# 基于表格及其上下文的对比学习预训练
# Context-based Table Pre-training with Contrastive Learning

金笑缘　指导老师：化柏林

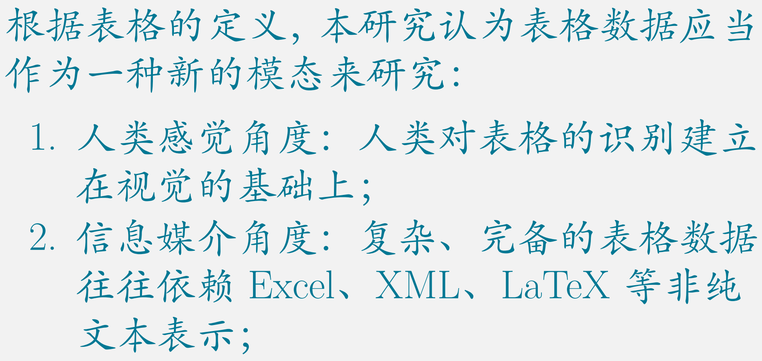
## 研究简介

表格数据是除了图像、文本等数据之外的另一种重要的数据格式。相比普通文本和数字，表格数据由于独特的二维对齐网格组织形式以及丰富的格式特征，具有更高的信息密度。已有的表格预训练模型主要基于表格文本的语言表示模型，没有充分利用表格的上下文信息，本研究尝试引入表格上下文信息用于加强表格表示模型，使用配对的表格-标题样本对作为正例样本进行对比学习中间预训练，以获得更好的表格特征表示用于下游任务。

本研究构造了基于科技文献文献 LaTeX 源代码的表格数据集 ar5iv，并使用自然语言方法对表格配对的标题进行数据增强。同时本研究提出了 TabCL（Tabel Contrastive Learning）模型，使用 TAPAS 表格编码模型和 BERT 文本编码模型获得两个模态的特征向量，并计算多正样本对比学习损失函数。获得预训练模型后在表格问答任务上进行微调，用于比较对比学习模型的性能。

本研究的主要贡献为：

- 提出表格作为一种独立模态的观点；
- 抓取 ar5iv 论文数据库的 HTML 源代码并抽取表格和标题，为表格领域增加了新的科技文献表格数据源；
- 创新地将多模态对比学习方法应用到表格-文本的多模态协同表示学习；
- 对标题文本进行了正样本数据增强实验，分析了不同增强方法对对比学习模型性能的影响。




## 观点：表格也是一种模态

模态（Modality）是一种广义的信息来源概念，可以从人类感觉、信息媒介、传感器等多种角度对模态进行分类。

目前有关表格的研究都是在自然语言处理即文本数据模态的框架下开展，近几年的研究方法大都建立在基于 Transformer 结构的 BERT 语言表示模型的基础上。但是事实上，根据表格的定义，本研究认为表格数据应当作为一种新的模态来研究：

1. 人类感觉角度：人类对表格的识别建立在视觉的基础上；
2. 信息媒介角度：复杂、完备的表格数据往往依赖 Excel、XML、LaTeX 等非纯文本表示；
3. 传感器角度：表格数据不同于纯文本，需要专门的解析器；
4. 数据组织形式角度：表格具有基于网格的二维对齐属性，可以灵活合并，单元格内容丰富且具有视觉特征。

## 模型方法：表格与文本编码器

本研究提出 TabCL 预训练模型，如左图所示：该模型为双塔编码器结构，基于 End-to-End 的对比学习框架，使用本文提出的多正样本对比学习损失函数。

为了获得更大的训练 BatchSize，本研究使用两个“小编码器”节约显存。使用 TAPAS-small 表格编码器获得表格 [CLS] 嵌入表示向量，DistilBERT 编码器获得表格上下文的 [CLS] 嵌入表示向量。DistilBERT 模型仅用 66M 参数对 BERT 实现了进行蒸馏压缩，TAPAS-small 是 Google 2019 年提出的弱监督表格问答预训练模型 TAPAS 的小参数版本。

## 模型方法：多正样本对比学习损失函数

InfoNCE 是经典的对比学习损失函数，它使用余弦相似度 $sim(p,q)=\frac{p\cdot q}{\|p\|\|q\|}$ 来衡量样本相似性。在每个 Batch 中，每个样本 $q$ 的损失 $\mathcal{L}_q$ 如公式1所示，其中 $k_i$ 是 $k$ 的 In-Batch 负样本，$k_+$ 是 $k$ 的正样本。

$$\mathcal{L}_q=-\log\frac{\exp\left(q\cdot k_+/\tau\right)}{\sum_{i=0}^{K}\exp\left(q\cdot k_i/\tau\right)} \tag{1}$$

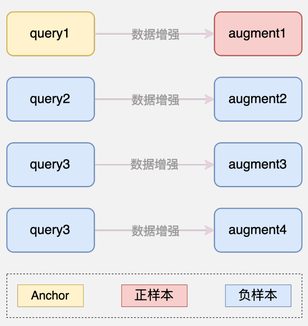



图 1：经典对比学习损失

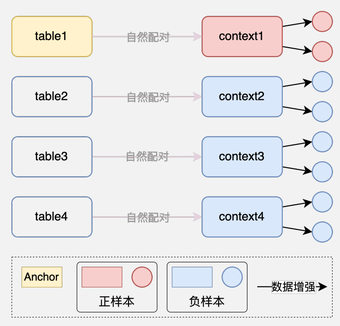



图 2：多正样本对比学习损失

相比传统的对比学习（图1），本研究（图2）中每个长度为 $N$ 的 Batch 由两种模态的数据构成：$N$ 个表格 table（记为 $t_i$）和 $N$ 个上下文文本 context（记为 $c_j$），$t_i$ 和 $c_i$ 为正样本对，$t_i$ 和 $c_j$ 为负样本对（$i\neq j$）。总损失函数为 $\mathcal{L}_{no-aug}$。

$$\mathcal{L}_{no-aug}=\sum_{i=0}^{N-1}-\log\frac{\exp\left(t_i\cdot c_i/\tau\right)}{\sum_{j=0}^{N-1}\exp\left(t_j\cdot c_j/\tau\right)} \tag{2}$$

在对标题增强时，可以将问题理解为上下文的多分类即可避免损失函数的计算问题，即判断上下文究竟属于哪个表格。使用交叉熵计算损失 $\mathcal{L}_{aug}$，其中 C 是 Batch 中所有的上下文向量构成的 $N_{context}\times Dim$ 矩阵，T 是 Batch 中所有的表格向量构成的 $N_{table}\times Dim$ 矩阵（$N_{table}\leq N_{context}$）。

$$\mathcal{L}_{aug}=f\left(\mathrm{C}\times\mathrm{T}^{\top},label\right) \tag{3}$$

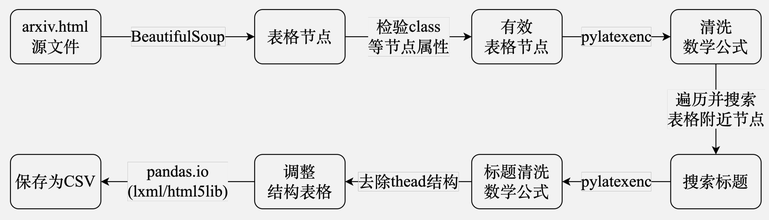
## 模型方法：数据增强

对比学习最重要的技术是使用数据增强技术制造正样本，本研究采用 python 第三方包 nlpaug 对表格标题进行词粒度的增强，经过筛选使用如下三种方法。

- Synonym（近义词）方法使用 WordNET 或 PPDB 等词典库来寻找近义词，通过替换来实现数据增强，速度较快；
- WordEmbs（词向量）方法采用 GloVe 等成熟的词向量模型来寻找最适合替换或插入的词语；
- BackTranslation（回译）方法指将原语言翻译为一种新的语种，再翻译回原语种获得增强后的问题，需要使用两个翻译模型，所以速度较慢但效果较好。

## 实验与数据

本实验共使用三个数据集，两个表格-标题预训练数据集 ar5iv 和 ToTTo，一个下游任务 SQA 问答数据集。其中 ar5iv 表格数据集由本文提出，通过 LaTeX 源码解析出 arXiv 文献表格，数据生成流程如下图所示。




实验主要由数据预处理、对比学习预训练、下游任务微调共三个部分构成。预处理部分主要包括生成、清洗表格标题，并进行数据增强。ToTTo 数据集采用词向量、回译、近义词替换三种增强方式；因专有名词和公式过多，ar5iv 数据集暂时只采用近义词替换增强方式。预训练部分共包括 7 个模型，5 个基于 ToTTo 数据集的模型，和 2 个基于 ar5iv 数据集的模型。下游任务微调用于测量上述每个预训练模型以及 TAPAS-small 原模型在 SQA 上的性能。

## 研究结果

对本实验预训练的 7 个模型以及 TAPAS 发布的 TAPAS-small 模型在 SQA 任务上分别进行 100 个 epoch 的微调并选出其中最高的验证集准确率，计算出对应的测试集准确率进行模型性能的比较，微调结果如表所示。

| 预训练模型 | 验证集准确率 | 测试集准确率 |
|---|---|---|
| TAPAS-small | 0.6115 | 0.6295 |
| ToTTo-Trans | 0.6071 | 0.6275 |
| ToTTo-Synonym | 0.5991 | 0.6265 |
| ToTTo-NoAug | 0.5885 | 0.6212 |
| ToTTo-AllAug | 0.5969 | 0.6182 |
| ToTTo-Embed | 0.5960 | 0.6162 |
| ar5iv-Synonym | 0.5983 | 0.6149 |
| ar5iv-NoAug | 0.6088 | 0.6225 |

基于 ToTTo 的五个模型中，使用回译数据增强的模型性能最好，使用词向量的模型性能最差且弱于不使用数据增强的模型；ar5iv 数据集上使用近义词数据增强也产生了副作用；由于训练不充分，所有对比学习均没有超过不使用对比学习的模型。

## 结论与讨论

本研究主要得出以下结论：

- 成功搭建了表格多模态对比学习预训练模型框架，并设计了多正样本对比损失函数；
- 由于数据量、批大小不足，导致模型在相对空间尚未定型，因此对比学习没有给模型带来提升；
- 回译增强方法可能效果最显著，不同数据集适合不同的增强方法不应该简单叠加，且数据量大增强效果更好。

在未来的研究中，首先可以探索新的下游任务并充分利用表格的整体性；其次应当扩充预训练数据集并实现更充分的调参；此外可以探索协同表示学习对另一个模态模型的影响，如是否能提升通用语言模型的效果。