基于自监督和视觉落地预训练的电影票房预测

秦超[1,3],Eunsoo Kim[2],和李博扬[3]/阿里巴巴集团和阿里巴巴—南洋理工大学联合研究院，新加坡[2]南洋商学院，南洋理工大学（*NTU*），新加坡[3]计算机科学与工程学院，*NTU*，新加坡chao0009@e.ntu.edu.sg，{eunsoo,boyang.li}@ntu.edu.sg

arXiv:2304.10311v1 [cs.MM] 20 Apr 2023

摘要—**电影制作投资具有较高风险，因为电影收入长期以来呈现长尾分布和双峰分布 [1]。准确预测票房收入有助于降低不确定性并鼓励投资。然而，如何为演员、导演以及与用户生成内容相关的关键词学习有效表征，仍是一个具有挑战性的开放问题。在这项工作中，我们研究了自监督预训练的影响，并提出将内容关键词与电影海报中的物体进行视觉接地，作为一种预训练目标。在包含 35,794 部电影的大型数据集上进行的实验表明，自监督训练和视觉接地能带来显著收益。具体而言，视觉定位预训练大幅提升了对包含内容关键词的电影的学习效果；与采用相同架构并经过微调的 BERT 模型相比，性能相对提升了 14.5%。**

索引词——**多模态学习、自监督学习、视觉接地、票房预测、电影收入预测**

一、引言

电影无疑是21st世纪人类文明中一种卓越的艺术形式。然而，电影制作的商业层面往往并不光鲜。统计数据 [1] 表明，票房收入长期以来呈现长尾和双峰分布：少数电影获得了大部分利润，而大多数电影勉强收支平衡。根据 Box Office Mojo[1],的数据，2019年，票房最高的前十部电影共获得132亿美元，占全球票房最高的前200部电影总收入的37.4%。疫情持续三年后，截至2022年11月，这一比例飙升至50.1%。行业的高风险促使制片人专注于超级英雄电影和续集，因为它们的结果相对可预测。无力承担高预算电影制作的小型制片公司，能分到的市场份额越来越少。

算法票房预测有望帮助制片人合理编制预算、降低风险，并鼓励对创意多样化内容的投资。该问题已引起广泛的研究兴趣 [2]–[15]。在本文中，我们研究了自监督预训练和视觉接地的影响。

演员和导演的明星影响力是决定票房收入的最重要因素之一，但每位演员和导演的数据有限。即使是多产的导演，在其职业生涯中通常也只拍摄不到30部电影。

[1]https://www.boxofficemojo.com/year/world/2019/

表 I：来自 TMDB 的用户生成关键词示例。

动作，犯罪学，联邦调查局，心理，飞机，机器人爱情，仇恨，高中，父女关系，法国巴黎，王国，改编自小说或书籍

职业生涯。同样，很少有现代演员在超过30部电影中担任主角。按照现代机器学习标准，这些数字被视为少样本设置。为了解决数据稀疏性问题，我们采用自监督预训练，促使网络在票房数据上进行训练之前学习数据分布。

票房的另一个重要但难以建模的方面是电影内容。电影情节是一种复杂的产物，具有多层次的语义 [16]–[19]，，即使是最先进的人工智能也难以理解。为了解决这一问题，我们利用来自 The MovieDatabase (TMDB)[2] 的用户生成内容关键词，将电影内容纳入票房预测问题中。表 I 展示了一些关键词示例。与传统的类型类别相比，这些关键词能够对内容进行更细粒度的分类，涵盖主题、情节、情感，甚至来源相关信息[3]。

为了准确理解这些关键词，我们进一步提出将关键词与视觉模态——电影海报相联系。在电影语境中，关键词的含义可能与日常用法略有不同。例如，动作 这一关键词可能与爆炸、汽车追逐或武术相关，而非其词典释义。机器人 通常指科幻或动画电影中的机器人，而不是装配线上的机器人。近期研究 [20], [21] 表明，将语言与视觉信号相联系能够改善表征。本文发现，这种效果同样存在，而且改进后的表征有助于提升票房预测效果。据我们所知，这是首篇将文本信息与视觉内容相联系以进行票房预测的论文。

总体而言，我们的研究凸显了自监督预训练和视觉接地在票房

[2]www.themoviedb.org[3]更多详情请参阅位于 https://www.themoviedb.org/bible/movie 的关键词贡献指南

预测中的有效性。与超参数数量相同的直接微调基线 BERT 模型相比，我们的模型将预测误差相对降低了 7.8%∼14.5%；而采用视觉接地进行预训练，性能最高可相对提升 2.1%。

本文的贡献如下。首先，我们提出了一种用于电影票房预测的自监督预训练方法，可利用文本信息与数值信息的组合。其次，我们证明，将用户生成的关键词视觉接地到电影海报上，能够显著改善预训练效果，这表明电影内容与海报之间具有良好的相关性。最后，我们构建了一个规模庞大、组织完善的电影票房预测数据集，并与研究界共享[4]。

II. 相关工作

预测电影成功。 先前的研究尝试预测多种商业和艺术成就指标，包括票房 [2]–[4], 投资回报率[5]–[7], IMDb 评分 [8], [9], 影评 [11], 以及奖项或奖项提名 [10]。近年来，随着机器学习的发展，深度网络逐渐受到研究界的关注 [12]–[15]。

在特征方面，除了常用的数值特征外， [2], [5]–[7], [11] 还利用了情感和主题等文本特征。具体而言，潜在狄利克雷分配 [5], [11] 得到的主题可视为一种内容特征。 [9], [15] 分别使用了 fastText [22]和 ELMO [23] 词嵌入。据我们所知，先前唯一使用视觉特征进行票房预测的研究是 [14],, 该研究在训练过程中融入了卷积神经网络提取的电影海报特征。相比之下，我们的工作利用海报中的物体，在预训练期间将内容关键词进行视觉接地，但在微调阶段不使用海报。

自监督多模态预训练。BERT 等预训练文本模型的成功 [24] 催生了一系列预训练多模态模型 [25]–[28],, 这些模型通常采用掩码语言建模（MLM）目标。与去噪自编码器类似，掩码语言建模目标会训练模型预测输入中被掩码的部分。这种看似简单的训练技术已被证明在广泛的下游应用中卓有成效。另一类研究（如 CLIP [29] 和 BLIP [30], ）采用一种预训练目标，用于区分正确与错误的图文配对。

认知科学中的一个经典问题——符号接地问题 [31], 关注的是词语如何作为指向其他概念和物体的指针而获得意义。将文本词元接地到视觉图像中，已在一些应用中取得成功 [20], [21], [32]–[36]。在这项工作中，我们使用电影海报作为文本词元——关键词的视觉接地来源。电影海报是一种广泛使用的视觉媒介，用于在电影上映前很久就为电影宣传。因此，我们利用单张海报中的物体来接地这些词元；每个词元都可以与多个物体相关联，反之亦然。与前述通过检索或生成与文本描述相关的图像来完成接地的研究相比，我们的任务中，关键词与海报之间的对应关系并非事先已知，必须以多实例方式加以发现。

[4]https://github.com/jdsannchao/电影票房预测

III. 方法

本节首先介绍所提出网络使用的特征，随后介绍预训练策略。

A. 特征

我们同时使用离散特征（如演员或导演）和实值特征（如电影预算）。离散词元的嵌入通过数据学习得到。对于实值特征，我们采用基于原型的数值嵌入 [37]。形式上，嵌入函数表示为 NE(x) :$\mathbb{R} \to \mathbb{R}^D$ ，它将一个实数 x 映射为一个 D 维向量，其分量

$$\mathrm{NE}_i(x) = \exp\left(-\frac{\|x - q_i\|_2}{\sigma^2}\right), \tag{1}$$

其中， $\{q_i\}_{i=0}^{D-1}$ 是指定区间内均匀间隔的数值，例如 $[-10, 10]$。在应用数词嵌入函数之前，我们根据特征是否呈长尾分布，使用对数归一化或最小-最大归一化对数值进行归一化。

我们将用于预测票房收入的特征大致分为四类：投资与营销、明星影响力、内容以及竞争与季节性。

投资与营销。制作预算通常是电影质量的一个指标。此处我们取以 10 为底的对数。此外，我们将发行公司作为一个词元，因为市场影响力更大的发行商可能会在更多银幕上发行电影，从而增加收入。

明星影响力。 我们的模型最多纳入两位导演、两位编剧和三位主演。每个人物都是一个独立的词元，其嵌入从头开始训练。我们还计算每个人物的盈利能力，定义为该人物此前作为主演参与的所有电影的平均收入。此外，我们还纳入每位演员在电影上映时的性别和年龄。

电影内容。 我们首先纳入电影类型和美国电影协会分级。此外，我们还纳入一个指标，用于表示一部电影是否属于某个系列作品。

受用户生成关键词作为内容描述词的成功启发， [38], 我们从 TMDB 收集用户生成关键词，共为 35,794 部电影收集到 7,700 个独特关键词。在这些关键词中，我们观察到许多罕见关键词和近义词，这可能会妨碍模型学习。对于罕见关键词，数据不足可能导致嵌入估计不准确。对于对比学习而言，同义词和近义词会带来问题，因为它会迫使

一个包含文本和数值特征的输入示例：

[CLS][PG-13]1.5678[类型][动作][科幻][关键词][盾牌][超级英雄][导演][Joss Whedon][演员][Chris Evans][SEP]

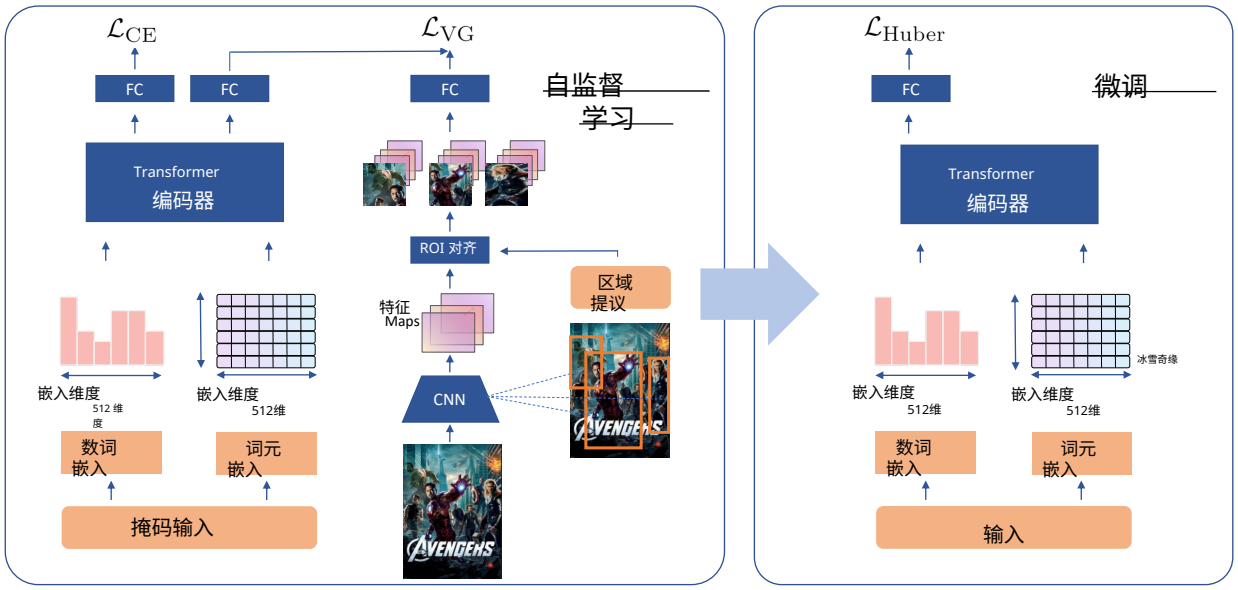

图1：自监督预训练和微调在票房预测任务上的整体流程。微调期间，词元嵌入保持冻结。

模型为两个含义相近的词学习出不相似嵌入。

为了解决这些问题，我们结合词汇相似度和共现统计对关键词进行聚类。为捕捉词汇信息，我们使用 fastText [22]计算的 300 维嵌入。接下来，我们构建电影－关键词词频－逆文档频率（TF-IDF）矩阵，以捕捉关键词的共现统计信息。基于 TF-IDF 矩阵，我们采用 [39] 技术为关键词构建嵌入。我们提取奇异向量的前 50 个维度来表示关键词。最终表征是将这两个向量拼接而成的 350 维向量。随后，我们进行平均链接凝聚聚类，并将得到的关键词簇作为电影的特征。附录 B 展示了详细的聚类结果。

竞争与季节性。为捕捉消费者偏好变化和节假日季节的影响，我们将电影上映的年份和月份作为离散词元。此外，我们还对上映窗口期的竞争强度进行建模。我们首先将当前电影上映前后两周内上映且类型相同的电影确定为竞争影片。随后，我们计算当前电影与每部竞争影片之间内容关键词重叠度的总和，重叠度通过 Jaccard 相似度计算。

B. 自监督预训练

图 1 展示了整体流程。在第一阶段，我们使用 MLM 和视觉接地目标对 Transformer 网络进行预训练。接下来，我们冻结词元嵌入，并在票房预测任务上对网络进行微调。现在，我们介绍预训练任务。

掩码字段预测。 我们采用了一种类似于掩码语言建模任务的预训练目标，该任务已被证明是一种有效的自然语言理解 [24] 和多模态理解[25]预训练方法。我们从每组输入特征中随机掩盖一个词元，包括类型、关键词、导演/编剧姓名和演员姓名。网络经过训练，以预测缺失的词元。预测采用交叉熵损失进行表述，我们将其记为 $\mathcal{L}_{CE}$。通过训练网络预测缺失字段，我们促使网络学习各项输入之间的相关性，从而有助于缓解数据稀缺问题。

结构化视觉接地。 电影内容无疑对票房至关重要，但理解用户生成的内容关键词颇具挑战。特别是，如前所述，随着关键词的含义与日常用法存在细微差异，内容关键词在电影语境中的含义也可能发生变化。

我们提出利用电影海报提供的视觉模态来对关键词进行接地。我们采用对比学习，鼓励海报与对应内容关键词之间具有较高的相似度，同时抑制错误配对的海报与关键词集合之间的相似度。首先，我们使用现成的网络 VinVL [40],对海报进行目标检测，但我们的方法并不局限于这一特定选择。我们将从 i^{th} 电影中提取的对象特征表示为 $\mathcal{Z}_i = \{\boldsymbol{z}_m\}_{m=1}^{M}$。请注意，我们使用下标 i 表示电影索引。我们还采用 Transformer 网络输出的关键词上下文化嵌入，记为 $\mathcal{X}_i = \{\boldsymbol{x}_k\}_{k=1}^{K}$。

我们定义海报与关键-



Keywords

based on true story | remake | class differences | caregiver | man in wheelchair | interracial friendship | social differences | rich man | quadriplegia | male nurse | handicapped | disabled person

图 2：电影*TheUpside*（2019年）的多个物体与关键词对齐

词语作为

$$\mathrm{sim}(\mathcal{X}_i, \mathcal{Z}_i) = \sum_{(\boldsymbol{x},\boldsymbol{z})\in\mathcal{X}_i\times\mathcal{Z}_i} \exp\left(\frac{\boldsymbol{x}^\top \boldsymbol{z}}{\|\boldsymbol{x}\|_2\|\boldsymbol{z}\|_2}\right), \tag{2}$$

其中，$\times$ 表示笛卡尔积，$\|\cdot\|_2$ 表示 L2 范数。为说明该定义，我们在图 2 中展示了一张海报及其对应的关键词。我们用关键词框的颜色表示簇归属（例如，“四肢瘫痪”和“残障人士”都属于红色簇）。我们观察到，一个簇可能对应多个物体，而一个物体也可能定位到多个簇。例如，“四肢瘫痪”簇由轮椅、轮胎和坐着的男人定位；坐着的男人则与红色簇和紫色簇相关。由于这种多对多关系，我们遵循 [41]，将两个集合之间的相似度定义为所有可能配对的相似度之和。

通过随机采样负样本对（i'，j'），我们将视觉定位损失 $\mathcal{L}_{\mathrm{VG}}$ 定义为

$$\mathcal{L}_{\mathrm{VG}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\mathrm{sim}(\mathcal{X}_i, \mathcal{Z}_i)}{\mathrm{sim}(\mathcal{X}_i, \mathcal{Z}_i) + \sum_{(i',j')}\mathrm{sim}(\mathcal{X}_{i'}, \mathcal{Z}_{j'})}\right) \tag{3}$$

其中 N 是训练集中的电影总数。

C. 票房预测微调

在微调阶段，我们训练网络预测票房收入。我们将所有输入位置的输出取平均值，并将其输入全连接层，从而生成预测值。票房收入呈长尾分布，我们使用对数正态分布对其进行近似。因此，我们取收入的以10为底的对数作为目标值。为了进一步减小异常值的影响，我们使用平滑L1损失（也称为Huber损失）训练网络，

$$\mathcal{L}_{\mathrm{Huber}} = \begin{cases} 0.5\,(y-\hat{y})^2, & \text{if } |y-\hat{y}| < 1 \\ |y-\hat{y}| - 0.5, & \text{otherwise} \end{cases}, \tag{4}$$

其中，y 是真实值，$\hat{y}$ 是预测值。

表 II：保留票房测试集上的性能比较。我们最优的模型与 $\mathrm{BERT}_{\mathrm{small}}$ 相比，准确率提高了 14.5%。

模型	测试集 Huber 损失(% improvement)	
仅使用数值特征		
随机森林	$0.3677_{(-3.5\%)}$	
文本和数值特征		
$\mathrm{BERT}_{\mathrm{small}}$ 微调	$0.3553_{(\mathrm{baseline})}$	
$\mathrm{BERT}_{\mathrm{medium}}$ 微调	$0.3446_{(2.5\%)}$	
我们的模型	聚类	关键词
随机初始化	$0.3290_{(7.4\%)}$	$0.3265_{(8.1\%)}$
+掩码语言模型预训练	$0.3109_{(12.5\%)}$	$0.3133_{(11.8\%)}$
+VG预训练	$0.3070_{(13.6\%)}$	$0.3109_{(12.5\%)}$
BERT嵌入初始化	$0.3137_{(11.7\%)}$	$0.3249_{(8.6\%)}$
+掩码语言模型预训练	$0.3102_{(12.7\%)}$	$0.3226_{(9.2\%)}$
+VG预训练	$0.3037_{(14.5\%)}$	$0.3182_{(10.4\%)}$

IV. 实验结果

A. 数据与实验设置

我们从 TMDB 收集了 35,794 部电影的元数据，涵盖 1920 年至 2020 年。每部电影在其首次上映期间的总票房数据均从 IMDbPro 抓取。我们使用分层抽样，按照 70/10/20 的比例将数据划分为训练集、验证集和测试集，并以“系列电影”作为分层标签。根据 §III-A 中的方法，我们将 7,700 个关键词聚为 1,414 个簇。簇的数量在验证集上进行调优。我们使用一个 4 层 Transformer，模型维度为 $d_{\mathrm{model}} = 512$，全连接层维度为 $d_{ff} = 512$，并采用 4 个注意力头。其架构与 $\mathrm{BERT}_{\mathrm{small}}$ 相同。更多超参数见附录 A。

B. 基线方法

我们介绍三种基线模型。第一种是随机森林（RF）。由于对离散特征进行独热编码会产生过多维度，因此我们只向 RF 输入数值特征。接下来，我们介绍小型和中型的预训练 BERT 模型，并在票房预测任务上对其进行微调。为了模拟经典的 BERT 输入，我们将所有输入词元拼接成一个句子，同时将数值特征四舍五入至小数点后一位，然后应用 BERT 分词器。最后，我们将票房预测任务上直接训练的随机初始化 $\mathrm{BERT}_{\mathrm{small}}$，以及使用演员、剧组成员、类型和关键词的预训练 BERT 嵌入的 $\mathrm{BERT}_{\mathrm{small}}$ 作为对比模型。当名称包含多个词时，我们使用预训练 BERT 嵌入的平均值。对于关键词簇，我们使用其中出现频率最高的关键词的嵌入。

C. 结果与讨论

在表 II 中，我们报告了所有模型在测试集上的 Huber损失，以及它们相对于 $\mathrm{BERT}_{\mathrm{small}}$ 的表现。预训练的 BERT 模型轻松优于 RF 基线，

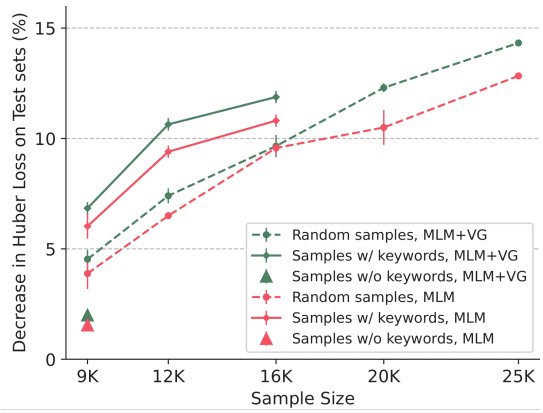

图 3：不同训练集规模下票房预测的测试损失。竖线表示标准差。具体数值见附录 E。

但其表现仍不如掩码语言模型和视觉 grounding 预训练。尽管更大的 $\text{BERT}_{\text{medium}}$ 优于 $\text{BERT}_{\text{small}}$，但其表现仍比我们采用掩码语言模型预训练的网络低 10% 以上（相对值）。预训练所用的电影数据与文本数据之间的领域差异，以及我们的特征工程，可能是造成性能差距的原因。

值得注意的是，对于两种嵌入初始化方式，VG预训练都在 MLM 的基础上取得了显著提升。即使使用经 BERT 预训练的词元嵌入，VG预训练仍能带来改进，这一事实印证了我们的假设：关键词在电影语境中可能具有专门语义，而视觉接地有助于捕捉这些专门语义。最后，最佳测试损失为 0.3037，相较于 $\text{BERT}_{\text{small}}$ 提升了 14.5%，这一结果由 MLM+VG 预训练取得。

内容关键词与扩展性。由于并非所有电影都带有用户提供的关键词，我们进一步研究了在有内容关键词和无内容关键词的电影上进行预训练的影响。我们将训练集划分为带有关键词的电影（25K 中有 16K）和不带关键词的电影（25K 中有 9K）。作为对比基线方法，我们还从整个训练集中随机抽取规模分别为 9K、12K、16K、20K 和 25K 的子集。当 MLM+VG 网络在不同训练集上进行训练时，我们在相同测试集上的损失如图 3 所示。我们注意到，在训练数据量相同的情况下，与随机混合的数据相比，在带有关键词的数据上训练时，MLM 和 VG 都表现出更强的泛化能力。这与我们的直觉一致，因为 MLM 利用关键词之间的相关性，而 VG 进一步通过视觉信息强化关键词。在附录中的图 6 里，我们考察了对于带有关键词的电影，VG 是否在 MLM 的基础上有所改进。我们观察到，随着训练数据增加，MLM+VG 相较于 MLM 的提升幅度扩大，这表明 VG 具有良好的扩展性，并且其有效性会随数据量增加而增强。

关键词聚类的影响。我们考察了关键词簇的影响。表 II 对比了采用关键词聚类（“聚类”）与使用原始关键词（“关键词



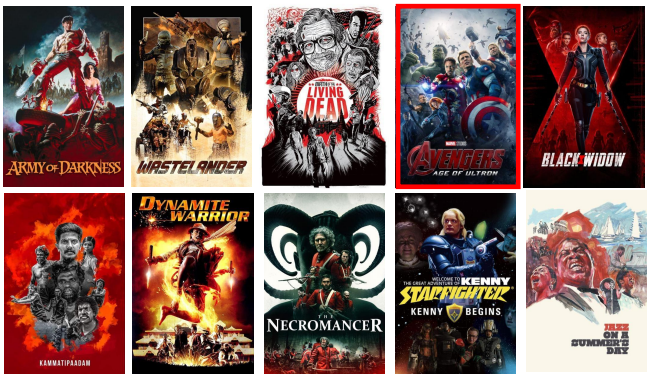

图 4：**上方**：在浪漫电影《一天》（2009）的语境下，使用关键词“爱”检索到的海报；《一天》（*2009*）；**下方**：在电影《复仇者联盟》（2012）的语境下，使用关键词“超级英雄”检索到的海报。《复仇者联盟》（*2012*）。

”）的结果。在大多数情况下，关键词簇能够提升性能，尤其是在使用预训练的 BERT 嵌入时。一个可能的原因是，近义词的 BERT 嵌入相似，模型难以区分，而聚类可以缓解这一问题。

D. 海报检索示例

我们对视觉接地的效果进行了定性考察。图 4 展示了在电影语境下与关键词最为相似的海报。前两行是以浪漫爱情电影《一天》（*2009*）为语境，针对关键词“爱”检索得到的结果。大多数海报都属于爱情片类型，画面中呈现出一对情侣相拥的场景。最下面十张海报是以电影《复仇者联盟》（*2012*）为语境，针对关键词“超级英雄”检索得到的结果。这些结果大多属于动作片，海报中央通常是一位英雄，周围环绕着其他人物。更多示例见附录 E。

V. 结论

票房收入受到大量相互交织因素的影响，这些因素往往难以观测，更不用说通过计算建模了。因此，票房预测的一项重要挑战是学习能够捕捉电影语义并与目标变量良好相关的表征。为此，

我们提出了一种利用掩码语言建模和视觉接地目标对 Transformer 网络进行预训练的方法，该方法显著提升了模型性能。我们希望这些研究结果能够为后续的多模态票房预测研究提供启发。

VI. 致谢

本研究部分由阿里巴巴集团通过阿里巴巴创新研究（AIR）计划和阿里巴巴-南洋理工大学新加坡联合研究院（编号：AN-GC-2020-011）、新加坡国家研究基金会奖学金（编号：NRF-NRFF13-2021-0006）以及南洋理工大学启动基金资助。

参考文献

[1] Pan, R. & Sinha, S. 流行度的统计规律：电影票房动态的普遍特性。《新物理学杂志》。（2010年）[2] Apala, K., Jose, M., Motnam, S., Chan, C., Liszka, K. & Gregorio, F. 利用社交媒体预测电影票房表现。IEEE/ACM 社交网络分析与挖掘进展国际会议（ASONAM 2013）。第 1209-1214 页（2013年）[3] Parimi, R. & Caragea, D. 电影上映前的票房成功预测。模式识别中的机器学习与数据挖掘国际研讨会。第 571-585 页（2013年）[4] Simonoff, J. & Sparrow, I. 预测电影票房收入：赢家与输家、大片与黑马影片。《Chance》。13，15-24（2000年）[5] Lash, M. & Zhao, K. 电影成功的早期预测：盈利能力取决于谁、什么以及何时。《管理信息系统杂志》。33，874-903（2016年）[6] Eliashberg, J., Hui, S. & Zhang, Z. 利用电影剧本评估票房表现：基于核的方法。《IEEE 知识与数据工程汇刊》。26，2639-2648（2014年）[7] Rajput, P., Sapkal, P. & Sinha, S. 利用双重情感分析预测票房收入。《国际机器学习与计算杂志》。7，72-75（2017年）[8] Cizmeci, B. & ¨Og¨ud¨uc¨u, S.. 利用社交媒体和因子分解机预测尚未上映电影的 IMDb 评分。2018 年第三届计算机科学与工程国际会议（UBMK）。第 173-178 页（2018年）[9] Shafaei, M., Lopez-Monroy, A. & Solorio, T. 利用文本、视觉和产品特征预测电影受欢迎程度。FLAIRS。（2019年）[10] Boccardelli, P., Brunetta, F., Vicentini, F. & Others. 什么是电影产业成功的关键？意大利电影产业关键成功因素研究。DIME。（2008年）[11] Kim, E., Ding, M., Wang, X. & Lu, S. 主题一致性重要吗？电影产业中的影评人与用户评论研究。《营销学杂志》。（2023年）[12] Quader, N., Gani, M. & Chaki, D. 七种机器学习分类技术在电影票房成功预测中的性能评估。2017 年第三届电气、信息与通信技术国际会议（EICT）。第 1-6 页（2017年）[13] Antipov, E. & Pokryshevskaya, E. 所有电影的票房收入都同样难以预测吗？基于随机森林的模型所提供的证据。《收益与定价管理杂志》。16，295-307（2017年）[14] Zhou, Y. & Yen, G. 演化深度神经网络以预测电影票房收入。2018 年 IEEE 计算进化大会。（2018年）[15] Kim, Y., Cheong, Y. & Lee, J. 利用深度学习模型根据剧情概要预测电影的成功。《第二届故事讲述研讨会论文集》。第 127-135 页（2019年）[16] Barot, C., Potts, C. & Young, R. 用于叙事话语生成的三方计划叙事模型。AIIDE。（2015年）[17] Cutting, J. 叙事理论与热门电影的动态。《心理学通报与评论》。23，1713-1743（2016年）[18] Li, B. 学习知识以支持领域无关的叙事智能。佐治亚理工学院博士论文。2015年。[19] Davis, N.、Li, B.、O'Neill, B.、Riedl, M. 和 Nitsche, M. 数字电影制作中的分布式创造性认知。第八届 ACM 创造力与认知会议论文集。（2011年）[20] Tan, H. 和 Bansal, M. Vokenization：通过情境化视觉基础监督改进语言理解。EMNLP。（2020年）。[21] Lu, Y.、Zhu, W.、Wang, X.、Eckstein, M. 和 Wang, W. 想象增强的自然语言理解。NAACL。（2022年）[22] Bojanowski, P.、Grave, E.、Joulin, A. 和 Mikolov, T. 利用子词信息丰富词向量。arXiv 预印本 arXiv:1607.04606。（2016年）[23] Sarzynska-Wawer, J.、Wawer, A.、Pawlak, A.、Szymanowska, J.、Ste- faniak, I.、Jarkiewicz, M. 和 Okruszek, L. 通过深度情境化词表示检测形式思维障碍。arXiv 预印本 arXiv: 1802.05365。（2018年）[24] Devlin, J.、Chang, M.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向 Transformer 预训练。arXiv 预印本 arXiv:1810.04805。（ 2018 年）[25] Lu, J.、Batra, D.、Parikh, D. 和 Lee, S. ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表示。NeurIPS。32（2019年）[26] Su, W.、Zhu, X.、Cao, Y.、Li, B.、Lu, L.、Wei, F. 和 Dai, J. VL-BERT：通用视觉语言表示的预训练。arXiv:1908.08530。（2019年）[27] Tan, H. 和 Bansal, M. LXMERT：从 Transformer 学习跨模态编码器表示。EMNLP-IJCNLP。（2019年）[28] Huang, Z.、Zeng, Z.、Liu, B.、Fu, D. 和 Fu, J. Pixel-BERT：通过深度多模态 Transformer 对齐图像像素与文本。arXiv 预印本 arXiv:2004.00849。（2020年）[29] Radford, A.、Kim, J.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J. 和其他人 从自然语言监督中学习可迁移的视觉模型。ICML。第8748–8763页（2021年）[30] Li, J.、Li, D.、Xiong, C. 和 Hoi, S. BLIP：通过自举式语言-图像预训练实现统一的视觉语言理解与生成。arXiv 预印本 2201.12086。（2022年）[31] Harnad, S. 符号接地问题。《Physica D：非线性现象》。42，335–346（1990年）。[32] Yang, Y.、Yao, W.、Zhang, H.、Wang, X.、Yu, D. 和 Chen, J. Z-LaVI：由视觉想象驱动的零样本语言求解器。arXiv 预印本 2210.12261。（2022年）[33] Zhang, L.、Chen, Q.、Siebert, J. 和 Tang, B. 通过句子可视化将半监督视觉特征整合到语言模型中。ICMI。（2021年）。[34] Kiros, J.、Chan, W. 和 Hinton, G. 图示性语言理解：通过图像搜索实现大规模视觉接地。ACL。（2018年）。[35] Liu, X.、Yin, D.、Feng, Y. 和 Zhao, D. 文本中未写明的内容：从视觉信号探索空间常识。ACL。（2022年）[36] Long, Q.、Wang, M. 和 Li, L. 生成式想象提升机器翻译。arXiv 预印本 arXiv:2009.09654。（2020年）[37] Jin, Z.、Jiang, X.、Wang, X.、Liu, Q.、Wang, Y.、Ren, X. 和 Qu, H. NumGPT：提升生成式预训练模型的数理能力。arXiv 预印本 2109.03137。（2021年）[38] Annalyn, N.、Bos, M.、Sigal, L. 和 Li, B. 利用用户生成内容标签，根据图书偏好预测人格。《IEEE 情感计算汇刊》。11，482–492（2018年）[39] Zhang, Y.、Li, B.、Liu, Y.、Wang, H. 和 Miao, C. 初始化至关重要：为神经推荐系统正则化流形信息初始化。ACM SIGKDD。（2021年）[40] Zhang, P.、Li, X.、Hu, X.、Yang, J.、Zhang, L.、Wang, L.、Choi, Y. 和 Gao, J. VinVL：重新审视视觉语言模型中的视觉表示。CVPR。（2021年）[41] Miech, A.、Alayrac, J.、Smaira, L.、Laptev, I.、Sivic, J. 和 Zisserman, A. 从未经筛选的教学视频中端到端学习视觉表示。CVPR。第9879–9889页（2020年）[42] He, K.、Gkioxari, G.、Doll´ar, P. 和 Girshick, R. Mask R-CNN。ICCV。（2017年）

附录

A. 实验设置与超参数

在视觉接地预训练过程中，我们随机为每部电影选择最多 6 个关键词，并为每张海报选择最多 20 个物体来计算相似度。每个物体的特征图维度为 $2048\times4\times4$，这是 VinVL 经过 ROI 对齐 [42] 和自适应平均池化层后的输出。随后，将特征图在空间维度上展平，并线性投影到 $\mathrm{R}^{d_{model}}$，其中 $d_{model}=512$。

在采用掩码语言模型（MLM）目标进行模型预训练时，我们使用的批量大小为 2048；在加入视觉定位目标（VG objective）时，将批量大小减小至 326。学习率为 3e-4。我们使用的优化器为 Adam，权重衰减系数设为 1e-4。在微调阶段，我们通过组合以下学习率 [1e- 3、3e-4、1e-4] 和批量大小 [328, 512, 1024]，在验证数据集上搜索最佳性能。

B. 聚类样本与唯一词元数量

表 III：聚类结果示例。簇的代表词是该集合中出现频率最高的关键词。

簇	元素
love	’爱’, ’爱过’, ’仇恨’, ’不开心’, ’等待’, ’开心’, ’感激’, ’幸运’, ’期待’, ’爱着’
超级英雄	’超级英雄’, ’女反派’, ’反派’, ’共生体’, ’助手’, ’超人类’, ’联手’, ’宿敌’, ’超级英雄’, ’超级反派’
心理	’心理’, ’精神病态的’, ’纵火狂’, ’精神变态的’,’有杀人倾向的’, ’精神错乱的’

表 IV：每种文本特征的唯一词元数量

特征名称	示例	No. 唯一词元
发行年份	发行年份范围从1920—2020年	100
发行月份	12个词元表示12个月份	12
MPAA	G, PG, PG-13级, R, NC17, 未评级, N.A.	7
制作公司	将小型制片公司分组（制作电影少于 10 部）归入“其他”	594
发行公司	小型制片公司分组（制作的电影少于 10 部）到“其他”	407
系列作品	是或否	2
模仿作品	是或否	2
类型	例如：剧情片、爱情片	18
关键词	例如：“友谊”	1414
剧组人员姓名	例如：“史蒂文·斯皮尔伯格”是一个词元	15333
演员姓名	例如“莱昂纳多·迪卡普里奥”是一个词元	20366

C. 来自 *TMDB* 的元数据

{’成人’: False, ’背景_路径’: ’c1BaOxC8bo5ACFYkYYxL0bBWRaq.jpg’, ’所属

’喜剧’}], ’homepage’: ’https://www.miramax.com/movie/four-rooms/’, ’

{'成人': False, '性别': 1, 'id': 3110, '已知_因_部门': '导演', '姓名':

D. 模型输入

特征名称	**词元 / 值**
发行年份	
（（是随机森林模型和Transformer模型的输入））	2012
发行月份（两者）	三月
MPAA（两者）	PG-13级
预算（两者）	7.892094608
制片人（两者）	Lionsgate
发行商（两者）	Lionsgate
N_个竞争者（两者）	0.693147181
竞争者相似度（两者）	0
系列作品（两者）	Yes
系列名称	饥饿游戏 0
人物数量（两者）	0.693147181
男性数量（两者）	0
女性数量（两者）	0
类型（两者）	[类型]
	冒险
	奇幻
	科幻
聚类	[聚类]
	改编重述
	社会主义
	西洋双陆棋
	诠释
导演	[导演]
导演1经验（两者）	Gary Ross
导演1盈利能力（两者）	1.098612289
导演2也相同	0.875335786
编剧	[编剧]
	Billy Ray
编剧1的经验（两者）	1.386294361
编剧1的盈利能力（两者）	0.86888262
编剧2类似	Gary Ross
	0
	0
演员	[演员]
演员1	詹妮弗·劳伦斯
演员1性别	女性
演员1年龄	22
演员1经验（两者）	0.693147181
演员1盈利能力（两者）	0.792347251
演员2也一样 ，演员3	乔什·哈切森
	Male
	20
	1.098612289
	0.928510042
	利亚姆·海姆斯沃斯
	Male
	22
	0
	0

E. 更多实验结果

调整损失权重。图5展示了预训练期间视觉定位损失权重对微调阶段测试集 Huber 损失的影响。具体而言，我们改变视觉定位损失与掩码语言模型

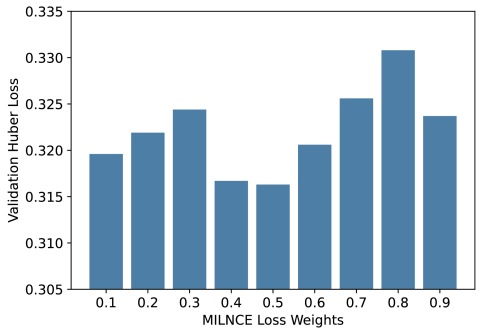

图5：联合使用掩码语言模型和视觉定位目标对模型进行预训练时，对 VG 损失权重的分析。

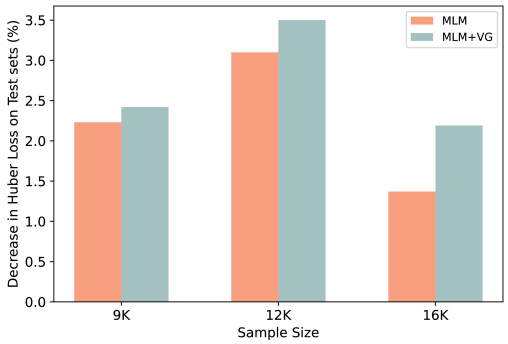

图6：在关键词数据与随机数据上训练时，Huber损失的下降情况；比较了样本量不同的掩码语言模型预训练模型和 MLM+VG 预训练模型。随着样本量增加，MLM 与 MLM+VG 之间的差距也随之扩大。

损失的权重，以考察其对测试集 Huber 损失的影响。根据我们的实验，为掩码语言模型损失和视觉定位损失设置相同的权重，可获得最低的测试集 Huber 损失。结果表明，电影上下文信息和视觉接地对关键词表示的贡献相同，凸显了视觉接地不可忽视的影响。

MLM 与 MLM 的详细比较+VG。表V展示了掩码语言模型和 MLM+VG 这两种预训练目标的详细对比数据。对于每个模型，我们进行了三次随机试验，然后计算平均值和标准差。这些数据反映在正文的图3中。

在图6中，我们展示了使用关键词数据训练相较于使用随机抽样数据训练所带来的相对测试损失改善；随机抽样的数据中包含带有关键词和不带关键词的电影。绿色柱表示 MLM+VG 模型的平均改善幅度，红色柱表示 MLM 模型的平均改善幅度。值得注意的是，随着样本量增加，MLM+VG 模型与仅使用 MLM 的模型之间的差距也随之扩大，这表明视觉接地的效果会随着数据量增加而增强。

表V：对于每个模型，我们进行三次随机试验，然后计算平均值和标准差。

	样本量	平均值测试集 Huber 损失（标准差）	
		MLM 预训练	掩码语言模型+视觉 grounding 预训练
随机样本	25k	$0.3097_{(0.0004)}$	$0.3044_{(0.0005)}$
	20k	$0.3180_{(0.0028)}$	$0.3116_{(0.0007)}$
	16k	$0.3213_{(0.0010)}$	$0.3201_{(0.0018)}$
	12k	$0.3322_{(0.0002)}$	$0.3290_{(0.0012)}$
	9k	$0.3415_{(0.0025)}$	$0.3392_{(0.0015)}$
使用关键词	16k	$0.3169_{(0.0012)}$	$0.3131_{(0.0009)}$
	12k	$0.3219_{(0.0007)}$	$0.3175_{(0.0011)}$
	9k	$0.3339_{(0.0020)}$	$0.3310_{(0.0012)}$
不使用关键词	9k	$0.3498_{(0.0014)}$	$0.3482_{(0.0002)}$



图9：在典型家庭与动画电影玩具总动员 *(1995)*的语境下，使用关键词“友谊”进行海报检索



图7：在非典型惊悚片*The Silence of the Limb(1991)*的语境下，使用关键词“心理”进行海报检索



图8：在典型战争片拯救大兵瑞恩 *(1998)*的语境下，使用关键词“战争”进行海报检索