martin Kersner, Beomsu Kim, Dongyoung Kim†, Sungjoo Ha†

HYPERCONNECT

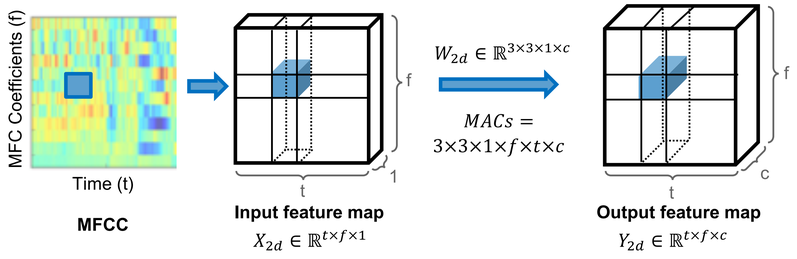
INTERSPEECH 2019

1. Motivation

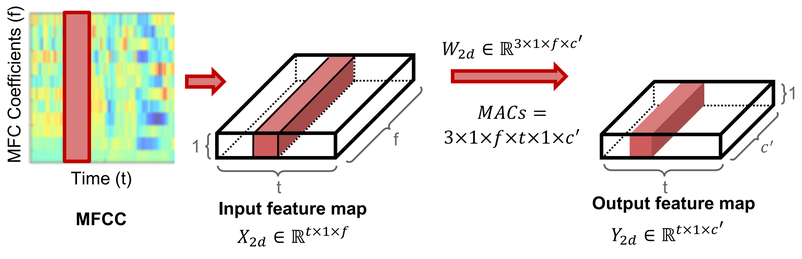
- To deploy keyword spotting (KWS) on mobile devices, it should be **fast** enough and **accurate** for real-time inference.
- However, previous studies are **either** fast or accurate.

Model	Acc. (%)	Time (*ms*)	FLOPs	Params
CNN-2	84.6*	**1.2**	1.5M	148K
Res15	**95.8***	424	1950.0M	239K

- Problem of 2D convolution.
  - Conventionally, MFCC is considered as input 2D image.
  - **Receptive fields** of conventional CNN are not enough to cover all features at specific time point.
  - It needs deeper or wider layers to cope with all frequency domains, which **slows down** the model.

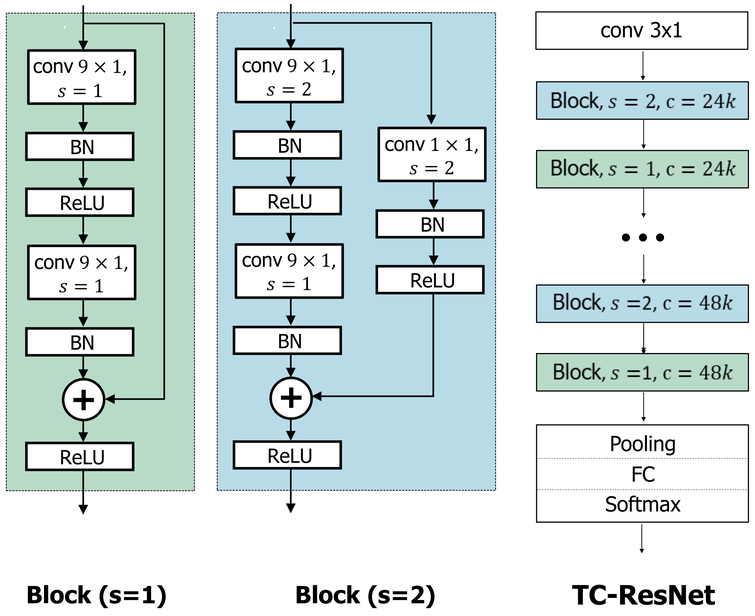


2. Temporal Convolution



- **Large receptive field of audio features.**
  - Enable a model to achieve better accuracy even with less computation.
- **Small footprint and low computational complexity.**
  - Suitable for fast inference.

3. Proposed model



4. Experimental Results

Model	Acc. (%)	Time (*ms*)	FLOPs	Params
CNN-1	90.7	32	76.1M	524K
CNN-2	84.6	**1.2**	1.5M	148K
DS-CNN-S	94.4	1.6	5.4M	24K
DS-CNN-M	94.9	5.2	19.8M	140K
DS-CNN-L	95.4	16.8	56.9M	420K
Res8-Narrow	90.1	47	143.2M	20K
Res8	94.1	174	795.3M	111K
Res15-Narrow	94.0	107	348.7M	43K
Res15	**95.8**	424	1950.0M	239K
TC-ResNet8	96.1	**1.1**	3.0M	66K
TC-ResNet8-1.5	96.2	2.8	6.6M	145K
TC-ResNet14	96.2	2.5	6.1M	137K
TC-ResNet14-1.5	**96.6**	5.7	13.4M	305K

** The accuracy is calculated on Google Speech Commands dataset and the speed is measured on Google Pixel 1.*

- **Accuracy improvement** regardless of the speed.
- Compared to the previous most accurate model.
  - TC-ResNet8 shows **385x faster** inference speed.
- Compared to the previous fastest model.
  - TC-ResNet8 shows **11.5%p accuracy improvement.**
- TC-ResNet **can adjust tradeoff** between accuracy and speed.
  - By modulating the number of layers and width multiplier (k).

- Ablation study shows **temporal convolution is a key component** for fast and accurate KWS.

2D-ResNet8	96.1	10.1	35.8M	66K
2D-ResNet8-Pool	94.9	3.5	4.0M	66K

- 2D-ResNet8: temporal convolution -> 2D convolution.
- 2D-ResNet8-Pool: + additional pooling after first convolution layer to reduce computation.

5. Conclusion

- State-of-the-art performance on Google Speech Commands.
  - In terms of both speed and accuracy.
  - By adopting **temporal convolution.**
- TC-ResNet facilitates **real-time KWS on mobile devices.**
  - Not estimating by FLOPs but measuring on mobile devices.
- **Implementation and benchmark tools are publicly released.**
  - https://github.com/hyperconnect/TC-ResNet

Paper

Github