Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Alex Graves[1] ALEX@IDSIA.CH
Santiago Fernández[1] SANTIAGO@IDSIA.CH
Faustino Gomez[1] TINO@IDSIA.CH
Jürgen Schmidhuber[1,2] JUERGEN@IDSIA.CH

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Galleria 2, 6928 Manno-Lugano, Switzerland

[2] Technische Universität München (TUM), Boltzmannstr. 3, 85748 Garching, Munich, Germany

Abstract

Many real-world sequence learning tasks require the prediction of sequences of labels from noisy, unsegmented input data. In speech recognition, for example, an acoustic signal is transcribed into words or sub-word units. Recurrent neural networks (RNNs) are powerful sequence learners that would seem well suited to such tasks. However, because they require pre-segmented training data, and post-processing to transform their outputs into label sequences, their applicability has so far been limited. This paper presents a novel method for training RNNs to label unsegmented sequences directly, thereby solving both problems. An experiment on the TIMIT speech corpus demonstrates its advantages over both a baseline HMM and a hybrid HMM-RNN.

1. Introduction

Labelling unsegmented sequence data is a ubiquitous problem in real-world sequence learning. It is particularly common in perceptual tasks (e.g. handwriting recognition, speech recognition, gesture recognition) where noisy, real-valued input streams are annotated with strings of discrete labels, such as letters or words.

Currently, graphical models such as hidden Markov Models (HMMs; Rabiner, 1989), conditional random fields (CRFs; Lafferty et al., 2001) and their variants, are the predominant framework for sequence labelling. While these approaches have proved successful for many problems, they have several drawbacks: (1) they usually require a significant amount of task specific knowledge, e.g. to design the state models for HMMs, or choose the input features for CRFs; (2) they require explicit (and often questionable) dependency assumptions to make inference tractable, e.g. the assumption that observations are independent for HMMs; (3) for standard HMMs, training is generative, even though sequence labelling is discriminative.

Recurrent neural networks (RNNs), on the other hand, require no prior knowledge of the data, beyond the choice of input and output representation. They can be trained discriminatively, and their internal state provides a powerful, general mechanism for modelling time series. In addition, they tend to be robust to temporal and spatial noise.

So far, however, it has not been possible to apply RNNs directly to sequence labelling. The problem is that the standard neural network objective functions are defined separately for each point in the training sequence; in other words, RNNs can only be trained to make a series of independent label classifications. This means that the training data must be pre-segmented, and that the network outputs must be post-processed to give the final label sequence.

At present, the most effective use of RNNs for sequence labelling is to combine them with HMMs in the so-called hybrid approach (Bourlard & Morgan, 1994; Bengio., 1999). Hybrid systems use HMMs to model the long-range sequential structure of the data, and neural nets to provide localised classifications. The HMM component is able to automatically segment the sequence during training, and to transform the network classifications into label sequences. However, as well as inheriting the aforementioned drawbacks of

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.