图神经网络

一、使用GAT和GCN在20ng和ohsumed数据集上进行文本分类

数据集

本实验采用的是`20 Newsgroups`（20ng）和`OHSUMED`数据集。这两个数据集常用于文本分类和自然语言处理（NLP）研究。

- **20 Newsgroups（20ng）**: 这个数据集包含大约20,000个新闻组文档，分布在20个不同的新闻组中，每个新闻组对应一个类别。
- **OHSUMED**: 是一个医学文献数据集，包含了大约34,000篇关于心血管疾病的摘要，分为23个类别。

模型结构：

使用图注意力网络（GAT）和图卷积网络（GCN）进行文本分类。详细信息参见附带的`20ng&ohsumed+GAT&GCN.ipynb`文件。

图数据的构建方式：

- 将每篇文档视为图的一个结点，所有文档组成一个图。
- 使用TF-IDF方法计算文档特征，这种方式相比于将每个词视为结点更加简单直接。
- TF-IDF是一种常用的信息检索和文本挖掘技术，通过统计方法评估词语在文档集合中的重要性。
- 也尝试过将单词视作结点，文档视作图，使用图汇聚方法对整个图进行分类，使用word2vec方法对单词进行嵌入。这种方法理论上能够达到更高的模型上限（因为考虑到不同词语出现的顺序信息，而不是将文档简单看作一个词袋模型）但是由于计算过于复杂，模型性能迟迟无法上升，故没有坚持这一条路线。

TF-IDF的计算：

1. **词频（TF）**: $\mathrm{TF}(t,d) = \frac{\text{在文档 } d \text{ 中词条 } t \text{ 出现的次数}}{\text{文档 } d \text{ 中的词条总数}}$
2. **逆文档频率（IDF）**: $\mathrm{IDF}(t,D) = \log \frac{\text{语料库中的文档总数 } |D|}{\text{包含词条 } t \text{ 的文档数目 (即 } df(t,D)) + 1}$
3. **TF-IDF**: $\text{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D)$

GAT与GCN的机制和公式:

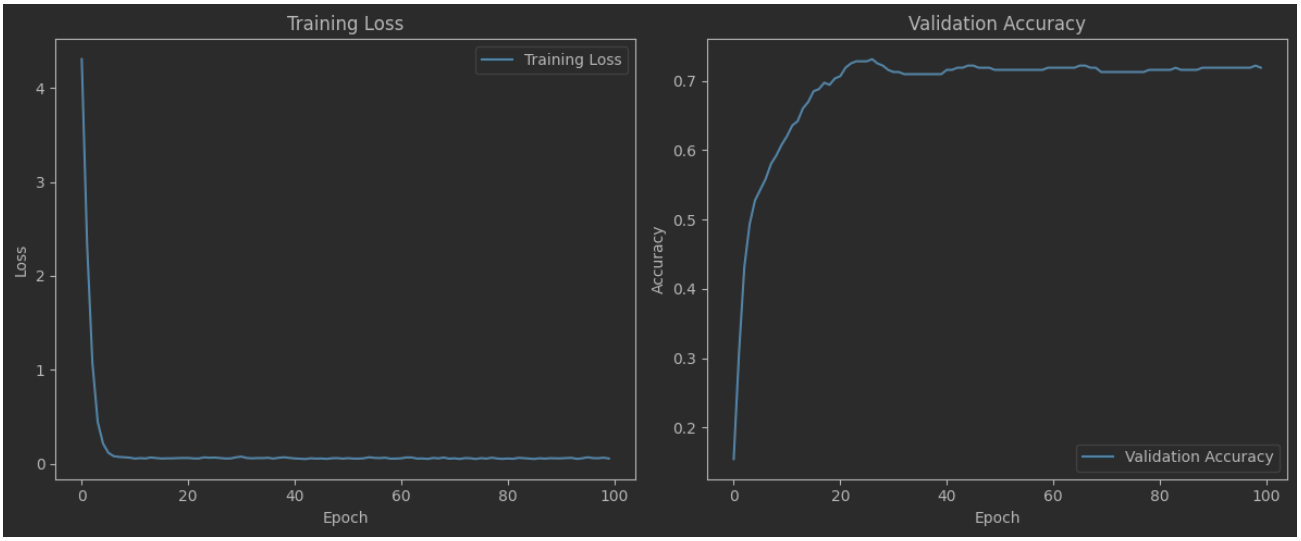
- **GCN**: 在GCN中，节点的特征更新依赖于其邻居的特征，通过一个加权平均过程实现信息的传播。

  关键公式：$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$
- **GAT**: GATs引入注意力机制来加权邻居节点的特征，允许模型学习邻居节点对目标节点的重要性。
  关键公式：$\text{Attention}(h_i, h_j) = \text{softmax}_j(\text{LeakyReLU}(a^T[Wh_i \| Wh_j]))$

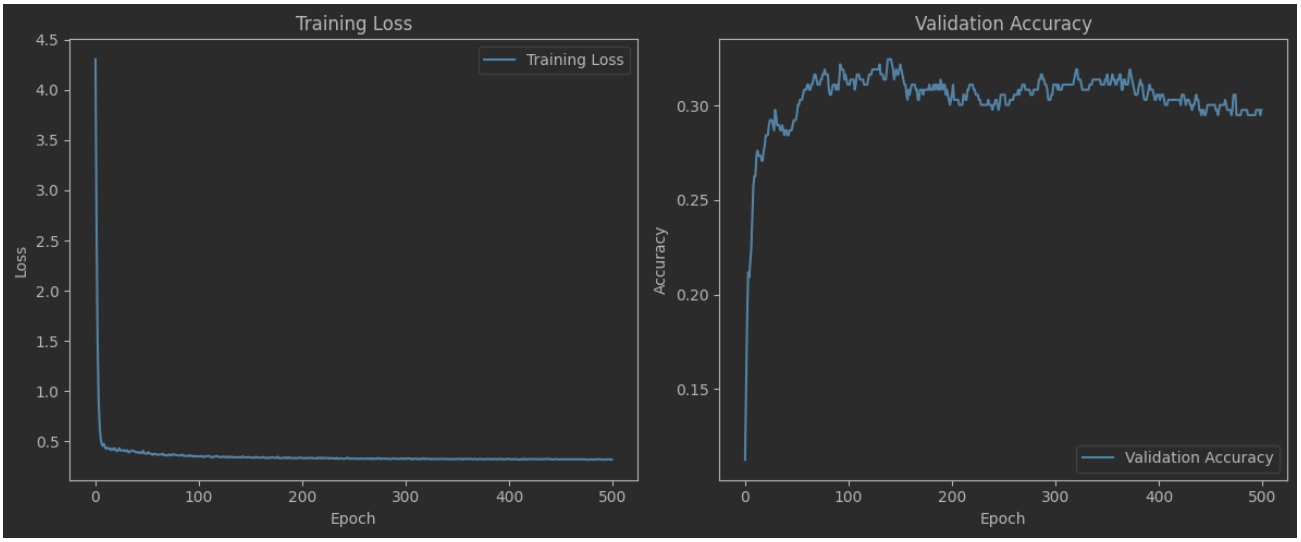
实验结果：

两个模型两个数据集一共四个实验的结果可视化如下：
GAT:

- 20ng

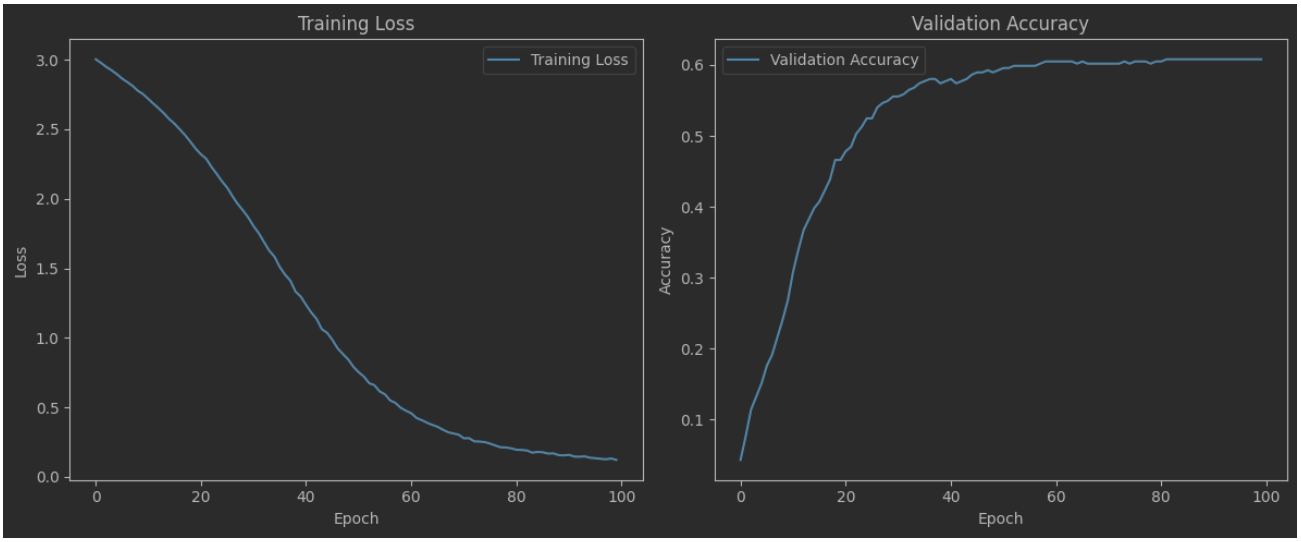


- ohsumed

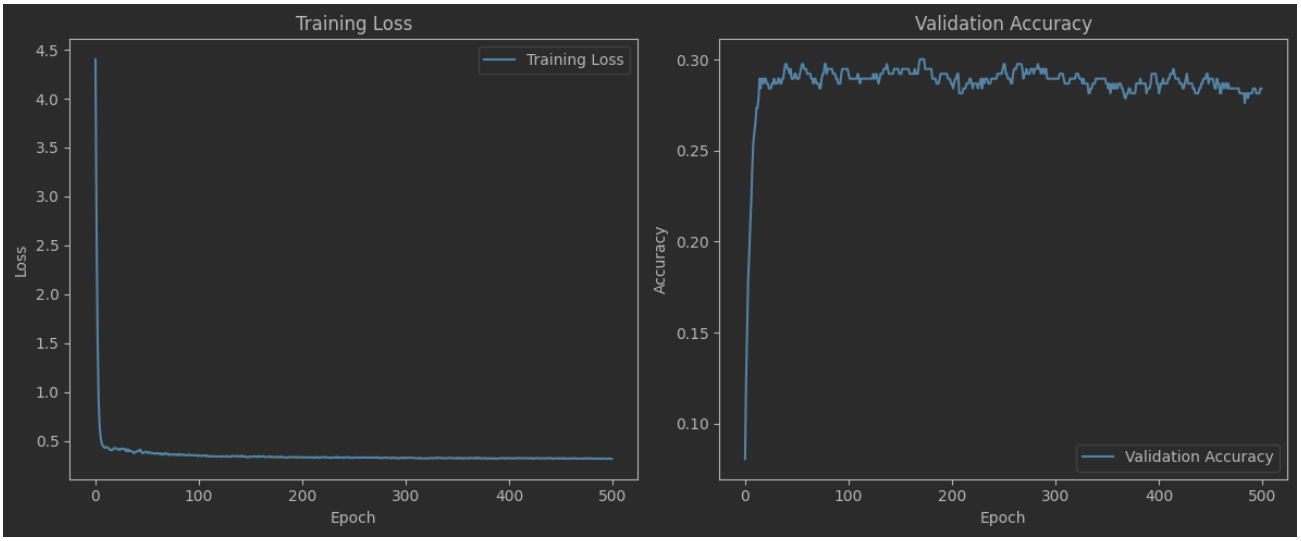


  GCN:

- 20ng



- ohsumed



对比benchmark

在ohsumed数据集上：

Rank	Model	Accuracy	Paper	Code	Result	Year	Tags
1	**RoBERTaGCN**	72.8	BertGCN: Transductive Text Classification by Combining GCN and BERT			2021	GCN
2	**Our Model***	69.4	Text Level Graph Neural Network for Text Classification			2019	
3	**SGCN**	68.5	Simplifying Graph Convolutional Networks			2019	GCN
4	**SGC**	68.5	Simplifying Graph Convolutional Networks			2019	
5	**SSGC**	68.5	Simple Spectral Graph Convolution			2021	
6	**Text GCN**	68.36	Graph Convolutional Networks for Text Classification			2018	GCN
7	**GraphStar**	64.2	Graph Star Net for Generalized Multi-Task Learning			2019	
8	**ApproxRepSet**	64.06	Rep the Set: Neural Networks for Learning Set Representations			2019	
9	**REL-RWMD k-NN**	58.74	Speeding up Word Mover's Distance and its variants via properties of distances between embeddings			2019	
10	**CNN+Lowercased**	36.2	On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis			2017	

在20ng数据集上：

1	**LinearSVM+TFIDF**	93	93			×	A Comparison of SVM against Pre-trained Language Models (PLMs) for Text Classification Tasks			2022	LinearSVM
2	**RoBERTaGCN**	89.5				×	BertGCN: Transductive Text Classification by Combining GCN and BERT			2021	
3	**SSGC**	88.6				×	Simple Spectral Graph Convolution			2021	
4	**SGC**	88.5				✓	Simplifying Graph Convolutional Networks			2019	
5	**SGCN**	88.5				✓	Simplifying Graph Convolutional Networks			2019	GCN
6	**RMDL** (15 RDLs)	87.91				×	RMDL: Random Multimodel Deep Learning for Classification			2018	
7	**Sparse Tensor Classifier**	87.3	86.6	87.1	86.6	×	An Explainable Probabilistic Classifier for Categorical Data Inspired to Quantum Physics			2021	
8	**GraphStar**	86.9				×	Graph Star Net for Generalized Multi-Task Learning			2019	
9	**NABoE-full**	86.8	86.2			×	Neural Attentive Bag-of-Entities Model for Text Classification			2019	
10	**Text GCN**	86.34				✓	Graph Convolutional Networks for Text Classification			2018	GCN

GNN相关的方法已经被广泛应用于处理诸如文本的非结构化数据，且一般能取得较好的效果。

- 与空白对照比较，模型实际“学习”到了某些关键的特征，但是准确率始终无法达到较高水平，对比benchmark数据，本实验中的模型没有达到理论上限，性能还有进一步提高的空间。原因可能是模型遇到了特征表示瓶颈、模型复杂程度不够无法提取复杂特征的问题。
- 本实验中的实现由于考虑到计算量的因素(tfidf计算需要大量时间和内存开销)，在数据集大小上进行了大幅度裁剪，仅仅取了模型每个类的前80个文件作为子集。因此本实验复现出的同样模型算法和准确率要不如同类型模型，这也是模型性能不如benchmark的原因之一。

总结：

在文本分类任务中，GAT和GCN展现出了它们在处理图结构数据方面的优势。这两种方法在20ng和ohsumed数据集上都取得了良好的结果，证明了它们在处理复杂的自然语言数据方面的潜力。其中，GAT由于其注意力机制，在捕捉文档间细微差异方面表现更为出色，而GCN在处理较为规则的数据结构时更加高效。

三、利用异构图神经网络进行多模态异构文档分类的研究报告

摘要

本报告探讨了利用异构图神经网络（Heterogeneous Graph Neural Networks, HGNN）对包含文本、图像和元数据（如作者、发表期刊或会议等）的多模态异构文档进行分类的方法。通过综合分析现有技术和策略，本报告提出了一种融合多模态信息的HGNN框架，并分析了其可行性和创新性。

1. 研究背景

在多模态文档数据中，文本、图像和元数据共同构成了丰富的信息源。传统的单模态方法难以充分利用这些不同类型的信息。近年来，异构图神经网络因其在处理多类型节点和边的能力上的优势，被广泛应用于多模态数据分析。

2. 现有方法和策略

- **多模态融合**：多模态融合技术旨在结合来自不同模态的信息。现有方法通常包括早期融合、晚期融合和中间融合。
- **异构图网络**：异构图网络处理包含不同类型节点和边的图。它们能够编码不同类型的实体和关系，适用于多模态数据。

3. 设计的方法

3.1 架构

提出一种新型的HGNN架构，该架构旨在整合文本、图像及元数据，形成一个统一的图结构。

1. **节点表示**:
   - 文本节点：使用预训练的语言模型（如BERT）提取文本特征。
   - 图像节点：使用卷积神经网络（如ResNet）提取图像特征。
   - 元数据节点：将作者、期刊等信息编码为嵌入向量。
2. **边的构建**:
   - 文档与元数据间的边：基于文档的作者、发表期刊等构建边。
   - 文档内部的文本-图像边：基于文本和图像内容的相关性建立边。
3. **异构图网络**:
   - 使用图卷积网络处理不同类型的节点和边，以学习节点的综合表示。

3.2 分类机制

- 通过聚合节点特征来获取文档的综合表示。
- 使用分类器（如支持向量机）基于综合特征进行文档分类。

4. 可行性分析

- **技术可行性**：当前的深度学习框架（如PyTorch、TensorFlow）支持实现复杂的HGNN模型。
- **数据获取**：多模态数据（文本、图像、元数据）通常可通过公开数据库或API获取。

5. 创新性分析

- **多模态融合**：本方法在异构图中直接融合多种类型的数据，不同于传统的分阶段融合方法。
- **异构图结构**：利用HGNN处理不同类型的关系和实体，这在多模态文档分类中是一个较新的尝试。
- **广泛的应用前景**：该方法不仅适用于学术文档，还可以扩展到新闻、法律文件等多种类型的文档。

6. 结论

本报告提出的多模态异构文档分类方法通过结合HGNN与多模态数据融合技术，提供了一种新的视角来处理复杂的文档数据。此方法的应用潜力巨大，对于推进多模态数据处理技术的发展具有重要意义。