E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)09-3068-15

论文引用格式:Yang B, Liu Y L, Liu Q and Zhu Y Y. 2025. TextLLM: a document multimodal large model based on dynamic resolution. Journal of Image and Graphics, 30(9):3068-3082(杨彪, 刘禹良, 刘强, 朱盈盈. 2025. TextLLM:基于动态分辨率的文档多模态大模型. 中国图象图形学报, 30(9):3068-3082)[DOI:10.11834/jig.240608]

TextLLM:基于动态分辨率的文档多模态大模型

杨彪[1],刘禹良[1],刘强[2],朱盈盈[1*]

1. 华中科技大学人工智能与自动化学院,武汉　430074; 2. 武汉金山办公软件有限公司,武汉　430074

摘　要:**目的**　文档智能旨在自动和智能地处理纸质文本信息,包括但不限于表格、表单和发票等,极大便利了信息的电子化管理。然而,传统深度学习方法往往专注于单一任务的优化,限制了它们在处理复杂多变的文档场景时的效能。此外,这些方法需要额外的光学字符识别(optical character recognition,OCR)工具来提取文档中的文字信息,这不仅增加了处理步骤,也可能引入额外的错误。多模态大模型的出现为免去OCR工具统一处理文档信息带来了希望,但是在处理高分辨率的文档图像和应对逐渐增加的视觉标记时,仍然面临着不小的挑战。本文提出一种基于动态分辨率的文档多模态大模型TextLLM,能免OCR工具处理高分辨率的文档图像。**方法**　基于最新的多模态大模型训练了一个能够处理动态分辨率的文档多模态大模型。在动态分辨率的基础上,提出一种动态特征压缩算法,设置动态的可学习压缩率来获得需要保留的特征长度,通过计算特征相似度来得到重要性特征,以此来聚合关键特征。更进一步,利用大语言模型的注意力机制捕捉与提示词相关的视觉特征部分,根据提示词的注意力分布图筛选出最相关的特征,并保留其周围相关特征。**结果**　实验在多个数据集上与6种先进方法进行了比较,TextLLM在多个文档理解基准测试中取得了显著的性能提升。在DocVQA、WTQ、ChartQA和TextVQA等数据集上的表现均优于对比模型,分别获得了82.4、37.6、70.8和65.3的分数。此外,在综合评测数据集OCRBench中,模型得分高达601,证明了其在多样化文本相关任务中的适应能力和整体效果。同时也在多个数据集中进行了消融实验以验证算法的有效性,消融实验验证了提出的动态算法能够改善模型效果。**结论**　本文提出基于动态分辨率的文档大模型TextLLM,并提出动态压缩特征和动态选择的算法来应对多场景的文档。实验结果表明,本文模型优于几种先进的文档大模型,兼具了高效性和准确性。

关键词:多模态;文档大模型;文档智能;动态分辨率;动态压缩率

TextLLM: a document multimodal large model based on dynamic resolution

Yang Biao[1], Liu Yuliang[1], Liu Qiang[2], Zhu Yingying[1*]

1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan　430074, China;

2. Wuhan Kingsoft Office Software Co., Ltd., Wuhan　430074, China

Abstract: **Objective** The advancement of document intelligence aims to realize the intelligent processing and interpretation of various document information, such as the processing of structured documents containing tables, forms, and invoices, among others, and text in natural scenes. Traditional deep learning methods excel at certain tasks, but because they are often optimized for a single task, they struggle to adapt to increasingly complex needs and diverse scenarios. When processing document text, optical character recognition (OCR) is usually required to extract text information. However,

收稿日期:2024-10-16;**修回日期**:2025-01-17;**预印本日期**:2025-01-24

*通信作者:朱盈盈　yyzhu@hust.edu.cn

基金项目:国家自然科学基金项目(62206103)

Supported by:National Natural Science Foundation of China (62206103)

the processing speed and accuracy decrease. Moreover, the various steps involved in handling the reading of the pipeline text can lead to the accumulation of errors. In addition, relying on ready-made OCR models/APIs introduces additional engineering complexity, limits connections between text and its surrounding context, and can increase computational costs. OCR-free solutions have attracted increasing attention recently to mitigate the disadvantages of external systems before they are understood. With the rise of multimodal large models, the field of document intelligence is undergoing a revolution. These models integrate textual and visual information to enable a more comprehensive and accurate understanding of document content, potentially eliminating reliance on OCR tools. However, these multimodal large models still face challenges, especially when it comes to processing high-resolution document images and dealing with the ever-growing number of visual markers. Previous similar working resolutions are limited by the input size of the encoder, and seeing specific small text is generally difficult. Reducing the image resolution will also greatly increase the number of tokens of the image, which is challenging for memory and space consumption. A large multimodal model based on dynamic resolution is proposed to overcome these challenges. This model is designed to handle high-resolution document images, eliminating the need for OCR tools while being flexible enough to accommodate ever-increasing visual tokens. **Method** In this study, we improve the latest multimodal large model and train a large document model on the basis of dynamic resolution, called TextLLM. This process mainly involves dynamic adjustment of the image, block processing, visual coding, feature compression, and screening to extract and retain useful information to the maximum extent. First, the area closest to the predefined scaling size is determined on the basis of the original image size for subsequent processing, including scaling processing, image slicing, and global information acquisition. Next, the image blocks are fed into the visual encoder for encoding processing, including window attention mechanisms to recognize and fuse information, and extract local regional visual features and global visual features. Then, the image resampling module is used to process the segmented features, and the compressed visual features are obtained on the basis of the attention mechanism. The compression rate is set dynamically, and discrete compression rates and a learnable parameter are defined to learn the compression rate dynamically. Gumbel softmax is used for end-to-end training to learn the compression rate. According to the compression ratio, the number of features after compression is obtained. The most important features are selected and sorted by calculating the similarity and importance, and these important features are used for further aggregation. Finally, the attention mechanism of the large language model is used to capture the visual features related to the prompt words, the most relevant features are selected according to the attention distribution map of the prompt words, and the surrounding relevant features are retained. The integrated application of these steps helps in effectively processing and utilizing image information and improving the efficiency and accuracy of information extraction. **Result** Experiments across multiple datasets compared TextLLM with the latest six methods, demonstrating considerably performance improvements in various document understanding benchmarks. It outperforms existing models on datasets such as DocVQA, WTQ, ChartQA, and TextVQA, achieving scores of 82.4, 37.6, 70.8, and 65.3, respectively. Furthermore, on the comprehensive OCRBench evaluation dataset, the model achieves a score of 601, proving its adaptability and overall strength in a variety of text-related tasks. Comparative experiments are performed across multiple datasets to verify the effectiveness of the algorithm, and results confirm that the proposed dynamic processing algorithm improved the model's performance. Several mainstream scenario images are selected for visualization in this paper to further demonstrate the model's capabilities in text-related tasks across various scenarios. The model is able to accurately recognize text in scene images and documents and answer questions on the basis of its understanding, showing its strong text processing capabilities and adaptability. Introducing dynamic feature compression ensures that the model can autonomously learn the compression rate and sample different compression rates during the learning process, which to some extent serves as data augmentation. Furthermore, incorporating visual feature selection enables the model to focus more on text-related features. Reducing features further allowed it to achieve improvements across all datasets. **Conclusion** In this study, we introduce an innovative document large model called TextLLM, which, on the basis of dynamic resolution, combines dynamic compression feature and dynamic selection algorithm. Extensive experimental validation shows that our model not only outperforms several state-of-the-art document large models but also achieves significant improvements in efficiency and accuracy.

Key words: multi-modality; large document model; document intelligence; dynamic resolution; dynamic compression ratio

0 引 言

在信息时代，文档智能技术正变得越发关键。它通过自动化处理和信息管理等功能，显著提升了文档处理的效率和准确性，这不仅极大地丰富了用户的文档管理体验，也促进了工作流程和决策过程的智能化(刘成林 等，2023；应自炉 等，2020)。文档智能的持续进步能实现对各类文档信息的深入处理和智能解读，包括但不限于结构化文档，如表格、表单、发票以及自然场景中的文本。这一趋势预示着，用户将享受到更智能化、更高效的文档处理方案，为信息管理和决策制定带来现代化的变革。近年来，学术界提出多种文档理解模型，大体可分为两类：一类是依赖额外光学字符识别(optical character recognition，OCR)工具的文档模型；另一类无须依赖OCR工具，直接对文档图像进行解析。

依赖额外光学字符识别工具的文档模型通过使用OCR技术提取文本及其位置信息，进而结合文本内容、位置信息和视觉元素来深入理解文档，更进一步会设置多种预训练任务来加强彼此的对应关系，完成视觉信息抽取(林泽柠 等，2023)等任务。UDOP(Tang等，2023)通过Transformer和统一的生成式预训练任务，整合了视觉、文本和布局信息，提升了模型的理解和处理能力。LayoutLM系列(Xu等，2020；Xu等，2021；Huang等，2022)则通过融合预先训练的文本、布局和视觉特征，采纳一个统一的模型框架和预训练目标来整体理解文档信息。DocLLM(Wang等，2023)进一步利用位置信息，将空间布局结构纳入考虑，利用分解的注意机制实现文本和空间之间的交叉对齐。

尽管OCR技术对于文档文本处理不可或缺，但其在提取文本时可能会降低处理效率并引入错误，如遗漏或误读文字。过分依赖现有的OCR系统也可能会增加项目的技术复杂度、维护工作量以及成本开销。鉴于此，学界出现了不依赖OCR系统的解决方案，以简化工程流程并提升处理的准确性和效率。Donut(Kim等，2022)率先提出一种基于Transformer并且无须额外OCR信息的端到端训练方法，通过阅读文本的预训练任务直接理解文字。Dessurt(Davis等，2022)则基于类似架构，结合了双向交叉注意力并采用了不同的预训练方法进行文档理解。Pix2Struct(Lee等，2023)通过学习将网页的屏幕截图解析为简化的HTML行预训练，更进一步引入了可变分辨率输入表示来更加灵活地集成语言和视觉输入。

这些先进的方法摆脱了对OCR工具的依赖，但为了达到最佳效果，通常需要针对特定任务进行精细调整。随着多模态大型模型的快速发展，文档智能领域正在经历一场变革。这些模型通过整合文本和视觉信息，提供了一种更为全面和精确的方式来理解文档内容，实现了在不依赖OCR工具的情况下，高效处理多样化的文档任务。为了解决文档相关任务，LLaVAR(Zhang等，2023)、mPLUG-DocOwl(Ye等，2023a)和UniDoc(Feng等，2023b)为了应对丰富的自然场景，都创建了多样的文档指令数据集，以增强微调阶段对富含文本的图像的理解。但是这些模型都采用CLIP(contrastive language-image pre-training)(Radford等，2021)的固定分辨率，无法适应高分辨率的文档需要。更进一步，为了解决高分辨率的文档问题，UReader(Ye等，2023b)设计了一个形状自适应裁剪模块裁剪图像，利用冻结的低分辨率视觉编码器处理裁剪后的低分辨率图像，从而达到高分辨率输入的目的。DocPedia(Feng等，2023a)选择在频域空间中处理视觉输入，由于频域空间特性，可以极大地增大图像分辨率。Vary(Wei等，2023)则使用双分支网络，两个分支分别处理高分辨率图像和低分辨率图像，然后两者特征进行融合，扩充了输入语言模型的词典。TextMonkey(Liu等，2024)则使用切块的思想扩充了分辨率，同时进一步使用滑动窗口注意力的机制减轻了切块带来的割裂问题，并且使用Token Resampler的机制减少了图像的视觉标记。

尽管在文档大模型方面取得了显著进展，但这些模型仍面临一些挑战。例如，TextMonkey和DocPedia等模型在处理小图像时，都采用固定分辨率的方法，可能会导致不必要的图像放大。此外，TextMonkey和UReader通过切片方法提高图像分辨率也会导致图像的token数量大幅增加，从而对显存和空间占用提出更高要求。虽然已有一些架构，如Q-former(Li等，2023)被提出来压缩视觉特征，但这些方法在处理高分辨率文档图像时，仍难以有效总结丰富的文本信息，而且TextMonkey的压缩机制是凭经验设置超参控制压缩率，压缩率并不一定是最

佳压缩率。

为了应对这些挑战，本文提出基于动态分辨率文档的文档多模态大模型TextLLM。该模型能动态处理高分辨率文档图像，摆脱对OCR工具的需求，统一地完成文档的各项任务。在动态分辨率的基础上，提出一种动态特征压缩算法，设置动态的可学习压缩率以获得需要保留的特征长度，通过计算相似度得到重要性特征，以此来聚合关键特征。更进一步，利用语言模型的注意力机制捕捉与提示词相关的视觉特征部分，根据提示词的注意力分布图筛选出最相关的特征，并保留其周围相关特征。实验结果表明，本文的模型优于几种最先进的文档大模型，兼具了高效性和准确性。

1 本文方法

文档大模型的目的是通过对大规模图像文本数据进行训练，使得模型能够理解文档图像，统一地完成各项文档任务。

本文所提出模型的总体框架如图1所示，包含动态分辨率切分、视觉编码器、滑动窗口注意机制、视觉重采样模块、动态特征压缩模块、大语言模型以及视觉特征选取器。TextLLM在训练时对输入图像进行动态处理，采用动态分辨率策略，根据原始图像的尺寸找到最接近448 × 448像素的倍数作为缩放大小，以确保图像在缩放过程中不失真。调整分辨率后的原始输入图像使用固定大小的滑动窗口切分为448 × 448像素的局部区域图像块。同时，原始输入图像的分辨率也调整为448 × 448像素，与视觉编码器的分辨率相匹配。经过该处理，局部区域图像块包含待处理的局部视觉信息，而调整后的完整原始图像则提供全局视觉信息。这些分辨率均为448 × 448像素的图像块将被输入到后续的网络视觉编码器、滑动窗口注意机制（Liu等，2021）以及视觉重采样机制。视觉编码器负责对不同图像块特征编码。切块的处理策略可能会导致文档分析中连接文本的碎片化，从而破坏语义连续性，因此引入了滑动窗口注意力机制增强CLIP模型的视觉处理能力。而视觉重采样机制则是一个图像映射机制，将视觉特征映射到了语言空间，采用的是基于Q-Former的范式，并且由于Q-Former的机制，会对图像做初步的特征压缩。更进一步，这些视觉特征会经过一个动态特征压缩模块，使用动态可学习的压缩率以减少图像中的冗余信息。此外，模型将进一步利用语言模型的注意力机制，筛选出与提示词相关的视觉特征，并保留最显著的特征。最后，这些经过处理的特征与输入问题相结合，通过大型语言模型（large language model，LLM）进行分析，得出所需的答案。最终，模型通过收集的高质量数据集进行了全面的多阶段训练，以此提升对文档的感知和理解能力以及特征压缩能力。

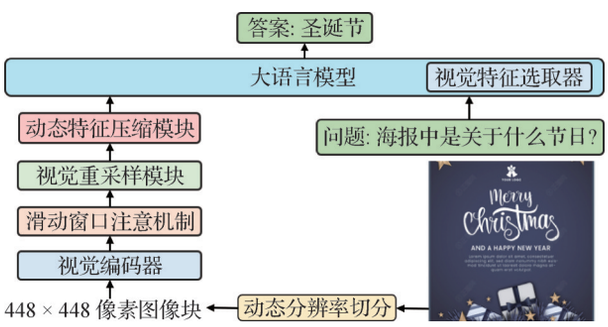

图1 TextLLM模型框架

Fig. 1 Framework of TextLLM model

1.1 动态分辨率切分

在实际的文档处理场景中，图像的尺寸和纵横比各不相同。因此将它们强行统一到一个固定的分辨率可能导致一定程度的图像扭曲，同时将低分辨率的图像拉伸到高分辨率也会造成过长的特征。因此动态分辨率切分目标是确定不同尺寸图像的切割方式，使得在保证足够特征编码长度的基础上，大致符合其原始的纵横比。

首先本文预定义了一个候选的grid池，对于切片数目N分解为m列和n行的grid，具体为

$$\boldsymbol{grid}_i = \{(m,n) \mid m \times n = N, m \in N, n \in N\} \quad (1)$$

因此，整体grid池就是1到最大切片数目P的集合。

如图2所示，给定分辨率为(H, W)的图像和预训练分辨率为(H_v, W_v)的CLIP，首先确定处理图像所需要的切片数目M，此过程表示为

$$M = \left\lceil \frac{H}{H_v} \right\rceil \times \left\lceil \frac{W}{W_v} \right\rceil \quad (2)$$

在实践中，本文发现仅仅依赖于M的因数分解来选择时，grd池的选择是有限的，并且在M为素数的情况下，只有一个选项可供选择。因此，本文选择的grid池为所有比M小的grid情况，即

$$\boldsymbol{grid} = \boldsymbol{grid}_1 \cup \boldsymbol{grid}_2 \cup \cdots \cup \boldsymbol{grid}_M \quad (3)$$

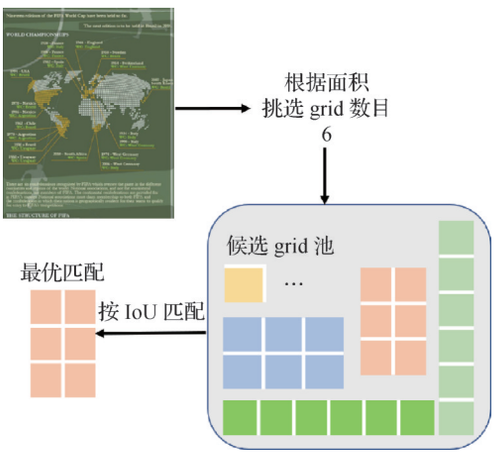

图2　动态分辨率grid选择示意图

Fig. 2　Dynamic resolution grid selection diagram

本文需要使用匹配算法来找到其中最优的grid，因此定义匹配的得分算法如下：使用与规定grid的交并比（intersection over union，IoU）判断重合程度，评分方式为

$$S(I,\boldsymbol{grid}) = IoU\left(\left(\frac{n_{\mathrm{h}}H}{w},n_{\mathrm{w}}\right),\left(n_{\mathrm{h}},n_{\mathrm{w}}\right)\right) \tag{4}$$

式中，n_{h}和n_{w}代表grid中的高宽尺寸，越高的S表示重合度越高，选取概率越大。此外，在分数相同的情况下，本文选取尺寸更大的grid，如分数相同的1×2和2×4的grid，会优先选取2×4的grid。对于少数极端面积相同的情况，如1×6、2×3的grid，考虑到文档大部分竖向分布，则选择高（n_{h}）更大的grid，这里选取1×6的grid。

根据其原始尺寸找到与预定义缩放比例大小最接近的grid，该预定义缩放大小为448×448的倍数。

即使原始图像尺寸小于448×448像素或不是448的倍数，都会在此算法上找到面积大小基本一致，且纵横比相近的缩放大小。然后利用固定大小的滑动窗口将调整至预定义大小的输入图像划分为较小的局部区域图像块，以获取局部视觉信息。为了进一步获取全局视觉信息，本文将原始图像的分辨率再次调整至与滑动窗口大小相同并送入视觉编码器中。

1.2　动态特征压缩机制

随着分辨率的提高，使用滑动窗口切块所生成的token数量显著增加，尤其在处理高分辨率图像时这一问题尤为突出。然而，鉴于大型语言模型对输入长度的限制和训练时间，优化token数量显得尤为关键。先前的研究TextMonkey已经成功验证了压缩token的可行性。这一方法不仅有效降低了输入标记符的数量，还在一定程度上提升了模型的处理效率和性能。但是它的压缩机制是经验性地设置超参来控制压缩率，压缩率并不一定是最佳压缩率，因此本文引入了动态特征压缩机制，自动学习最优压缩率，动态压缩切片的特征。

本文定义了n个离散的压缩率，这一设计旨在为模型压缩尺度的选择提供灵活性，使其能够在训练过程动态选择压缩力度，即

$$\boldsymbol{C} = \{C_1, C_2, \cdots, C_n\} \tag{5}$$

式中，C_i是其中的离散压缩率。更进一步考虑到不同尺寸的图像存在的特征分布可能不同，因此对于不同图像的压缩率是分别计算的，会有P个压缩率集合，为了简写，只展示其中某个压缩率的结果。

为了在训练中能可导且动态地选择压缩率，引入了一个可学习的参数，代表每个压缩率的选择概率，具体为

$$\boldsymbol{p} = \{p_1, p_2, \cdots, p_n\} \tag{6}$$

式中，p_i是第i个压缩率的选择概率。

由于Gumbel Softmax（Jang等，2016）的可导性，整个采样过程可以直接嵌入到梯度下降的训练框架中，并且由于它是根据概率采样，在训练过程中也可能会采样到一些小概率压缩率，在一定程度上起到数据增强的作用。在训练过程中，压缩率的选择策略如下：模型会根据当前的概率分布进行采样，选取一个压缩率来压缩视觉特征。这个过程模拟了在实际应用中可能遇到的各种压缩情况，从而使得模型在训练时就能够学习到如何在不同的压缩率下保持最佳性能，训练过程中压缩率选择为

$$r = L\times C_i, i = f_{\mathrm{GumbelSoftmax}}(p) \tag{7}$$

式中，L是输入特征长度，r为压缩后的特征长度。

在测试过程中，由于不需要采样过程，只需要挑选最大概率的压缩率即可，具体为

$$r = L\times C_i, i = f_{\mathrm{Argmax}}(p) \tag{8}$$

最终得到的压缩结果长度为r。接下来的特征压缩过程来源于TextMonkey中的TokenResamper算法。如图3所示，初始未曾压缩的特征维度为（B，L，D），B是批大小，L是特征长度，D是隐藏宽度，具体实现流程如下：

1）计算每个特征与其他特征之间的相似度。这种相似度是通过向量空间中的距离实现的，这里使

用的是余弦相似度。相似度值越高，说明两个特征之间越相似。

2）将相似度转化为重要性分数。具体而言，可以通过 1 减去相似度值获得该特征的重要性分数，即它的独特性。

3）在计算出重要性分数后，根据这些分数选择出最重要的 r 个特征。然后，将这些重要的特征按照它们最初的顺序进行排序，特征维度为（B，r，D）。

4）为了避免直接丢弃其他特征造成的信息丢失，利用选出的重要特征作为查询，采用交叉注意力机制进一步聚合所有特征，最后特征维度为（B，r，D）。交叉注意力机制通过关注所有特征之间的关系，能够有效地整合不同 token 的特征信息，从而增强模型对整体输入的理解。

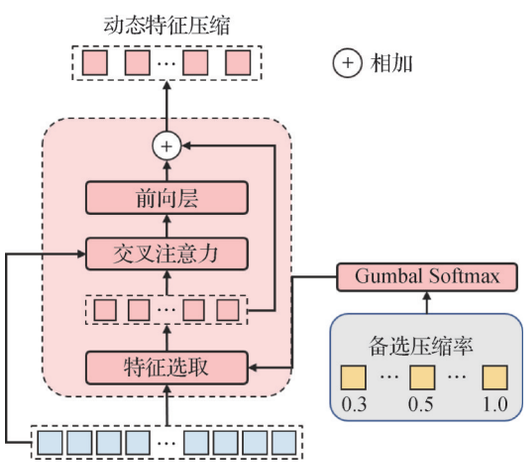

图 3　动态特征压缩模块

Fig. 3　Dynamic feature compression module

1.3　视觉特征选取器

之前的研究（Chen 等，2024）发现，在大语言模型的前两层注意力分布图中，图像 token 的注意力是较为稀疏的。因此本文利用语言模型的注意力机制捕捉与提示词相关的视觉特征。首先将提取的视觉特征和文本特征输入到前两层语言模型中，在模型处理过程中，语言模型会生成一个注意力分布图。在第 2 层的 Transformer 的多头注意力层中，得到查询和键值的乘积，这就是对应的注意力矩阵。这个矩阵展示了每一对 tokens 之间影响程度的分数，行表示查询 token，列代表键 token。

本文根据这个注意力分布图计算图像 token 的注意力，具体为

$$\varepsilon_{\text{img}}^{j}=\sum_{i=1}^{n}\alpha_{\text{img}}^{ij} \tag{9}$$

式中，n 为输入问题 query 的长度，α_{img}^{ij} 代表第 i 个问题 token 对于第 j 个图像 token 的注意力值，因此 $\varepsilon_{\text{img}}^{j}$ 表示的是第 j 个图像 token 所获得的注意力值。

该注意力值反映提示词对各个视觉特征的关注程度，也显示了哪些视觉特征在理解提示词的上下文时被认为是最重要的。模型从注意力分布图中选择与提示词相关性最高的 top70% 的视觉特征。为了保证特征的连续性和上下文信息的完整性，设置了一个保存窗口，保留这些特征前后各 5 个特征。最后将这些特征重新输入语言模型中进行联合训练。这一策略能捕捉特征之间的关联性，确保最终保留的特征能够更全面地反映出视觉信息的上下文。

1.4　训练目标

大语言模型的输入包含了图像和文字，输入图像 $\boldsymbol{I}$ 经过图像编码器以及本文的特征压缩等压缩特征会得到图像特征 $\boldsymbol{x}^{v}=\{x_1^{v},x_2^{v},\cdots,x_n^{v}\}$，其中 n 是图像 token 的长度。同样地，输入的问题 query 以及答案经过编码和向量化也会得到文本特征 $\boldsymbol{x}^{t}=\{x_1^{t},x_2^{t},\cdots,x_m^{t}\}$，其中 m 是文本 token 的长度。图像和文本标记被连接为最终输入序列 $\boldsymbol{X}$，长度为 $n+m$。

将输入序列 $\boldsymbol{X}$ 输入模型中，模型以自回归方式输出答案，根据先前标记预测下一个标记，形式化为

$$p\left(\boldsymbol{X}_t\left|\boldsymbol{X}_{<t}\right.\right)=f_{\text{Softmax}}\left(LLM\left(\boldsymbol{X}_{<t}\right)\right) \tag{10}$$

因此，在训练过程中，模型训练目标为最大化下一个 token 的概率，具体为

$$L=\max\sum_{t=1}^{L}\log p\left(\boldsymbol{X}_t\left|\boldsymbol{X}_{<t}\right.\right) \tag{11}$$

综上，本文方法利用动态分辨率模块处理大小迥异的文档图像，使用动态特征压缩模块做压缩图像特征，进一步使用视觉特征选取器选择与问题更相关的图像特征。最后复用大语言模型的损失模块对齐语言模型的训练，保证了与大语言模型的一致性。

2　实验与分析

2.1　实验数据

为了优化模型的训练效果，本研究采用两阶段训练策略。在第 1 阶段，首先通过详尽的文档描述

数据集，为模型提供了丰富的信息和上下文进行初步特征的压缩。第2阶段则利用了更为广泛的下游文档数据，进一步丰富了模型的训练样本，提升了其泛化能力和视觉特征选取能力。这一策略既满足了模型对丰富数据的需求，也实现了对大规模数据集的有效利用。为了提高训练效率，本文数据大多采取对话格式（Liu等，2023a），减少了训练的轮次，并且相同grid放到同一批训练。

2.1.1 第1阶段数据

第1阶段的数据可以为模型提供丰富的信息和上下文进行初步特征的压缩。mPLUG-DocOwl 1.5（Hu等，2024a）提出一个规模为4 M左右的文档预训练数据集，涉及了跨5个领域的统一结构学习，包括自然图像、文档、表格、图表和网页，还涉及结构感知解析任务和多粒度文本定位任务。本文使用部分筛选后的数据作为第1阶段的训练数据，其中主要有以下几个任务，能较为详细地描述图像中的所有内容。

文档解析通过添加换行符和空格表示文本的行和水平距离。图表解析将图表转化为表格以保留其数学特性，并使用Markdown代码表示。表格解析则利用标准化的行和列结构，通常采用HTML或Markdown代码表示。自然图像解析结合了图像中的物体和场景文本，形成目标解析序列。多粒度文本定位则关注将文本与图像中的具体位置相关联，以增强视觉文档理解的结构感知能力。

在处理规模达到4 M的数据集时，模型训练的迭代过程可能会异常耗时。考虑到数据集中可能包含一些对模型学习贡献有限的简单样本，本文使用了一种提高训练效率的策略——不采用完整的数据集进行训练，而是利用一个预先训练好的模型对数据进行初步筛选。具体而言，借助初始模型的预测能力，对全量数据的问题提前进行回答，通过对预测结果和真值结果对比，对数据集中的样本进行难度评估，进而挑选出难度适中且具有代表性的样本。这种方法不仅能够缩短训练时间，还能在一定程度上增强模型的泛化能力。通过筛选出更具挑战性的样本，可以更加高效地利用计算资源，并且确保模型在处理复杂样本时的性能。因此通过这种方式，模型能够更加专注于从那些对学习最有价值的样本中提取信息，从而在训练过程中实现效率的提升。

2.1.2 第2阶段数据

在第2阶段，训练过程中进一步扩充了模型的训练数据，从更广泛的下游文档数据中汲取信息，以提升其泛化问答能力和视觉特征选取的准确性。这一阶段的数据来源多元，包含了各个领域的文档、自然图像、表格和图表，使得模型在多样化数据中获得更全面的训练。训练数据如图4所示，涵盖了多个领域，为更好地展示效果部分，部分图像进行了裁剪。

1）文档数据。采用各行各业的文档数据，包括UCSF（University of California, San Francisco）行业文档库、慈善组织的报告以及信息提取数据集等。这些数据主要来源于IIT-CDIP（Lewis等，2006）、DocVQA（Mathew等，2021）、InfoVQA（Mathew等，2022）、DeepForm（svetliehnaya，2020）和KLC（Kleister Charity）（Stanisławek等，2021）等知名数据集。

2）表格数据。整合了来自维基百科的表格图像数据，主要依托于WikiTableQuestions（WTQ）（Pasupat和Liang，2015）数据集。

3）图表数据。模型还特别针对包含坐标轴信息的图表数据进行了优化，主要来源包括ChartQA（Masry等，2022）数据集。

4）自然图像数据。引入了带有文字信息的自然图像数据，例如街景照片、书籍截图等。这部分数据主要来自TextVQA（Singh等，2019）、COCOText（Veit等，2016）、TextOCR（Singh等，2021）、HierText（Long等，2022）、STVQA（Biten等，2019）和MLT（Nayef等，2019）等数据集。

5）网页截图数据。利用完整屏幕截图数据，这些数据主要来源于VisualMRC（Tanaka等，2021）数据集。

在这些原始问答数据的基础上，还使用了一系列扩充内容，包括TextHawk（Yu等，2024）合成的数据、TextMonkey（Liu等，2024）的结构化数据，以及LayoutLLM（Luo等，2024）的推理数据。这些新增内容为模型提供了更多样化和详尽的信息，进一步丰富了模型的训练样本，使其能够更好地理解和回答各种类型的问题。通过整合这些丰富多样的数据源，模型的训练使用了共计250 K个数据，不仅扩展了数据集的规模，也提升了模型在问答任务和视觉特征选取方面的表现。这种综合性的训练方法有助于模型在多样化数据中建立更加全面和准确的知识表示，提高其泛化能力和应对复杂问题的能力。

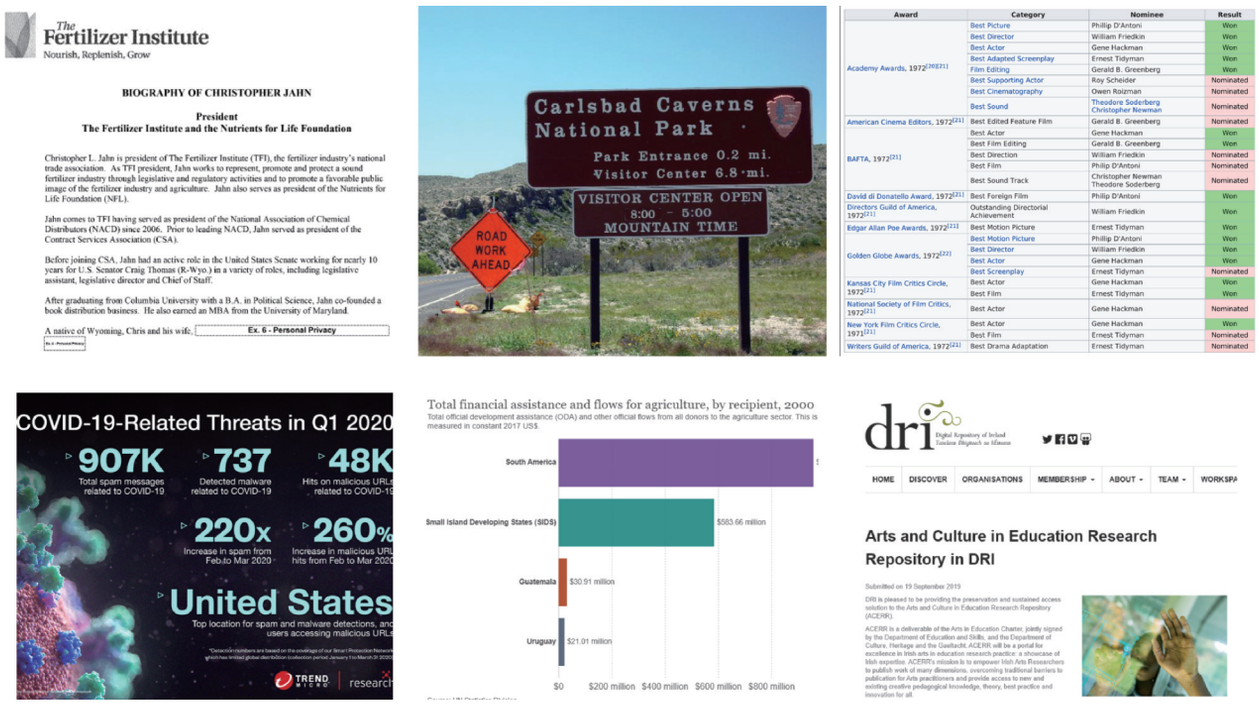

图4　训练数据图像样例(为了更好地展示效果部分,部分图像经过裁剪)

Fig. 4　Training data image samples (some images are cropped for better display effects)

最后,为了与推理数据的格式保持一致,还额外引入了单轮对话数据进行微调,这些数据仅包含原始的单轮问答对话,总计 238 K 个样本。这些对话数据的引入进一步完善了模型的训练内容,使其单轮问答场景中得到更加精细的调整和优化。

2.2　实验设置

在实验中使用了 8 卡 A800 GPU,显存为 80 GB,实验中使用了目前应用广泛的深度学习框架 PyTorch。初始化模型选用 Qwen-VL(Bai 等,2023),这是已经预训练完成的多模态模型。为进一步优化显存的使用,数据格式设置为 BFloat16,这种格式不仅能显著节省显存,还能在一定程度上加速计算过程。此外,为了充分利用数据的可用范围,模型的最大填充长度设置为 6 400。

在本实验的两个训练阶段中,均采用了 ZeRO-2 策略,以有效地节省模型训练所需的显存。ZeRO-2 通过切分优化器状态和梯度,使得训练过程中能够以更低的内存开销训练更大规模的模型。在训练过程中,模型的批大小设置为 128。优化器方面,本文选择了 AdamW 优化器。此外,模型采用了一个余弦学习率调度方案,以使学习率在训练过程中逐渐减小,从而帮助模型更好地收敛。在 AdamW 优化器的参数配置中,将 $\beta 1$ 和 $\beta 1$ 分别设置为 0.9 和 0.95,这一选择有助于稳定训练过程并加快收敛速度。同时,本文还引入了权重衰减因子,设置为 0.1,以进一步抑制过拟合现象。为了优化模型的初期学习过程,模型的预热时间设置为总训练迭代次数的 2%。这一策略在早期逐步提高学习率,使得模型在开始阶段以更平稳的方式进行训练,从而提高整体性能。

在两个阶段的训练过程中,第 1 阶段完成初步特征的压缩,第 2 阶段则更进一步进行视觉特征选取。因此,第 1 阶段仅包含动态特征压缩机制,而第 2 阶段则额外包含了视觉特征选取机制。第 1 阶段学习率设置为 5×10^{-5},使用详尽的文档描述数据集训练,训练轮次为 1 轮。第 2 阶段学习率设置为 2×10^{-5},使用下游数据进行训练,训练轮次为 2 轮,训练得到了 TextLLM。同时为了更好地符合测试单轮问答的场景,在第 2 轮基础上额外进行了单轮问答的训练,训练轮次为 2 轮,学习率为 2×10^{-5},训练得到了 TextLLM*。通过训练策略和参数设置,实现模型的高效训练和较好的泛化能力。

在动态分辨率设置上,模型设置最大的切分块数 P 为 9,避免过大的切分造成的模型填充长度过长。在动态特征压缩机制上,设置了 20 个离散压缩率,其中最大压缩率为 30%。在视觉特征选取中,模型会选取大语言模型的第 2 层之后的注意力图,并

且计算每一个问题的token到每一个视觉token的注意力进行训练阶段的裁剪。

在测评阶段，为了全面而系统地评估文档大模型的能力，本文选择了多个具有代表性的数据集，共计4个文档相关的数据集：DocVQA、InfoVQA、DeepForm和KLC。这些数据集涵盖了各种文档任务，能够有效反映模型在处理不同类型文档时的表现。此外，还纳入了表格问答数据集WTQ和图表数据集ChartQA，以评估模型在处理结构化数据和图形信息时的能力。WTQ专注于从表格中提取信息，而ChartQA则测试模型在理解和解读图表数据方面的能力。为了进一步扩展评估的范围，本文还包括了自然图像问答数据集TextVQA，以测试模型在与自然场景图像相关的问答任务中的表现。

最后，为了综合评估模型的OCR能力，本文还使用了专门设计的OCRBench（Liu等，2023b）数据集。该数据集的引入能够对模型在文本识别和理解方面进行更为细致的评价。为了确保与其他方法进行更公平的比较，在评测模型性能时严格采用了原始数据集所定义的标准指标。这一选择不仅有助于提供更为客观的评估结果，也使得不同方法之间的比较变得更加一致和可靠。具体而言，对于DocVQA和InfoVQA数据集，使用了平均标准化编辑相似度（average normalized levenshtein similarity，ANLS）作为评估指标。在DeepForm和KLC数据集中，采用了F1分数作为评估指标。对于WTQ数据集，选择了准确率指标。在ChartQA数据集中，采用了放松准确率指标（relaxed accuracy measure）。最后，对于TextVQA数据集，使用了VQA Score作为评估标准。同时，OCRBench数据集则采用了准确率指标，专门用于评估模型在光学字符识别任务中的表现。通过以上多样化的标准指标，旨在对模型在多个数据集和任务上的表现进行客观和深入的分析，从而为后续研究提供有价值的参考。

2.3 实验结果

2.3.1 定量实验

为了确保公平地比较本文提出的算法与其他流行的算法，选择了多种当前主流的多模态大模型进行对比，这些模型的视觉编码长度均在1 000左右。具体而言，选取了以下6种方法进行评估：mPLUG-DocOwl（Ye等，2023a）、UReader（Ye等，2023b）、Qwen-VL（Bai等，2023）、Monkey（Li等，2024）、TextMonkey（Liu等，2024）以及mPLUG-DocOwl2（Hu等，2024b）。这些方法均在各自的文献中进行了详尽的实验证明，展示了它们在多种文档处理任务中的优秀性能。在比较过程中，各个模型的性能指标均来自于相应的文献，缺省值用“-”表示，以确保数据的透明性和可比性。

如表1所示，模型在大部分数据集上的表现优于其他模型，表明其在处理各种文档信息时的优势。特别是在不同领域的具体任务中，TextLLM*在DocVQA中取得了82.4分，在WTQ中获得了37.6分，在ChartQA中达到了70.8分，而在TextVQA中则获得了65.3分。这些结果不仅体现了本文模型在多个文档相关任务中的出色表现，还展示了模型在处理文档、表格、图表和自然场景文本的能力与灵活性。更进一步，在综合评测数据集OCRBench中，本文模型获得了601分，展示了模型在不同文字相关任务之间的适应性。在某些数据上略低于其他模型，这可能是由于数据量较少，没有大规模覆盖对应的数据集分布。最后通过比较TextLLM和TextLLM*的结果，本文还验证了单轮数据微调的重要性。通过这种微调，模型能够更好地契合问答任务的特定需求，从而实现较为一致的性能提升。此外，通过对模型性能的分析，发现现有模型在海报理解InfoVQA和涉及复杂版式的长文档关键信息抽取的相关数据KLC上仍然存在挑战，并且有进一步改进空间。

2.3.2 定性分析

为了展示动态分辨率选择的效果，本文选取了一些图及其对应的grid进行展示，grid已表示在图像上。如图5所示，动态分辨率切分能确定不同尺寸图像的切割方式，使得在保证足够特征编码长度的基础上，大致符合其原始的纵横比。此外，在分数相同的情况，本文选取尺寸更大的grid，如图5右上角选取了2×2的grid而非1×1的grid。

为了更进一步展示模型在各个场景中与文字相关的任务性能，本文选取了几个主流场景的图像进行了可视化，测试结果如图6所示。本文模型能够准确地识别场景图像以及文档中的文本并且根据理解回答问题，展现出其强大的文本处理能力和适应性。

具体而言，图6（a）展示了表单中的关键信息抽取，本文模型能够准确理解大分辨率表单数据中的

表 1　各类方法在数据集上的指标对比

Table 1　Comparison of metrics for various methods on datasets

模型	大小/B	token 数	DocVQA	InfoVQA	DeepForm	KLC	WTQ	ChartQA	TextVQA	OCRBench
mPLUG-DocOwl	7	~841	62.2	38.2	42.6	30.3	26.9	57.4	52.6	–
UReader	7	~841	65.4	42.2	49.5	32.8	29.4	59.3	57.6	–
Qwen-VL	9	256	65.1	35.4	4.1	15.9	21.6	65.7	63.8	488
Monkey	9	1 280	66.5	36.1	40.6	32.8	25.3	65.1	64.3	514
TextMonkey	9	768	73.0	28.6	59.7	**37.8**	31.9	66.9	65.9	561
mPLUG-DocOwl2	8	324	80.7	46.4	**68.8**	37.5	36.5	70.0	**66.7**	–
TextLLM	9	~760	81.1	49.2	62.7	37.3	**38.4**	69.5	66.5	598
TextLLM*	9	~760	**82.4**	**50.2**	64.6	37.6	37.6	**70.8**	65.3	**601**

注：加粗、下划线字体分别表示各列最优、次优结果，“–”表示对应文献中未对该项指标提供相关信息。“*”表示经过了单轮数据微调。各数据集的评价指标如下：DocVQA 和 InfoVQA 使用 ANLS，DeepForm 与 KLC 使用 F1 分数，WTQ 和 OCRBench 使用准确率，ChartQA 使用放松准确率，TextVQA 使用 VQA Score。

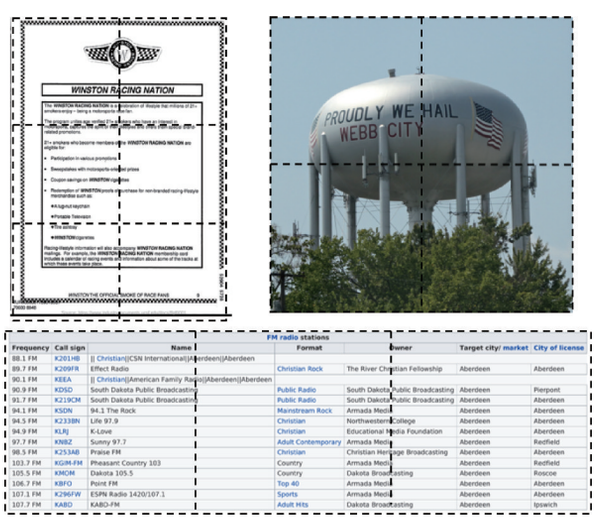

图 5　动态 grid 选择的可视化

Fig. 5　Visualization of dynamic grid selection

各个 key-value 对。这一能力不仅体现了模型在处理结构化数据时的高效性，还反映了其在信息提取任务中的精确性。图 6(b)则呈现了模型在扫描文档时的能力，能够识别并提取细小的答案。这一过程展示了 TextLLM 在面对复杂文本时的敏锐度，能够迅速定位并获取关键信息。图 6(c)展示了模型在图表数据中的表现，成功找到了众多数据中最小的一列，表明其在图形识别和数据分析方面的出色能力。图 6(d)涉及到表格中的信息理解，模型能够综合理解表格内的信息，并给出关于“谁比赛更多”的准确答案。这一特性强调了模型在处理多维度信息时的综合分析能力，使其在实际应用中更加可靠。最后，图 6(e)展示了自然图像中与文本相关的问答任务，模型在复杂的自然背景下依然能够关注到具体的文字信息，展现了其在动态场景中的适应性。这些不同类型的视觉内容充分体现了 TextLLM 在辨别和理解视觉与文本信息方面的卓越能力，凸显了其在广泛场景中的应用潜力。

2.4　消融实验

设计模型的各个组件是本文的主要创新点，能动态地处理多种类型的文档图像，同时兼顾高效性和有效性。为了验证各个设计组件对模型整体性能的贡献，本文特别设计了多组消融实验进行对比分析。这些消融实验能系统性地评估每个组件在模型运行中的重要性和有效性，通过逐步剔除或替换特定组件，能够深入了解各部分在整个系统中的作用。为了加快实验进程，在消融实验过程中，本文选择不使用单轮数据微调，而是仅采用两阶段的数据进行微调。其中不同组件的消融实验中的平均 token 数是使用 DocVQA 上的数据集取平均得到，浮点运算次数(floating point operations，FLOPs)则根据理论值计算。

2.4.1　patch 数目的消融实验

动态分辨率处理中的 patch 数量决定了模型能够处理的最大分辨率范围。因此，本文对模型能够处理的最大 patch 数目进行了消融实验，如表 2 所示，通过逐步增加 patch 的数量，观察到模型性能随之提升的趋势相对一致。在动态分辨率处理的过程中，增加 patch 数目使模型能够选择更适宜的尺寸进行处理，模型能够更灵活地适应不同分辨率的输入，

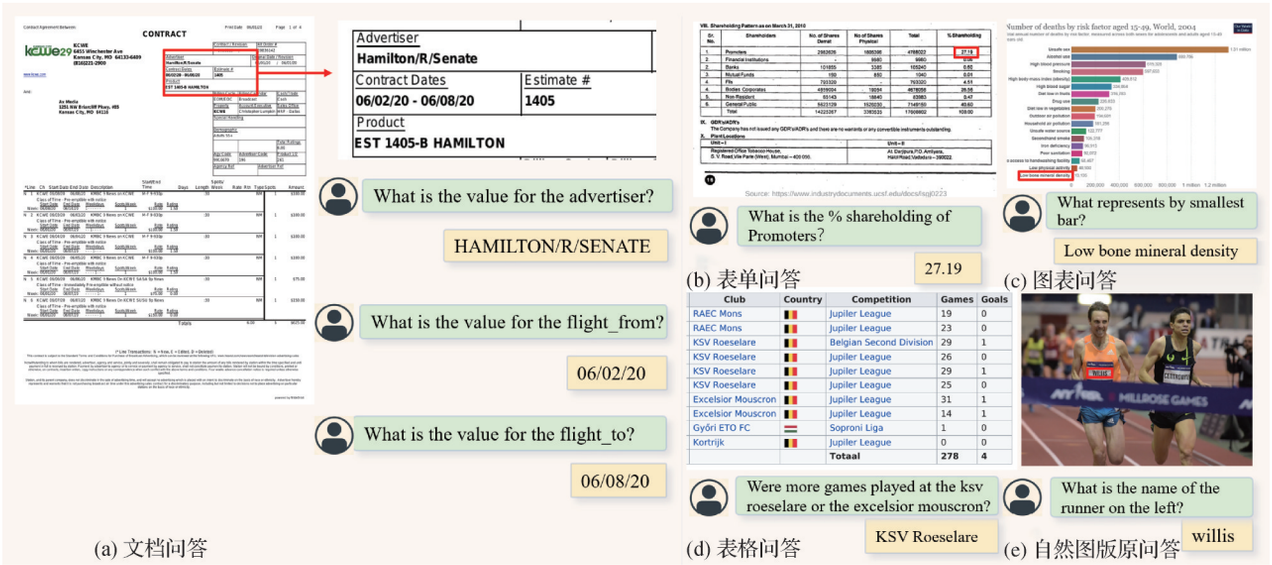

图6　模型可视化结果，答案位置在原始图像用红框标注

Fig. 6　The visualization results of the model, answer locations are marked with red boxes in the original images ((a) document question answering; (b) form question answering; (c) chart question answering; (d) table question answering; (e) natural image text-based question answering)

表2　patch数目的消融实验

Table 2　Ablation study on the number of patches

patch 数	DocVQA	ChartQA	InfoVQA	WTQ
4	77.3	68.2	44.3	35.8
6	79.4	69.1	45.3	35.0
9	**80.8**	**69.3**	**49.0**	**38.2**

注：加粗字体表示各列最优结果。

从而在保持计算效率的同时，提高处理精度。这一发现凸显了在设计模型时，适当调整处理最大分辨率的重要性，这对于最大限度地提升模型性能至关重要。

2.4.2　不同组件的消融实验

如表3所示，逐渐提升分辨率会导致token数的增加，也会间接导致FLOPS的增加，并且过度放大图像分辨率到1 344会导致信息过分冗余，效果反而不如固定分辨率896。而动态分辨率能自适应地选取合适的尺寸，每幅图像都有合适的token长度表征，因此模型取得了一致的提升，尤其是在InfoVQA上取得了4.4的增益。更进一步，本文尝试了固定压缩率，选取了经验性的50%作为压缩率，虽然token数目减少了近40%，但是由于压缩率并不参与优化，在测试结果上性能存在一定程度的衰减。通过引入动态特征压缩，模型能自主学习压缩率，并且在训练过程中使用采样方式采样不同压缩率，这在一定程度上起到数据增强的作用，因此，在动态特征压缩技术的助力下，模型在进一步压缩30%的token数量的基础上，在各个指标上也取得了较为一致的提升，在WTQ上相比于不压缩的方法取得了1.4的提升，在DocVQA上取得了0.6的提升。更进一步，通过加入视觉特征选取，模型能更加关注到文本相关的特征，在进一步减少特征的同时，在DocVQA上取得了81.1的结果，在ChartQA上取得了69.5的结果，在InfoVQA上取得了49.2的结果，在WTQ上取得了38.4的结果。通过一系列的消融实验，本文充分验证了模型中各个组件的有效性和贡献。

2.4.3　采样方式的消融实验

直接梯度估计(straight-through estimator, STE)是一种用于训练神经网络的技巧，特别是在处理非连续或离散变量时，如量化和二值化模型。在神经网络中，某些操作(如Argmax)可能导致梯度无法有效传播，因为它们是不可导的。STE则一定程度上解决了这一问题，它通过直接将梯度回传到上一个网络来使得能够使用反向传播算法训练这些网络。本文采用的Gumbel Softmax通过引入Gumbel噪声，使得离散采样过程变得可导，从而可以使用标准的反向传播进行训练，提供了更好的性能和灵活性。如表4所示，使用Gumbel Softmax的方式相较于STE

表 3　不同组件的消融实验

Table 3　Ablation study of different components

分辨率设置	压缩方式	DocVQA	ChartQA	InfoVQA	WTQ	平均 token 数	FLOPs
固定 448	不压缩	65.1	65.7	35.4	21.6	256	63.9
固定 896	不压缩	79.8	69.0	44.6	36.9	1 280	330.2
固定 1 344	不压缩	74.7	65.5	44.2	32.2	2 560	687.2
动态分辨率	不压缩	80.2	**69.8**	49.0	36.8	1 816	476.4
动态分辨率	固定 50% 压缩	79.9	68.6	48.4	36.2	1 036	265.2
动态分辨率	动态特征压缩	80.8	69.3	49.0	38.2	773	196.2
动态分辨率	动态特征压缩 + 视觉特征选取	**81.1**	69.5	**49.2**	**38.4**	**760**	**192.8**

注：加粗字体表示各列最优结果（第 1 行是基准数据，仅作为参照，不进行优劣对比）。

的方式在各个数据集都有一致的提升。因为 Gumbel Softmax 使用的是基于概率采样的方式，在采样过程中一定程度起到了特征增强的作用，而单纯使用 STE 可能导致模型压缩率不能被充分训练到。

表 4　采样方式的消融实验

Table 4　Ablation study on sampling methods

采样方式	DocVQA	ChartQA	InfoVQA	WTQ
STE	77.0	68.3	46.0	35.7
本文	**80.8**	**69.3**	**49.0**	**38.2**

注：加粗字体表示各列最优结果。

2.5　讨论

本文对不同分辨率设置了不同的压缩率（此处压缩率意义为压缩后长度/压缩前长度）。如图 7 所示，压缩率越低代表压缩层长度越小。实验表明，即使经过初步压缩，图像平均压缩率仍然能降至 50% 以下（即长度减少超过 50%），部分切片的压缩率甚至达到 35%（即长度减少 65%）。这一结果显示，未来模型设计可进一步增大初始压缩力度。此外，图 7（b）显示模型在最优压缩率附近存在多个可比性能的压缩率，表明模型对压缩率参数具有一定弹性。

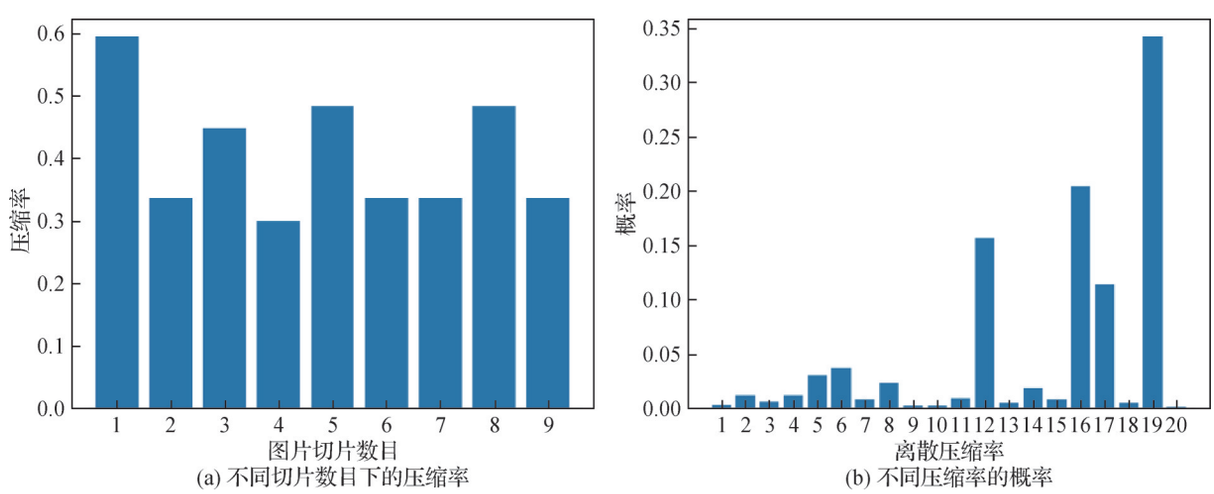

图 7　压缩率的统计结果

Fig. 7　Statistics of compression ratio

((a)compression ratio under different numbers of slices; (b)the probability of different compression ratios)

3　结　论

本研究提出并实现了一种基于动态分辨率的文档多模态大模型 TextLLM，旨在克服传统深度学习方法在处理文档信息时面临的挑战。TextLLM 不仅免去了对 OCR 工具的依赖，而且针对处理高分辨率的文档图像和逐渐增加的视觉标记提供了有效的解

决方案。通过改进最新的多模态大模型，并训练一个基于动态分辨率的文档大模型，实现了一种动态特征压缩算法，使得模型能够根据动态的可学习压缩率来保留所需的特征长度，从而实现了特征的聚合和提取。此外，通过利用语言模型的注意力机制，模型成功地捕捉了与提示词相关的视觉特征部分，通过筛选出最相关的特征并保留其周围相关特征，进一步提高了文档信息处理的准确性和效率。实验中，本文方法在DocVQA、WTQ、ChartQA和TextVQA等数据集上的表现均优于现有模型，分别获得了82.4、37.6、70.8和65.3的分数。未来，将继续优化模型，并探索更多可能的改进方向，以进一步推动文档智能处理技术的发展，以满足日益复杂和多场景的需求。

参考文献（References）

Bai J Z, Bai S, Yang S S, Wang S J, Tan S N, Wang P, Lin J Y, Zhou C and Zhou J R. 2023. Qwen-VL: a frontier large vision-language model for understanding, localization, text reading, and beyond [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2308.12966.pdf

Biten A F, Tito R, Mafla A, Gomez L, Rusiñol M, Jawahar C V, Valveny E and Karatzas D. 2019. Scene text visual question answering// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4291-4301 [DOI: 10.1109/ICCV.2019.00439]

Chen L, Zhao H Z, Liu T Y, Bai S, Lin J Y, Zhou C and Chang B B. 2024. An image is worth 1/2 tokens after layer 2: plug-and-play inference acceleration for large vision-language models//Proceedings of the 18th European Conference on Computer Vision – ECCV 2024. Milan, Italy: Springer: 19-35 [DOI: 10.1007/978-3-031-73004-7_2]

Davis B, Morse B, Price B, Tensmeyer C, Wigington C and Morariu V. 2022. End-to-end document recognition and understanding with dessurt//Proceedings of 2022 European Conference on Computer Vision. Tel Aviv, Israel: Springer: 280-296 [DOI: 10.1007/978-3-031-25069-9_19]

Feng H, Liu Q, Liu H, Tang J Q, Zhou W G, Li H Q and Huang C. 2023a. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. Science China Information Sciences, 67(12): #220106 [DOI: 10.1007/s11432-024-4250-y]

Feng H, Wang Z J, Tang J Q, Lu J H, Zhou W G, Li H Q and Huang C. 2023b. UniDoc: a universal large multimodal model for simultaneous text detection, recognition, spotting and understanding [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2308.11592.pdf

Hu A W, Xu H Y, Ye J B, Yan M, Zhang L, Zhang B, Zhang J, Jin Q, Huang F and Zhou J R. 2024a. mPLUG-DocOwl 1.5: unified structure learning for OCR-free document understanding//Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, USA: Association for Computational Linguistics: 3096-3120 [DOI: 10.18653/v1/2024.findings-emnlp.175]

Hu A W, Xu H Y, Zhang L, Ye J B, Yan M, Zhang J, Jin Q, Huang F and Zhou J R. 2024b. mPLUG-DocOwl2: high-resolution compressing for OCR-free multi-page document understanding [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2409.03420.pdf

Huang Y P, Lv T C, Cui L, Lu Y T and Wei F R. 2022. LayoutLMv3: pre-training for document AI with unified text and image masking// Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: Association for Computing Machinery: 4083-4091 [DOI: 10.1145/3503161.3548112]

Jang E, Gu S X and Poole B. 2016. Categorical reparameterization with gumbel-softmax [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/1611.01144.pdf

Kim G, Hong T, Yim M, Nam J, Park J, Yim J, Hwang W, Yun S, Han S and Park S. 2022. OCR-free document understanding transformer//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 498-517 [DOI: 10.1007/978-3-031-19815-1_29]

Lee K, Joshi M, Turc I, Hu H X, Liu F Y, Eisenschlos J, Khandelwal U, Shaw P, Chang M W and Toutanova K. 2023. Pix2Struct: screenshot parsing as pretraining for visual language understanding//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #780

Lewis D, Agam G, Argamon S, Frieder O, Grossman D and Heard J. 2006. Building a test collection for complex document information processing//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA: Association for Computing Machinery: 665-666 [DOI: 10.1145/1148170.1148307]

Li J N, Li D X, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #814 [DOI: 10.5555/3618408.3619222]

Li Z, Yang B, Liu Q, Ma Z Y, Zhang S, Yang J X, Sun Y B, Liu Y L and Bai X. 2024. Monkey: image resolution and text label are important things for large multi-modal models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26753-26763 [DOI: 10.1109/CVPR52733.2024.02527]

Lin Z N, Wang J P and Jin L W. 2023. Visual information extraction deep learning method: a critical review. Journal of Image and Graphics, 28(8): 2276-2297 (林泽柠, 汪嘉鹏, 金连文. 2023. 视觉信息抽取的深度学习方法综述. 中国图象图形学报,

28(8)：2276-2297)[DOI：10.11834/jig.220904]

Liu C L, Jin L W, Bai X, Li X H and Yin F. 2023. Frontiers of intelligent document analysis and recognition：review and prospects. Journal of Image and Graphics, 28(8)：2223-2252 (刘成林, 金连文, 白翔, 李晓辉, 殷飞. 2023. 文档智能分析与识别前沿：回顾与展望. 中国图象图形学报, 28(8)：2223-2252)[DOI：10.11834/jig.221112]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023a. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans：34892-34916 Curran Associates Inc. #57

Liu Y L, Li Z, Huang M X, Yang B, Yu W W, Li C Y, Yin X C, Liu C L, Jin L W and Bai X. 2023b. OCRBench：on the hidden mystery of OCR in large multimodal models [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2305.07895.pdf

Liu Y L, Yang B, Liu Q, Li Z, Ma Z Y, Zhang S and Bai X. 2024. TextMonkey：an OCR-free large multimodal model for understanding document [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2403.04473.pdf

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer：hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada：IEEE：9992-10002 [DOI：10.1109/ICCV48922.2021.00986]

Long S B, Qin S Y, Panteleev D, Bissacco A, Fujii Y and Raptis M. 2022. Towards end-to-end unified scene text detection and layout analysis//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA：IEEE：1039-1049 [DOI：10.1109/CVPR52688.2022.00112]

Luo C W, Shen Y F, Zhu Z Q, Zheng Q, Yu Z and Yao C. 2024. LayoutLLM：layout instruction tuning with large language models for document understanding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA：IEEE：15630-15640 [DOI：10.1109/CVPR52733.2024.01480]

Masry A, Do X L, Tan J Q, Joty S and Hoque E. 2022. ChartQA：a benchmark for question answering about charts with visual and logical reasoning//Findings of the Association for Computational Linguistics：ACL 2022. Dublin, Ireland：Association for Computational Linguistics：2263-2279 [DOI：10.18653/v1/2022.findings-acl.177]

Mathew M, Bagal V, Tito R, Karatzas D, Valveny E and Jawahar C V. 2022. InfographicVQA//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA：IEEE：2582-2591 [DOI：10.1109/WACV51458.2022.00264]

Mathew M, Karatzas D and Jawahar C V. 2021. DocVQA：a dataset for VQA on document images//Proceedings of 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA：IEEE：2199-2208 [DOI：10.1109/WACV48630.2021.00225]

Nayef N, Patel Y, Busta M, Chowdhury P N, Karatzas D, Khlif W, Matas J, Pal U, Burie J C, Liu C L and Ogier J M. 2019. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia：1582-1587 [DOI：10.1109/ICDAR.2019.00254]

Pasupat P and Liang P. 2015. Compositional semantic parsing on semi-structured tables//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). Beijing, China：Association for Computational Linguistics：1470-1480 [DOI：10.3115/v1/P15-1142]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]：PMLR：8748-8763

Singh A, Natarajan V, Shah M, Jiang Y, Chen X L, Batra D, Parikh D and Rohrbach M. 2019. Towards VQA models that can read//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA：IEEE：8317-8326 [DOI：10.1109/cvpr.2019.00851]

Singh A, Pang G, Toh M, Huang J, Galuba W and Hassner T. 2021. TextOCR：towards large-scale end-to-end reasoning for arbitrary-shaped scene text//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA：IEEE：8798-8808 [DOI：10.1109/CVPR46437.2021.00869]

Stanisławek T, Graliński F, Wróblewska A, Lipiński D, Kaliska A, Rosalska P, Topolski B and Biecek P. 2021. Kleister：key information extraction datasets involving long documents with complex layouts//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland：Springer：564-579 [DOI：10.1007/978-3-030-86549-8_36]

Svetlichnaya S. 2020. DeepForm：understand structured documents at scale, 2020 [EB/OL]. [2024-10-01]. https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--VmlldzoyODQ3Njg

Tanaka R, Nishida K and Yoshida S. 2021. VisualMRC：machine reading comprehension on document images//Proceedings of 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania：AAAI：13878-13888 [DOI：10.1609/aaai.v35i15.17635]

Tang Z N, Yang Z Y, Wang G X, Fang Y W, Liu Y, Zhu C G, Zeng M, Zhang C and Bansal M. 2023. Unifying vision, text, and layout for universal document processing//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada：IEEE：19254-19264 [DOI：10.1109/CVPR52729.2023.01845]

Veit A, Matera T, Neumann L, Matas J and Belongie S. 2016. COCO-text：dataset and benchmark for text detection and recognition in

natural images [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/1601.0714.pdf

Wang D S, Raman N, Sibue M, Ma Z Q, Babkin P, Kaur S, Pei Y L, Nourbakhsh A and Liu X M. 2023. DocLLM: a layout-aware generative language model for multimodal document understanding//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics: 8529-8548 [DOI: 10.18653/v1/2024.acl-long.463]

Wei H R, Kong L Y, Chen J Y, Zhao L, Ge Z, Yang J R, Sun J J, Han C R and Zhang X Y. 2023. Vary: scaling up the vision vocabulary for large vision-language models//Proceedings of the 18th European Conference on Computer Vision – ECCV 2024. Milan, Italy: Springer [DOI: 10.1007/978-3-031-73235-5_23]

Xu Y, Xu Y H, Lv T C, Cui L, Wei F R, Wang G X, Lu Y J, Florencio D, Zhang C, Che W X, Zhang M and Zhou L D. 2021. LayoutLMv2: multi-modal pre-training for visually-rich document understanding//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [s.l.]: Association for Computational Linguistics: 2579-2591 [DOI: 10.18653/v1/2021.acl-long.201]

Xu Y H, Li M H, Cui L, Huang S H, Wei F R and Zhou M. 2020. LayoutLM: pre-training of text and layout for document image understanding//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [s.l.]: Association for Computing Machinery: 1192-1200 [DOI: 10.1145/3394486.3403172]

Ye J B, Hu A W, Xu H Y, Ye Q H, Yan M, Dan Y H, Zhao L C, Xu G H, Li C L, Tian J F, Qi Q, Zhang J and Huang F. 2023a. mPLUG-DocOwl: modularized multimodal large language model for document understanding [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2307.02499.pdf

Ye J B, Hu A W, Xu H Y, Ye Q H, Yan M, Xu G H, Li C L, Tian J F, Qian Q, Zhang J, Jin Q, He L, Lin X and Huang F. 2023b. UReader: universal OCR-free visually-situated language understanding with multimodal large language model//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: Association for Computational Linguistics: 2841-2858 [DOI: 10.18653/v1/2023.findings-emnlp.187]

Ying Z L, Zhao Y H, Xuan C and Deng W B. 2020. Layout analysis of document images based on multifeature fusion. Journal of Image and Graphics, 25(2): 311-320 (应自炉, 赵毅鸿, 宣晨, 邓文博. 2020. 多特征融合的文档图像版面分析. 中国图象图形学报, 25(2): 311-320) [DOI: 10.11834/jig.190190]

Yu Y Q, Liao M H, Wu J H, Liao Y X, Zheng X Y and Zeng W. 2024. Texthawk: exploring efficient fine-grained perception of multimodal large language models [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2404.09204.pdf

Zhang Y Z, Zhang R Y, Gu J X, Zhou Y F, Lipka N, Yang D Y and Sun T. 2023. LLaVAR: enhanced visual instruction tuning for text-rich image understanding [EB/OL]. [2024-10-01]. https://arxiv.org/pdf/2306.17107.pdf

作者简介

杨彪,男,硕士研究生,主要研究方向为计算机视觉。
E-mail: hust_byang@hust.edu.cn
朱盈盈,通信作者,女,工程师,主要研究方向为计算机视觉和机器学习。E-mail: yyzhu@hust.edu.cn
刘禹良,男,研究员,主要研究方向为计算机视觉。
E-mail: ylliu@hust.edu.cn
刘强,男,工程师,主要研究方向为计算机视觉和自然语言处理。E-mail: liuqiang2@wps.cn