Karlsruhe Institute of Technology

interACT

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany

Exploring CTC-Network Derived Features with Conventional Hybrid System

Thai-Son Nguyen, Sebastian Stüker, Alex Waibel

ASR with CTC Model

- Using LSTMs, the training with CTC criterion can efficiently model the dependencies between a small number of units (e.g., phonemes or characters) and speech frames
- The CTC criterion automatically handles possible alignments between a label sequence and the speech frames
- Eliminating complex steps in the conventional hybrid system, e.g., HMM topology definition, CD phonemes, and frame-wise alignment
- *Good performance on many thousands hours of speech data but severely overfit with less training data [1]*
- *Bad performance without incorporating language model during inference*
- *Optimizing for CTC decoding is hard, e.g. with or without incorporating the priors*

CTC Alignment

- The CTC posteriors have peaky behaviors in which *blank* has the highest probability in almost all frames, except for short peaks where regular labels dominate
- **Do the phone probabilities assigned by the CTC model still correlate to the fixed labels of a traditional Viterbi alignment?**

Our Approach

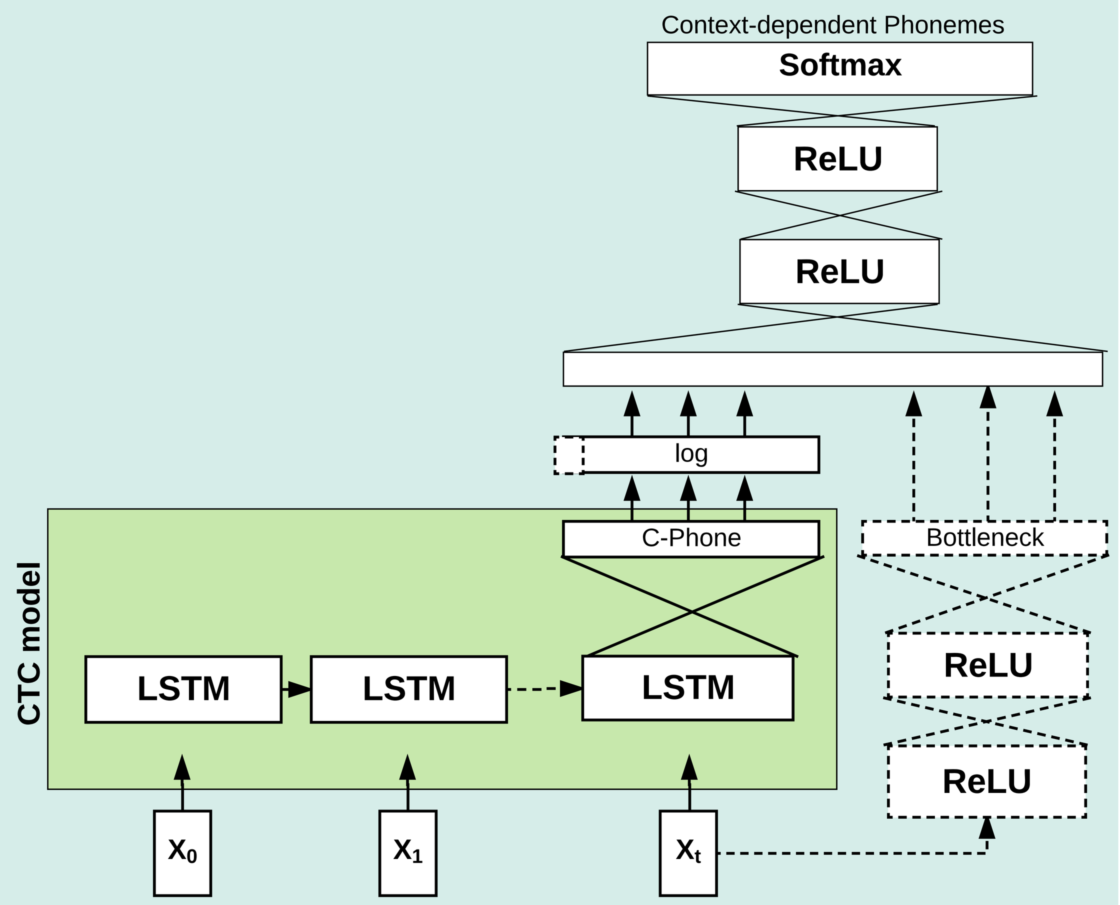
- *We train CTC model with phone labels and use CTC posterior probabilities as input features (so-called C-Phone) in hybrid HMM/ANN system.*
- *To benefit from the strengths of the CTC network at label discrimination on the one side and the highly optimized decoding stack of conventional hybrid systems on other size*
- *Taking advantages of combining different features e.g., i-vectors, bottleneck features for further improving phonemes classification performance*

Related Works

- The posterior output of MLP was originally proposed as input features to Tandem GMM models [2]
- When the multiple HMM states per phone and CD states were introduced, bottleneck features [3] a small layer in the middle of the MLP, was used for instead.

Extracting & Using C-Phone

- Train LSTMs on speech data with CTC-loss criterion
- Transform the posterior output of the LSTMs into the log domain with or without using the elimination of the *blank* node
- Can also the logits instead of the log features
- Possibly concatenated with the bottleneck features



Experimental Setups

- Training data includes 300 hours of the Switchboard-1 Release 2 (LDC97S62)
- CTC modeling with Bi-directional LSTM with 5 layers of 320 units on input features of 40 filter-bank co-efficients with 45 English phonemes as labels
- FFNN architecture of 7 layers of 1600 units for all hybrid HMM/ANN models
- Evaluated on Hub500 evaluation data (LDC2002S09)
- Used 4-gram language model from Fisher corpus and the training data
- Using Essen [4] and Janus Recognition Toolkit (JRTK) [5] for decoding

C-Phone Performance

Model	Features	Window	Hub5'e (SWB)
FFNN	FBank	11	22.4 (15.8)
CTC	FBank	-	19.9 (14.1)
FFNN	C-Phone-P	-	-
	C-Phone-L	1	19.3 (13.7)
	C-Phone-L	7	19.0 (13.6)
	C-Phone-L	11	**18.9 (13.5)**
	C-Phone-L	15	19.3 (13.8)
	C-Phone-NB	1	19.3 (13.8)
	C-Phone-NB	7	19.0 (13.6)
	C-Phone-NB	11	19.1 (13.6)
	BNF	1	22.7 (16.0)
	BNF	7	21.8 (15.3)
	BNF	11	21.5 (15.1)
	BNF	15	21.5 (15.1)
	fMLLR-BNF	11	21.0 (14.6)
GMM	C-Phone-L	1	20.9 (15.7)
	C-Phone-L	11	20.0 (14.5)
	BNF	11	22.1 (15.7)

- *The CTC posteriors contains excellent features for classifying CD phonemes labeled in the fixed alignment*
- *The probability of the blank does not carry useful information*
- *The FFNN systems trained on C-Phone outperform FBank by a large margin and also clearly outperform CTC system*

Features Combination

+Features	Window	Hub5'e (SWB)
FBank	1/1	23.0 (17.7)
	3/3	18.9 (13.6)
	5/5	19.1 (13.7)
	1/5	19.1 (13.7)
BNF	1/1	18.4 (13.1)
	2/2	18.2 (12.9)
	3/3	18.4 (13.1)
	5/5	18.6 (13.3)
	1/5	18.5 (13.1)
fMLLR-BNF	1/1	**18.1 (12.8)**
	2/2	18.2 (13.0)
	3/3	18.2 (13.1)
	5/5	18.3 (13.2)
	1/5	18.2 (12.9)

- *BNF features can supplement C-Phone and result in a better recognition performance*
- *We have not tested yet with speaker adaptive features such as i-vectors*

Extracting with Uni-directional LSTM

Model	+Features	Window	Hub5'e (SWB)
CTC		-	25.4 (17.8)
FFNN		1	35.4 (28.2)
		11	25.5 (18.6)
	FBank	2/2	25.4 (18.8)
	FBank	3/3	24.8 (18.2)
	BNF	1/1	21.9 (15.8)
	BNF	2/2	21.2 (15.3)
	BNF	3/3	21.0 (15.0)

- *We only achieve some small improvement using C-Phone to complement the bottleneck features*

Conclusions

- *A feed-forward network system using our proposed CTC-network derived features with cross-entropy training outperforms a strong CTC baseline by a margin of 5% rel.*
- *With the same model, we achieved further improvements of 9% rel. when combining them with bottleneck features*
- *We are examining the gain when performing sequence training as well as the performance of the presented systems on different training data sets*

References

- [1] Pundak, Golan, and Tara N. Sainath. "Lower Frame Rate Neural Network Acoustic Models." Interspeech. 2016.
- [2] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems". ICASSP 2000.
- [3] Frantisek Grzl, Martin Karafit, Stanislav Kontr, and Jan Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings". ICASSP 2007.
- [4] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "Eesen: End-to-end speech recognition using deep RNN models and wfst-based decoding". ASRU 2015.
- [5] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ri Es, and Martin Westphal. "The karlsruhe VERBMOBIL speech recognition engine" ICASSP 1997.