ABR Laboratory
Artificial Brain Research

Fast Intent Classification for Spoken Language Understanding

Akshit Tyagi et al., arXiv, Dec 2019

(Amazon and University of Massachusetts)

Presenter: Gwenaelle Cunha Sergio

Artificial Brain Research Lab., School of Electronics Engineering,
Kyungpook National University

29-Jan-2020

Previous Works

- Previous works attempt to modify the model architecture in order to reduce computational complexity
- Examples: Regularization, model distillation, compression
- **Limitation**: *accuracy loss*

Proposed Model

***BranchyNet** scheme to reduce complexity and latency while retaining accuracy in SLU systems by inserting exit points throughout the model.*

Allows early decision making when possible

Entropy threshold

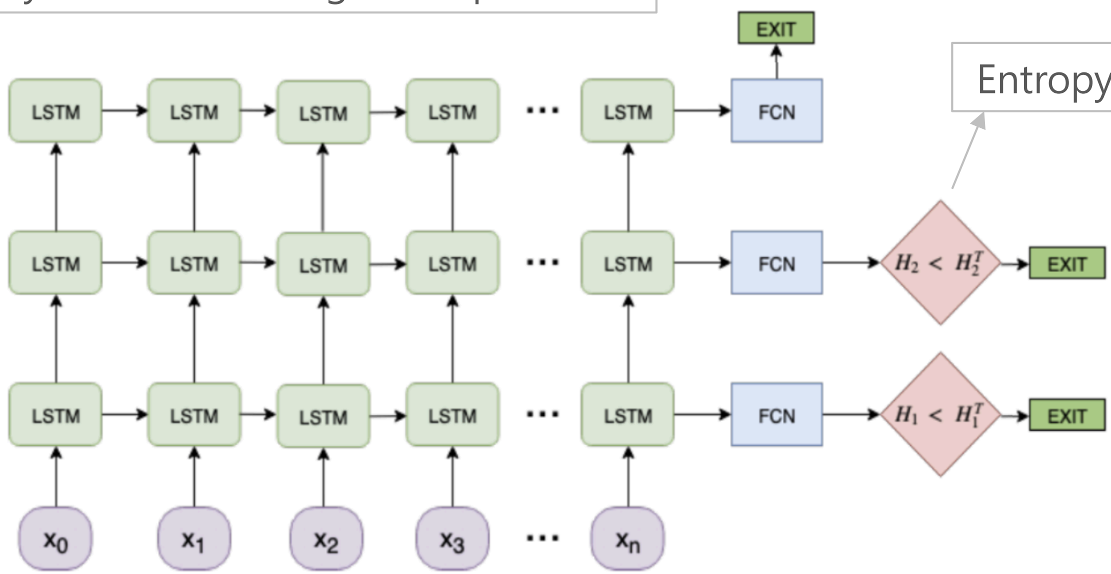

Fig. 2. Stacked LSTM model with early exiting strategy. The model can exit at each LSTM layer. FCN implies a fully connected layer.

Proposed Model

- Candidate architectures: 3-layer DNN and Stacked LSTM
- **Advantage**: Requires minimal modification
  - Exit points are added at each hidden layer in DNN
  - Allows the model to make "a decision as soon as it is confident in its prediction"

Proposed Model

- *Loss function*: weighted sum of cross entropy losses from every exit point.

$$L = \sum_{n=1}^{N} \alpha_n L_n$$

- **α_n**: linearly decreasing function
  - More weight given to early branches: "improves the accuracy of the later branches due to the added regularization"
  - Encourages early exit by encouraging "the learning of discriminative representatio ns in earlier layers"

$$\alpha_n = r_l + \frac{r_u - r_l}{n}, \quad n = 1, .., N$$

r_l: range (lower bound)
r_u: range (upper bound)
(values not specified in paper)

- Early exit: $H_n < H_n^T$

$$\text{entropy}(\boldsymbol{y}) = \sum_{c \in \mathcal{C}} y_c \log y_c$$

H_n^T : entropy threshold (defined after training for each exit point)
H_n : entropy at point n
y : vector containing computed probabilities for all possible class labels
C : set of all possible labels

Results

- The introduction of BranchyNet in DNN and Stacked LSTM does not lead to accuracy loss.
- **Boost in performance** due to its *regularization effect* and *tailored representations* from each layer with exit points.

Model	F1(Macro)	Acc.(%)
DNN	0.48	88.5
DNN + BranchyNet	0.55	89.6
Stacked LSTM	0.65	92.8
Stacked LSTM + BranchyNet	0.66	93.2

Table 1. Performance of DNN models on the FSPS dataset with and without the BranchyNet mechanism

- Reduced computational complexity in # of parameters and FLOPS.

References

[1] Tyagi, Akshit, et al. *"Fast Intent Classification for Spoken Language Understanding."* arXiv preprint arXiv:1912.01728 (2019). [This paper: BranchyNet for Intent Classification in SLU]

[2] Teerapittayanon, Surat, Bradley McDanel, and Hsiang-Tsung Kung. *"Branchynet: Fast inference via early exiting from deep neural networks."* 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016. [BranchyNet Original Paper]

Thank you!