End-to-end speech recognition primarily uses encoder-decoder or CTC models, mostly using LSTMs or a LSTM+CNN combination.

- We explore purely convolutional CTC models for lexicon-free conversational speech recognition, which are much faster than recurrent models.

- Unlike most previous work [1, 2] we focus on 1-D convolutions. TDNNs are closely related to our work.

### Model & Experimental Setup

**Neural “encoders”** map input sequences to hidden states $h_t$ and a softmax layer maps $h_t$ to a distribution over frame level CTC labels $\pi_t$.

**Connectionist Temporal Classification (CTC)**

- We use the standard CTC collapsing function $B(\pi)$ which removes the special blank symbols and consecutive repetitions.

\[ p(x) = \sum_{i \in B^{-1}(x)} \prod_{t=1}^{n} p(\pi_t|h_t) \]

- We present results using a greedy decoding approach as well as beam search with a $n$-gram character LM. If $x = B(\pi)$,

\[ \hat{t} = \underset{t \in \pi}{\arg\max} \ p(x)[t] \prod_{i=t}^{n} p(\pi_i|h_i) \]

We use the decoding algorithm in [4].

**Long Short-Term Memory (LSTM) Baseline**

- 5-layer 320 hidden unit bidirectional LSTM network with dropout between consecutive layers. Every two consecutive input frames are concatenated to reduce time resolution.

**CNN Model**

- 1-D convolutions across time only. Following [5], we use residual connections (“ResBlocks (RBs)” and batch normalization.

**Experimental Setup**

- We use the 300h Switchboard corpus for training, and report results on the 4k utterance Switchboard dev set, and Eval2000 setup consisting of Switchboard (SWB) and Callhome (CH) utterances.

- All models trained on a single Titan X GPU, with two CPU threads in TensorFlow.

**CNN Architecture**

- 1-D CNNs train and decode significantly faster than LSTMs for speech recognition with CTC.

- For the same number of weights, deeper networks with smaller filters perform best.

- CNNs are only 0.2% behind LSTMs on the Switchboard test set, but are a larger 1.1% behind on Callhome, indicating over-fitting.

- CNNs respond less to language model based beam-search decoding.

- Very deep ResNet-style CNNs [5] (50+ layers) are needed to match LSTM performance.

### Future Work

- Better regularization techniques for CNN architectures to prevent over-fitting.

- Analysis of larger all-CNN systems on a word level CTC architecture.

- Response of all-convolutional systems to non-CTC architectures, and different decoding schemes [6]

---

**References**


