

CLAIMS

1. A computer-implemented method for training a model, comprising:
obtaining a training sample including an input sequence of observations and a target sequence of symbols having length different from the input sequence of observations;
feeding the input sequence of observations into the model to obtain a sequence of predictions;
shifting the sequence of predictions by an amount with respect to the input sequence of observations;
and
updating the model based on a loss using a shifted sequence of predictions and the target sequence of the symbols.
2. The method of claim 1, wherein the sequence of predictions is shifted forward with respect to the input sequence of observations to generate the shifted sequence of predictions and wherein the model is unidirectional.
3. The method of claim 2, wherein the model is a recurrent neural network based model.
4. The method of claim 2, wherein the loss is CTC (Connectionist Temporal Classification) loss.
5. The method of claim 2, wherein shifting the sequence of predictions comprises:
adjusting so that the lengths of the shifted sequence of predictions and the input sequence of observations are the same.
6. The method of claim 2, wherein shifting the sequence of predictions and updating the model using the shifted sequence of predictions are performed at a predetermined rate.
7. The method of claim 6, wherein the predetermined rate ranges from approximately 5 to 40%.
8. The method of claim 2, wherein the amount to shift is fixed.
9. The method of claim 2, wherein the amount to shift is determined probabilistically within a predetermined range.
10. The method of claim 1, wherein the model is a neural network based model having a plurality of parameters, wherein feeding the input sequence comprises conducting a forward-propagation through the neural network based model, and wherein updating the model comprises performing a back-propagation through the neural network based model to update the plurality of parameters.

11. The method of claim 1, wherein the model comprises an end-to-end speech recognition model, each observation in the input sequence of the training sample represents an acoustic feature and each symbol in the target sequence of the training sample represents a phone, a context dependent phone, a character, a word-piece, or a word.
12. A computer system for training a model, by executing program instructions, the computer system comprising:
a memory storing the program instructions;
processing circuitry in communications with the memory for executing the program instructions, wherein the processing circuitry is configured to:
obtain a training sample including an input sequence of observations and a target sequence of symbols having length different from the input sequence of observations;
feed the input sequence of observations into the model to obtain a sequence of predictions;
shift the sequence of predictions by an amount with respect to the input sequence of observations; and
update the model based on a loss using a shifted sequence of predictions and the target sequence of the symbols.
13. The computer system of claim 12, wherein the sequence of predictions is shifted forwardly with respect to the input sequence of observations to generate the shifted sequence of predictions and wherein the model is unidirectional.
14. The computer system of claim 13, wherein the processing circuitry is also configured to:
adjust to shift the sequence of predictions so that the lengths of the shifted sequence of predictions and the input sequence of observations are the same.
15. The computer system of claim 13, wherein shifting of the sequence of predictions and updating of the model using the shifted sequence of predictions are performed at a predetermined rate.
16. The computer system of claim 12, wherein the model is a recurrent neural network based model.
17. A computer program product for training a model, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a computer to cause the computer to perform a method comprising:
obtaining a training sample including an input sequence of observations and a target sequence of symbols having length different from the input sequence of observations;
feeding the input sequence of observations into the model to obtain a sequence of predictions;

shifting the sequence of predictions by an amount with respect to the input sequence of observations;
and

updating the model based on a loss using a shifted sequence of predictions and the target sequence of the symbols.

18. The computer program product of claim 17, wherein the sequence of predictions is shifted forwardly with respect to the input sequence of observations to generate the shifted sequence of predictions and wherein the model is unidirectional.

19. The computer program product of claim 18, wherein shifting the sequence of predictions comprises:
adjusting so that the lengths of the shifted sequence of predictions and the input sequence of observations are the same.

20. The computer program product of claim 18, wherein shifting the sequence of predictions and updating the model using the shifted sequence of predictions are performed at a predetermined rate.

21. The computer program product of claim 17, wherein the model is a recurrent neural network based model.

22. A computer program product for decoding using a model, the computer program product comprising a computer readable storage medium having program instructions embodied therewith, the program instructions executable by a computer to cause the computer to perform a method comprising:

feeding an input into the model to obtain an output;

wherein the model is trained by:

obtaining a training sample including an input sequence of observations and a target sequence of symbols having length different from the input sequence of observations;

feeding the input sequence of observations into the model to obtain a sequence of predictions;

shifting the sequence of predictions by an amount with respect to the input sequence of observations;

and

updating the model based on a loss using a shifted sequence of predictions and the target sequence of the symbols.