Text Adaptive Detection For Customizable Keyword Spotting

Yu Xi[1], Tian Tan[2], Wangyou Zhang[1], Baochen Yang[1], Kai Yu[1]
[1]Shanghai Jiao Tong University [2]AISpeech Ltd, Suzhou China

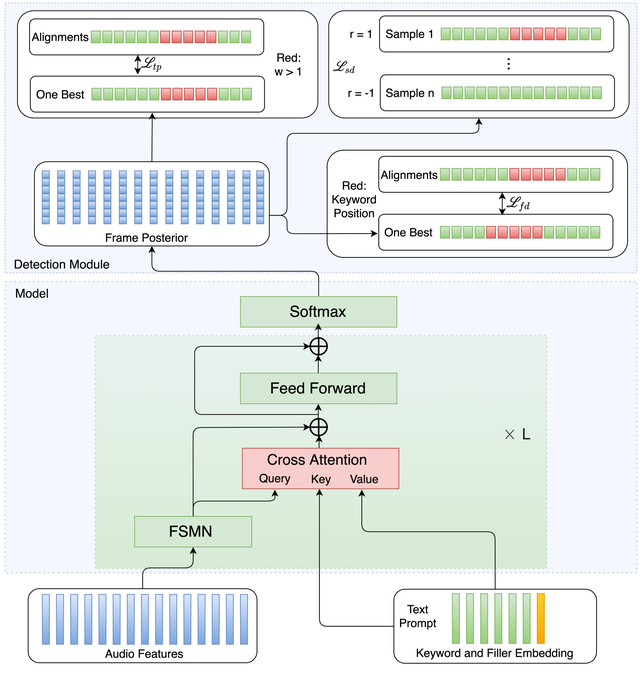
Highlights

- Text adaptive detection framework
- Customize arbitrary wake words
- Address the loss-metric mismatch
- 16.88% relative improvement of F1-score

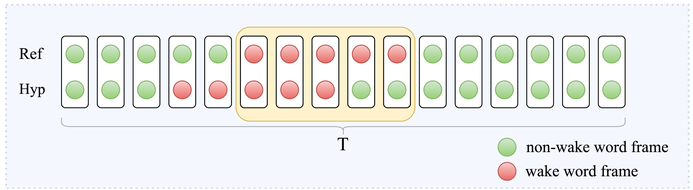
Always-on keyword spotting

- Wake word detection
- Wake up smart devices via predefined keywords
  - OK Google
  - Hey Siri
  - Alexa

Text Adaptive Detection Framework



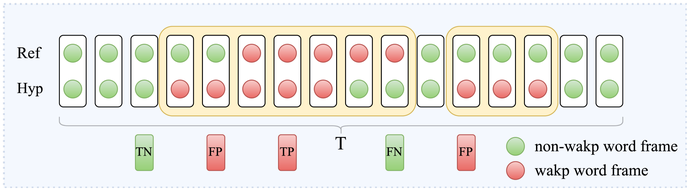
Modified Weighted Cross-entropy Loss



$$\mathcal{L}_{tp} = -\Big(w \sum_{t \in W} \log p_t + \sum_{t \notin W} \log p_t\Big).$$

- A weight $W > 1$ applied to wake word frames
- $W = 1$, the loss degenerate to the standard CE

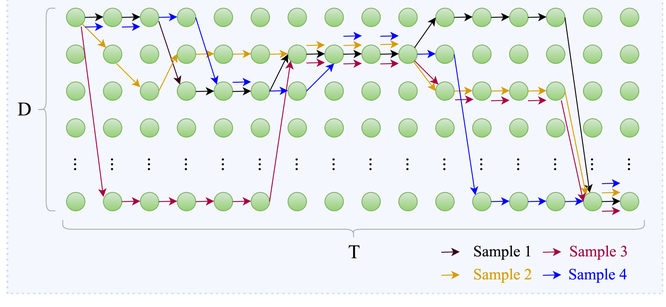
Frame-level Detection Loss



$$\mathcal{L}_{fd} = -\Big(\sum_{t \in \{TP,TN,FN\}} \log p_t^l + \sum_{t \in \{FP\}} \log\left(1 - p_t^o\right)\Big)$$

- Treat each frame as a binary classification(detection) problem

Sequence-level Detection Loss



- Detect whether the keyword appears in the hypothetical samples sequence or not.

Experimental Setups

- Dataset
  - WSJ (a keyword spotting version of WSJ)
- Positive test dataset
  - The keyword must appear at least five times in WSJ test dataset
  - The keyword is not polyphonic.
- Negative test dataset
  - Choose 20 utterances that do not contain any of the keywords.
- Test metric
  - Micro f1-score

Results

Model	#Param. (#MACs)	Dataset		
		Dev93	Ev92	Ev93
Baseline	270K (8.47M)	0.692	0.762	0.844
Larger Baseline	4000K (122.82M)	0.780	0.825	0.883
Text Prompt ($\mathcal{L}_{tp}$, $w = 15$)	268K (7.64M)	0.794	0.890	0.898
+ Frame level ($\mathcal{L}_{tp}$, $\mathcal{L}_{fd}$)		0.805	0.887	0.910
+ Sequence level ($\mathcal{L}_{tp}$, $\mathcal{L}_{sd}$)		0.832	0.901	0.912
+ Frame + Sequence ($\mathcal{L}_{tp}$, $\mathcal{L}_{fd}$, $\mathcal{L}_{sd}$)		**0.850**	**0.909**	**0.917**

- Compared with baseline:
  - Achieves a relative improvement of 22.83%, 19.22%, and 8.59% on dev93, ev92, and ev93 with less parameters and lower computational cost.
- Compared with larger baseline:
  - Still get a relative improvement of 9.03%, 10.14%, 3.85%.

Summary

- Proposed a text adaptive detection framework for always-on keyword spotting task.
- Our method outperforms the baseline by a significant margin for the comparable model size.