Recurrent Neural Networks (RNNs) are very powerful sequence models that do not enjoy widespread use because it is extremely difficult to train them properly. Fortunately, recent advances in Hessian-free optimization have been able to overcome the difficulties associated with training RNNs, making it possible to apply them successfully to challenging sequence problems. In this paper we demonstrate the power of RNNs trained with the new Hessian-Free optimizer (HF) by applying them to character-level language modeling tasks. The standard RNN architecture, while effective, is not ideally suited for such tasks, so we introduce a new RNN variant that uses multiplicative (or "gated") connections which allow the current input character to determine the transition matrix from one hidden state vector to the next. After training the multiplicative RNN with the HF optimizer for five days on 8 high-end Graphics Processing Units, we were able to surpass the performance of the best previous single method for character-level language modeling – a hierarchical non-parametric sequence model. To our knowledge this represents the largest recurrent neural network application to date.

1. Introduction

Recurrent Neural Networks (RNNs) form an expressive model family for sequence tasks. They are powerful because they have a high-dimensional hidden state with non-linear dynamics that enable them to remember and process past information. Furthermore, the gradients of the RNN are cheap to compute with backpropagation through time. Despite their attractive qualities, RNNs failed to become a

Appearing in Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

mainstream tool in machine learning due to the difficulty of training them effectively. The cause of this difficulty is the very unstable relationship between the parameters and the dynamics of the hidden states, which manifests itself in the "vanishing/exploding gradients problem" (Bengio et al., 1994). As a result, there has been surprisingly little research on standard RNNs in the last 20 years, and only a few successful applications using large RNNs (Robinson, 2002; Pollastri et al., 2002), including a recent notable application of RNNs as a word-level language model (Mikolov et al., 2010).

Recently, Martens (2010) developed a greatly improved variant of Hessian-Free optimization (HF) which was powerful enough to train very deep neural networks from random initializations. Since an RNN can be viewed as an extremely deep neural network with weight sharing across time, the same HF optimizer should be able to train RNNs.

Fortunately, Martens & Sutskever (2011) were able to show that this is indeed the case, and that this form of non-diagonal, 2nd-order optimization provides a principled solution to the vanishing gradients problem in RNNs. Moreover, with the addition of a novel damping mechanism, Martens & Sutskever (2011) showed that the HF optimizer is robust enough to train RNNs, both on pathological synthetic datasets known to be impossible to learn with gradient descent, and on complex and diverse real-world sequence datasets.

The goal of the paper is to demonstrate the power of large RNNs trained with the new Hessian-Free optimizer by applying them to the task of predicting the next character in a stream of text. This is an important problem because a better character-level language model could improve compression of text files (Rissanen & Langdon, 1979) and make it easier for people with physical disabilities to interact with computers (Ward et al., 2000). More speculatively, achieving the asymptotic limit in text compression requires an understanding that is “equivalent to intelligence” (Hutter, 2006). Good compression can be achieved by exploiting simple regularities such as the vocabulary and the syntax of the relevant languages and the shallow associations exem-