**Character-Level Language Modeling with Deeper**

**Self-Attention**

**Rami Al-Rfou∗ Dokook Choe∗ Noah Constant∗**

**Mandy Guo∗ Llion Jones∗**

Google AI {rmyeid, choed, nconstant, xyguo, llion}@google.com

**Abstract**

LSTMs and other RNN variants have shown strong performance on character- level language modeling. These models are typically trained using truncated back- propagation through time, and it is common to assume that their success stems from their ability to remember long-term contexts. In this paper, we show that a deep (64-layer) transformer model [42] with fixed context outperforms RNN variants by a large margin, achieving state of the art on two popular benchmarks—1.13 bits per character on text8 and 1.06 on enwik8. To get good results at this depth, we show that it is important to add auxiliary losses, both at intermediate network layers and intermediate sequence positions.

**摘要**

LSTM和一些RNN的其他变体在字符级语言建模方面有出强大的表现。这些模型通常使用截断的反向传播进行训练，并且通常假设它们的成功源于它们拥有长期记忆的能力。在本文中，我们展示了具有固定上下文的深（64层）变换器模型[42]大大优于RNN变体，在两个流行的基准测试中实现了最先进的技术 - 在text8上每个字符1.13位，在enwik8上为1.06，在这个深度上获得良好的结果，我们表明在中间网络层和中间序列位置添加辅助损失是很重要的。