TinySpeech: Attention Condensers for Speech Recognition on Edge Devices

Alexander Wong, Mahmoud Famouri, Maya Pavlova, and Siddharth Surana

VIP Research Group University of Waterloo & DarwinAI Corp., Waterloo

Objective

Design self-attention mechanisms for building low-footprint, highly-efficient deep neural networks for on-device speech recognition on edge devices.

Introduction

Advances in deep learning have led to state-of-the-art performance across a multitude of speech recognition tasks. Nevertheless, the widespread deployment of deep neural networks for on-device speech recognition remains a challenge, particularly in edge scenarios where the memory and computing resources are highly constrained (e.g., low-power embedded devices) or where the memory and computing budget dedicated to speech recognition is low (e.g., mobile devices performing numerous tasks besides speech recognition). In this study, we leveraged two complementary strategies to build **TinySpeech**: low-footprint, low precision deep neural networks tailored specifically for limited-vocabulary speech recognition.

- *Attention Condensers*: A new self-attention mechanism designed for selective attention based on joint local and cross-channel activation relationships captured via condensed embeddings.
- *Machine-Driven Design Exploration*: Incorporating the new attention condensers to automatically determine the macroarchitecture and microarchitecture designs of the final TinySpeech networks [1]

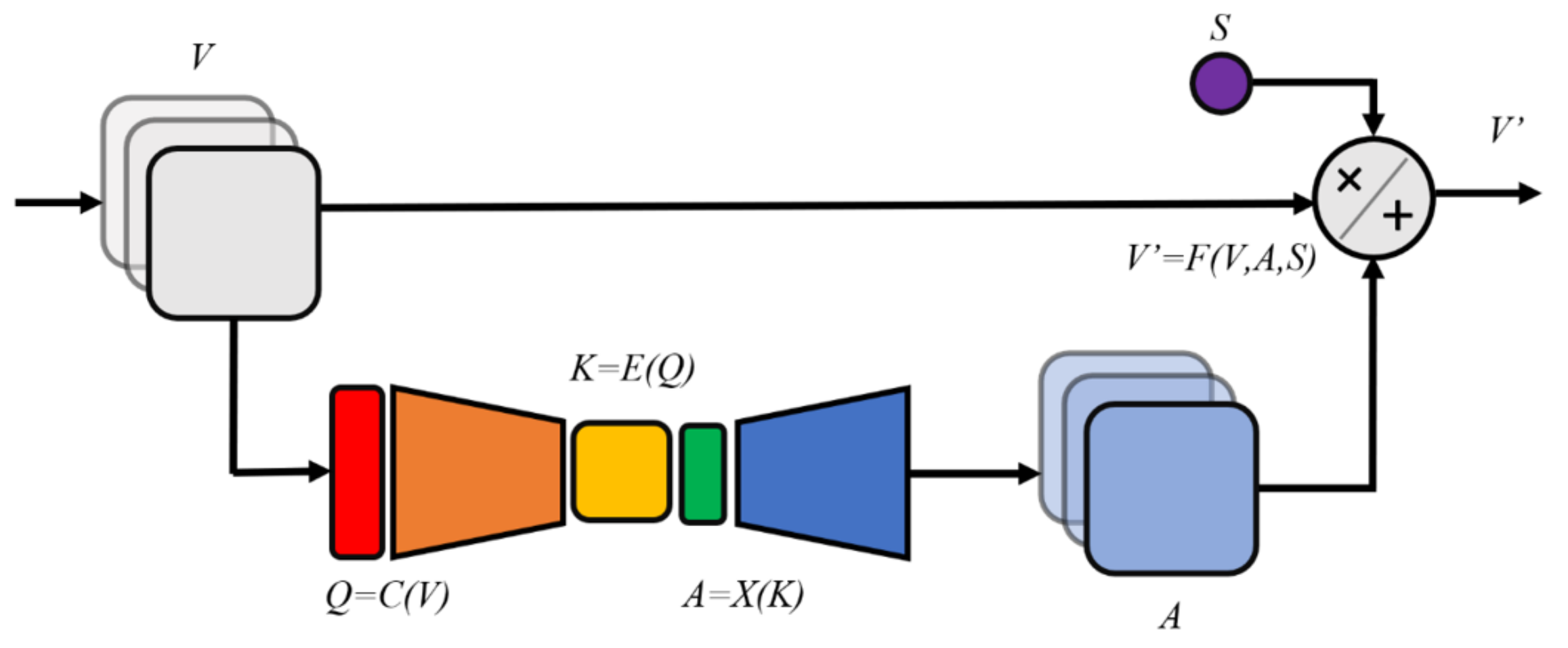

Figure 1:Attention Condenser Architecture.

Attention Condenser

Attention condensers are new self-attention mechanisms that learn and produce a condensed embedding characterizing joint local and cross-channel activation relationships. Attention condensers perform selective attention, with a greater emphasis on activations in close proximity of strong activations.

The attention condenser module is a stand-alone module while existing self-attention and channel-wise attention mechanisms are designed to augment network architectures to improve accuracy at the expense of some complexities.

The attention condenser module consists of:

- A condensation layer C
- An embedding structure E
- An expansion layer X
- A selective attention mechanism F

TinySpeech

Attention Condensers

- Condensation Layer: Max Pooling
- Embedding Structure: Grouped + Pointwise Convolutions
- Expansion Layer: Unpooling (Upsampling)

Machine-Driven Design Exploration

- Generative Synthesis with constraints including:
  1) Validation accuracy greater than 90%
  2) 8-bit weight precision
  3) Number of parameters < 15k
  4) TensorFlow support Lite for TinySpeech-M

Input Pipeline

- Google Speech Commands dataset
- Band-Pass Filter
- Sliding Window
- Mel-Frequency Cepstrum Coefficients

Important Result

Attention Condensers: new self-attention mechanisms based on characterized joint local and cross-channel activation relationships.

TinySpeech Architecture

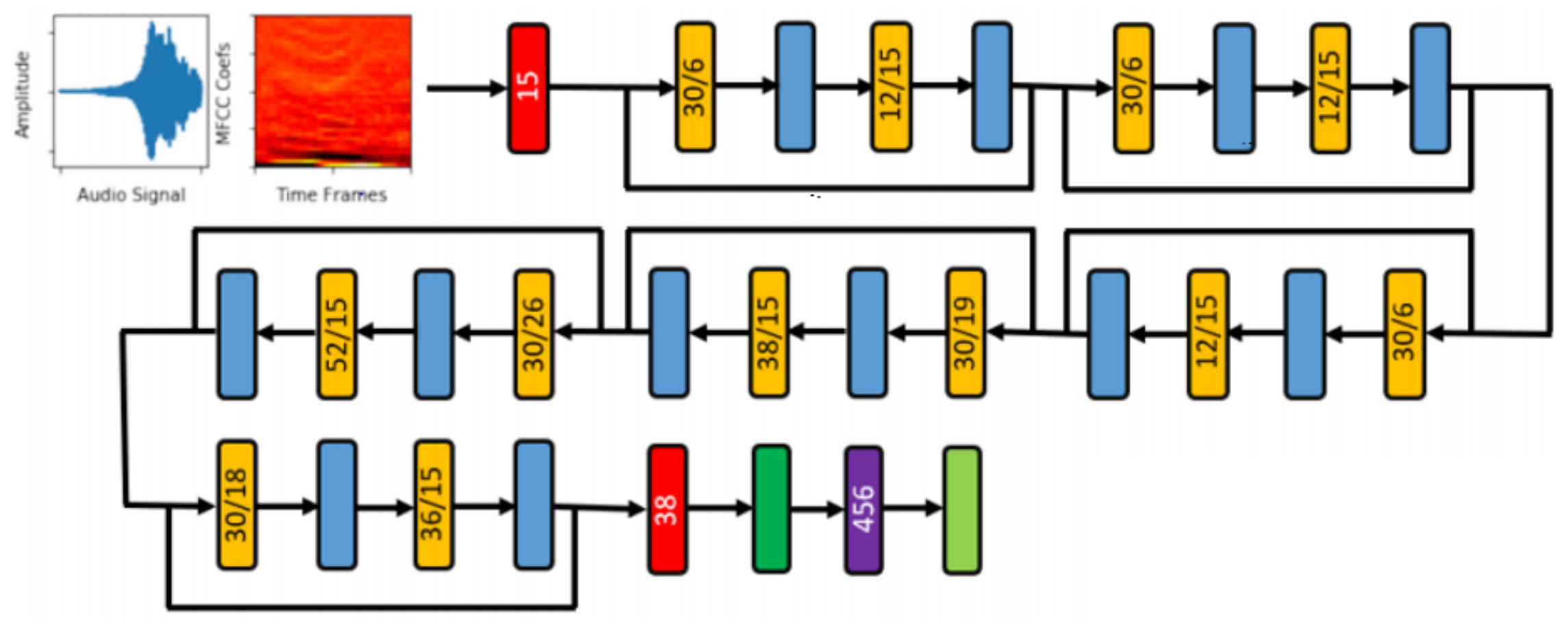

Figure 2:TinySpeech-X

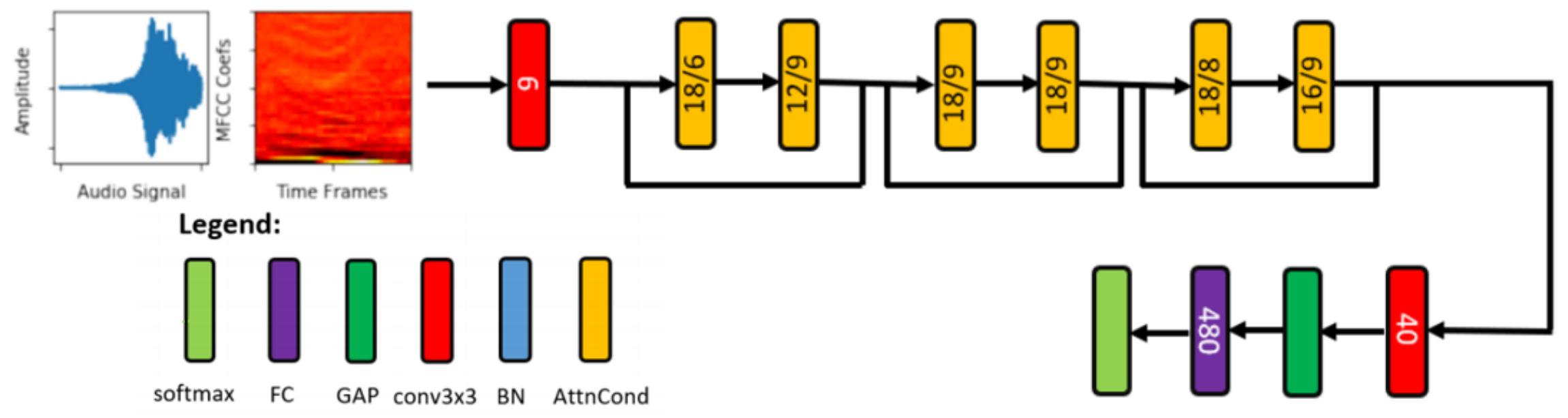

Figure 3:TinySpeech-M

Results

Model	Accuracy	Params	Mult-Adds
trad-fpool13[2]	90.5	1370K	125M
tpool2[2]	91.7	1090K	103M
TDNN[3]	94.2	251K	25.1M
res15-narrow[4]	94.0	42.6K	160M
PONAS-kws2[5]	94.3	131K	168M
TinySpeech-X	**94.6**	10.8K	10.9M
TinySpeech-Y	93.6	6.1K	6.5M
TinySpeech-Z	92.4	**2.7K**	**2.6M**
TinySpeech-M	91.9	4.7K	4.4M

Table 1:TinySpeech networks in comparison to state-of-the-art methods

Conclusion

In this study, we introduce the notion of attention condensers for building highly-efficient and high-performance deep neural networks for on-device speech recognition for edge scenarios. By jointly modeling local and cross-channel activation relationships within a unified condensed embedding, attention condensers can act as self-contained, stand-alone modules that can be leveraged within a deep neural network architecture to reduce the quantity of larger stand-alone convolution modules needed to achieve a high level of accuracy.

References

[1] A. Wong, M. J. Shafiee, B. Chwyl, and F. Li. Ferminets: Learning generative machines to generate efficient neural networks via generative synthesis. 2018.

[2] T. N. Sainath and C. Parada. Convolutional neural networks for small-footprint keyword spotting. 2015.

[3] S. Myer and V. S. Tomar. Efficient keyword spotting using time delay neural networks. 2018.

[4] R. Tang and J. Lin. Deep residual learning for small-footprint keyword spotting. 2018.

[5] R. Dahyot A. Anderson, J. Su and D. Gregg. Performance-oriented neural architecture search. 2020.

Contact Information

- Web: http://www.darwinai.ca
  mahmoud@darwinai.ca
  mspavlova@uwaterloo.ca

DARWIN AI