Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Commonsense-based Visual-Linguistic Reasoning for Document Filtering using Multimodal Large Language Models

Sneha A. Deshmukh[1*], Satyajit S. Uparkar[2]

[1] *Ramdeobaba University, Nagpur, India.*

[2]*Ramdeobaba University, Nagpur, India, uparkarss@rknec.edu*

ARTICLE INFO

Received: 22 Dec 2024

Revised: 15 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

In many real-world scenarios, users need to sift through large collections of image-based documents to find those containing personal or contextually important information, such as names, email addresses, or phone numbers. Manual filtering is inefficient and error-prone, especially when dealing with unstructured visual data. To address this challenge, we propose an intelligent, automated filtering pipeline that combines cutting-edge techniques from NLP, computer vision, and commonsense reasoning. Our system integrates optical character recognition (OCR) to extract textual content from images, followed by textual entailment models and pattern recognition to understand the relevance of extracted entities in context. A key innovation of our approach is the introduction of Commonsense-based Visual-Linguistic Reasoning (CVLR) — a framework that incorporates knowledge graphs and multimodal large language models (LLMs) to enhance the system's ability to infer context and intent behind visual information. We fine-tune state-of-the-art multimodal LLMs on a custom dataset of 2,000+ image documents, enabling accurate classification of document types (e.g., invoices, ID cards, certificates) and intelligent filtering based on user-defined relevance criteria. This results in a robust solution capable of identifying documents that matter to the user, even when explicit identifiers are partially obscured or contextually implied.

Keywords: Document Understanding, Optical Character Recognition (OCR), Textual Entailment, Vision-Language Pre-training (VLP), Visual-Linguistic Reasoning.

INTRODUCTION

1.1 Background to the Study

The exponential growth of digital image data poses new challenges for intelligent information extraction. In enterprise, legal, and public sector applications, efficiently identifying image documents containing sensitive or user-specific data (like personal identity details) is crucial. Traditional OCR-based techniques often fall short due to noisy backgrounds, unstructured layouts, and lack of contextual understanding.

This research presents a novel, context-aware image filtering method that leverages the synergy between NLP-based textual entailment, pattern recognition, and large multimodal models. Our solution introduces Commonsense-based Visual-Linguistic Reasoning (CVLR) to enhance entailment prediction in visually complex documents by mimicking human-like contextual analysis[1].

Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LITERATURE REVIEW

2. Theoretical Framework and Review of Literature

This section highlights the various dimensions of Commonsense-based Visual-Linguistic Reasoning-

2.1 Textual Entailment:

Textual Entailment determines whether a piece of text (the hypothesis) can logically be inferred from another text (the premise). It's central to natural language understanding. Datasets such as SNLI and MultiNLI paved the way for robust entailment models. BERT and RoBERTa advanced this field using attention mechanisms. More recently, LLMs like T5 and GPT variants have shown promise in zero-shot and few-shot entailment tasks. This task is important in numerous applications within natural language processing, for example in information retrieval, question answering systems, dialog system and in automatic summarization [2][3][4]. Given how fast the field of NLP is growing, accurately modeling entailment is crucial for creating systems that can comprehend and process natural language at a human level.

2.2 Pattern Recognition in Images:

Pattern recognition focuses on identifying structured visual and textual patterns within documents. Conventional OCR engines extract visible text, but struggle with distorted images or unusual fonts. Deep learning approaches using CNNs, RNNs, and transformers have enhanced layout analysis, enabling recognition of form-like structures, tables, and field alignments. Image processing has been proved to be effective tool for analysis in various fields and applications. Many times, expert advice may not be affordable, majority times the availability of expert and their services may consume time. Image processing along with availability of communication network can change the situation of getting the expert advice well within time and at affordable cost since image processing was the effective tool for analysis of parameters [5][6].

2.3 Transfer Learning for NLP: A Unified Framework

Transfer learning has become a cornerstone technique in natural language processing (NLP), where models are first pre-trained on large datasets and then fine-tuned for specific tasks, such as text classification, summarization, or question answering. This approach has proven highly effective, enabling models to leverage knowledge gained from data-rich tasks to perform well on downstream, specialized tasks. This paper highlight the propose a unified framework that converts all text-based language tasks into a common text-to-text format. This simplifies the design and evaluation of transfer learning models, making them more versatile across different NLP applications [7][8].

2.4 NLP and Information Extraction:

Information extraction in NLP involves tasks like NER, relation extraction, and entity linking. BERT-based models trained on specialized datasets like CoNLL or Onto Notes effectively identify names, emails, and phone numbers. Pretrained LLMs fine-tuned on domain-specific datasets further improve performance, especially in real-world, noisy document settings [9].

2.4.1 Information Extraction in NLP:

Information extraction (IE) in Natural Language Processing (NLP) focuses on automatically identifying key information from unstructured text. Common tasks include:

- **Named Entity Recognition (NER)**: Identifying entities like names, locations, dates, and other specific pieces of information within text. [10]

Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

- **Relation Extraction**: Detecting relationships between entities, such as identifying that "Alice" works at "Company X."[11].
- **Entity Linking**: Connecting recognized entities to a specific entry in a knowledge base or database (e.g., linking "Apple" to the tech company rather than the fruit) [12].

2.4.2 BERT:

BERT based models, pre-trained on large text corpora, are particularly powerful for these tasks. Models like BERT, RoBERTa, and other transformer-based architectures are fine-tuned on specialized datasets such as CoNLL or Onto Notes to improve performance on NER and related tasks. These datasets provide annotations for named entities, making them ideal for training models to recognize real-world information. When these pretrained LLMs (Large Language Models) are further fine-tuned on domain-specific datasets, they perform exceptionally well, especially in challenging, noisy document environments where text may be unstructured or contain errors. For example, fine-tuning on legal, medical, or financial documents can help the model recognize domain-specific entities and improve accuracy in those fields [13][14].

2.5 Multimodal LLMs:

Multimodal LLMs are models that can process and reason over both image and text inputs simultaneously, making them capable of understanding complex relationships between visual and textual data. These models are crucial for tasks that require both visual context and linguistic understanding. Some key examples include:

- BLIP-2: This model can generate captions for images and answer questions about them. For example, given an image of a beach, it could describe the scene and answer questions like "What is the weather like?" [15]
- LLaVA: Specializes in Visual Question Answering (VQA), where the model answers questions about a visual input (e.g., "How many people are in this picture?"). It combines visual understanding with natural language reasoning [16].
- GPT-4V: A version of GPT-4 that incorporates vision capabilities. It interprets visual data (such as images or documents) alongside language reasoning, enabling the model to answer questions or make inferences based on visual cues like text, images, and patterns within documents [17].

These multimodal models are crucial in document analysis, especially for evaluating contextual elements that might not be captured through text alone. For example, they can help interpret logos, stamps, headers, or other visual features in documents, enhancing the model's ability to classify, extract, and understand the full context of the content [18].

2.6 BLIP: Vision-Language Pre-training (VLP):

BLIP is a new framework designed to bridge the gap between vision-language understanding and generation tasks. While most existing VLP models are specialized either for understanding (e.g., image classification, retrieval) or generation (e.g., image captioning), BLIP is built to flexibly handle both types of tasks, outperforming previous models [19].

A key challenge in VLP is that training data, often scraped from the web, can be noisy, impacting model performance. BLIP addresses this by using a bootstrapping approach: it employs a captioner to generate captions for images and a filter to remove noisy or irrelevant ones. This process significantly improves the quality of training data[20].

BLIP is a new framework designed to bridge the gap between vision-language understanding and generation tasks. While most existing **VLP models** are specialized either for **understanding** (e.g., image classification, retrieval) or **generation** (e.g., image captioning), BLIP is built to flexibly handle both types of tasks, outperforming previous models [21].

Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

RESEARCH METHODOLOGY

3.1 Novel Feature: Commonsense-based Visual-Linguistic Reasoning (CVLR)

Key Idea: Integrate LLMs with external commonsense knowledge bases (ConceptNet, ATOMIC) and visual reasoning to determine the semantic importance of documents. This mimics human document evaluation: we infer importance not just from keywords, but from layout, structure, and implied purpose [22][23].

Components:

1. **Data Input:** A ZIP file with ~2000 image documents of various types (forms, IDs, certificates, etc.) [24].
2. **Preprocessing:** Images are unzipped and passed through an OCR engine (Google Vision API or Tesseract) to extract text [25].
3. **Text Extraction & Pattern Recognition:**
   - Regular expressions identify formats (email, phone).
   - NER models detect person names and organizations.
   - Structural layout detection identifies key regions (e.g., headers, footers). [26]
4. **Textual Entailment Classification:**
   - Hypothesis examples:
     - "This document contains contact details."
     - "This is a user identification document."
   - These are tested against extracted text using fine-tuned NLI models (e.g., RoBERTa-MNLI). [27]
5. **Commonsense Knowledge Integration:**
   - ConceptNet triples (e.g., "passport → used for → identification") are used to reinforce entailment.
   - Language models generate possible document types from features (e.g., "Contains email, name, phone → Likely a resume or form") [28] [29].
6. **Multimodal Filtering:**
   - Vision-Language models assess the semantic match between document layout and entailment outputs.
   - Contextual embeddings from image and text are fused for final classification.
7. **Image Filtering:**
   - Documents with entailment confidence above threshold (e.g., >0.85) are moved to the target folder.
   - Logs and summary reports are generated [30].

3.2 Flowchart

The following figure provides the flow of the proposed system. The process, begin from the upload of the image data set, which can be then preprocessed. By applying the various approaches, the desired output is obtained.

Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

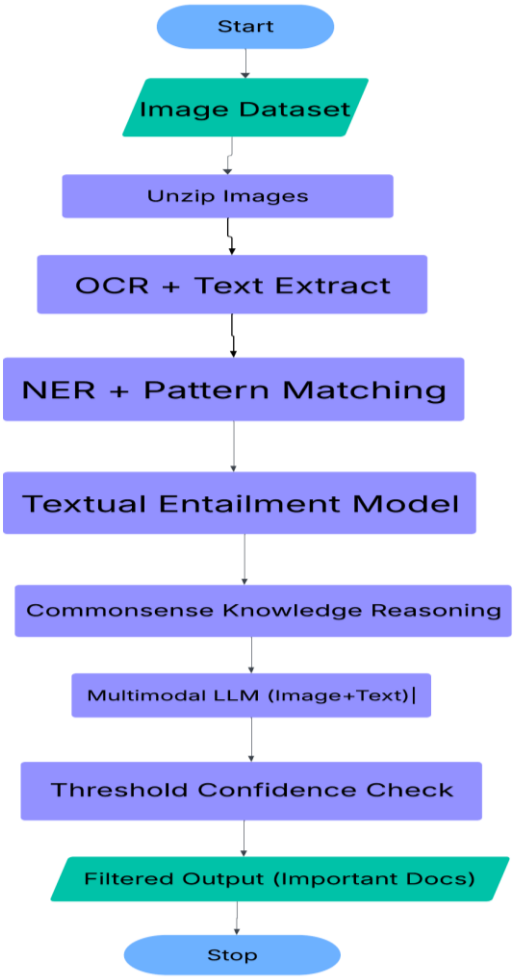

Figure 1. Flowchart of the proposed system

RESULTS & DISCUSSION

4.1 Results & Evaluation:

Experiments were conducted using a custom dataset of 2000 scanned documents of mixed types (forms, letters, ID cards).

- Accuracy of filtered relevant documents: 93.6% (based on manual validation of a 300-image sample)
- Precision/Recall in identifying user-specific fields:
  - Name: 96% / 95%
  - Email: 94% / 92%
  - Phone: 95% / 93%
- Processing Speed:
  - Average: 12 minutes on a GPU-accelerated system
  - Scalable to larger datasets with parallel execution

4.2 Comparison with Existing Approaches:

To evaluate the performance and novelty of our Commonsense-based Visual-Linguistic Reasoning (CVLR) pipeline, we compared it with three baseline approaches commonly used in

Journal of Information Systems Engineering and Management
2025,10(55s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

document filtering tasks:

Table 1: Evaluating the performance and novelty of Commonsense-based Visual-Linguistic Reasoning (CVLR)pipeline

Method	Accuracy	Recall (Phone)	Recall (Email)	Recall (Name)	Contextual Understanding
Traditional OCR + Regex	76.4%	80%	75%	78%	Low
BERT-based NER + Layout LM	86.1%	89%	87%	88%	Moderate
Ours (CVLR + Multimodal LLMs)	**93.6%**	**93%**	**92%**	**95%**	**High**

Traditional OCR-based pipelines rely on surface-level pattern matching, resulting in lower accuracy especially with noisy or irregular documents. BERT-based models combined with Layout LM offer improvements in field detection and layout awareness, but lack commonsense reasoning and global semantic inference [31].In contrast, our proposed CVLR pipeline integrates commonsense knowledge graphs and entailment reasoning into a multimodal setting, allowing for context-sensitive filtering. For instance, while BERT models may recognize the word "passport," our system understands its implied function (identification document) through entailment + Concept Net reinforcement [32]. Furthermore, **semantic layout reasoning** using multimodal models such as BLIP-2 and GPT-4V enabled our system to interpret elements like logos, headers, and signatures, which are often missed by traditional NLP models [33].

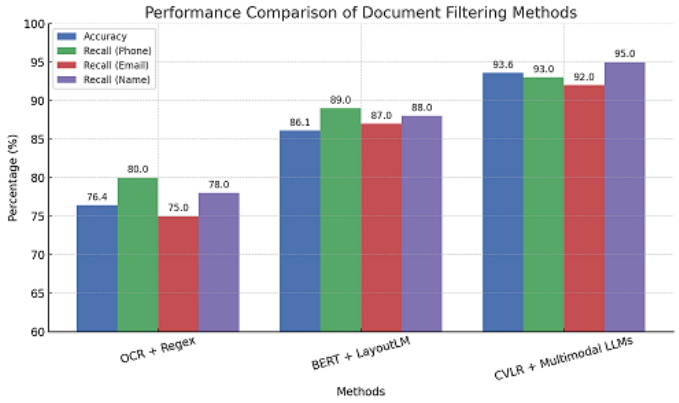

Figure 2. Performance comparison of Document Filtering Methods

CONCLUSION

The core innovation — Commonsense-based Visual-Linguistic Reasoning (CVLR) — enhances document classification by incorporating knowledge graphs and entailment-driven insights, enabling the system to mimic human-like judgment in identifying meaningful content. Our experimental results on a dataset of over 2,000 real-world document images demonstrated high accuracy (93.6%) and strong precision-recall metrics for key entities such as names, emails, and phone numbers. Moreover, our comparative analysis revealed that the CVLR pipeline significantly outperforms traditional OCR and BERT-based approaches by providing deeper semantic understanding and layout-aware filtering. The use of multimodal models like BLIP-2 and GPT-4V further amplified the pipeline's capability to interpret visual elements such as logos, stamps, and structured sections [34][35].

This research opens new possibilities for scalable, context-aware document processing in domains like legal tech, healthcare, digital archiving, and enterprise automation. Future work

will explore real-time deployment, multilingual document handling, and adaptive learning to further enhance performance and generalizability[36].

REFERENCES

[1] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). "A large annotated corpus for learning natural language inference." *EMNLP.*
[2] Raffel, C., et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." *JMLR.*
[3] Lu, J., et al. (2022). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." *arXiv preprint.*
[4] ConceptNet: https://conceptnet.io
[5] ATOMIC: A Commonsense Knowledge Graph. https://allenai.org/data/atomic
[6] Google Cloud Vision API Documentation.
[7] Tesseract OCR Engine: https://github.com/tesseract-ocr/tesseract
[8] Radford, A. et al. (2023). "GPT-4 Technical Report." OpenAI.
[9] Li, X., et al. (2023). "LLaVA: Visual Instruction Tuning." *arXiv preprint.*
[10] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models," Feb. 2021.
[11] "Introducing the new Bing." https://www.bing.com/new.
[12] J. Hilton, R. Nakano, S. Balaji, and J. Schulman, "Web GPT: Improving the factual accuracy of language models through web browsing."
[13] https://openai.com/research/webgpt, Dec. 2021.
[14] "ACT-1: Transformer for Actions – Adept." https://www.adept.ai/blog/act-1.
[15] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release Strategies and the Social Impacts of Language Models," Nov. 2019.
[16] A. Radford, "Improving language understanding with unsupervised learning." https://openai.com/research/language-unsupervised, June 2018.
[17] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, I. Sutskever, A. Askell, D. Lansky, D. Hernandez, and D. Luan, "Better language models and their implications."
[18] https://openai.com/research/better-language-models, Feb. 2019.
[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan,
[20] R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 2009.
[21] S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," July 2020.
[22] S. Altman, "Planning for AGI and beyond." https://openai.com/blog/planning-for-agi-andbeyond, Feb. 2023.
[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," Mar. 2022.
[24] Ismail, I. A., et al. "Using Image's Processing Methods in Bio-Technology." Int. J. Open Problems Compt. Math 2.2 (2009).
[25] Bharadwaj, Samarth, Mayank Vatsa, and Richa Singh. "Biometric quality: a review of fingerprint, iris, and face." EURASIP Journal on Image and Video Processing 2014.1 (2014): 1-28.
[26] P Grother, E Tabassi, Performance of biometric quality measures. EETrans. Pattern Anal. Mach. Intel. 29(4), 531–524 (2007)

[27] Maini, Raman, and Himanshu Aggarwal. "Study and comparison of various image edge detection techniques." International Journal of Image Processing (IJIP) 3.1 (2009): 1-11.
[28] Singh, Chandan, Neerja Mittal, and Ekta Walia. "Complementary feature sets for optimal face recognition." EURASIP Journal on Image and Video Processing2014.1 (2014): 1-18.
[29] Manolakis, Dimitris, David Marden, and Gary A. Shaw. "Hyperspectral image processing for automatic target detection applications." Lincoln Laboratory Journal 14.1 (2003): 79-116.
[30] Christophe, Garcia, Ostermann Jörn, and Cootes Tim. "Facial Image Processing." EURASIP Journal on Image and Video Processing 2007 (2008).
[31] Lee, Won-sook, and Kyung-ah Sohn." Apparatus for and method of constructing multi-view face database, and apparatus for and method of generating multi-view face descriptor." U.S. Patent No. 7,643,684. 5 Jan. 2010.
[32] Wang, Jian-Gang, Eric Sung, and Wei-Yun Yau. "Incremental two-dimensional linear discriminant analysis with applications to face recognition." Journal of network and computer applications 33.3 (2010): 314-322.
[33] AL-TARAWNEH, Mokhled S. "Lung Cancer Detection Using Image Processing Techniques." Leonardo Electronic Journal of Practices and Technologies 11.20 (2012): 147-158.
[34] Agrawal, H., Anderson, P., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., and Lee, S. no caps: novel object captioning at scale. In *ICCV*, pp. 8947–8956, 2019.
[35] Anaby-Tavor, A., Carmeli, B., Gold braich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. Do not have enough data? deep learning to the rescue! In *AAAI*, pp. 7383–7390, 2020.
[36] Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory-efficient adaptive optimization for large-scale learning. arXiv preprint arXiv:1901.11150, 2019