Chapter 7
Connectionist Temporal Classification

This chapter introduces the connectionist temporal classification (CTC) output layer for recurrent neural networks (Graves et al., 2006). As its name suggests, CTC was specifically designed for temporal classification tasks; that is, for sequence labelling problems where the alignment between the inputs and the target labels is unknown. Unlike the hybrid approach described in the previous chapter, CTC models all aspects of the sequence with a single neural network, and does not require the network to be combined with a hidden Markov model. It also does not require presegmented training data, or external post-processing to extract the label sequence from the network outputs. Experiments on speech and handwriting recognition show that a BLSTM network with a CTC output layer is an effective sequence labeller, generally outperforming standard HMMs and HMM-neural network hybrids, as well as more recent sequence labelling algorithms such as large margin HMMs (Sha and Saul, 2006) and conditional random fields (Lafferty et al., 2001).

Section 7.1 introduces CTC and motivates its use for temporal classification tasks. Section 7.2 defines the mapping from CTC outputs onto label sequences, Section 7.3 provides an algorithm for efficiently calculating the probability of a given label sequence, Section 7.4 derives the CTC loss function used for network training, Section 7.5 describes methods for decoding with CTC, experimental results are presented in Section 7.6 and a discussion of the differences between CTC networks and HMMs is given in Section 7.7.

7.1 Background

In 1994, Bourlard and Morgan identified the following reason for the failure of purely connectionist (that is, neural-network based) approaches to continuous speech recognition:

There is at least one fundamental difficulty with supervised training of a connectionist network for continuous speech recognition: a target function must be defined, even though the training is
Connectionist Temporal Classification

Fig. 7.1 CTC and framewise classification networks applied to a speech signal. The coloured lines are the output activations, corresponding to the probabilities of observing phonemes at particular times. The CTC network predicts only the sequence of phonemes (typically as a series of spikes, separated by ‘blanks’, or null predictions, whose probabilities are shown as a grey dotted line), while the framewise network attempts to align them with the manual segmentation (vertical lines).

In other words, neural networks require separate training targets for every segment or timestep in the input sequence. This has two important consequences. Firstly, it means that the training data must be presegmented to provide the targets. Secondly, since the network only outputs local classifications, the global aspects of the sequence, such as the likelihood of two labels appearing consecutively, must be modelled externally. Indeed, without some form of post-processing the final label sequence cannot reliably be inferred at all.

In Chapter 6 we showed how RNNs could be used for temporal classification by combining them with HMMs in hybrid systems. However, as well as inheriting the disadvantages of HMMs (which are discussed in depth in Section 7.7), hybrid systems do not exploit the full potential of RNNs for long-range sequence modelling. It therefore seems preferable to train RNNs directly for temporal classification tasks.

Connectionist temporal classification (CTC) achieves this by allowing the network to make label predictions at any point in the input sequence, so long as the overall sequence of labels is correct. This removes the need for presegmented data, since the alignment of the labels with the input is no longer important. Moreover, CTC directly outputs the probabilities of the complete label sequences, which means that no external post-processing is required to use the network as a temporal classifier.

Figure 7.1 illustrates the difference between CTC and framewise classification applied to a speech signal.