Evaluating the Impact of OCR Accuracy on Downstream Document Summarization

Gael Joao
Supervised By Dr. Steven James and Dr. Benjamin Rosman
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

Abstract

Optical Character Recognition (OCR) plays a critical role in digitizing documents, yet its impact on downstream tasks like summarization remains under-explored. This study investigates how OCR accuracy affects the performance of document summarization models. The goal is to evaluate the robustness of summarization models when exposed to varying levels of OCR noise (distortions and blur) and as such we compare two OCR systems, TrOCR and DONUT, and assess their outputs using clean and noisy document images. The resulting text is fed into a summarization model like PEGASUS to analyse performance degradation and use the data collected to investigate whether TrOCR is more robust compared to DONUT in such scenarios. This research highlights the importance of OCR robustness for real-world document processing pipelines.

Introduction

Industries increasingly rely on Optical Character Recognition (OCR) for processing and summarizing large volumes of documents, particularly historical records (Jatowt et al., 2019; van Strien et al., 2020). OCR is used to convert image-based or scanned documents into machine-readable text and is widely used across various domains, including healthcare and finance. However, OCR systems are computationally intensive and often require extensive preprocessing because they tend to be inflexible across different languages or document types and errors introduced by OCR can affect later processing stages (Kim et al., 2022). One of the key challenges in this domain lies in the accuracy of OCR outputs, especially when the input documents are affected by noise such as distortions and blurs. These imperfections degrade the quality of the extracted text, which in turn impacts the performance of downstream tasks like document summarization.

This research addresses the intersection of OCR and machine learning by investigating how different types and levels of noise affect the performance of transformer-based summarization models. In particular, we fix the summarization stage and systematically introduce synthetic noise during the OCR stage to evaluate the robustness of various state-of-the-art OCR models. This will help determine which models are more resilient to degraded input and why they outperform alternatives under challenging conditions.

Despite the growing importance of OCR and document summarization, few studies have examined these components in tandem. The impact of noisy inputs caused by factors such as poor print quality, font variations, or corrupted text on summarization accuracy remains under-explored. Previous research (Liu and Lapata, 2019; Lewis et al., 2020) suggests that denoising strategies during pre-training can improve summarization performance, but a more systematic evaluation of OCR robustness and its downstream effects is still needed.

While numerous OCR technologies exist, many lack an end-to-end architecture, rely heavily on CNN backbones, or are not designed for joint understanding with summarization tasks. Furthermore, most current approaches treat OCR and summarization as independent stages, which limits its scalability for long, complex document layouts. To address these challenges, this research leverages deep learning and transformer-based models. Transformers are particularly well-suited for modelling long-range dependencies and can maintain contextual understanding even when the sentence structure is disrupted by OCR noise (Appalaraju et al., 2021). Their integration with CNN-based encoders enables end-to-end learning, improved generalization, and enhanced layout understanding. On the summarization side, transformer architectures provide abstractive capabilities and can be fine-tuned for domain-specific applications, such as legal or medical document processing.

The plan is to evaluate the robustness of two transformer-based OCR models, TrOCR and DONUT, under the noisy conditions mentioned above. The extracted text from each model will be summarized using the PEGASUS model, and the quality of the summaries will be assessed using ROUGE scores to determine which OCR model performs better in degraded scenarios.

A clean dataset of document images with corresponding ground-truth summaries will be prepared. These images will then be synthetically altered to simulate noise, using Gaussian blur and scribbles, creating multiple variants for each image (clean, blurred and scribbled).

This pipeline will allow for a comparative analysis of how OCR degradation impacts downstream summarization, revealing which model is more robust to noise and better pre-

serves semantic content under noisy conditions.

Background and Related Work

Optical Character Recognition Systems

OCR systems have evolved from rule-based algorithms (Smith, 2007) to deep learning models (Cohan et al., 2018; Zhang et al., 2020; Appalaraju et al., 2021), driven by advances in document interpretation. Built on text detection and recognition, OCR still faces challenges in accurate digitalization (Li et al., 2023). Earlier methods utilized CNNs for image understanding and RNNs for text generation, and while now outdated, modern end-to-end models build on Tesseract's core principles of segmentation and sequential character recognition.

Formally, OCR as a supervised sequence-to-sequence problem aims to learn a function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ that maps document images $x \in \mathcal{X}$ to their corresponding structured markup $y \in \mathcal{Y}$ (Blecher et al., 2023):

- The encoder (Swin Transformer) encodes the input into latent embeddings: $z \in \mathcal{R}^{d \times N}$, where d is the latent dimension and N is the number of patches
- The decoder (mBART) decodes z into a sequence of tokens $\hat{y} \in \mathcal{Y}$ in an auto-regressive fashion, then maps the output to the size of the vocabulary v, resulting in the logits $\ell \in \mathcal{R}^v$

From a machine learning perspective, the goal is to find the function f^* that minimizes the expected loss over the data distribution $\mathcal{D}$:

$$f^* = \arg\min_f \mathbf{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(f(x), y)\right]$$

where $\mathcal{L}$ is the cross-entropy loss over tokens in the generated sequence.

One essential feature across the various reports is the use of computer vision for structured information extraction in visually rich documents (DocTR) (Liao et al., 2023). It was mentioned earlier the impact of evolution in OCR, from rule-based systems (e.g.: Tesseract) to modern deep learning models (e.g.: TrOCR, DocTR) and the latter will be the main approach, but alternatives are also discussed where appropriate.

Rule-Based Systems Rule-based systems where developed with the goal of improving the weakness of commercial OCR, which at the time would only show good performance on prints of the highest quality, from the models developed at the time we have Tesseract by Smith, 2007 (Smith, 2007), techniques for automatically correcting words in text introduced by Kukich, 1993 (Kukich, 1993). Kukich, 1993 (Kukich, 1993) used dictionary-based approaches to identify and correct misspellings such as edit distance between strings, and keyboard distance. Young et al., 1991 (Young et al., 1991), on the other hand, focused on syntactic rules and pattern-matching approaches for query interpretation and correction in document retrieval systems.

Tesseract introduced algorithms that were new at the time to outperform the existent commercial OCR engines.The model receives an input image and converts it to black and white, then analyses its layout to detect text lines and words, segmenting words into characters, classifying those characters using adaptive classifiers, and then verifying recognized words with dictionaries and post-processing rules for final output. The adaptive classifier embedded in the Tesseract can recognize damaged characters easily, and as a result the training time is significantly reduced. The key strength of the Tesseract was its unusual choice of features; it would extract the shape and contours of each character and convert them into a polygonal approximation; however, the model experiences a performance drop when dealing with complex and noisy layouts.

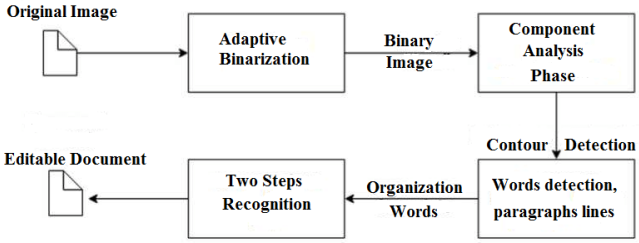


Figure 1: Tesseract OCR architecture

Deep Learning Models Deep learning models are capable of learning the semantics and contextual relationships within natural language, both at sentence/document level, and multimodal understanding. For instance, models such as DONUT, TrOCR, and discourse-aware attention models will differ in their architectures and learning strategies, including attention-based mechanisms, multimodal transformers, visual transformers, OCR independence, reinforcement learning and de-noising auto-encoders (Cohan et al., 2018; Kim et al., 2022; Liao et al., 2023).

Deep learning models such as TrOCR (Li et al., 2023) would not use CNN as the backbone, instead TrOCR uses pre-trained image transformers following the implementation presented by Dosovitskiy et al., 2020 (Dosovitskiy et al., 2020), achieving then state-of-the-art results on handwritten, printed and scene text image datasets with no complex pre/post-processing steps.

This study classifies systems into two main types: rule-based and deep learning. Rule based systems are fast and interpretable, more robust to OCR noise in some cases, but they are less accurate or fluent (Chen and Bansal, 2018), and deep learning models can be fine-tuned for OCR noise, but in compensation it requires a large amount of labelled data and they are sensitive to OCR noise without preprocessing (Zhang et al., 2020; Yu et al., 2021). Both systems will experience common errors in OCR which affects the performance of their respective models: noise sensitivity, character unrecognition (confusing "0" with "O"), punctuation drops or insertions, incorrect character order (especially in cursive handwriting), detection errors, and recognition errors (Smith, 2007; Li et al., 2023).

Technical analysis of models

TrOCR (Transformer-Based OCR) TrOCR is a transformer-based OCR model that segments document images into 16×16 pixel patches, encodes them using a Vision Transformer (ViT), and decodes the visual embeddings into text sequences via a cross-attention mechanism (Li et al., 2023). The model is trained end-to-end using cross-entropy loss to associate visual patterns with textual outputs, enabling accurate transcription of document content.

DONUT (Document Understanding Transformer) DONUT is an OCR-free transformer model that processes document images using a Swin Transformer encoder and generates structured text directly through a decoder inspired by BART (Kim et al., 2022). It bypasses explicit character recognition by learning visual-text alignment from image-text pairs, optimized via sequence-level cross-entropy loss.

PEGASUS PEGASUS is a transformer-based summarization model that employs Gap-Sentence Generation (GSG), where key sentences are removed from a document and used as target summaries (Zhang et al., 2020). Its encoder processes the modified input, and the decoder reconstructs the missing content using self-attention and cross-attention mechanisms. Fine-tuning involves minimizing cross-entropy loss between predicted and reference summaries.

Document Summarization

Document summarization primarily follows two paradigms: extractive summarization, which selects key fragments directly from the source text, and abstractive summarization, which paraphrases and compresses content to produce fluent, informative summaries(Zhang et al., 2020). Hybrid approaches combine both, often identifying salient sentences before refining them through paraphrasing, as demonstrated by Chen and Bansal, 2018 (Chen and Bansal, 2018). The emergence of transformer-based models has significantly advanced abstractive summarization, particularly for long-form documents. Notably, BART (Lewis et al., 2020) introduces noise to input sequences and learns to reconstruct the original text, while PEGASUS (Zhang et al., 2020) masks key sentences and trains the model to generate them from the remaining context, enabling high-quality summaries validated through human evaluation. These models exemplify the shift toward pre-trained, self-supervised architectures that emphasize document understanding and dynamic attention to relevant content.

OCR–Summarization Interface

Despite the growing adoption of Optical Character Recognition (OCR) in heritage digitization efforts (Terras, 2011), OCR outputs often contain transcription errors, ranging from character-level misrecognitions to full-word or sentence distortions, that propagate into downstream tasks such as summarization (van Strien et al., 2020; Li et al., 2023). These errors, including substitutions, insertions, deletions, and real-word confusions, disproportionately affect short words and first characters (Kukich, 1993; Jatowt et al., 2019; Young et al., 1991), and can severely degrade the lexical and syntactic integrity of summarization inputs (Khurana et al., 2023). While post-processing techniques and multimodal end-to-end architectures have been proposed to mitigate these effects (Xu et al., 2021; Appalaraju et al., 2021), most systems still treat OCR and summarization as disjointed components, leading to fragmented performance and limited robustness to input noise (Kim et al., 2022). These limitations underscore the need to systematically assess how input distortions, particularly blur and character-level noise, affect the performance of transformer-based summarization models.

Research Gap and Motivation

Transformer-based models such as BART and PEGASUS (Lewis et al., 2020; Zhang et al., 2020) consistently achieve state-of-the-art performance in document summarization, particularly under noisy conditions introduced by OCR. Their ability to model long-range dependencies and recover disrupted context makes them more resilient to spelling errors, incomplete sentences, and structural distortions. Recent advancements include models fine-tuned for OCR tasks like LayoutLMv2 and TrOCR (Xu et al., 2021; Li et al., 2023), as well as OCR-free architectures such as DONUT and DocFormer (Appalaraju et al., 2021; Kim et al., 2022), which process raw images directly. These approaches enable robust summarization pipelines through post-correction, noise-aware training, and multimodal integration, though domain-specific fine-tuning remains critical for handling extreme OCR degradation. However, there remains a notable gap in large-scale datasets that pair scanned documents with ground-truth OCR outputs and corresponding reference summaries, limiting the ability to jointly evaluate OCR and summarization performance. High-quality reference summaries and human evaluation are essential for assessing factual consistency, particularly in abstractive summarization(Yu et al., 2021; Zhang et al., 2024; Durmus et al., 2020). While benchmark pipelines typically involve OCR processing, summarization, and evaluation using metrics such as ROUGE (Cohan et al., 2018), few studies address this full pipeline. Moreover, although datasets like NEWSROOM, SROIE, PubLayNet, and XSUM support individual components (Grusky et al., 2018; Li et al., 2023; Zhong et al., 2019), benchmarks for noisy document domains and multimodal learning remain underdeveloped.

Given these limitations, it becomes essential to understand how OCR-induced distortions affect downstream summarization quality. This motivates our central research question: What impact do varying levels of distortions and blur have on the summarization performance of transformer-based models?

Methodology

Research Design

This study addresses the intersection of OCR and machine learning by analysing summaries generated by transformer-based models under noisy conditions. The summarization

stage was fixed, while synthetic distortions and blur were introduced during OCR to evaluate model robustness. The central hypothesis posited that TrOCR would outperform DONUT in text extraction accuracy under noise due to its architecture and contextual modelling capabilities. The empirical workflow involved applying both models to a shared dataset of scanned document images, measuring OCR output fidelity using Character Accuracy Rate and Word Error Rate. The transcribed text was then passed to a summarization model, enabling a controlled assessment of document understanding performance across varying noise levels. The experimental design followed a three-stage pipeline: OCR processing, summarization, and evaluation.

Dataset and Noise Simulation

To evaluate model robustness under noisy conditions, distortions were introduced by applying random scribbles and varying levels of blur[1] to the original document images. These distortions were generated using a custom Python script that accepts an integer input to control the intensity of scribbles and blur. The analysis began with simpler one-page document layouts and was later extended to more complex multi-page formats. A clean benchmark dataset was assembled using a custom set of scanned PDFs paired with reference summaries sourced exclusively from the XSUM dataset, which provides concise, single-sentence summaries for news articles and contains approximately 226,000 document-summary pairs.

Level	Scribbles (*strokes/line*)	Blur (*intensity*)
Medium	5	1.2
High	8	1.4
Extreme	12	2

Table 1: Experiment distortion intensity levels

OCR Models

We started by implementing Optical Character Recognition using two transformer-based models: TrOCR and DONUT. TrOCR is a transformer OCR model with a Vision Transformer (ViT) encoder and mBART decoder, while DONUT is an OCR-free, end-to-end document understanding model. Both models were initialized with pre-trained weights from Hugging Face and tested on clean and synthetically distorted document images. The evaluation focused on comparing their robustness to noise and their effectiveness in extracting or generating text under varying input conditions.

Summarization and Evaluation

In our implementation, we fixed the summarization stage while varying the OCR inputs according to different levels of induced noise. A single summarization model was used across all tests. The evaluation of summarization quality relied on automated metrics such as ROUGE scores (Zhang et al., 2020), which is a software package used for evaluating automatic summarization in language processing. It mainly measures overlap between the generated text and the reference text (using recall, precision and F1 score), and it is a common practice applied by previous researchers (Cohan et al., 2018; Lewis et al., 2020).

Our analysis involved comparing summaries generated from clean document inputs with those produced under noisy conditions. We then correlated OCR error rates with the observed decline in summarization quality, enabling a quantitative assessment of how input degradation impacts downstream performance, and also enabling us to assess and compare the relative robustness of each model.

Experiments

Experimental Setup

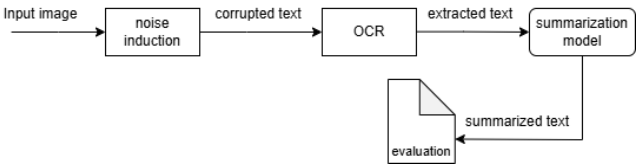


Figure 2: Experiment pipeline

All experiments were conducted using PyTorch 1.10.2 and executed on a single NVIDIA GPU (CUDA 11.3) using Python 3.9. Pretrained transformer-based models were obtained from the Hugging Face library without any fine-tuning to maintain comparability under zero-shot conditions. TrOCR-large-handwritten and DONUT-base were employed for OCR, while PEGASUS-xsum was used for summarization. Image distortions were simulated using Python-based augmentation scripts that applied random scribbles and blur at three levels of intensity (medium, high, and extreme) as shown in Table 1. Evaluation was performed on thirty manually curated document images with corresponding ground-truth text and summaries, composing a total of 300 experiments.

Evaluation and Testing Procedure

Each document was processed under four noise levels (clean + three distortions) by both OCR models independently. TrOCR and DONUT produced text outputs from the same image set, ensuring a fair comparison. For each OCR output, accuracy was evaluated against the ground-truth reference text using the following metrics:

- **Character Accuracy Rate (CAR)**: proportion of correctly recognized characters over the total.
- **Word Error Rate (WER)**: normalized edit distance (substitutions, insertions, deletions) between OCR and reference text.
- **Character Error Rate (CER)**: similar to WER but computed at the character level.

The resulting transcriptions were then passed into the PEGASUS summarizer. Summarization quality was evaluated using the metrics bellow to compare generated summaries

[1]Blur: loss of sharpness in an image.

to their ground-truth references:

- **ROUGE-1**: overlap of unigrams (single words)
- **ROUGE-2**: overlap of bigrams (two consecutive words)
- **ROUGE-L**: longest common subsequence
- **BERTScore-F1**: tells us about semantic similarity

Results and Discussion

OCR Performance Comparison

Model	WER ↓	CER ↓	CAR ↑
TrOCR (Clean)	0.28	0.075	0.92
TrOCR (Medium)	0.45	0.23	0.77
TrOCR (High)	0.61	0.37	0.63
TrOCR (Extreme)	0.74	0.53	0.47
DONUT (Clean)	1.82	0.92	0.08
DONUT (Medium)	1.59	0.83	0.17
DONUT (High)	1.93	0.88	0.12
DONUT (Extreme)	2.09	1.03	0

Table 2: OCR performance across different scribbles levels. Lower WER/CER and higher CAR indicate better accuracy.

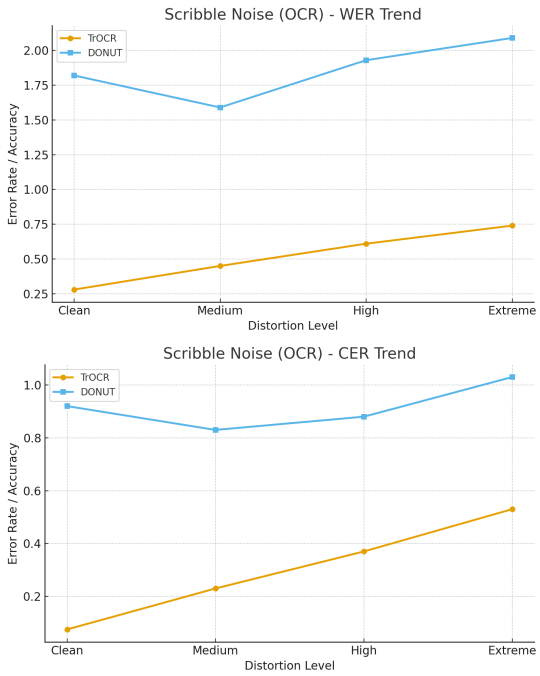


Figure 3: Graphs showing WER and CER trend across varying levels of scribbles.

TrOCR maintains relatively low WER (0.28 → 0.74) and CER (0.075 → 0.53) across noise levels, showing gradual degradation. DONUT, however, starts with much higher error rates (WER ≈ 1.8) and fluctuates inconsistently across noise conditions, with CER exceeding 1.0 at extreme scribbles. Despite TrOCR's lexical precision, CAR for both models declines as noise increases; TrOCR drops from 0.92 → 0.47, while DONUT saturates near zero.

TrOCR's transformer-based decoder with language priors enables more robust token recovery even under occlusion, while DONUT's visually grounded encoder suffers when scribbles obscure text contours. This confirms TrOCR's encoder-decoder structure with image-transformer encoding provides greater robustness to visual corruption.

Model	WER ↓	CER ↓	CAR ↑
TrOCR (Clean)	0.28	0.075	0.92
TrOCR (Medium)	0.29	0.08	0.92
TrOCR (High)	0.3	0.09	0.91
TrOCR (Extreme)	0.42	0.2	0.8
DONUT (Clean)	1.82	0.92	0.08
DONUT (Medium)	0.96	0.85	0.15
DONUT (High)	2.2	1.13	0
DONUT (Extreme)	2.6	1.12	0

Table 3: OCR performance across different blur levels. Lower WER/CER and higher CAR indicate better accuracy.

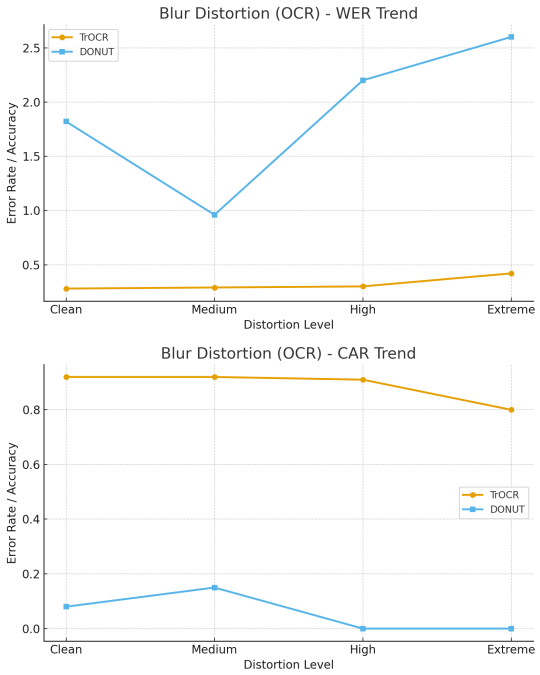


Figure 4: Graphs showing WER and CAR trend across varying levels of blur.

TrOCR exhibits remarkable stability up to medium blur (WER ≈ 0.28 – 0.30; CAR ≈ 0.91 – 0.92), only degrading sharply at extreme blur (WER = 0.42). DONUT behaves irregularly: its WER briefly improves at medium blur (1.82 → 0.96) but worsens drastically thereafter (2.6 at extreme

blur). CAR collapses to 0, implying near-total text loss.

TrOCR's textual attention layers appear resistant to Gaussian blur up to moderate levels, due to its text-centric pretraining. DONUT's performance volatility indicates over-reliance on sharp visual cues.

Summarization Performance

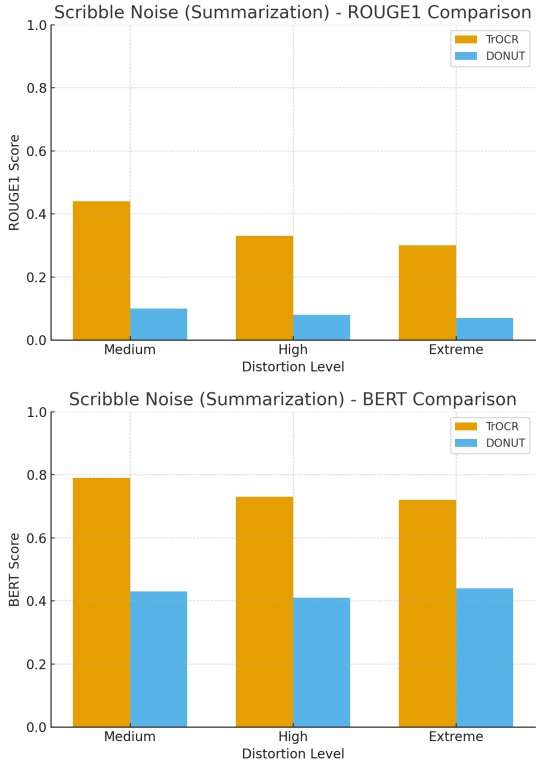


Figure 5: Bar graph showing ROUGE-1 and BERTScore summarization performance under scribble corruption 5.

TrOCR's summarization metrics degrade steadily but remain strong (ROUGE-1: 0.44 $\rightarrow$ 0.30; BERTScore: 0.79 $\rightarrow$ 0.72). DONUT's summarization collapses (ROUGE-1 $\approx$ 0.1 $\rightarrow$ 0.07), indicating weak downstream semantic consistency once OCR text deteriorates. Interestingly, DONUT's BERTScore ($\approx$ 0.43 – 0.44) remains slightly stable, suggesting some retained semantic alignment despite severe lexical corruption.

While TrOCR's higher text fidelity sustains the PEGASUS summarizer's input quality, DONUT's visually embedded features may preserve minimal conceptual information even with poor transcription. TrOCR maintains high ROUGE scores across medium and high blur (ROUGE-1: 0.51–0.52; BERT: 0.81) and only dips under extreme blur (ROUGE-1 = 0.41). DONUT's scores are uniformly low (ROUGE-1 $\leq$ 0.09), and further deteriorate under extreme blur (ROUGE-1 = 0.04, BERT = 0.39).

Summarization pipelines remain viable with TrOCR-derived text under moderate blur. DONUT's output, even when intelligible, fails to convey sufficient lexical detail for meaningful abstraction.

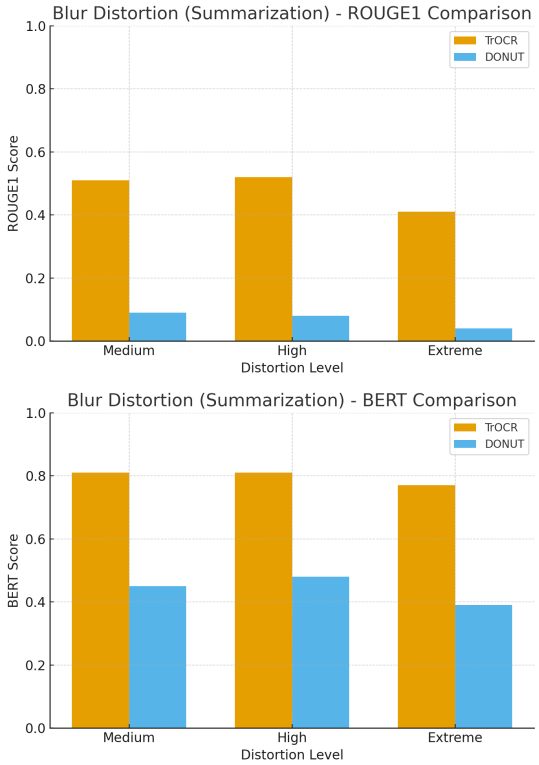


Figure 6: Bar graph showing ROUGE-1 and BERTScore summarization performance under Gaussian blur corruption 6.

Qualitative Analysis

Example outputs revealed that TrOCR primarily suffered from punctuation misinterpretation and word merging under extreme blur, while DONUT occasionally produced hallucinated phrases absent from the source. At moderate noise levels, both systems preserved the documents' overall semantic structure, though with varying degrees of lexical fidelity. These observations, summarized in Table 4, highlight complementary strengths: TrOCR maintains superior character-level accuracy, whereas DONUT's multimodal design allows partial semantic recovery under visual degradation. When coupled with a fixed summarization stage, PEGASUS further reinforces semantic abstraction, mitigating some of the OCR-induced noise.

Limitations and Future Work

This study is limited by its focus on single-page documents and the use of synthetic noise, which may not fully capture the variability and complexity of real-world document degradation. Both OCR models were evaluated in zero-shot mode without fine-tuning, potentially underutilizing their adaptation capacity. Furthermore, summarization quality was assessed solely through automated metrics (ROUGE and BERTScore), which, while informative, may overlook subtleties of factual accuracy, readability, and coherence.

Future work should expand the dataset to encompass a

Observation	Explanation
TrOCR's WER and CER increase linearly with noise	Reflects predictable degradation under controlled distortion, indicating stable response to input perturbation.
DONUT's metrics fluctuate non-linearly	Its vision-based encoder-decoder architecture is more sensitive to spatial occlusion and scribble intensity.
BERTScore declines more gradually than ROUGE	Semantic similarity remains more robust than lexical overlap, suggesting that meaning preservation persists despite text errors.
Blur affects DONUT more than scribbles	Visual blur disrupts spatial feature encoding, while scribbles may still preserve character edge information useful for recognition.

Table 4: Observed trends across OCR and summarization metrics for TrOCR and DONUT.

broader range of document domains, multi-page layouts, and naturally degraded samples. Pipeline efficiency can be improved by integrating PyTesseract to process full-page images rather than line-by-line segments, reducing computational overhead. Another promising direction is the joint fine-tuning of OCR and summarization models on noisy inputs, enabling the development of end-to-end noise-aware systems. Finally, incorporating human evaluation of summary quality and factual consistency would strengthen the interpretability and reliability of the findings.

Conclusion

This research examined the impact of OCR degradation on transformer-based summarization pipelines using TrOCR and DONUT. By inducing controlled noise in scanned document images, we observed how OCR inaccuracies propagate into downstream summarization. TrOCR demonstrated superior transcription accuracy, whereas DONUT exhibited partial semantic resilience under severe noise. These findings highlight a trade-off between textual precision and contextual robustness in multimodal document understanding. The study underscores the importance of noise-resilient architectures and motivates further exploration of OCR-free or jointly trained models for real-world document processing tasks.

Acknowledgments

This work was conducted under the guidance of Dr. Steven James and Dr. Benjamin Rosman, whose support and insight were instrumental to the success of this research.

Appendix

GitHub Repository

The pipeline used for the experiments in this research is available at: souoGael/Research-Pipeline

Additional visualizations of the experimental results are presented here for completeness.

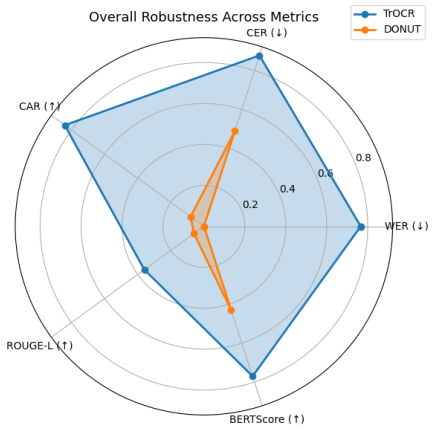


Figure 7: Radar graph showing the overall robustness across all metrics.

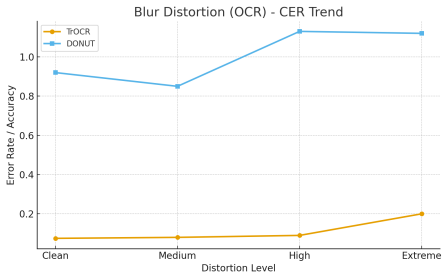


Figure 8: Graph of CER trend across varying levels of scribbles.

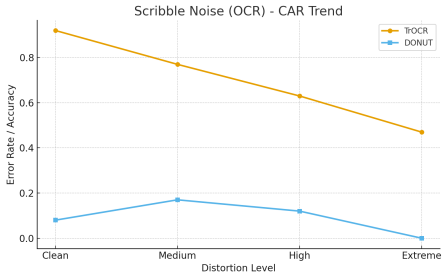


Figure 9: Graph of CAR trend across varying levels of scribbles.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
TrOCR (Medium)	0.44	0.24	0.39	0.79
TrOCR (High)	0.33	0.14	0.25	0.73
TrOCR (Extreme)	0.30	0.12	0.24	0.72
DONUT (Medium)	0.10	0.01	0.06	0.43
DONUT (High)	0.08	0.02	0.07	0.41
DONUT (Extreme)	0.07	0.01	0.06	0.44

Table 5: Evaluation of PEGASUS summarization metrics for TrOCR and DONUT across varying levels of scribbles.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
TrOCR (Medium)	0.51	0.28	0.43	0.81
TrOCR (High)	0.52	0.30	0.46	0.81
TrOCR (Extreme)	0.41	0.20	0.35	0.77
DONUT (Medium)	0.09	0.02	0.08	0.45
DONUT (High)	0.08	0.03	0.06	0.48
DONUT (Extreme)	0.04	0	0.04	0.39

Table 6: Evaluation of PEGASUS summarization metrics for TrOCR and DONUT across varying levels of blur.

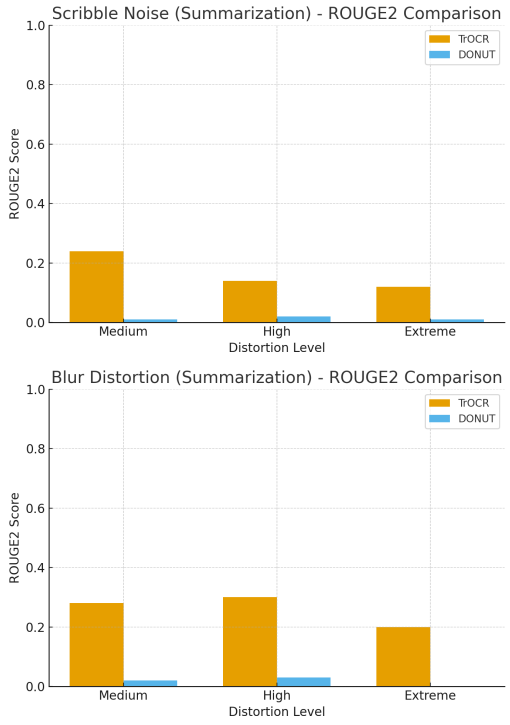


Figure 10: Graphs of ROUGE-2 comparison across varying levels of blur.

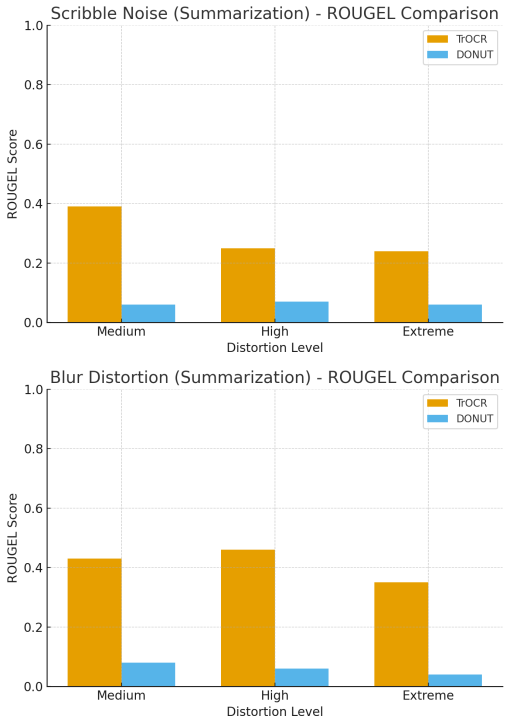
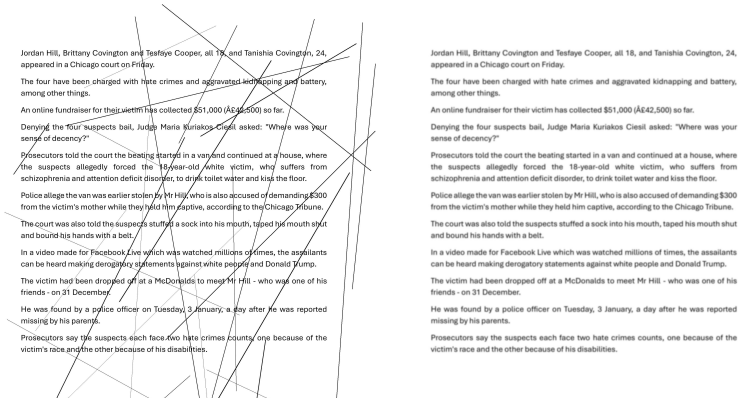


Figure 11: Graphs of ROUGE-L comparison across varying levels of blur.

(a) (b) (c) (d)

Figure 12: The figures above illustrate the overall research pipeline. The first image shows the original sample, followed by the model's visual interpretation of that sample, processed line by line. The final two images present the extracted outputs generated by the respective models — TrOCR and Donut.



(a) Scribbles (b) Gaussian blur

Figure 13: The figures above represent the sample at a medium level of distortion.

References

[Appalaraju et al., 2021] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., and Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

[Blecher et al., 2023] Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. (2023). Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

[Chen and Bansal, 2018] Chen, Y.-C. and Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 675–686.

[Cohan et al., 2018] Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[Durmus et al., 2020] Durmus, E., He, H., and Diab, M. (2020). Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

[Grusky et al., 2018] Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

[Jatowt et al., 2019] Jatowt, A., Coustaty, M., Nguyen, N.-V., Doucet, A., et al. (2019). Deep statistical analysis of ocr errors for effective post-ocr processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 29–38. IEEE.

[Khurana et al., 2023] Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. volume 82, pages 3713–3744. Springer.

[Kim et al., 2022] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. (2022). Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

[Kukich, 1993] Kukich, K. (1993). Techniques for automatically correcting words in text. In *Proceedings of the 1993 ACM conference on Computer science*, page 515.

[Lewis et al., 2020] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7871. Association for Computational Linguistics.

[Li et al., 2023] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2023). Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

[Liao et al., 2023] Liao, H., RoyChowdhury, A., Li, W., Bansal, A., Zhang, Y., Tu, Z., Satzoda, R. K., Manmatha, R., and Mahadevan, V. (2023). Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19584–19594.

[Liu and Lapata, 2019] Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

[Smith, 2007] Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

[Terras, 2011] Terras, M. M. (2011). The rise of digitization. In *Digitisation perspectives*, volume 39, pages 3–20. SensePublishers.

[van Strien et al., 2020] van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., Mcgillivray, B., Colavizza, G., et al. (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *ICAART 2020-Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, volume 1, pages 484–496. SciTePress.

[Xu et al., 2021] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al. (2021). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Lin-*

guistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics.

[Young et al., 1991] Young, C. W., Eastman, C. M., and Oakman, R. L. (1991). An analysis of ill-formed input in natural language queries to document retrieval systems. In *Information Processing and Management*, volume 27, pages 615–622. Elsevier.

[Yu et al., 2021] Yu, W., Lu, N., Qi, X., Gong, P., and Xiao, R. (2021). Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.

[Zhang et al., 2020] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

[Zhang et al., 2024] Zhang, T., Ladhak, F., Durmus, E., Liang, P., Mckeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 39–57.

[Zhong et al., 2019] Zhong, X., Tang, J., and Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE.