智能感知认知实践实践项目：语言模型

涂宇清 522030910152

1 简介

在本次实验中，我使用了多种模型结构（如 RNN、LSTM、GRU、Transformer 等），通过消融实验观察模型在不同超参数下的表现，最优化测试集上的 PPL，得出每种模型在最优性能下的超参数配置。在实验过程中，我还使用了 Weight and Biases（图1.1）监控实验过程中的各项指标，为后续实验分析提供了不少便利。此外，我还尝试分析在默认参数配置下，Transformer 性能不如 LSTM 的原因，并提出了可行的改进方法。

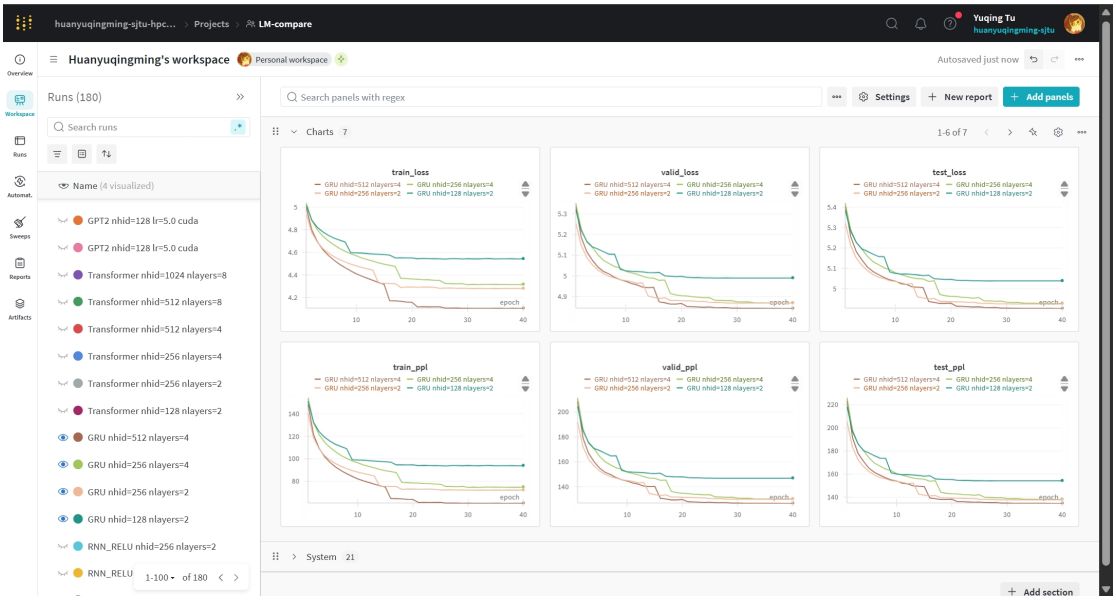
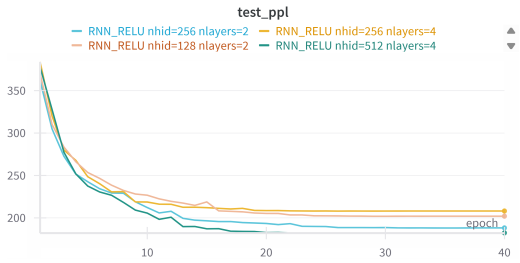


图 1.1　Weight and Biases 页面

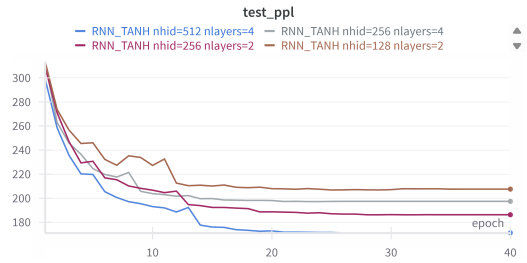
2 消融实验

2.1 模型结构

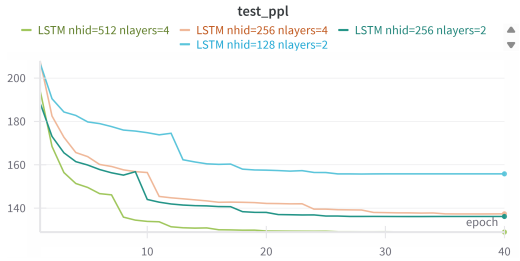
在本节中，将讨论不同模型中隐藏层数量与维度对模型性能的影响。除了模型结构外，其他超参数均保持默认。实验结果如图2.1所示。



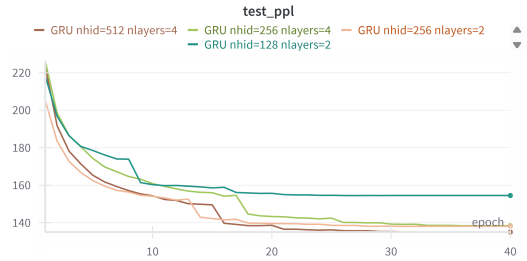
(a) RNN 模型 +ReLU 激活函数：4-512 达到了最佳性能，测试集上的 PPL 为 182.30



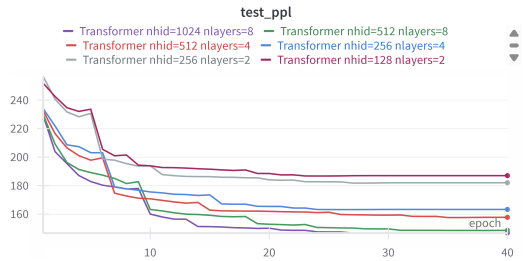
(b) RNN 模型 +tanh 激活函数：4-512 达到了最佳性能，测试集上的 PPL 为 171.26



(c) LSTM 模型：4-512 达到了最佳性能，测试集上的 PPL 为 128.93



(d) GRU 模型：4-512 达到了最佳性能，测试集上的 PPL 为 134.97



(e) Transformer 模型：8-1024 达到了最佳性能，测试集上的 PPL 为 146.77

图 2.1　不同模型中隐藏层数量与维度对模型性能的影响

上述模型参数量均在 60M 以下。由实验结果可以看出，LSTM 模型在 4-512 的配置下达到了最佳性能。

在 GRU、LSTM 这些基于 RNN 的模型中，控制隐藏层维度不变，增加隐藏层数量反而会导致模型性能小幅下降；控制隐藏层数量不变，增加隐藏层维度能

使得模型性能上升。而在 Transformer 模型中，增加隐藏层数量和维度均能使得模型性能上升。

2.2 词嵌入维度

词嵌入维度是影响模型性能的重要因素之一。词嵌入维度过小可能导致信息丢失，而过大则可能导致模型过拟合。在本节中，将讨论不同模型中词嵌入维度对模型性能的影响。除了词嵌入维度外，其他超参数均使用节2.1中最佳的参数。实验结果如图2.2所示。

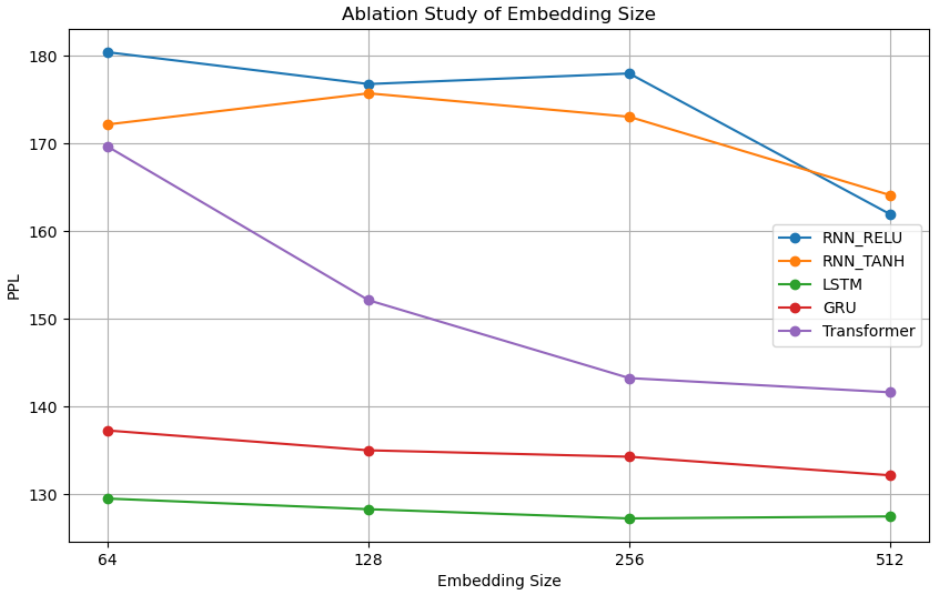


图 2.2　不同模型中词嵌入维度对模型性能的影响

可以看出，随着词嵌入维度的增加，PPL 整体呈下降趋势。这是因为更高的词嵌入维度可以捕捉到更多的语义信息，从而提高模型的性能。然而，过大的词嵌入维度也可能导致模型过拟合，因此在实际应用中需要根据具体情况进行调整。

2.3 Dropout Rate

Dropout 是一种常用的正则化方法，可以有效防止模型过拟合。在本节中，将讨论不同模型中 Dropout Rate 对模型性能的影响。除了 Dropout Rate 外，其他超参数均使用节2.1中最佳的参数。实验结果如图2.3所示。

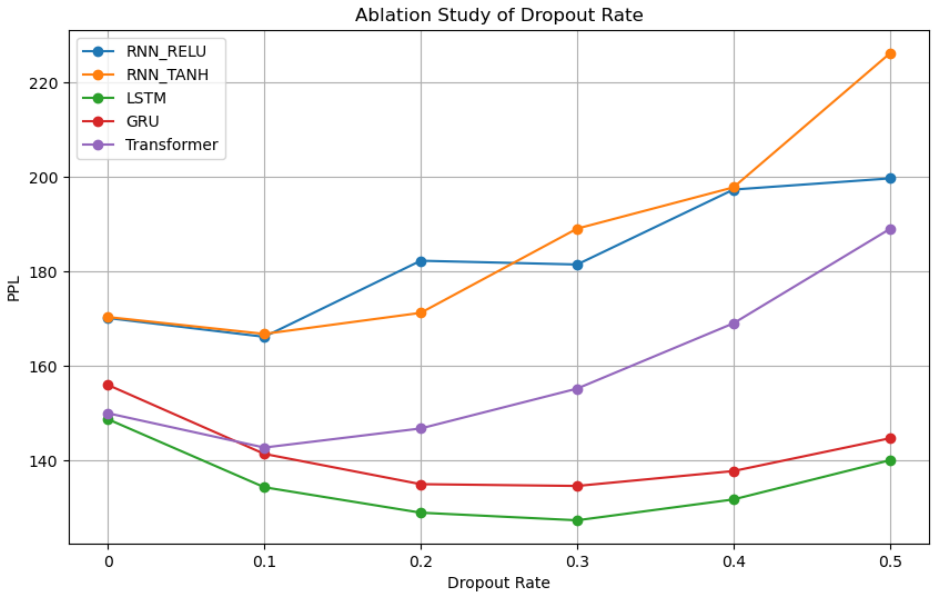


图 2.3　不同模型中 Dropout Rate 对模型性能的影响

可以看出，当 dropout rate 在一个中等区间时，模型性能较好。当 dropout rate 过大时，会导致模型丧失太多信息，从而影响模型的性能。当 dropout rate 过小时，模型可能会过拟合训练数据，从而导致性能下降。

3 为何 Transformer 性能不如 LSTM

在此前的消融实验中，我们发现在超参数一致的情况下，Transformer 的性能始终不如 LSTM。这是很奇怪的，因为 Transformer 模型在理论上应该比 LSTM 模型更强大。在查看参数配置情况后，发现序列长度`bptt`设置为了 35，这实在是有点小，无法让 Transformer 充分发挥其对长序列建模的优势。

因此，我们尝试将`bptt`调高后进行实验，实验结果如图3.1所示。

可以看出，随着序列长度的增加，Transformer 的性能显著提升，而 LSTM 的性能逐渐下降。当序列长度大于 100 时，Transformer 的性能明显优于 LSTM。这是因为 Transformer 能够更好地捕捉长距离依赖关系。

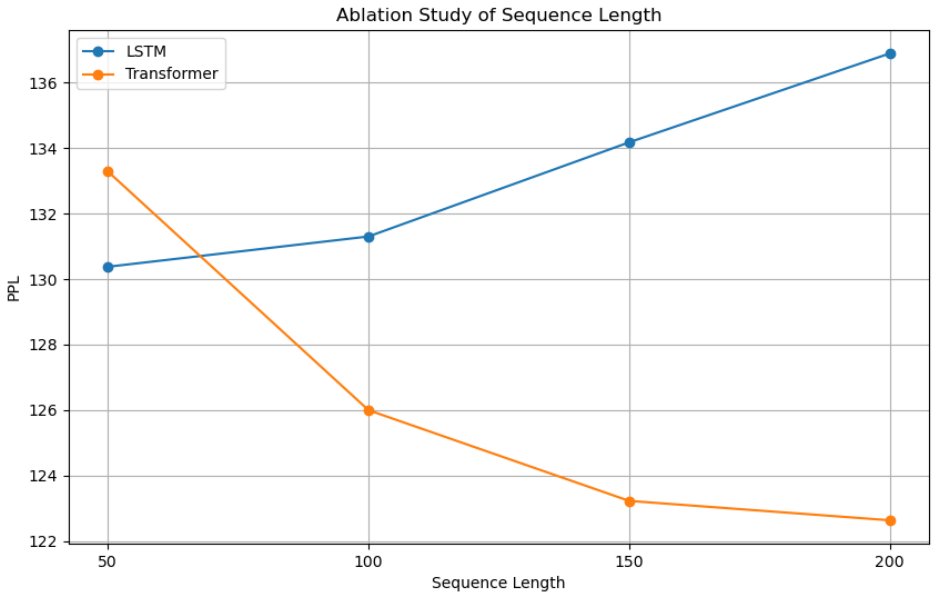


图 3.1　不同模型中序列长度对模型性能的影响

4 更先进的语言模型

在本节中，我们将尝试使用更先进的语言模型 GPT-2 进行实验。GPT-2 是一个基于 Transformer Decoder 的自回归语言模型。由词嵌入与位置嵌入组成的嵌入层、多个 Transformer Decoder 层、线性层与 Softmax 组成。其训练目标是最大化下一个 token 的条件概率。

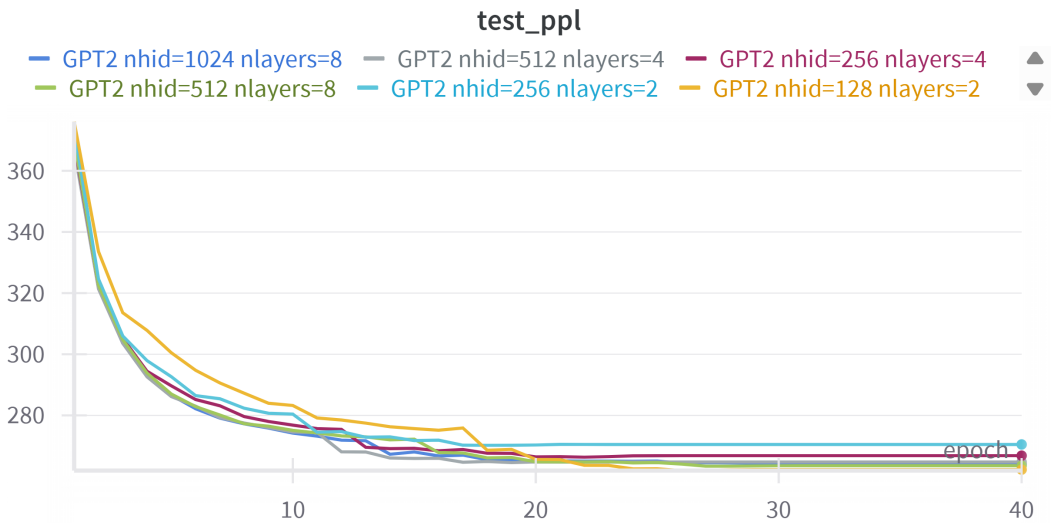


图 4.1　GPT-2 实验结果

在实验过程中，我们观察到 GPT-2 的性能未达到预期。在训练集、验证集和测试集上，其 PPL 均显著高于其他模型。进一步分析 GigaSpeech 数据集发现，该数据集以简短的口语句子为主，并包含一定比例的标记错误与语义模糊的词汇。由于 GPT-2 采用基于自回归机制的单向上下文建模方式，这类输入特征可能限制了其有效捕捉长程上下文和语言结构的能力，进而影响了整体建模效果。

5 结论

在本次实验中，通过对多种语言模型（RNN、LSTM、GRU、Transformer 等）进行消融实验，探究了不同超参数对模型性能的影响，并利用 Weight and Biases 监控实验指标，最终得出了每种模型在最优性能下的超参数配置。实验表明，LSTM 模型在特定配置下表现出色，而 Transformer 模型在初始超参数配置下性能不如 LSTM，但通过调整序列长度等超参数后，性能显著提升且在长序列建模方面优势明显。同时，尝试使用更先进的 GPT-2 模型进行实验，虽然其在特定数据集上的表现未达预期，但也为后续研究提供了方向。总体而言，本次实验深入分析了不同语言模型的性能特点及影响因素，为实际应用中模型的选择与优化提供了有益的参考。

6 代码及复现方式

详情见GitHub。