**[Semi-supervised Sequence Learning](https://arxiv.org/abs/1511.01432)**

**Andrew M. Dai**

Google Inc.

adai@google.com

**Quoc V. Le**

Google Inc.

qvl@google.com

**Abstract**

We present two approaches that use unlabeled data to improve sequence learning with recurrent networks. The first approach is to predict what comes next in a sequence, which is a conventional language model in natural language processing. The second approach is to use a sequence autoencoder, which reads the input sequence into a vector and predicts the input sequence again. These two algorithms can be used as a “pretraining” step for a later supervised sequence learning algorithm. In other words, the parameters obtained from the unsupervised step can be used as a starting point for other supervised training models. In our experiments, we find that long short term memory recurrent networks after being pretrained with the two approaches are more stable and generalize better. With pretraining, we are able to train long short term memory recurrent networks up to a few hundred timesteps, thereby achieving strong performance in many text classification tasks, such as IMDB, DBpedia and 20 Newsgroups.

**摘要**

我们提出了两种使用未标记数据来改进循环网络序列学习的方法。第一种方法是预测序列中接下来的内容，这是自然语言处理中的传统语言模型。第二种方法是使用序列自动编码器，它将输入序列读入为向量并再次预测输入序列。这两种算法可以用作后期监督序列学习算法的“预训练”步骤。换句话说，从无监督步骤获得的参数可以用作其他监督训练模型的起点。在我们的实验中，我们发现在使用这两种方法预训练后的长期短期记忆循环网络更稳定并且能够更好地扩展。通过预训练，我们能够训练LSTM长达几百步长，从而在许多文本分类任务中可实现强大的性能，例如IMDB，DBpedia和20个新闻组。