Bidirectional LSTM-HMM Hybrid System For Polyphonic Sound Event Detection

Vienna Deep Learning Meetup

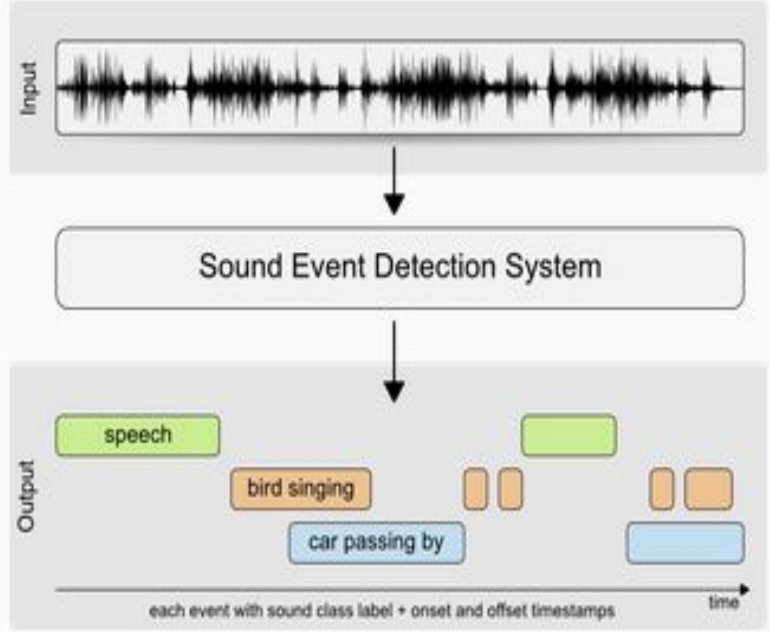
Anahid Jalali

Outline

- Motivation and task description
- Brief Overview on Hidden Markov Models
- Brief Overview on Recurrent Neural Networks
- Vanishing Gradient Problem
- Long Short Term Memory
- Hybrid system BLSTM-HMM
- Conclusion

Sound Event Detection

- Sound contains important information, used in various applications;
  - Life Log
  - Monitoring Systems
  - Environmental context understanding
  - ...
- The objective of SED systems is to understand various sounds by identifying the beginning and ending of a sound event to identify the label of the sound.



Overview of sound event detection system.
http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio

Sound Event Detection

- Two SED scenarios:
  - Monophonic
  - Polyphonic
- Most typical approach used for Polyphonic SED is Hidden Markov Model where emission probability distribution is represented by Gaussian Mixture Models (GMM-HMM), using Mel Frequency Cepstral Coefficient as features.

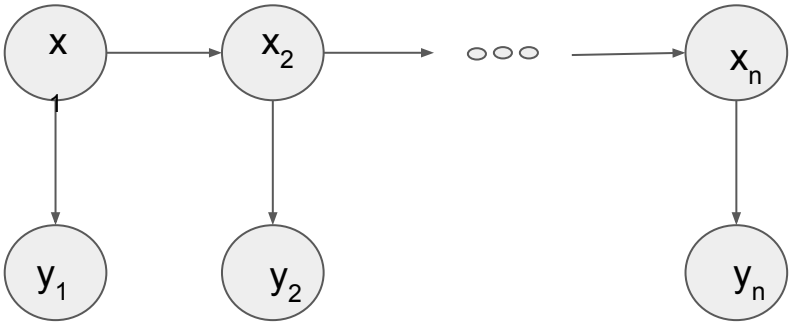
Hidden Markov Models

- A strong tool for modeling sequential data, representing probability distribution over sequences of observations.
- Goal is to make a sequence of decisions where a particular decision may be influenced by earlier decisions.
- HMM defines two properties:
  - Observation at a specific time was generated by some process whose state is hidden from the observer.
  - The state of this hidden process satisfies the markov property.
- Joint Probability Distribution:

$$P(S, Y) = P(s_1)\, P(y_1 | s_1) \prod_{t=2}^{T} P(s_t | S_{t-1})\, P(y_t | s_t)$$

Hidden Markov Models

- HMM graphical model of the joint distribution probability is represented by a Bayesian Network



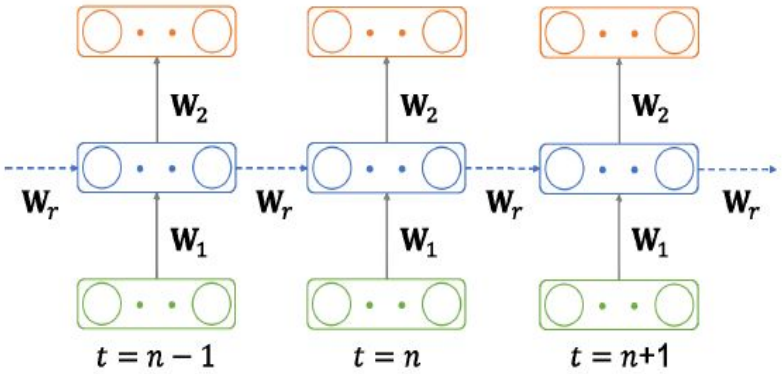
Motivation

- Inspired by the BLSTM-HMM used in speech recognition.
- Good performance in speech recognition
- Allowance of the smoothing of the framewise output without post-processing
- No need for thresholding

- T. Hayashi, S. Watanabe, T. Toda, T. Hori, J.L. Roux, K. Takeda, 2016, “Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection”, Proceeding of Detection and classification of acoustic scenes and events 2016, page 35-39.

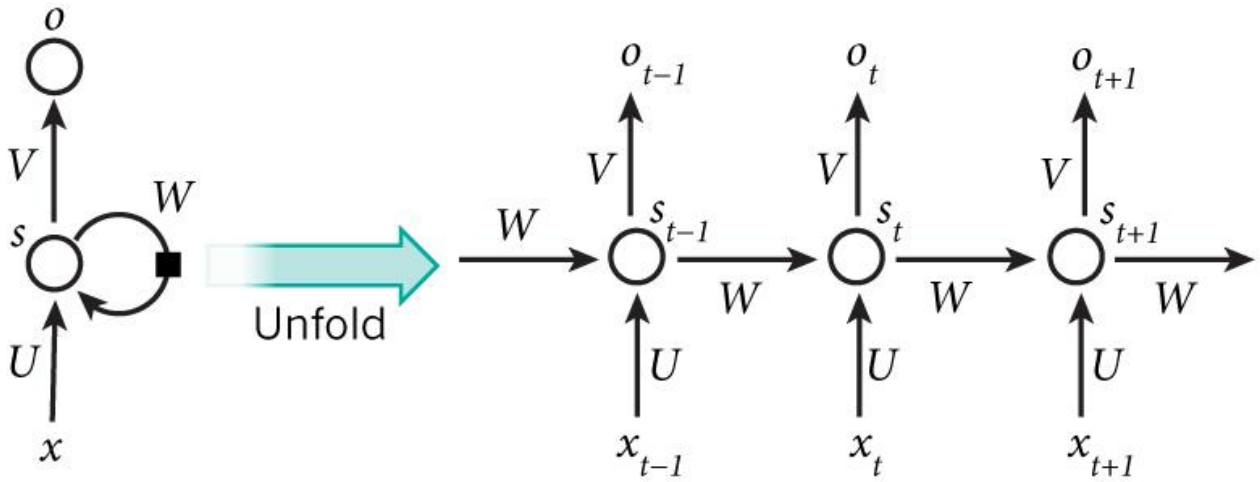
Recurrent Neural Network

- Popular models of Deep Learning
- Used for training on time series data such as audio or text
- A layered Neural Network with a feedback structure
- Can propagate information from previous time steps to the current time step
- Hidden layers in RNN serves as a memory function



Architecture of RNN [1]

Recurrent Neural Network



http://d3kbpzbmcynnmx.cloudfront.net/wp-content/uploads/2015/09/rnn.jpg

Recurrent Neural Network

- Mathematically representing RNN

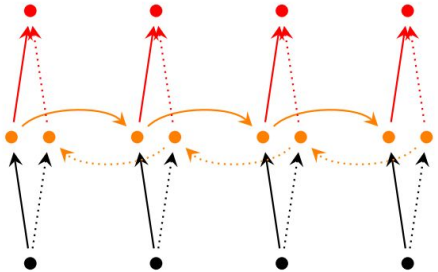
$$\mathbf{h}_t = f(\mathbf{W}_1\mathbf{x}_t + \mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{b}_1), \quad (1)$$
$$\mathbf{y}_t = g(\mathbf{W}_2\mathbf{h}_t + \mathbf{b}_2), \quad (2)$$

RNN Formulas taken from [1]

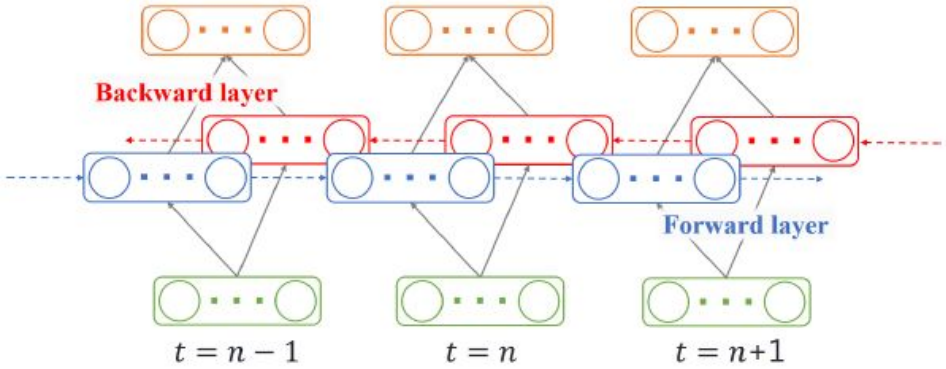
- W_i represents the input weight matrix
- B_i denotes the bias vector of the first layer
- W_r represents the recurrent weight matrix
- f and g are activation functions of the hidden layer and output layer, respectively.

Bidirectional RNN

- Has feedback from both previous and future time periods
- Hidden layer connected to the following time period is called forward layer
- Hidden layer connected to the previous time period is called backward layer
- Propagates information from both past and the future
- Ability to understand and exploit the full context in an input sequence



Architecture of BRNN
http://d3kbpzbmcynnmx.cloudfront.net/wp-content/uploads/2015/09/bidirectional-rnn.png



Architecture of BRNN [1]

Vanishing Gradient Problem

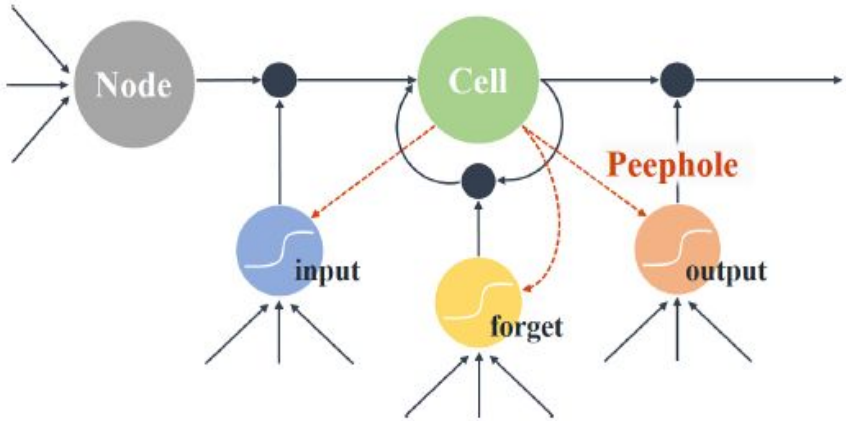
Problem: The gradients of the network's output w.r.t the parameters in the prior layers become significantly small.

Cause: Choice of the activation function is very important, calculation of the gradients at each point results a very small gradient at the end → vanishing gradient.

Solution: using Rectified Linear Unit (ReLu) activation function, Long Short Term Memories and Gated Recurrent Units, weight initialization

Long Short Term Memory

- An effective solution to the Vanishing Gradient Problem
- Allows the memorization of long term context information
- Characteristics of LSTM architecture:
  - Memory cells
  - Input gate
  - Forget gate
  - Output gate
- Each gate has a value between 1 and 0
- Value 1 → gate is open
- Value 0 → gate is closed



Architecture of LSTM [1]

Long Short Term Memory

- In an LSTM the hidden layer output in equation 1 is replaced by:

$$
\begin{aligned}
\mathbf{g}_t^I &= \sigma(\mathbf{W}^I \mathbf{x}_t + \mathbf{W}_r^I \mathbf{h}_{t-1} + \mathbf{s}_{t-1}), && (3)\\
\mathbf{g}_t^F &= \sigma(\mathbf{W}^F \mathbf{x}_t + \mathbf{W}_r^F \mathbf{h}_{t-1} + \mathbf{s}_{t-1}), && (4)\\
\mathbf{s}_t &= \mathbf{g}_t^I \odot f(\mathbf{W_1} \mathbf{x}_t + \mathbf{W}_r \mathbf{h}_{t-1} + \mathbf{b}_1) + \mathbf{g}_t^F \odot \mathbf{s}_{t-1}, && (5)\\
\mathbf{g}_t^O &= \sigma(\mathbf{W}^O \mathbf{x}_t + \mathbf{W}_r^O \mathbf{h}_{t-1} + \mathbf{s}_{t-1}), && (6)\\
\mathbf{h}_t &= \mathbf{g}_t^O \odot \tanh(\mathbf{s}_t), && (7)
\end{aligned}
$$

LSTM Formulas taken from [1]

- W denotes input weight matrices
- W_r denotes recurrent weight matrices

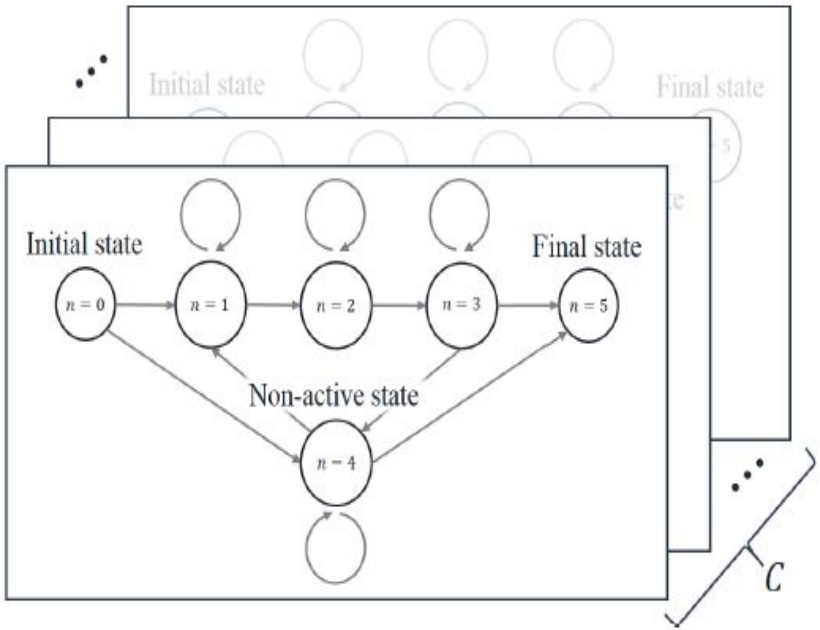
DCASE2016 task 2 Dataset

There exists 11 categories for the provided sound events (Audio index c):

- Clearing throat
- Coughing
- Door knock
- Door slam
- Drawer
- Human laughter
- Keyboard
- Keys (put on table)
- Page turning
- Phone ringing
- Speech

BLSTM-HMM Hybrid System

- The extension was in order to handle the multi label classification
- A three state left-to-right HMM with a non-active state for each sound event is built.
- n = 0, n = 5 and n = 4 represent the initial state, final state, and non-active state, respectively.
- The non-active state represents not only the case where there is no active event, but also the case where other events are active.



Hidden Markov Model for each sound event [1]

BLSTM-HMM Hybrid System

- Using Bayes' theorem, HMM state emission probability is approximated as follows:

$$P(x_t \mid s_{c,t} = n) = \frac{P(s_{c,t} = n \mid x_t)\, P(x_t)}{P(s_{c,t} = n)}$$

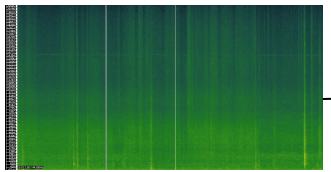
- HMM state posterior $P(s_{c,t} = n \mid x_t)$ is calculated using a BLSTM-RNN.

BLSTM-HMM Hybrid System

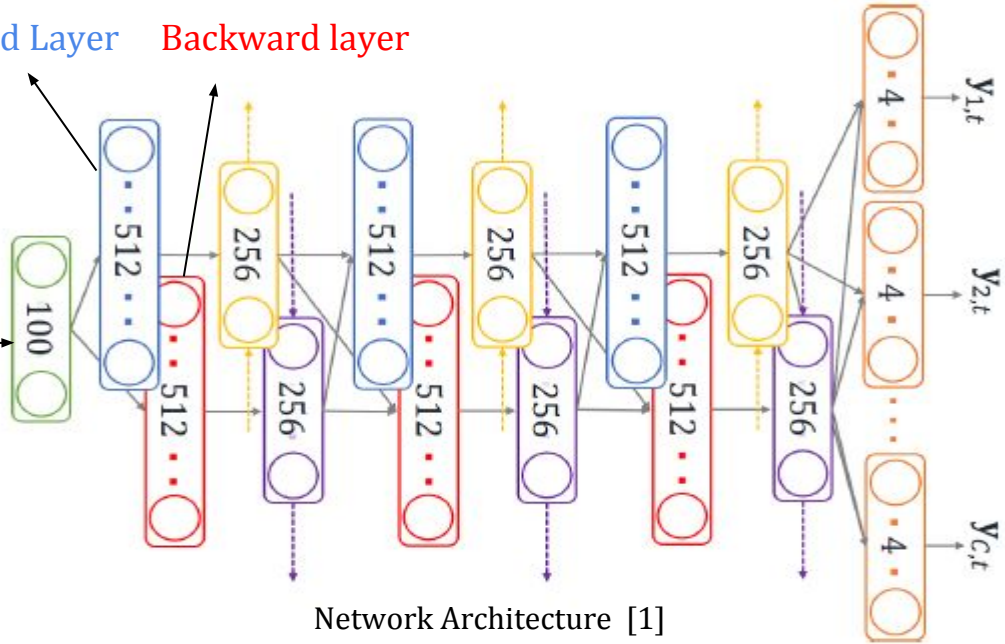
- All the values of the posterior $P(s_{c,t}|x_t)$ is calculated using a softmax operation.
- HMM state prior $P(s_{c,t})$ is calculated by counting the number of occurrence of each HMM state.
- The network was optimized using back-propagation through time (BPTT) with Stochastic Gradient Descent (SGD) and dropout under the cross-entropy for multi-class multi-label objective function.

BLSTM-HMM Hybrid System

- The network has three hidden layers which consist of an LSTM layer, a projection layer, and the number of output layer nodes is C x N.



An audio spectrogram



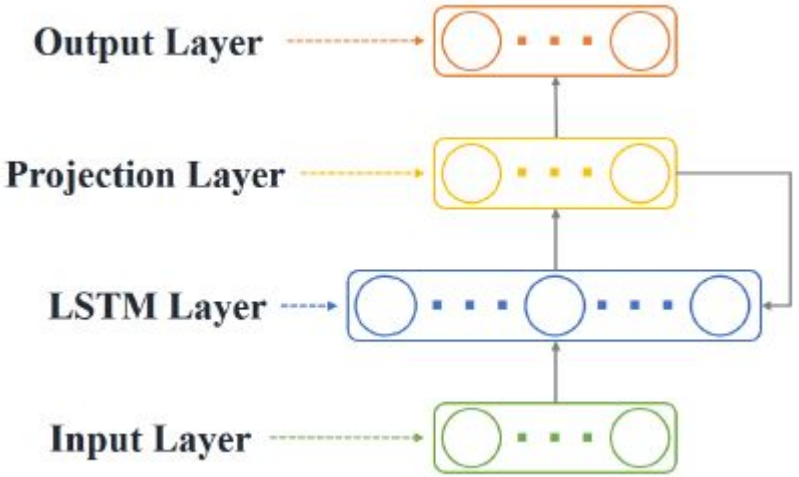
Network Architecture [1]

Projection Layer

- A technique to reduce the computational complexity of deep recurrent network structures.
- A linear transformation
- In previous equations 3-6 h_{t-1} will be replaced by p_{t-1} and the following equation will be added:

$$P_t = W_I h_t$$

- Further Reading on understanding the projection layer: [3, 4]



Architecture of LSTMP [1]

Conclusion

- The proposed method was applied to the DCASE2016 challenge task 2.
- This method was then compared to baseline non negative matrix factorization and the standard BLSTM-RNN which outperformed them both in polyphonic and monophonic sound event detection task
- achieving an average F-score of 67.1% and error rate 64.5% on the evaluation.

Thank You For Your Attention

References

[1] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J.L. Roux, K. Takeda, 2016, “Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection”, Proceeding of Detection and classification of acoustic scenes and events 2016, page 35-39.

[2] Z. Ghahramani, “An Introduction To Hidden Markov Models and Bayesian Networks”, International journal of pattern recognition and Artificial Intelligence, 15(1):9-42.

[3] H. Sak, A. Senior, and F. Beaufays, “Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, ArXiv e-prints arXiv:1402.1128, 2014.

References

[4] H. Sak et. al., “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in Proc. IEEE INTERSPEECH, 2014.