

Building a state of the art speech recogniser

Moritz Wolter

Thesis submitted for the degree of
Master of Science in Mathematical
Engineering

Thesis supervisor:

Prof. dr. ir. Patrick Wambacq

Assessor:

Prof. dr. ir. Johan Suykens

Mentor:

Ir. Vincent Renkens

© Copyright KU Leuven

Without written permission of the thesis supervisor and the author it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to the Departement Computerwetenschappen, Celestijnenlaan 200A bus 2402, B-3001 Heverlee, +32-16-327700 or by email info@cs.kuleuven.be.

A written permission of the thesis supervisor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Preface

I would like to thank everybody who kept me busy the last year, especially my promotor and my assistants. I would also like to thank the jury for reading the text. My sincere gratitude also goes to my family for supporting me through my studies.

Moritz Wolter

Contents

Preface	i
Abstract	iii
List of Figures and Tables	iv
List of Abbreviations and Symbols	v
1 Literature Study	1
1.1 The classical approach to speech recognition	1
1.2 Methods for Deep-Network based speech recognition	1
1.3 Listen, Attend and Spell	2
Bibliography	5

Abstract

The `abstract` environment contains a more extensive overview of the work. But it should be limited to one page.

List of Figures and Tables

List of Figures

- 1.1 The LAS architecture [1, page 3]. BLSTM blocks are shown in red.
LSTM blocks in blue and attention nets in green. 4

List of Tables

List of Abbreviations and Symbols

Abbreviations

Con- vNet	Convolutional neural network
MSE	Mean Square error
PSNR	Peak Signal-to-Noise ratio

Symbols

42	“The Answer to the Ultimate Question of Life, the Universe, and Everything” according to [?]
c	Speed of light
E	Energy
m	Mass
π	The number pi

Chapter 1

Literature Study

The first contains a general introduction to the work. The goals are defined and the modus operandi is explained. TODO: describe the problem.

1.1 The classical approach to speech recognition

1.1.1 Hidden Markov Models

1.1.2 Gaussian mixture models

1.1.3 DNN

1.2 Methods for Deep-Network based speech recognition

1.2.1 Preprocessing

f-bank features

f-bank (filter banks) features are one option or raw data. time domain or frequency domain?

1.2.2 Stochastic gradient descent

When training networks on very large training sets, working with the full data set to compute the current gradient becomes very inefficient. As a remedy it is good practice in machine learning to work with so called mini-batches. A mini-batch includes a random subset of the training data set. This procedure is known as randomized gradient descent. With an added momentum term it can be formalized as [3, page 4]:

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\partial C}{\partial w_{ij}(t)} \quad (1.1)$$

C is the cost, which is computed by comparing the current network output to the desired output. $\alpha \in (0, 1)$ is a momentum coefficient. The weight of a connection

from unit i in the layer under consideration to unit j in the layer below is given by the expression w_{ij} .

1.2.3 Dropout

1.2.4 Classical layer architecture

1.2.5 Convnets

When looking for a signal in different parts of a recording it is not always advisable to relearn recognition of the signal in different locations. With a conventional fully connected structure every sample on the time axis gets its own input weight. For shifted version of this signal the network will not be able to reuse weights it used to recognize the same signal at another time point in the recording. Convolutional neural nets aim to solve this problem [2, page 6], while preserving essential ordering information.

1.2.6 RNNs

1.2.7 Tensor-flow

1.3 Listen, Attend and Spell

The Listen Attend and Spell architecture (LAS) is the main Idea around which this thesis revolves. This section is based on [1]. The las-network consists of two mayor parts, the listener and the speller. The listener is a pyramidal recurrent neural net. It accepts filter bank spectra \mathbf{x}_n as inputs and produces high level output features \mathbf{h}_m . The speller in turn accepts the features as input and outputs distributions over Latin character sequences \mathbf{y}_p . An overview of the las-achrcitecture is given in figure 1.1.

1.3.1 The listener

The listener shown in figure 1.1 on the bottom, consists of Bidirectional Long Short Term Memory RNN (BLSTM) blocks. These blocks are arranged in a pyramidal structure, such that the time resolution is cut in half in every layer. This operation reduces the length U of the high level features \mathbf{h} . Without this compression the following attend and spell operation has a hard time extracting the relevant information. Additionally the compression reduces the problem complexity, which speeds up the training process significantly [1, page 4].

1.3.2 Attend and spell

The speller takes the features and produces a distribution over Latin character sequences as output. The computation of this output involves the context vector \mathbf{c}_i , the decoder state s_i , the features \mathbf{h} and the previous output y_i . The index i denotes time, $i - 1$ is used to refer to results from the last time step.

These values are computed using [1, page 4]:

$$s_i = \text{RNN}(s_{i-1}, y_{i-1}, c_{i-1}) \quad (1.2)$$

$$\mathbf{c}_i = \text{AttentionContext}(s_i, \mathbf{h}) \quad (1.3)$$

$$P(y_i | \mathbf{x}, y_{<i}) = \text{CharacterDistribution}(s_i, c_i) \quad (1.4)$$

The state follows from a recurrent neural net (RNN) made of a two layer LSTM. The attention mechanism, called AttentionContext above, computes a new context vector once every time step. This computation starts with the determination of the scalar energy \mathbf{e}_i , which is later transformed into

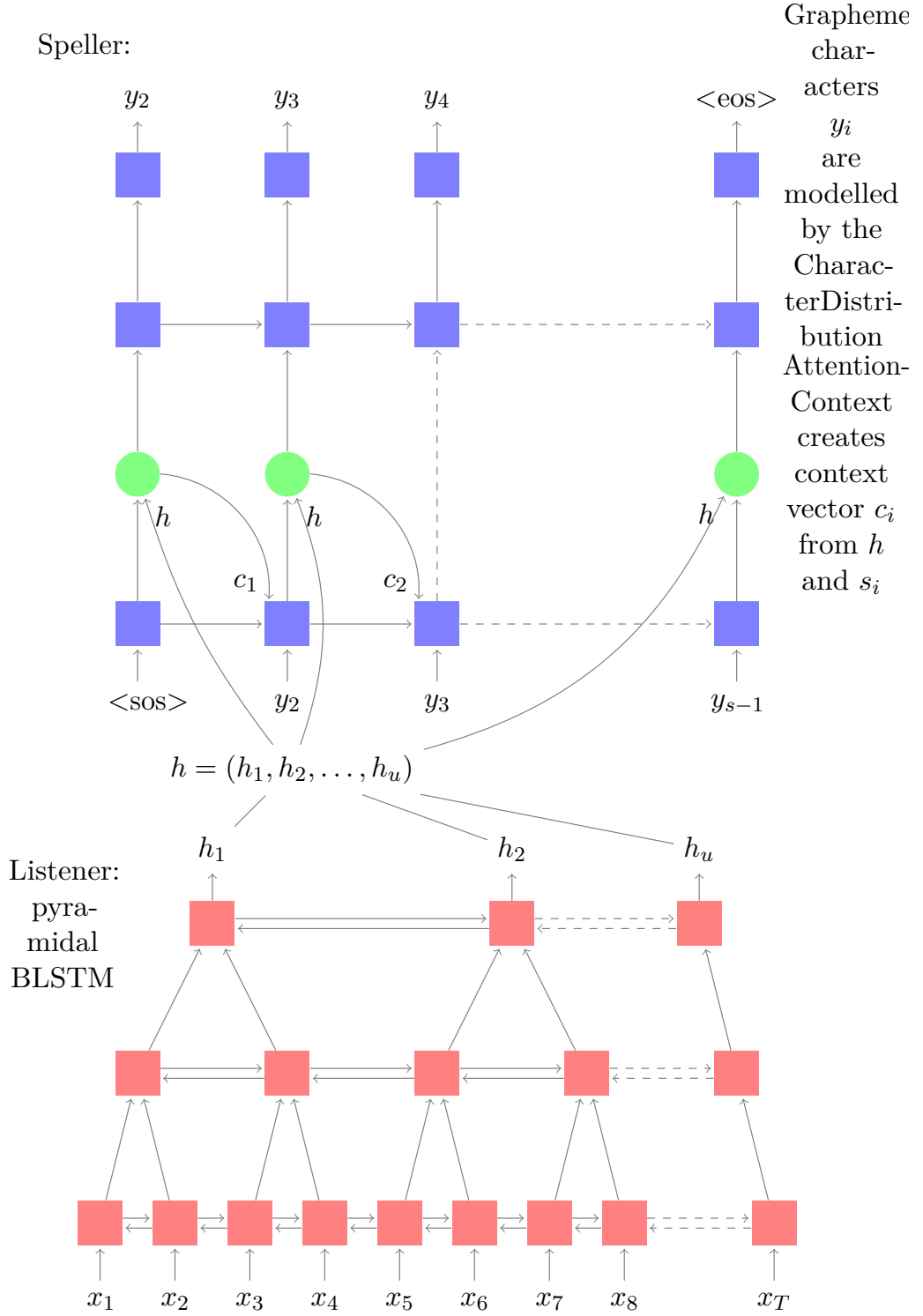


Figure 1.1: The LAS architecture [1, page 3]. BLSTM blocks are shown in red. LSTM blocks in blue and attention nets in green.

Bibliography

- [1] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. *arXiv preprint*, pages 1–16, 2015.
- [2] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. pages 1–28, 2016.
- [3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Master thesis filing card

Student: Moritz Wolter

Title: Building a state of the art speech recogniser

UDC: 621.3

Abstract:

In the past machine learning relied heavily on algorithms designed by experts to solve a specific task. Which lead to highly sophisticated algorithms, which could be grasped only by small groups of people. The human brain however does not work this way, although specialized areas exist, these areas consist of similar building blocks. Artificial neural networks attempt to mimic this layout. Similar algorithmic structures are used for a wide variety of tasks. This thesis deals with the application of neural networks in speech recognition. Replacing the various subsystems by one integrated network based approach.

Thesis submitted for the degree of Master of Science in Mathematical Engineering

Thesis supervisor: Prof. dr. ir. Patrick Wambacq

Assessor: Prof. dr. ir. Johan Suykens

Mentor: Ir. Vincent Renkens