Recep Oğuz Araz
ELEC390
Engin Erzin
03.10.2020

Report II: High Level Understanding of the Architecture

Review

DS2 is an end-to-end learning system, trained with Connectionist Temporal Classification (CTC) function.

Predicts speech transcriptions from audio.

Variable length audio sequences directly mapped to variable length transcriptions.

1. RNN encoder-decoder architecture with attention
   - performs well in predicting phonemes or graphemes.
2. CTC Loss Function coupled with RNN to model temporal information.
   - Performs well in end-to-end grapheme transcription, for phonemes it needs lexicon

???Before DS2, CTC-RNN required pretrained alignments from GMM-HMM.

Model Architecture

Spectrogram to transcription

Let a single utterance $x^{(i)}$ and label $y^{(i)}$ be sampled from a training set X = $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots\}$.

Each utterance, $x^{(i)}$, is a time-series of length $T^{(i)}$, where every time-slice is a *vector* of audio

Features: $x_t^{(i)}$, $t = 0,1,\ldots,T^{(i)}$

We use a <u>spectrogram of power normalized audio clips</u> as the features to the system, so $x_{t,p}^{(i)}$

denotes the power of the p'th frequency bin in the audio frame at time t.

The goal of the RNN is to convert an input sequence $x^{(i)}$ into a final transcription $y^{(i)}$.

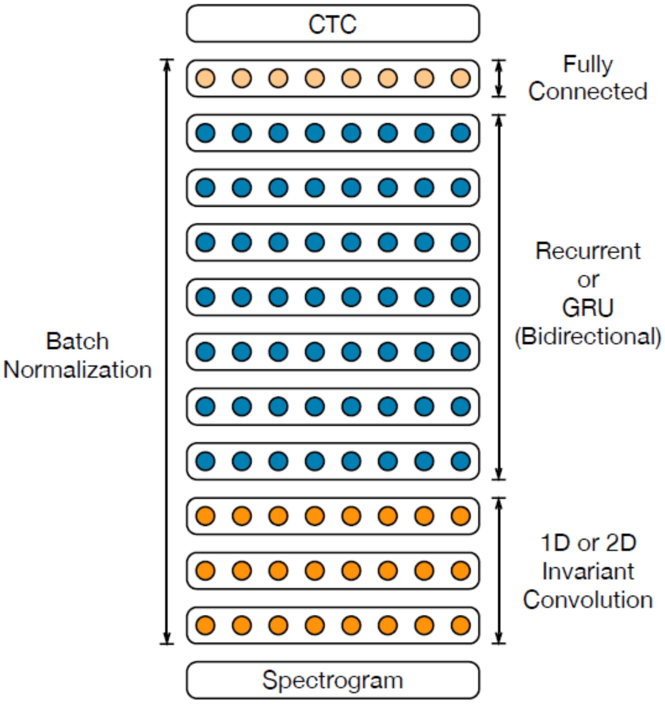
Notation: use x to denote a chosen utterance and y the corresponding label.

The outputs of the network are the graphemes.

At each output time-step t, RNN makes a prediction over characters, $p(l_t \mid x)$ where l_t is either a character or blank. l_t is in {a,b,c,ç,d,…,boşluk,kesme işareti, blank}.

For non-linearity, clipped ReLu used.

In some layers sub-sample by striding the convolution.


CTC
Fully Connected
Recurrent or GRU (Bidirectional)
Batch Normalization
1D or 2D Invariant Convolution
Spectrogram

The output layer is softmax, computing a probability distribution over characters. $p(l_t=k|\ x)$

Trained with CTC loss function.

Parameter Update with backpropagation through time algorithm.

Connectionist Temporal Classification

RNNs require pre-segmented training data and post processing to transform their outputs into label sequences.

Standard NN objective functions are defined separately for each point in the training sequence. That is, RNNs can only be used make a series of independent label classifications. So, training data must be pre-segmented and network outputs must be post-processed.

CTC models all aspects of the sequence within a single network architecture.

- Prediction of sequences of labels from noisy, unsegmented input data.

Recep Oğuz Araz
ELEC390
Engin Erzin
03.10.2020

Interprets the network outputs as a probability distribution over all possible label sequences, conditioned on a given input sequence. Given this distribution, an objective function can be derived that directly maximizes the probabilities of the correct labeling.

Differentiable objective function, standard bptt algorithm applicable.

Labelling unsegmented data sequences → temporal classification

Independent labelling of each time-step of the sequence → framewise classification

Our use → CTC

S = training examples drawn from fixed distribution D_{XxZ}.

The input space X = (R^m)* is the set of all sequences of m dimensional real valued vectors.

The target space Z = L* is the set of all sequences over the finite alphabet L of labels. Elements of L* are called label sequences.

Each example in S consists of a pair of sequences (**x,z**). **z** is at most as long as **x.**

Goal: Use S to train a temporal classifier to classify unseen sequences in a way that minimizes some error measure.

Label Error Rate given S' LER of a temporal classifier h as the normalized edit distance between its classificiations and the targets on S'

If we can transform the network outputs into a conditional probability distribution over label sequences, the network can be used as a classifier by selecting the most probable label sequence for a given input sequence.

The CTC network predicts only the sequence of labels, without an attempt on aligning the with any frames. (CTC paper figure 1)