TabComp: A Dataset for Visual Table Reading Comprehension

Somraj Gautam, Abhishek Bhandari, Gaurav Harit
Indian Institute of Technology Jodhpur
{gautam.8,bhandari.1,gharit}@iitj.ac.in

Abstract

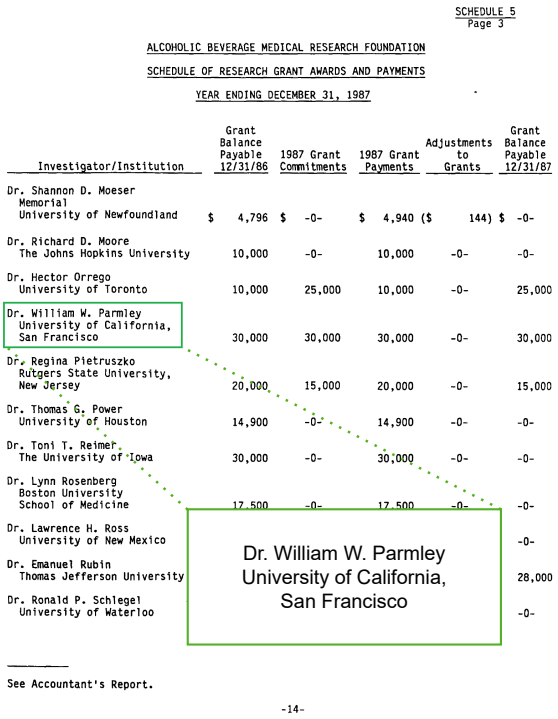
Reaching a human-level understanding of real-world documents necessitates effective machine reading comprehension, yet recent developments in this area often struggle with table images. In response, we introduce the Visual Table Reading Comprehension (TabComp) dataset, which includes table images, questions, and generative answers designed to evaluate OCR-free models. Unlike general Visual Question Answering (VQA) datasets, TabComp uniquely focuses on table images, fostering the development of systems which obviate the use of optical character recognition (OCR) technology, which often struggles with complex table layouts. Our findings reveal that current OCR-free models perform poorly on TabComp, highlighting the need for robust, specialized models for accurate table reading comprehension. We propose TabComp as a benchmark for evaluating OCR-free models in table reading comprehension and encourage the research community to collaborate on developing more effective solutions. The code and data are available at - https://github.com/dialabiitj/TabComp/

1 Introduction

The ability to automatically read and comprehend textual and visual information from documents has become increasingly critical as vast amounts of structured data are embedded within document images, particularly in the form of tables. However, extracting and interpreting information from table images poses significant challenges, as traditional text-processing systems are often ill-equipped to handle the complexities of table layouts, formatting, and multi-modal content. This challenge emphasizes the need for table reading comprehension, which not only reads the content but also interprets and answers questions from tabular data.

Table reading comprehension is a generative Machine Reading Comprehension (MRC) task that integrates natural language understanding (NLU) and natural language generation (NLG) capabilities. However, extracting relevant information from tables remains challenging, especially when dealing with scanned or photographed documents where Optical Character Recognition (OCR) may not be reliable.

In recent years, there has been a growing interest in developing OCR-free models that can perform downstream tasks in document images. These models have shown promising results in various doc-



Q: Who is the Investigator at the University of California?
A: The investigator at the University of California is Dr. William W. Parmley.

Figure 1: Sample example from our TabComp dataset. The dataset consists of Table images of documents with questions (**Q**) and their generative answers (**A**). The image was sourced from Task 1 of the DocVQA (Mathew et al., 2021) dataset: https://rrc.cvc.uab.es.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5788–5795

ument understanding tasks, including layout analysis, form understanding, VQA, and information extraction. Despite these achievements, their potential in table reading comprehension has not been thoroughly investigated.

Our extensive experiments with current OCR-free models Donut (Kim et al., 2022), UReader (Ye et al., 2023) revealed that they underperform on our dataset. This underperformance highlights a gap in existing solutions, as these models struggle with the complexity of the table reading task, which requires understanding table structures, element relationships, and context.

To bridge this gap and enhance the OCR-free models' capabilities in understanding and generating, we present TabComp in two formats: (1) a question-answering format for a model like Donut and (2) an instruction-tuning format for a universal model like UReader, where all the downstream tasks are reorganized into instruction-tuning format (Dai et al., 2023). Notably, existing works such as (Deng et al., 2024) and (Zheng et al., 2024a) have explored solving table VQA tasks through prompting and fine-tuning, respectively. This paper provides an analysis of the challenges and opportunities within TabComp, enriched by insights from these recent studies

Our main contributions are as follows:

- While most existing Visual Question Answering (VQA) models struggle with queries related to tables, charts, and figures, we introduce TabComp, a custom dataset which emphasizes a generative question-answering task specifically focused on table images.

- TabComp does not incorporate an accompanying OCR. It is currently the only dataset focused on enabling models to read and comprehend the text within tables from document images and generate contextually accurate answers.

- We identify and highlight the shortcomings of existing models that prevent them from performing effectively on our dataset.

2 Related Work

2.1 Dataset containing few words

Visual Question Answering (VQA) on images containing text has been an area of intensive study (Antol et al., 2015), (Goyal et al., 2017). Recently, several VQA datasets featuring text in images, annotated using optical character recognition (OCR), have been released. For example, VizWiz-VQA (Gurari et al., 2018) includes questions from blind individuals who took pictures using their mobile phones. TextVQA (Singh et al., 2019), STVQA (Biten et al., 2019), and EST-VQA (Wang et al., 2020) are crowd-sourced datasets focusing on daily scenes. Other datasets target specific image types, such as OCR-VQA (Mishra et al., 2019) with book covers, FigureQA (Kahou et al., 2017), and DVQA (Kafle et al., 2018) with diagrams and charts. Our dataset distinguishes itself by featuring images of tables with a higher volume of text plus non-dependence on the OCR machine, emphasizing the development of Natural Language Understanding (NLU) on tables where multiple pieces of text and visual content are presented together.

2.2 Generative answers

One of the recent works, VisualMRC by Tanaka et al. (Tanaka et al., 2021), represents a significant milestone in integrating natural language understanding (NLU) and natural language generation (NLG) with visual data analysis. In this work, the authors introduced a dataset containing images, questions, and generative answers, requiring models to interpret textual content extracted from images and contextualize it within the visual scene. Along similar lines, WebSRC (Chen et al., 2021), another dataset, expands this vision by focusing on structured reading comprehension within web-based images and reasoning tasks, few works like TabFact (Chen et al., 2020), (Pasupat and Liang, 2015), (Zheng et al., 2024b) used for table understanding and QA task. These tasks challenge models to demonstrate robust NLU capabilities to accurately parse and comprehend text in its visual and contextual surroundings while leveraging NLG to generate coherent, contextually appropriate responses.

While this work marks a significant advancement in the simultaneous enhancement of NLU and NLG capabilities, it is important to note that it relies on an OCR-dependent dataset and method. This dependency on Optical Character Recognition (OCR) technology implies that the efficacy of the models is fundamentally linked to the precision and reliability of the OCR system utilized, potentially introducing various challenges and constraints.

2.3 OCR-free Models

Recently, OCR-free document understanding and Visual Question Answering (VQA) models Donut (Kim et al., 2022), and UReader (Ye et al., 2023) present a significant evolution in handling multimodal data. These models are designed to interpret and analyze visual documents without relying on Optical Character Recognition (OCR) to answer the question, thereby preserving the original visual context and nuances that OCR might distort or overlook. In the VQA domain, these models directly interpret the text within images, enabling them to effectively comprehend and answer questions. Donut utilizes a transformer-based architecture to directly interpret the visual layout and text in document images. UReader, on the other hand, combines OCR capabilities with advanced NLU to interpret documents.

3 Dataset Creation

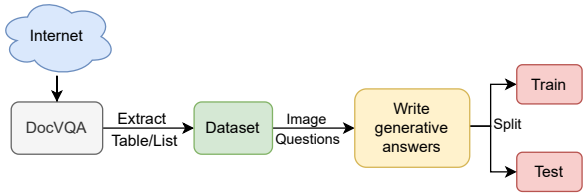


Figure 2: **Dataset creation process:** We extracted images of tables/lists from DocVQA dataset, along with their corresponding questions and answers. Subsequently, we replaced each short answer with more generative responses. Following this modification, we divided the dataset into training and test sets in a 70:30 ratio. The contents of each set are detailed later in this section.

As mentioned earlier, OCR-free models struggle with table images due to complex structures and diverse content. Thus we create a specialized dataset focused solely on table images, enhancing the performance of OCR-free models in this domain.

3.1 Table image collection

According to VisualMRC (Tanaka et al., 2021), the DocVQA dataset does not primarily focus on answer generation, which led us to develop a tailored dataset. We chose to utilize the DocVQA dataset due to its comprehensive collection of document images. From this dataset, we manually extracted table images, which were predominantly from industrial documents containing both handwritten and printed text. These table images belonged to a variety of semi-structured document images, adding to the complexity and diversity of our dataset. Notably, these images differ from VisualMRC in two aspects:

1. **Industrial Documents:** Our dataset includes images from industrial documents, which are typically not present in VisualMRC.

2. **Handwritten Document Contents:** Unlike VisualMRC, our dataset incorporates handwritten text, adding another layer of challenge for OCR-free models.

3.2 QA Pairs extraction and ground truth

Following the extraction of table images, the next step involved the extraction of QA pairs corresponding to each image. Upon reviewing each question and answer, we observed that while the questions were often lengthy and detailed, the answers maintained a concise format similar to the SQuAD-like format. We replaced the original SQuAD-like answers with more detailed ones to better fit our reading comprehension tasks. This manual annotation ensured the answers were accurate and provided a deeper understanding, enhancing the dataset's utility for training OCR-free models.

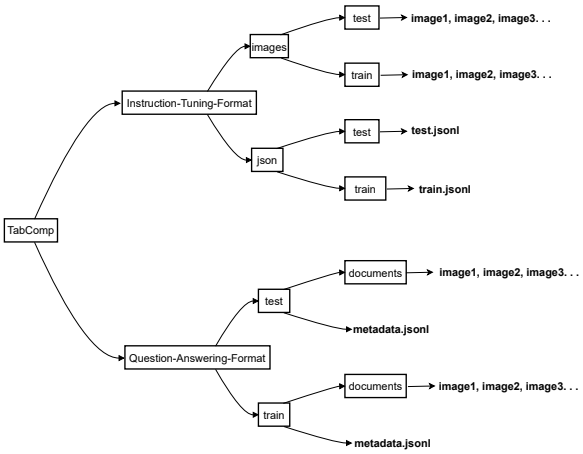


Figure 3: TabComp Structure: The dataset is presented in two formats.

4 Experimental Setup

The dataset was employed to fine-tune the existing donut-base and donut-proto models. We conducted the fine-tuning process for a total of 300 epochs, ensuring that the models had ample opportunity to learn and adapt to the nuances of the dataset. In addition to fine-tuning, we also trained the donut-base and donut-proto models in end-to-end settings.

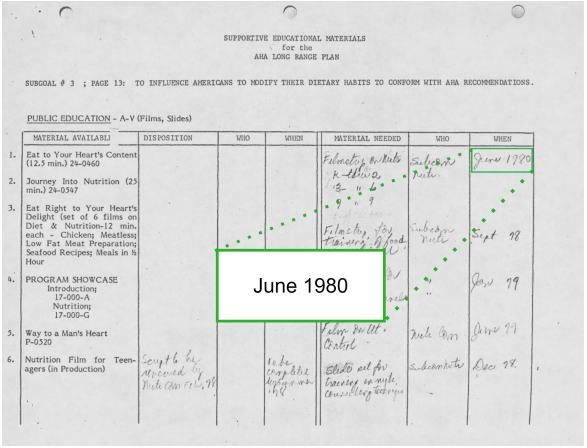
	VisualMRC	TabComp
# Images	10,197	3,318
# Questions	30,562	19,610
# Unique questions	29,419	15,164
% Unique answers	91.82	91.87
Avg. len. questions	10.55	10.32
Avg. len. answers	9.53	11.48

Table 1: Statistics of TabComp and VisualMRC datasets, the average length of questions and answers are measured by tokenizing them with NLTK.

The image sizes used during the training process were tailored to each model's specifications. For the donut-base model, the images were resized to 2560x1920 pixels, while for the donut-proto model, the images were resized to 2048x1536 pixels. Moreover, the swin window sizes for the donut-base and donut-proto models were set to 10 and 8, respectively. These configurations were adopted as per the recommendations outlined in the Donut paper.

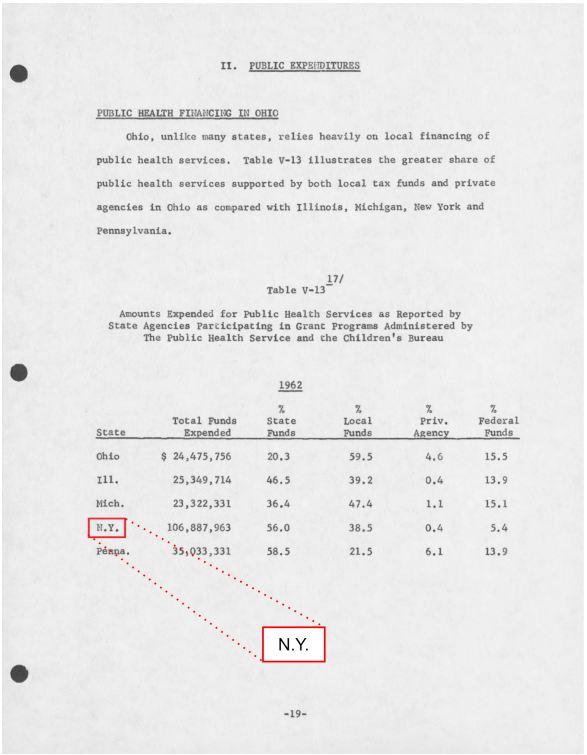
5 Results and Analysis

Table 2 shows the performance of different OCR-free models on our TabComp dataset, comparing fine-tuned and end-to-end configurations. The test set is in instruction-tuning format and is used for UReader inference. Upon examination, it is evident that in the case of Donut, end-to-end models outperform fine-tuned models for tasks with complex, task-specific interdependencies (such as generating answers or information extraction). This superiority is particularly noticeable in generating answers, where end-to-end models excel. End-to-end models can understand the intricate connections between various stages of a task because they are trained on the entire task (Generation + Extraction) as a unified process. Additionally, they are more flexible when applied to new tasks or domains since they don't depend on fixed steps or predefined features during the process. In contrast, fine-tuned models may experience catastrophic forgetting (French, 1999), where adapting to new tasks can result in losing previously acquired knowledge. This challenge can hinder their performance on tasks that require generating answers. Since UReader performs well on table VQA tasks without fine-tuning (Ye et al., 2023), we used it as-is for inference with our dataset.



Q: What is the date of the first film?
GT: The date of the first film is June 1980.
Donut-base fine-tuned: The date of the first film is June 1980.(✓)
UReader: 1927 (✗)

Figure 4: Inference output of Donut-base (fine-tuned) and UReader on handwritten document image from the test set.



Q: as per table v-13, in which state is total funds expended the highest
GT: In table V-13, total funds expended are highest in N.Y.
Donut-base fine-tuned: The total funds expended the highest year 1963 is $34,475,75..(✗)
UReader: 27 (✗)

Figure 5: Incorrect Example generated by both Donut-base fine-tuned and UReader model.

Models	Fine-tuned	End-to-end	B-1	B-2	B-3	B-4	R-L	BERTScore	Meteor	CIDEr
Donut-base	✓	✗	66.60	55.59	48.84	42.69	37.29	83.38	60.14	69.32
	✗	✓	55.12	41.36	34.28	28.59	32.24	85.06	47.19	54.01
Donut-proto	✓	✗	29.82	15.30	09.73	06.49	17.84	73.26	19.80	23.56
	✗	✓	62.69	49.32	41.63	34.87	37.02	87.74	56.49	66.03
UReader	✗	✗	42.01	35.86	31.59	28.14	37.64	88.04	20.71	210.77

Table 2: Performance of TabComp on OCR-free VQA models: B implies BLEU, and R-L implies ROUGE-L.

6 Conclusion

We present a custom table reading comprehension dataset, carefully curated and manually checked to ensure high confidence in its generative answers. The dataset includes a diverse array of table images extracted from industrial documents with both handwritten and printed text, providing a comprehensive resource for training and evaluating advanced models. Our fine-tuning experiments demonstrated the need for a more robust OCR-free model to accurately understand and interpret table images, highlighting TabComp's potential for improving OCR-free VQA systems. Moreover, by making our dataset available to the research community, we aim to foster further advancements in the field of table reading comprehension and document analysis.

Limitations

Unlike the VisualMRC dataset, TabComp lacks text localization information, which is essential for identifying the answer's position within the document image for OCR-based models. While a study by (Kim et al., 2023) offers methodologies for OCR-free models to potentially address this limitation, the lack of specified regions of interest remains a constraint.

Ethics Statement

Our dataset utilizes document images from the publically available DocVQA dataset. We ensured that the use of the DocVQA dataset complies with its intended usage policies and respects all conditions set forth by the original data providers. The author performed manual QA annotation to transform existing structured answers into more generative formats suitable for advanced document understanding tasks. The annotations were conducted with a commitment to maintaining the integrity of the original data and without introducing any biases or alterations that could misrepresent the information presented in the images.

Acknowledgment

We thank Abhirama Subramanyam Penamakuri for his feedback and suggestions. We also acknowledge the anonymous reviewers for their constructive feedback. Additionally, we thank the metareviewer for their careful consideration and insightful comments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. *Preprint*, arXiv:2101.09465.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 407–426.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. *Preprint*, arXiv:2111.15664.

Geewook Kim, Shuhei Yokoo, Sukmin Seo, Atsuki Osanai, Yamato Okamoto, and Youngmin Baek. 2023. On text localization in end-to-end ocr-free document understanding transformer without text localization supervision. In *International Conference on Document Analysis and Recognition*, pages 215–232. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *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)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *Preprint*, arXiv:2101.11272.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. 2023. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *Preprint*, arXiv:2310.05126.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024a. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024b. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

A Appendix

A.1 Output Example

Table 3 illustrates several outputs produced by the model. Among the four models examined, the donut-base model stands out as the best performer, generating the highest number of exact correct answers. In comparison, the donut-proto model tends to repeat tokens, resulting in non-sequitur sentences. This observation leads to the conclusion that the donut-proto model is overfitting and would benefit from additional training data to improve its performance and generalization. Despite these differences, it is noteworthy that both models, when used in an end-to-end setting, exhibit the generation capability.

A.2 Computational setup

The dataset was employed to fine-tune the existing donut-base and donut-proto models. The fine-tuning process was carried out using two Nvidia A30 GPUs, each equipped with 24GB of memory. This setup provided the necessary computational power to handle the extensive training tasks efficiently. For the donut-base model, the fine-tuning process took approximately 2.5 days to complete. On the other hand, the donut-proto model required about 2 days for fine-tuning. The difference in training durations can be attributed to the varying complexities and configurations of the models.

A.3 Suitable Metrics

To evaluate the quality of the generated answers, we used a variety of widely adopted metrics:

- BLEU (Bilingual Evaluation Understudy) (Brown et al., 2020) measures the precision of N-grams between the generated and reference texts. The score is calculated by taking the geometric mean of N-gram precisions with a brevity penalty to discourage overly short translations:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

  Where p_n is the precision of N-grams of length n. w_n is the weight assigned to N-grams of length n, typically equal across all N-grams. BP is the brevity Penalty that penalizes shorter generated sequences.

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) evaluates the longest common subsequence (LCS) between the generated and reference texts. It captures both precision and recall, with a combined F1 score:

$$ROUGE - L = \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

  β is typically set to 1, giving equal weight to precision (P) and recall (R).

- BERTScore (Zhang et al., 2019) leverages BERT embeddings to measure semantic similarity between generated and reference texts. It computes the precision and recall based on cosine similarities of word embeddings, followed by the F1 score:

$$F_{BERT} = \frac{2 \cdot P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

  where, P_{BERT} is $Precision_{BERT}$ and R_{BERT} is $Recall_{BERT}$.

- METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005) aligns the generated and reference texts based on unigram matches (including stemming and synonyms). It emphasizes recall, with a weighted F1 score and penalty for disjoint word chunks:

$$METEOR = (1 - Penalty) \cdot F_{mean}$$

  $Penalty = 0.5(\frac{\#chunks}{m})^3$, which penalizes cases where the generated text has many disjoint chunks.

- CIDEr (Consensus-based Image Description Evaluation) (Vedantam et al., 2015) measures the consensus between generated and reference texts using TF-IDF weighting, rewarding N-grams that are important yet rare across multiple references:

$$\text{CIDEr} = \frac{1}{m}\sum_{j=1}^{m}\sum_{n=1}^{N}\frac{T(g_n) \cdot T(r_n)}{\|T(g)\| \cdot \|T(r)\|}$$

  Where m is the number of reference texts, g_n and r_n are the N-grams in the generated and reference texts. $T()$ is the TF-IDF score of N-grams x_n.

Questions	Answers
who is the speaker?	The speaker is Miss Midred Kauman. The supplier is LRM. The speaker is Dr. T. Turner.
who is the supplier?	The supplier is M/A/R/C. The speaker by the speaker by the a faxances1% of 12. The supplier of supplier of 25, 25, 25, 25, 25, also simply represented by a fax.

Table 3: Samples showcasing different cases of generated answers for the same question across different models. Green indicates correct answers generated by the Donut-base model, blue represents incorrect answers (instances of catastrophic forgetting) generated by fine-tuned and inference models, and red highlights non-sequitur sentences (instances of overfitting) generated by the Donut-proto model.

Metric	Purpose	Methodology
BLEU	Machine Translation	n-gram precision
ROUGE	Document Summarization	n-gram recall
METEOR	Machine Translation	n-gram with synonym matching
CIDEr	Image Captioning	tf-idf weighted n-gram similarity

Table 4: Summary of Metrics Used for Generative Answers Evaluation.

Feature	DONUT	UReader
Input Type	Document Images	Document Images, optionally with Text Inputs
Model Type	Vision-Language Model	Multimodal Vision-Language Model
Architecture Base	Vision Transformer (ViT) + Transformer Decoder	Vision Transformer (ViT) + Transformer Encoder-Decoder
OCR Usage	Not required; operates without OCR	Optional; can operate with or without OCR
Task Specialization	Document Text Recognition, Key-Value Extraction, Document Layout Analysis	Document Classification, Information Extraction, Question Answering, Table Parsing
Adaptation	End-to-End, Fine-tuning	low-cost instruction tuning
Pre-training	Uses document images for vision pre-training	Uses both document images and text for pre-training

Table 5: Key Architectural Differences between DONUT and UReader Models