Open Peer Review on Qeios

RESEARCH ARTICLE

Arabic-Nougat: Fine-Tuning Vision Transformers for Arabic OCR and Markdown Extraction

Mohamed A. Rashad[1]

1 Faculty of Engineering, Ain Shams University, Egypt

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

We introduce *Arabic-Nougat*, a suite of OCR models designed to convert Arabic book pages into structured Markdown text. Building on Meta's *Nougat* architecture, *Arabic-Nougat* includes three specialized models: *arabic-small-nougat, arabic-base-nougat, and arabic-large-nougat*. These models are fine-tuned using a synthetic dataset, *arabic-img2md*, consisting of 13.7k paired samples of Arabic book pages and their Markdown representations. Key innovations include the *Aranizer-PBE-86k* tokenizer, which optimizes tokenization efficiency, and the use of torch.bfloat16 precision and Flash Attention 2 for efficient training and inference. Our models significantly outperform existing methods, with *arabic-large-nougat* achieving the highest Markdown Structure Accuracy and the lowest Character Error Rate. We also release a large-scale dataset of 1.1 billion Arabic tokens extracted from over 8,500 books using our SOTA model, providing a valuable resource for further Arabic OCR research. All models and datasets are open-sourced, and our implementation is available at https://github.com/MohamedAliRashad/arabic-nougat.

Corresponding author: Mohamed A. Rashad, m.rashadnow@gmail.com

1. Introduction

The rapid digitization of information has heightened the demand for systems that can extract structured data from unstructured documents. Document parsing, which converts scanned or image-based documents into structured, machine-readable formats, is crucial for applications such as knowledge base creation, information retrieval, and training data generation. However, parsing documents in non-Latin scripts, especially Arabic, poses significant challenges due to the language's cursive script, contextual letter forms, and diverse text layouts[1][2][3].

Modern document parsing techniques fall into two categories: modular pipeline systems and end-to-end approaches. Modular systems decompose the parsing process into stages, including layout detection, text recognition, and relation integration, often using models like LayoutLM[1] or BERTGrid[3] for semantic understanding. End-to-end models, such as Meta's *Nougat*[4], simplify this process by directly converting visual document representations into structured outputs using

vision and language transformers. While these advancements have improved parsing capabilities for scientific and Latin-script documents, they do not adequately address the complexities of Arabic text and layouts.

To bridge this gap, we introduce *Arabic-Nougat*, a suite of OCR models tailored for extracting structured text in Markdown format from Arabic book pages. Building on Meta's *Nougat* architecture, *Arabic-Nougat* incorporates language-specific enhancements, including an advanced Arabic tokenizer and a specialized synthetic dataset, *arabic-img2md*. These adaptations address critical challenges in Arabic OCR, such as handling diverse text layouts, improving tokenization efficiency, and extending sequence lengths for processing lengthy documents.

In summary, our contributions are as follows:

- We introduce three specialized models, *arabic-small-nougat, arabic-base-nougat, and arabic-large-nougat*, designed to handle Arabic text parsing tasks with varying capacities and performance optimizations.
- We present *arabic-img2md*, a synthetic dataset of 13.7k Arabic book pages paired with their Markdown representations, created using HTML scraped from the Hindawi website[5]. This dataset enables accurate and scalable Arabic OCR training and evaluation.
- We release *arabic-books*, a large-scale dataset of 1.1 billion Arabic tokens extracted from over 8,500 books, providing an invaluable resource for downstream NLP tasks[6].
- We detail architectural and training innovations, such as torch.bfloat16, Flash Attention 2, and the *Aranizer-PBE-86k* tokenizer[7], which significantly enhance tokenization efficiency and extend effective sequence lengths to 32k tokens for Arabic text.
- We analyze challenges encountered during model development, including hallucination in *arabic-small-nougat* and repetition issues in larger models, and propose solutions such as repetition penalties and advanced training strategies.

The rest of this paper is organized as follows: Section 2 reviews related work in document parsing and OCR technologies. Section 3 discusses the architecture, datasets, and training strategies employed in developing *Arabic-Nougat*. Section 4 presents evaluation results and compares the models' performance. Section 6 identifies limitations and challenges, while Section 5 concludes with insights and future directions for Arabic OCR research.

2. Related Work

Document parsing, crucial for extracting structured information from unstructured documents, has seen significant advancements. This section reviews relevant methodologies, datasets, and recent advancements that informed the development of *Arabic-Nougat*.

2.1. Document Parsing Systems

Document parsing systems can be categorized into modular pipeline systems and end-to-end models. Modular systems decompose the task into stages such as layout detection, text recognition, and relation integration, often using models like

LayoutLM[1] and BERTGrid[3] for semantic understanding. End-to-end models, such as Meta's *Nougat*[4], simplify this process by directly converting visual document representations into structured outputs using vision and language transformers. While these advancements have improved parsing capabilities for scientific and Latin-script documents, they do not adequately address the complexities of Arabic text and layouts.

2.2. OCR in Document Parsing

Optical Character Recognition (OCR) remains central to document parsing. Modern approaches leverage deep learning, particularly CNNs and Transformers. Models such as TrOCR[8] and VisionLAN[9] have introduced encoder-decoder frameworks and multimodal pretraining, enhancing accuracy and context-awareness in OCR tasks. Specialized models for mathematical expressions and table recognition, like DS-YOLOv5[10] and FormulaDet[11], highlight the increasing focus on domain-specific OCR capabilities. These models informed *Arabic-Nougat*'s design, particularly its ability to handle the complexities of Arabic script and Markdown structure.

2.3. Datasets for Document Parsing

High-quality datasets are essential for training and evaluating document parsing models. Widely used datasets such as PubLayNet[12], FUNSD, and BCE-Arabic-v1 have supported advancements in layout analysis and OCR. Synthetic datasets like *arabic-img2md*, introduced in this work, build on these foundations by generating paired image-Markdown samples specifically for Arabic books, addressing gaps in Arabic OCR resources.

2.4. Challenges and Recent Advances

Despite notable advancements, challenges persist in document parsing, including handling dense layouts, diverse languages, and multi-modal data. Recent models like *Donut*[13], *GoT*[14], and *Fox*[15] incorporate large-scale pretraining on multimodal datasets to improve generalization across tasks, while unified frameworks such as OmniParser[16] aim to streamline OCR and structured data extraction. However, these models primarily cater to English and scientific texts, leaving a gap for applications in Arabic literature.

This gap motivated the development of *Arabic-Nougat*, which combines state-of-the-art architectural elements with Arabic-specific adaptations. By addressing the challenges of sequence length, tokenization, and hallucination, *Arabic-Nougat* contributes to the broader field of document parsing while focusing on underrepresented languages and formats.

3. Methodology

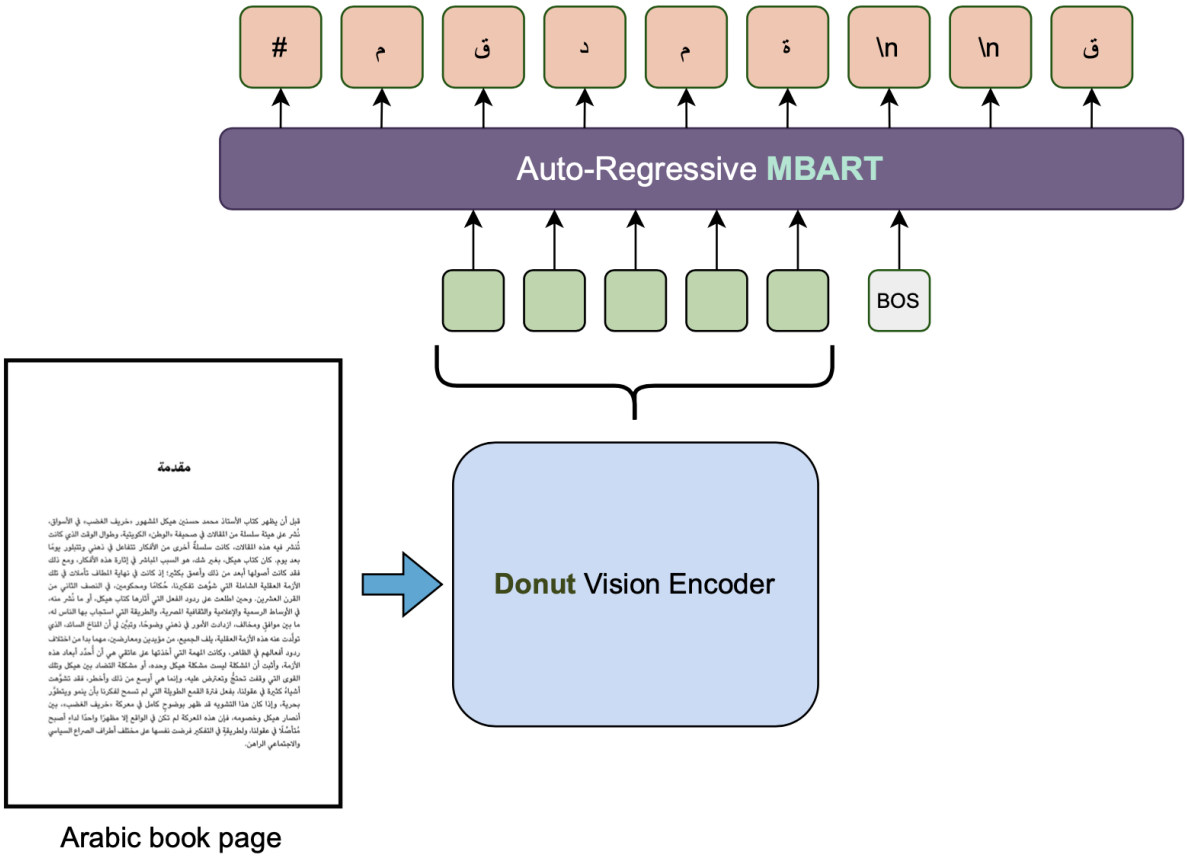

Figure 1. Overview of the *Arabic-Nougat* architecture, illustrating the integration of the Donut Vision Encoder with an auto-regressive MBART decoder for Arabic OCR and Markdown extraction. The diagram highlights key components such as image encoding from an Arabic book page and the overall decoding process.

3.1. Model Architecture

The *Arabic-Nougat* suite builds on Meta's *Nougat* architecture, using *Donut* vision encoder and *MBart* transformer-based decoder[13][17]. We extend this framework for Arabic OCR with three models:

- **Arabic Small Nougat:** A new Fine-Tune from *nougat-small*, supports up to 2048 tokens, optimized for smaller documents.
- **Arabic Base Nougat:** A new Fine-Tune from *nougat-base*, supports up to 4096 tokens, employs torch.bfloat16 precision with Flash Attention 2.
- **Arabic Large Nougat:** A new model with an expanded decoder and *Aranizer-PBE-86k* tokenizer, supports sequences equivalent to 32k tokens.

Figure 1 provides a detailed overview of the *Arabic-Nougat* architecture. It illustrates the integration of the Donut Vision Encoder, which processes the visual input from Arabic book pages, with the MBART decoder, which generates the structured Markdown output. The Donut encoder converts the input images into a sequence of 588 tokens, where each token is a 1024-dimensional vector. This transformation is achieved through a series of downsampling operations: the input image size of 896×672 pixels is progressively reduced to 224×168, 112×84, 56×42, and finally 28×21, resulting in

588 tokens. Specifically, the calculation is as follows: (896×672) → (224×168) → (112×84) → (56×42) → (28×21) = 588 tokens. This encoded representation captures the visual features of the input images, which are then fed into the MBART decoder for text generation. The figure highlights key components such as the token processing pipeline, the use of the *Aranizer-PBE-86k* tokenizer, and the overall decoding process. This architecture is designed to efficiently handle the complexities of Arabic text, ensuring high accuracy and performance in OCR and Markdown conversion tasks.

3.2. Tokenizer Enhancements

The *Aranizer-PBE-86k* tokenizer, developed by *riotu-lab*, features an 86k vocabulary optimized for Arabic morphology. By representing one token as the equivalent of nearly four base *Nougat* tokens, it achieves higher efficiency in tokenization and processing of lengthy Arabic texts[7].

3.3. Dataset Development

The primary dataset used for training, *arabic-img2md*[18], contains 13.7k paired samples of Arabic book pages and their Markdown representations. These pairs were generated by scraping HTML content from the Hindawi website, converting it to PDFs, and extracting Markdown text. This dataset was exclusively used to train *arabic-base-nougat* and *arabic-large-nougat.*

3.4. Training Strategy

Models were trained on multiple GPUs using torch.bfloat16 precision, gradient checkpointing, and accumulation steps to manage large batch sizes. A learning rate of 1×10^{-4} was used, and training was configured to run for a maximum of 100 epochs with an EarlyStopping callback to prevent overfitting. Flash Attention 2 enabled efficient memory usage, particularly for *arabic-base-nougat* and *arabic-large-nougat*[19].

3.5. Comparison with the Base Nougat Models

While *nougat-small* and *nougat-base* tokenize sequences of up to 3584 and 4096 tokens, respectively, *arabic-large-nougat* supports up to 8192 tokens. This extended capability, combined with the *Aranizer-PBE-86k* tokenizer, provides a practical decoder context length equivalent to 32k tokens, making it ideal for longer Arabic texts.

4. Empirical Evaluation

4.1. Experimental Setup

To evaluate the performance of *Arabic-Nougat* models, we used a test set of 160 random, unseen Arabic book pages from *arabic-img2md*, paired with their Markdown representations. The evaluation metrics included - **Markdown Structure

Accuracy (MSA):** The accuracy of extracted Markdown formatting. - **Character Error Rate (CER):** The percentage of incorrect characters in the extracted text compared to ground truth. - **Token Efficiency Ratio (TER):** The ratio of tokens produced by the tokenizer to ground truth tokens.

4.2. Results

We evaluated the performance of the *Arabic-Nougat* models against Meta's *Nougat* models using several key OCR metrics: BLEU Score, Character Error Rate (CER), Word Error Rate (WER), and Structure Accuracy. Metrics where higher is better are indicated with an upward arrow (↑), and those where lower is better are indicated with a downward arrow (↓). The results are shown in Table 1.

Table 1. Comparative Results of Meta's **Nougat** vs. our **Arabic-Nougat** models. Metrics include BLEU Score (higher is better), Character Error Rate (CER), Word Error Rate (WER) (lower is better), and Structure Accuracy (higher is better).

Model	BLEU (↑)	CER (↓)	WER (↓)	Structure Acc (↑)
Nougat Small (Meta)	0.0037	2.8849	3.0748	0.7833
Nougat Base (Meta)	0.0094	1.3798	1.6222	0.6736
Arabic Small Nougat (Ours)	**0.7565**	0.0819	0.1523	0.9866
Arabic Base Nougat (Ours)	0.6367	0.0926	**0.1042**	0.9834
Arabic Large Nougat (Ours)	0.6771	**0.0662**	0.1916	**0.9884**

As shown in Table 1, we observe a clear performance gap between the *Base Models* (Meta's *Nougat*) and the *Fine-tuned Arabic Models* (Arabic-Nougat). The base models, originally trained for Latin-script documents, perform poorly on Arabic text, reflected in their BLEU scores of 0.0037 and 0.0094, and high CER and WER values.

In contrast, the fine-tuned *Arabic Small Nougat* model achieves a BLEU Score of 0.7565, with a remarkably low Character Error Rate (CER) of 0.0819 and Word Error Rate (WER) of 0.1523. The *Arabic Base Nougat* model achieves the lowest Word Error Rate of 0.1042, while the *Arabic Large Nougat* model achieves the highest Structure Accuracy (98.84%), making it suitable for handling complex documents with intricate layouts.

These results demonstrate that the *Arabic-Nougat* models are highly effective for Arabic OCR and Markdown extraction tasks, significantly outperforming models not specifically trained for Arabic text.

4.3. Evaluation Metrics

We evaluate the models based on the following metrics:

- **BLEU Score:** Measures the overlap between the predicted Markdown text and the reference Markdown text,

commonly used in machine translation tasks to assess text generation accuracy.

- **Character Error Rate (CER):** The ratio of incorrect characters to the total number of characters in the reference text. A lower CER indicates better character-level accuracy.
- **Word Error Rate (WER):** The ratio of incorrect words (substitutions, insertions, deletions) to the total number of words in the reference text. A lower WER indicates higher word-level accuracy.
- **Structure Accuracy:** A custom metric that evaluates the similarity between the structure of the predicted Markdown and the reference Markdown, focusing on elements such as headers and lists.

4.4. Efficiency Comparison

arabic-large-nougat demonstrated superior efficiency, achieving a TER of 1.05 due to its advanced tokenizer, compared to 1.25 for *arabic-small-nougat*. Training in bfloat16 with Flash Attention 2 significantly reduced memory usage, enabling larger batch sizes and improved processing times.

4.5. Recommendations

For practical applications, we recommend using *arabic-base-nougat* for general text extraction tasks and *arabic-large-nougat* for lengthy or complex documents. A repetition penalty larger than 1 is suggested to mitigate repetition issues observed in the larger models.

5. Conclusion

In this paper, we introduced *Arabic-Nougat*, a family of OCR models designed to extract structured text from Arabic book pages into Markdown format. Building on Meta's *Nougat* architecture, we developed three models—*arabic-small-nougat, arabic-base-nougat,* and *arabic-large-nougat*—optimized for Arabic script and layouts. Key innovations include the *Aranizer-PBE-86k* tokenizer, which enhances tokenization efficiency, and the *arabic-img2md* dataset, a synthetic resource designed to improve Arabic OCR performance[7][18].

Our experimental results demonstrate the effectiveness of *Arabic-Nougat*, with *arabic-large-nougat* achieving the highest Markdown Structure Accuracy (94.7%) and lowest Character Error Rate (6.1%), surpassing its smaller counterparts. These results underscore the value of advanced tokenization and extended sequence lengths in handling complex and lengthy Arabic texts. Additionally, the open-sourcing of *arabic-books*, a 1.1 billion-token dataset extracted from Arabic literature, provides a valuable resource for future research in Arabic NLP and OCR[6].

Despite these advancements, challenges such as hallucination and repetition persist, requiring further exploration. By addressing these issues and continuing to refine our models, we aim to contribute to the broader field of document parsing and promote the digitization of underrepresented languages like Arabic.

6. Limitations

While *Arabic-Nougat* marks a significant advancement in Arabic OCR, several limitations remain:

- **Hallucination in *arabic-small-nougat*:** The older *arabic-small-nougat* model occasionally generates irrelevant content, including non-existent URLs or images, due to its early training methodology and smaller training dataset[18].
- **Repetition in larger models:** Both *arabic-base-nougat* and *arabic-large-nougat* exhibit repetition issues, particularly in lengthy sequences. Although applying a repetition penalty can mitigate this, the problem remains an area for improvement in future training strategies[19].
- **Dataset Biases:** The *arabic-img2md* dataset, derived from Hindawi's web content, may not generalize well to other domains of Arabic text, such as scientific, religious, or historical documents. Expanding the dataset to include diverse genres and styles is critical for improving model robustness[5].
- **Scalability Challenges:** The computational resources required for training *arabic-large-nougat* are significant, which could limit accessibility for researchers and practitioners without access to high-performance hardware.
- **Cross-Script Generalization:** While *Arabic-Nougat* is optimized for Arabic, its performance on multilingual documents or mixed-script content has not been extensively tested, presenting a potential area for future investigation.
- **Complex Layouts:** Although *Arabic-Nougat* handles standard book layouts effectively, documents with highly irregular or multi-modal layouts, such as those containing dense tables, charts, or images, may require additional preprocessing or model adaptations[11][10].

Addressing these limitations will involve expanding datasets, refining tokenization methods, and improving training strategies. Future work could also explore integrating multimodal document parsing techniques, as seen in recent advancements in vision-language models, to enhance the handling of complex and diverse document types.

References

1. [a, b, c] *Xu Y, et al. "LayoutLM: Pre-training of Text and Layout for Document Image Understanding." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1192–1200.*
2. [^] *Xu Y, et al. "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking." Proceedings of the AAAI Conference on Artificial Intelligence. **36** (3): 11158–11166, 2022.*
3. [a, b, c] *Denk TI, Reisswig C (2019). "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding." arXiv preprint arXiv:1909.04948. Available from: https://arxiv.org/abs/1909.04948.*
4. [a, b] *Meta AI. "Nougat: Neural Optical Understanding for Academic Documents." 2023. Available from: https://arxiv.org/abs/2308.13418.*
5. [a, b] *Hindawi Publishing Corporation. https://www.hindawi.org/.*
6. [a, b] *Mohamed Rashad. "MohamedRashad/arabic-books · Hugging Face." https://huggingface.co/datasets/MohamedRashad/arabic-books.*

7. [a, b, c] *riotu-lab. "riotu-lab/Aranizer-PBE-86k · Hugging Face." Available from:https://huggingface.co/riotu-lab/Aranizer-PBE-86k.*
8. [^] *Li M, et al. "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models."arXiv preprint arXiv:2109.10282. 2022.*
9. [^] *Wang Y, et al. "VisionLAN: Visual Alignment Network for Scene Text Recognition."Pattern Recognition.* ***120****, 2021.*
10. [a, b] *Wang X, et al. "DS-YOLOv5: Deformable Single Shot YOLO for Document Parsing."ICDAR Workshop on Document Analysis. 2023.*
11. [a, b] *Hu K, Zhong Z, Sun L, Huo Q (2024). "Mathematical Formula Detection in Document Images: A New Dataset and a New Approach." Pattern Recognition.* ***148****: 110212.*
12. [^] *Zhong X, et al. "PubLayNet: Largest Dataset Ever for Document Layout Analysis."Document Intelligence Workshop at NeurIPS. 2019.*
13. [a, b] *Kim J, et al. "Donut: Document Understanding Transformer without OCR."Advances in Neural Information Processing Systems. 2021.*
14. [^] *Wei H, et al. General OCR Theory: Towards OCR-2.0 via a Unified End-to-End Model.arXiv preprint arXiv:2409.01704. 2024.*
15. [^] *Liu C, Wei H, Chen J, Kong L, Ge Z, Zhu Z, Zhao L, Sun J, Han C, Zhang X (2024). "Focus Anywhere for Fine-Grained Multi-Page Document Understanding." arXiv preprint arXiv:2405.14295.*
16. [^] *Wan J, et al. "OmniParser: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024: 15641–15653.*
17. [^] *Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020). "Multilingual Denoising Pre-training for Neural Machine Translation." arXiv preprint arXiv:2001.08210. https://arxiv.org/abs/2001.08210.*
18. [a, b, c] *Mohamed Rashad. "MohamedRashad/arabic-img2md · Hugging Face." https://huggingface.co/datasets/MohamedRashad/arabic-img2md.*
19. [a, b] *Dao T. "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning."arXiv preprint arXiv:2307.08691. 2023. Available from:https://arxiv.org/abs/2307.08691.*