arXiv:2305.01054v1 [cs.DB] 1 May 2023

CHIC：用于视觉问答的企业文档

Ibrahim SOULEIMAN MAHAMOUD[1, 2*]、Micka¨el Coustaty[2,1]、Aur´elie Joseph[1,2]、VincentPoulain d'Andecy[1,2] 和 Jean-Marc Ogier[1,2]

[1*]L3i，拉罗谢尔大学，Michel Cr´epeau 大道，拉罗谢尔，17000，法国。[2]研究团队，YOOZ，Fleming 街 1 号，拉罗谢尔，17000，法国。

*通讯作者。电子邮箱：ibrahim.souleiman mahamoud@univ-lr.fr；贡献作者：mickael.coustaty@univ-lr.fr；aurelie.joseph@getyooz.com；vincent.poulaindandecy@getyooz.com；jean-marc.ogier@univ-lr.fr；

摘要

无纸化倡议日益普及，数字文档被大量使用，这促使一些公司寻求自动处理每日数千份文档的方法。为此，他们采用自动信息检索（信息检索），以便从大型数据集中快速提取有用信息。要实现有效的信息抽取方法，首先需要合适的数据集。尽管公司拥有足以满足自身需求的数据，但也需要一个公共数据库，以便比较最先进方法之间的成果。现有文档公开数据包括 DocVQA[2]和 XFUND [10]，，但它们无法充分满足公司的需求。XFUND 仅包含表单文档，而公司会使用多种类型的文档（即表单等结构化文档，也包括发票等半结构化文档，以及电子邮件等非结构化文档）。与 XFUND 相比，DocVQA 涵盖多种类型的文档，但其中只有 4.5% 是企业文档（如发票、采购订单等）。这 4.5% 的文档无法满足公司对文档多样性的需求。我们提出了 CHIC，这是一个面向公众的视觉问答数据集。该数据集涵盖不同类型的企业文档，从这些文档中提取的信息能够满足公司的实际需求。

关键词：数据集、视觉问答、多模态

1 引言

私营企业每天都要处理大量文档。这些文档来自各种来源（例如公司内部流程、分包商的文档以及行政文档）。如果不使用自动化系统，处理如此大量的文档需要投入大量精力和人力。因此，YOOZ、Esker 等公司致力于提供自动文档处理解决方案。这些解决方案利用信息检索（IR），从大型数据库中的文档提取有用信息并进行分类。这使公司能够节省时间和资金。为了提供成功的信息检索解决方案，首先必须验证其是否符合行业的要求。

这些工业约束对应着实际需求，针对这些约束提供解决方案将使最终用户能够高效地使用这些方案。

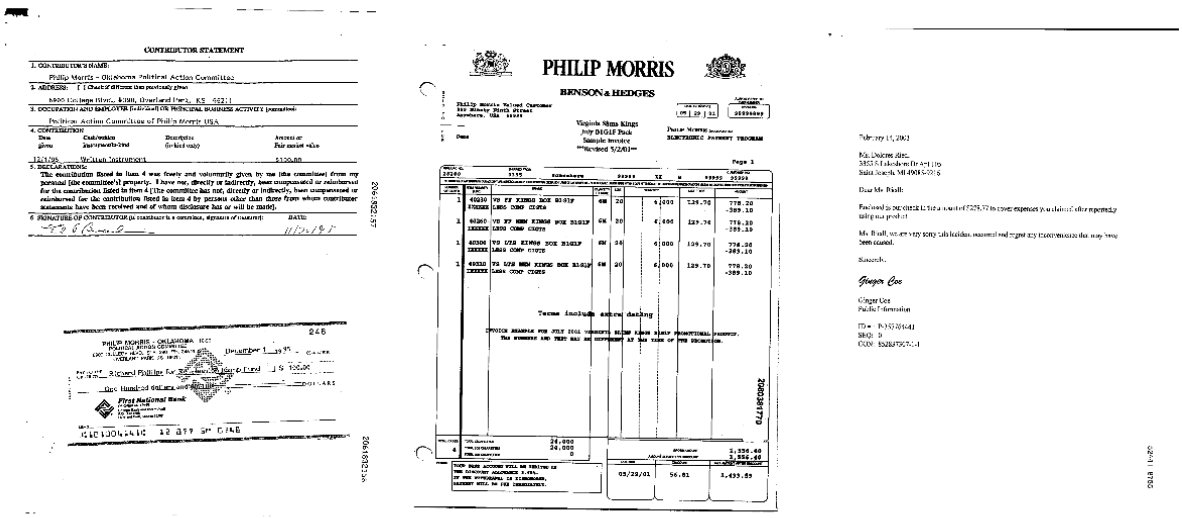


图 1：我们在此情境中处理的不同类型的文档

在这些约束中，例如可以提到与文档处理速度相关的约束，处理时间必须控制在几秒以内。这样终端用户就能快速获得结果，因为如果自动处理一份文档需要几分钟，人工处理也要花费同样的时间，这样就没有效率了。这些信息抽取解决方案必须快速，同时保持良好的实际性能；不能为了节省处理时间而牺牲结果质量。尽可能获得准确的结果也很重要，因为人工纠正这些错误需要付出更多精力和时间。信息检索解决方案应在达到较高置信度时提供结果，否则应提示终端用户处理该文档。对于某些情况，人工标注比逐一检查提取的信息、查找错误并加以纠正更快。满足所有这些约束后，还需要有适当的公开数据，以便比较最先进模型的性能，并通过指标判断这些约束是否得到满足。

公开数据对于信息抽取方法非常有用，因为它可以比较不同方法之间的结果，还可以了解这些方法在这类数据上的局限性。这些数据具有两个特征：第一是文档的类型，第二是元素 to 被提取。提供信息检索解决方案的公司通常专注于一种文档类型，因为处理任何类型的文档并提取任何类型的信息都需要多个平台、多种方法以及大量工作。在本文中，我们将重点关注企业文档，例如发票、采购订单等；我们已在图1中展示了这些不同的文档。我们可以将它们分为三类。结构化文档（例如身份证件、支票、表格）的特点是文档版式具有高度相似性，因为同一类型的文档始终采用相同的版式。半结构化文档（例如发票和采购订单）可以在不同文档之间改变结构，但会遵循构成这些文档的元素的特定组织方式。最后，非结构化文档（如电子邮件）不遵循预定义版式，彼此之间的差异也很大。下文中，我们将把上述文档集合称为企业文档。除了文档类型不同之外，这些公司还面临多种语言的文档。这种多语言特性是必要的，因为提供信息检索解决方案的公司与来自不同国家、使用不同语言的多家公司合作。我们将重点关注

欧洲语言，例如英语、德语、意大利语、法语和西班牙语。语言的选择对应于这些公司处理的绝大多数文档。这些文档中需要为公司提取的相关信息多种多样。为了涵盖所有需要提取的信息类型，我们采用了视觉问答（VQA）方法。这些方法的优点在于能够通过问题提取大量信息，并且终端用户易于使用。

在最先进技术领域，我们可以看到 DocVQA 和 XFUND 等数据集。我们将在相关工作部分2中说明它们（DocVQA、XFUND）的局限性，以及 C HIC 相较于这些最先进的数据集有何贡献。

CHIC 是一个新的视觉问答数据集，能够满足这些公司的需求，并为它们提供包含日常使用文档的数据集，且提取的信息与预期信息相似。

我们的贡献总结如下：

- 我们提出了一个新数据集，其中包含 774 份公司文档和 3384 个带标注的问答。
- CHIC 是首个包含多语言文档的企业数据集。文档语言包括英语、德语、法语、意大利语和西班牙语。
- CHIC 不仅是适用于这些文档的强大数据集，也适用于其中的问题和答案。该数据集考虑了这些公司在问题多样性方面的需求，以便提取任何类型的相关信息。

2 相关工作

在本节中，我们将介绍现有的最先进数据。我们将首先逐一介绍这些数据，然后结合我们的研究背景说明其局限性。

2.1 DocVQA

DocVQA 是首个用于企业文档问题与答案的数据集。DocVQA 是一个包含 12 767 份不同类型和版式文档的数据集。这些文档包括手写文档、表单、广告等。要从中提取信息在该文档中，他们标注了50,000个问题和答案。该数据集非常适合从任何类型的文档中提取任何类型的信息。尽管 DocVQA 在可用问题和答案的数量以及文档方面做出了重大贡献，但其在工业环境中的应用仍然存在局限性。

DocVQA 的局限性可归纳为两点。第一，DocVQA 中的企业文档（例如发票、采购订单等）非常少。为得出这一结论，我们检索了符合我们场景的文档类型。为了提高精确度，我们手动提取了 DocVQA 中的企业文档；完成这项工作耗费了 20 人时。我们发现有 586 份文档可用于企业场景，并将这一子集称为 DocVQA-Corporate。尽管 DocVQA-Corporate 仅占 DocVQA 所含全部文档的 4.5%，但数量实在太少。另一方面，这一数量足以构建一个关于企业文档的问题与答案数据集。然而，在分析 DocVQA-Corporate 中的文档类型后，我们发现它们并不满足在企业文档场景下提供公开数据集所需的所有要求。DocVQA-Corporate 中的文档版式类型几乎都相同，因此文档间多样性不足。我们发现，其中大约有 20 种不同的文档布局，并且这些版式在 DocVQA-Corporate 中多次重复。这一观察使我们得出结论：DocVQA-Corporate 中的文档远不足以构成适用于工业场景的数据集，因为它没有考虑到企业文档输入流中可能出现的多样性需求。第二点涉及从 DocVQA-Corporate 中提取的信息。当我们检查从 DocVQA-Corporate 检索到的响应类型时，发现检索到的信息主要是字符串（占响应的 40%）。虽然字符串属于公司所需的信息，但在这一类响应中的占比仍不足 20%（即审视客户对信息抽取的期望时）。DocVQA-Corporate 中这些字符串响应

对公司而言几乎毫无用处（也就是说，在我们的案例中，这些字符串响应只有 38% 是有用的）。其余问题则与公司的期望相符（也就是说，我们能找到金额、参考编号等信息）。除这两项主要问题外，还存在其他局限性。首先，DocVQA 文档只有英语版本，而在工业场景中，经常会遇到多语言文档。其次，DocVQA 中标注的答案只包含文本，没有标明答案的精确位置。精确定位答案有助于算法学习答案的正确位置，而不是简单地返回找到的第一个匹配项。这一点很有必要，因为很多情况下，文档中的多个位置都可能出现答案，而算法通常会将找到的第一个答案选为正确答案并据此学习（例如，10 可能是总金额，也可能是日期中的数字，或对应其他数值）。上述局限性表明，DocVQA 和 DocVQA-Corporate 都无法完全满足工业场景下构建公开数据集的所有要求。

2.2 XFUND

XFUND 是一个包含 1393 份文档表单（结构化文档）的数据集，这些文档表单包含由人类标注的键值对。在该数据集中，键值对的含义例如：如果文档中有“出生日期：2023 年 1 月 13 日”，那么“出生日期”将被标注为键，并与值“2023 年 1 月 13 日”关联。1393 份文档中约有 30,000 个经过标注的键值对。这些键值对从 7 种语言（中文、日语、西班牙语、法语、意大利语、德语和葡萄牙语）的多份文档中提取。尽管 XFUND 确实作出了贡献，但它存在一些局限性，使我们无法在企业文档环境中使用它

我们可以先从文档类型以及该数据集中标注的键值信息两方面总结 XFUND 的局限性。第一点涉及 XFUND 文档：这些文档都是结构化文档。顾名思义，这类结构化文档中的词语通常以相同的方式组织（即以键值形式组织）。尽管该公司处理表单文档，也会处理其他类型的文档（如引言中所述），包括半结构化文档和非结构化文档。为了符合其日常工作流程中遇到的文档类型，提供具有多种结构类型的文档是有意义的。遗憾的是，XFUND 未能满足最终用户对公共数据集多样性的要求。第二点涉及 XFUND 中提取的键值信息。在 XFUND 的标注中，一个词要么是键，要么是值，要么是标题，其他所有内容都归为“其他”。在这些标注为“其他”的词语中，有许多可能对公司有用（例如，表格顶部没有对应键的表格编号）。遗憾的是，如果公司想用 XFUND 提取某些“其他”词语，就必须投入时间为这些词语添加标注。这些局限性使 XFUND 无法成为企业数据集，而只是一个满足特定需求的表单数据集。

2.3 RVL-CDIP

RVL-CDIP 是一个包含 400,000 份文档的数据集，这些文档被分为 15 个类别（例如信件、广告、文件夹、表单、电子邮件、手写文档和发票）。在标注的类别中，有些类别反映了公司经常处理的文档类型，但也包括 RVL-CDIP 中存在的其他文档类型。尽管该数据集包含企业文档，但仍存在一个局限性。

RVL-CDIP 旨在用于文档分类。尽管文档分类对公司而言是一项挑战，但最先进技术拥有大量公开数据来应对这一挑战（[1],[14]）。在信息检索中，首先检索到能反映我们所关注情境的文档当然很重要，但从这些文档中提取有用信息也同样重要。遗憾的是，这一限制意味着这些文档无法以当前状态使用，必须经过标注才能投入使用。

3 CHIC

3.1 问题与答案问题的表述

在最先进技术中，我们发现了若干与问答相关的问题，例如自然场景图像上的问答 [15]、文本上的问答 [13] 以及视频上的问答 [12]。本节将介绍在企业文档情境下预测答案的问题表述。我们将重点说明如何根据问题**a** 和问题**q** ，从文档**d** 中提取信息并给出答案，方法依据模型参数 a（见公式1）。接下来，我们将描述如何表示q， d, a。

问题 $\mathbf{q}$ 是一个由若干词组成的序列，即 $q = [q_1, \ldots, q_n]$，其中每个 q_t 都是第t个问题词，q_n ='?'标记问题的结束。问题之间存在上下文关联，在本例中，关联的上下文就是文档。在我们的场景中，我们使用文档的图像 i 和文本 t d。如果原始文档是 PDF，我们会将其转换为图像，并使用光学字符识别（OCR）工具从图像中提取文本。答案是从文档文本中提取的文本，一个答案可以由一个或多个词组成。除了文本之外，答复还包含其在图像中的精确位置。因此，一个答案由若干词和一个边界框 $a =$ （$[a_1, ..., a_2]$, $bounding\ box$）组成。

$$\hat{a} = \arg\max_{a \in \mathcal{A}} P(a|x, q;\ \theta) \qquad (1)$$

定义了问题、文档和答案之后，下面我们将介绍 CHIC。

3.2 新 CHIC 数据集概述

CHIC 数据集包含 774 份使用 5 种语言（英语、法语、德语、西班牙语和意大利语）的多语言文档。更准确地说，我们有 774 张图像，以及从这些图像中提取的文本。我们借助 Abby FineReaderOCR 提取文本。CHIC 中的图像分辨率很高，为 2560 * 3332 像素。CHIC 尽可能模拟了公司中可能出现的不同类型文档，其中包含结构化、半结构化和非结构化文档。随后，我们针对这些文档标注了 3384 个问题及答案。下文将介绍我们标注 CHIC 的步骤。

3.3 标注协议

为了构建符合公司要求的文档，我们首先需要找到与这一背景相符的文档。遗憾的是，现有最先进技术中的 XFUND 数据集仅包含一种类别的文档，但这些文档是多语言的；而 RVL-CDIP 则包含种类繁多的文档。要找到合适的文档，就需要充分利用这两类数据集的优势，构建我们的新数据集。我们使用的文档提取自两个公共数据库：RVL-CDIP[1]和 XFUND[10]。使用它们有两方面的优势。首先，RVL-CDIP 能提供半结构化或非结构化文档，且这些文档的版式各不相同，与 DocVQA-Corporate 不同。其次，从 XFUND 提取的文档是多语言表单，而 RVL-CDIP 中仅包含英语文档。总计，我们收集了 770 份包含 CHIC 的文档，其中 655 份来自 RVL-CDIP，115 份来自 XFUND。下文将重点关注从这些文档中提取的信息。

问题和答案的标注协议如下。首先，我们对问题进行标注，然后标注这些问题的答案，最后检查响应是否符合预期。要标注问题，首先需要了解终端用户需要哪种类型的问题。为此，我们收集了关于需要提取哪类信息以及应使用哪类问题的预期。我们的指示是让问题类型多样化：除了标准问题（例如，当文档中出现“净额”一词时，询问净额是多少）之外，还要添加其他类型的问题。我们设计了一些需要理解文档内容之外信息的问题，例如“表格 I 顶部的金额是多少？”或“文档引用顶部的数字是

多少？”我们还设计了一些需要对文档中的元素进行算术和比较运算的问题，例如“表格中金额最大的元素叫什么？”这些其他类型的问题旨在获取无法通过标准问题提取的信息。对这些问题完成标注后，我们接着标注问题的答案。这些问题采用一种特殊的方式进行标注：我们要求每位标注员标注其他标注员的问题，而不是自己的问题。这样可以让标注员了解他人对自己问题的理解，并从这种理解中提取必要的信息。问题和答案全部完成后，我们要求每位标注员检查问题的答案是否符合自己的预期。如果答案不符合预期，我们就尝试重新表述问题，或一起找出正确答案。所有这些标注都可以通过我们自行开发的工具完成。

为了完成所有这些标注，我们没有找到任何符合需求的标注工具，因此自行开发了一款标注工具。该工具旨在易于使用且直观。我们只纳入了必不可少的功能。其中一部分用于查看文档图像，另一部分用于标注问题及其答案。答案标注既包括答案文本，也包括答案在文档中的位置，即边界框。此标注工具可在 [9]中获取。

我们在图2 中总结了各项必要工作及所需人力。我们总共投入约 137 人时来提供 CHIC。

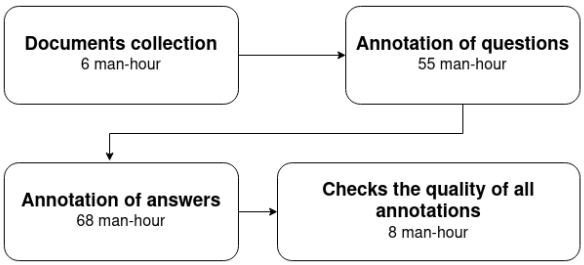


图 2：CHIC 标注的步骤及其所需时间

3.4 统计与分析

在本节中，我们将分析 CHIC 是否满足引言中所述的工业约束。

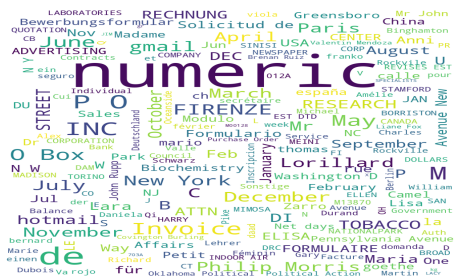
我们首先对 CHIC 文档进行一些分析。我们在该小节中已经看到，收集的文档来自不同的数据库，其结构也符合公司的预期。企业文档在版式和所含词语方面都不同于其他类型的文档。我们观察到，CHIC 中 93% 的文档包含 100 至 400 个词，只有约 30 份文档包含超过 500 个词。将这些统计数据与另一类文档（例如科学文档 [11], ）进行比较，后者每份文档平均包含 1500 个词。这表明，CHIC 中的企业文档所包含的词语尽管平均只有 150 个，却分布在整份文档中，而非仅仅以词块的形式排列（如报纸文章、科学论文等）。接下来，我们将详细分析这些词语，看看哪些词语被提取出来，以及它们有何用处。

图??展示了文档中提取的词语和问题中的词语分布。首先，我们来看问题中的词语。CHIC 中的问题大多与日期、金额或其他数字有关，这些都是终端用户的主要需求所涉及的内容。这些要素对公司至关重要，因为它们是文档中的关键信息。能够提取这些信息会带来诸多好处。例如，可以自动查找日期或金额准确的发票。此外，还可以将金额相关信息提供给会计师，以便其轻松开展会计工作。这些问题还有许多其他用途。通过这些问题，我们将从文档中提取答案。

在分析答案的分布时，词云未考虑数字和字母数字答案，因为这类答案彼此差异很大。为了呈现这类答案，我们将所有此类答案都转换为“数字”词。观察散点图可以发现，大多数



(a) 问题词云



（b）答案词云**图3**

答案都是数字或字母数字形式；答案中还有文档类型、公司名称（当然也包括这些文档的供应商名称）等。

总而言之，在 CHIC 中，我们拥有各种不同类型的问题，而这些问题的答案符合终端用户的预期。

4 实验

在本节中，我们将介绍所选用的方法，以提供一个 CHIC 基线；随后解释所使用的指标，最后分析所得结果。

4.1 视觉问答模型

为了评估当前最先进方法在 CHIC 数据集上的性能，我们选取了两类方法作为基线：预训练方法和非预训练方法。大多数方法可分为两类：一类是通过无监督方式在数据上进行预训练，以尽可能获取先验知识（即受到人类学习机制的启发；人类学习新知识时总是依赖已获得的先验知识）。其他方法不使用这些先验知识，但非常善于充分利用当前数据，尽可能提取知识，以取得良好的性能。对于预训练方法，我们选择了 Donut [4]，LayoutLMv3 [3],和 LayoutLMv2 [5]。对于非预训练方法，我们选择了 QALayout[7]和 QANet [6]。这两类方法将帮助我们了解，在 CHIC 上使用预训练方法或非预训练方法有多重要。

基于 LayoutLM 的模型是预训练方法，例如 LayoutLMv2[5] 或 LayoutLMv3[3]。微软提出这些方法，用于多种文档理解任务（文档分类、问答等）。其预训练基于文本、图像和版面特征（即文档中文字的位置）。预训练的目标是隐藏文档中的特定部分，并在上下文中找出该部分。LayoutLMv2 与 LayoutLMv3 的主要区别在于掩码方法和所使用的输入特征。我们先介绍 LayoutLMv2。模型的输入包括从图像中提取的文本、图像特征，以及与文档中文字位置相关的版面信息。所用图像特征是将图像分成 4 个部分；每个部分仅包含图像的四分之一。特征与文字在文档中的位置相关联，并使用顺序位置（即文字相对于通过光学字符识别提取的文本的位置）。每个词的边界框特征既对应文字的位置，也对应图像各部分的位置。LayoutLMv2 的预训练目标有三个：掩码视觉语言建模（MVLM）、文本-图像对齐（TIA）和文本-图像匹配（TIM）。LayoutLMv3 [3] 使用的特征与 LayoutLMv2 大致相同，二者的区别在于图像特征。LayoutLMv3 不会将图像分成 4 部分，而是将其划分为一系列大小均匀的 P × P 图像块（即预训练模型中的 P 为 16）。LayoutLMv3 的预训练目标包括掩码视觉语言建模（MVLM）、掩码图像建模（MIM）和词块对齐（WPA）。

与基于 LayoutLM 的方法相比，Donut[4]仅基于图像。它不像大多数模型那样使用通过光学字符识别提取的信息，而是仅基于

图像。作者展示了该方法在正确答案检测和文本提取方面的效率，并将其与传统光学字符识别进行了比较。传统方法在识别手写词语或不够清晰的文字时往往会出错。为取得这些成果，他们首先让模型学习“如何阅读”，并为此使用了数百万份文档。他们从左上到右下解析文档，目标是根据先前的上下文，最小化下一个词元预测的交叉熵。学会阅读后，他们又学习了如何理解文档。为此，他们使用了分类、文档信息抽取和文档视觉问答等任务。

为了比较预训练方法与未预训练模型的能力，我们采用了 QANet [6] 和Qalayout [7]。

QANet[6] 是一种仅基于文本的问题答案模型。它表明，将注意力机制与卷积模型结合使用，可以提升运行速度，并获得优于循环神经网络（RNNs）的性能。

QALayout 是对 QANet 的改进。除了使用 QANet 的特征外，它还使用图像和版面信息等其他特征。此外，它还使用协同注意力机制，以便更好地在多种模态上运用注意力机制。

4.2 评估指标

为了评估模型的性能，我们选择了一些指标：精确匹配、ANLS、F1 和 IOU。我们可以将这些指标分为两类：严格指标不允许答案中出现任何错误（例如，精确匹配得分），而宽松指标则会考虑答案中的错误，例如遗漏某个单词等。我们选择这些指标，是因为它们各自满足特定的评估需求，并且结合起来能够全面反映各种方法的性能。

平均归一化莱文斯坦距离（ANLS）衡量两个字符（q_k，p_k）之间的距离，见公式2。 NL（q_k，p_k）是真实值（GT）与预测之间的归一化莱文斯坦距离。然后设置阈值 $\tau = 0.5$ ，以过滤大于 τ 的 NL 值，并返回一个分数 0。

$$ANLS = \max_{1..3} s(q_k, p_k)$$
$$s(q_k, p_k) = \begin{cases} (1 - NL(q_k, p_k) & \textbf{if} \quad NL(q_k, p_k) < \tau \\ 0 & \textbf{if} \quad NL(q_k, p_k) >= \tau \end{cases} \tag{2}$$

$$precision = \frac{1 * same_word}{tail(p_k)} \qquad recall = \frac{1 * same_word}{tail(q_k)}$$
$$f1 = \frac{2 * precision * recall}{(precision + recall)} \tag{3}$$

指标 F1 分数的说明见 3。其中，same word 统计真实值与预测之间相同词语的数量。

精确匹配分数只有 0 或 100：如果预测答案与预期答案完全一致，则得 100 分，否则得 0 分。

4.3 CHIC 实验

在表格中，我们列出了在 CHIC 和 DocVQA 数据集上测试的基线方法。首先，我们展示了所使用方法的名称及其参数数量，然后列出了这些方法分别在图形处理器和中央处理器上的执行时间（以秒为单位），即说明了这些实验所使用的机器类型。最后，我们给出了这些方法在 CHIC 数据集和 DocVQA 上的各项指标。

我们首先从时间性能开始分析。在工业场景中，找到执行时间与性能之间的恰当平衡至关重要。为了获得这一指标，我们计算了每种方法在同一台机器的图形处理器和中央处理器上的平均响应时间。我们选择图形处理器和中央处理器，是因为在推理时，一些公司可能会使用中央处理器运行这些方法。 α 表示光学字符识别提取图像中文本所需的执行时间。该方法需要文本才能提取信息，因此会先使用光学字符识别提取文本，然后利用该文本和其他特征来预测答案。 α 参数因所用光学字符识别工具的类型而异。在我们的场景中，我们使用了 Abby FineReader，平均而言，从一张图像中提取文本需要 2 秒。时间性能结果表明，速度最快的方法是 QANet，最慢的是 Donut。

方法	预训练	参数	时间（秒）		CHIC			DocVQA
			GPU	CPU	F1	EM	ANLS	ANLS
人类			24		83.78	79.14	90.25	98.1
Donut	Yes	176 M	1.23	22.60	48.19	44.8	53.36	71.66
LayoutLMv3	Yes	133M	0.39 $+\alpha$	1.94 $+\alpha$	63.16	59.85	68.01	78.35
LayoutLMv2	Yes	200M	0.21 $+\alpha$	1.65 $+\alpha$	65.67	61.50	69.69	73.41
QALayout	No	10M	0.25 $+\alpha$	1.80 $+\alpha$	59.05	46.27	63.19	53.60
QANet	No	2M	0.04 $+\alpha$	0.16 $+\alpha$	36.98	26.08	39.12	48.90

表格 1：此表介绍了基线方法、参数数量以及每个问题的平均处理时间。α 是光学字符识别从图像中提取文本所需的平均时间。在我们的场景中，我们使用了 Abby OCR，处理时间约为 2 秒。

在 CHIC 数据集上，兼顾时间性能和正确答案数量的方法是 LayoutLMv2。

使用这些方法分析 CHIC 上反复出现的错误时，我们发现它们可以归为三类主要错误。第一类与光学字符识别（OCR）造成的错误有关。这类错误又可分为两种：第一种是由于 OCR 未能正确识别某个词而导致的错误提取。另一种错误则与 OCR 提取词语的顺序有关，这一顺序会对预测产生重大影响。例如，如果文档中有“日期：12/02/2003”这一词组，OCR 有时可能会先识别出“日期”，再识别出中间隔着其他内容的一个词，等等，最终的顺序可能是“日期 F72535 01/01/2004 358 12/02/2003”。最后，系统会选择日期一词旁边的词，因此有时会选错日期，或者选中其他数字。另一类错误是将答案与其他值混淆，因为它们旁边有相同的关键词搜索词。例如，如果问题是“工作编号是多少？”，在某些情况下，由于预测答案是“工作”这一关键词后面的第一个词，系统可能会选择“工作参考编号 AF12321”，而“工作编号”可能出现在文本中稍靠后的位置。

4.4 实验设置

所有实验均在同一台机器上完成。该机器配有 4 块 48GB 的 NVIDIA RTX A6000 图形处理器、256 GB 内存和 4 TB SSD。每项实验均在单块图形处理器上进行。我们测试的方法具有不同的参数规模和内存使用需求。针对这一点，我们为 QANet 和 QALayout 方法选择了 32 的批量大小，然后为 LayoutLMv3 和 LayoutLMv2 方法选择了 16 的批量大小，最后为 Donut 选择了 1 的批量大小。训练这些方法时，我们使用相同的 Adam 优化器，初始学习率为 1e-3，并在性能停滞时逐渐降低学习率。我们将训练轮数设为 200，并设置了 5 轮的提前停止。

5 结论

我们提出了一个用于企业文档视觉问答的新型公开数据集。为了构建该数据集，我们充分考虑了企业级数据集所需满足的所有要求。CHIC 包含 774 份文档，涵盖多种语言和不同的版式（即半结构化、非结构化和结构化文档）。我们从这些文档中标注了 3384 个不同语言的问题及其答案。这些问题也具有多样性，包括需要理解数值之间关联的问题（例如，文档中的最大金额是多少？）、需要理解文档中不同词语之间关联的问题（例如，表格 I 第 3 行中的金额是多少？），以及涉及键值理解的其他问题（例如，jobnumber 是什么？）。该数据集将有助于比较不同模型在企业文档视觉问答任务上的表现。

致谢

本研究由 LabCom IDEAS 资助，项目编号为 ANR-18-LCV3-0008；同时获得法国 ANRT 机构（CIFRE 项目）和 YOOZ 公司的资助。

参考文献

[1] Adam W Harley、Alex Ufkes 和 Konstanti- nos G Derpanis。在*2015*年 第*13*届国际文档分析与识别会议（*ICDAR*）上发表《用于文档图像分类和检索的深度卷积网络评估》，第 991–995 页。IEEE，2015年。

[2] Minesh Mathew、Dimosthenis Karatzas 和 C. V. Jawahar。DocVQA：面向文档图像视觉问答的数据集，2020年。

[3] Yupan Huang、Tengchao Lv、Lei Cui、Yutong Lu 和 Furu Wei。LayoutLMv3：通过统一的文本与图像掩码进行文档人工智能预训练，2022年。

[4] Geewook Kim、Teakgyu Hong、MoonbinYim、Jeongyeon Nam、Jinyoung Park、Jinyeong Yim、Wonseok Hwang、Sangdoo Yun、Dongy- oon Han 和 Seunghyun Park。无OCR文档理解变换器，2021年。

[5] Yang Xu、Yiheng Xu、Tengchao Lv、Lei Cui、Furu Wei、Guoxin Wang、Yijuan Lu、Dinei Florencio、Cha Zhang、Wanxiang Che、Min Zhang 和 Lidong Zhou。LayoutLMv2：面向富视觉文档理解的多模态预训练，2020年。

[6] Adams Wei Yu、David Dohan、Minh-Thang Luong、Rui Zhao、Kai Chen、Mohammad Norouzi 和 Quoc V.Le。《QANet：结合局部卷积与全局自注意力进行阅读 g comp理解 ，2018年。

[7]Ibrahim SOULEIMAN MAHAMOUD、Micka¨el Coustaty、Aur´elie Joseph、Vincent Poulain d’Andecy 和 Jean-Marc Ogier。《QALayout：基于多模态注意力的问答版面，用于企业文档上的视觉问答》。载于 Seiichi Uchida、Elisa Barney 和 V´eronique Eglin 编，《文档分析系统》，第 659–673 页，Cham，2022年。Springer International Publishing。

[8] Sai Yelagandula 和 Sasikanth Mandati。

[9]Ibrahim SOULEIMAN MAHAMOUD, Micka¨el Coustaty, ur´elie Joseph, Vincent Poulain d’Andecy 和 Jean-Marc Ogier。CHIC 数据的公开访问：https://gitlab.com/contributions3/data/chic

[10]Yiheng Xu、Tengchao Lv、Lei Cui、Guoxin Wang、Yijuan Lu、Dinei Florencio、Cha Zhang 和 Furu Wei。《XFUND：用于多语言视觉丰富表单理解的基准数据集》。载于 计算语言学协会会议论文集：*ACL 2022*年，第 3214–3224 页，都柏林，爱尔兰，2022年5月。计算语言学协会。

[11] Sara Lafia、Werner Kuhn、Kelly Caylor 和 Libby Hemphill。多层次细节的研究主题映射。*Patterns*，2:100210，2021年02月。

[12] Yaoyao Zhong、Junbin Xiao、Wei Ji、Yicong Li、Weihong Deng 和 Tat-Seng Chua。视频问答：数据集、算法与挑战，2022年。

[13] Zahra Abbasiantaeb 和 Saeedeh Momtazi。从信息检索和深度神经网络视角探讨基于文本的问答：一篇综述，2020年。

[14]Lulu Wan、George Papageorgiou、Michael Seddon 和 Mirko Bernardoni。长篇法律文档分类，2019年。

[15]Ankan Bansal、Yuting Zhang 和 Rama Chellappa。图像集上的视觉问答，2020年。