CS-FIG: An Unsupervised Classification on Figures From CS Articles

Sunku Bhanu Kedhaar Nath
Z1974769@students.niu.edu

Abstract: The vast and growing volume of scientific literature necessitates advanced methods for efficient information retrieval and comprehension. This paper introduces CS-FIG, a novel dataset and framework aimed at the unsupervised classification of figures in computer science articles. Leveraging figures extracted from the ArXiv dataset using PDFFigures 2.0, we address key limitations of existing datasets, notably the scarcity of labeled architectural diagrams and the ambiguity in figure types. CS-FIG enriches this dataset with comprehensive metadata, enabling precise classification through a hybrid approach combining Google ViT, Swin Transformer models, and a custom model for processing textual captions. We employ an ensemble method that integrates model predictions with enhancements from GPT-4.0 and human annotations to refine labels. Preliminary results demonstrate the efficacy of our approach in classifying complex figures, facilitating faster and more accurate retrieval of architectural diagrams. This framework not only serves as a valuable tool for researchers but also sets the stage for further advancements in automated document analysis.

1. Introduction

The proliferation of digital scientific literature has made accessing and assimilating information from research articles increasingly challenging, particularly within the rapidly evolving field of computer science. Figures, especially architectural diagrams, play a crucial role in conveying complex information succinctly. However, the effective retrieval and comprehension of these figures are hindered by the lack of structured labeling and classification, which are essential for systematic review and analysis. Existing datasets that facilitate figure classification in scientific papers, such as ACL-FIG and SCI-CAP, are limited by either the quantity of specific figure types or the absence of detailed labels that delineate figure categories. These limitations complicate the training of models capable of automated figure classification and recognition, particularly in an unsupervised setting where labeled data are scarce or inconsistent.

To address these challenges, we introduce CS-FIG, a framework designed for the unsupervised classification of figures from computer science articles. Our approach harnesses the power of advanced machine learning models, including Google ViT and Swin Transformer, integrated via an ensemble method to enhance the accuracy and reliability of figure classification. Additionally, we introduce a new baseline ensemble model that combines the strengths of ViT and Swin Transformer for image-based tasks with the LLaMA model for classification using captions and labels. This ensemble model serves as the best-in-class solution, attaining superior accuracy and setting a new standard in model architecture for state-of-the-art document analysis.

This initiative is not merely an academic exercise but a foundational step towards improving the accessibility and navigability of scientific knowledge. In developing CS-FIG, we extracted a comprehensive dataset of figures from articles in the ArXiv repository, spanning papers published from 2015 onwards. This dataset was enriched with metadata extracted using PDFFigures 2.0, and refined through a novel labeling process that combines automated model predictions with human annotations and adjustments via GPT-4.0. Our results indicate that CS-FIG significantly outperforms existing approaches in classifying architectural diagrams and other complex figures, thereby facilitating a more efficient review of scientific literature. The subsequent sections of this paper detail the methodologies employed in the dataset construction, describe the architecture of the machine learning models used, including the new ensemble model, and discuss the implications of our findings for both academic research and practical applications in the field of document analysis.

2. Related Work

The classification of figures in scientific documents has garnered increasing attention as the volume of digital publications grows. Early efforts in this domain have primarily focused on traditional machine learning techniques using feature engineering to classify simple graphic elements in documents [4]. Recent advancements have shifted

toward deep learning approaches due to their superior capability in handling complex image data and unstructured text.

Image Classification in Documents: Vision Transformer (ViT) [1] and Swin Transformer [7] are among the leading models for image analysis tasks. ViT, which applies the principles of transformers directly to patches of an image [1], has demonstrated significant improvements in understanding the global context of images. The Swin Transformer, by introducing a hierarchical structure, efficiently scales to larger models and datasets, showing enhanced performance in various benchmarks [**?**].

Textual Information Processing(In-Progress): The integration of textual context into the classification process has seen less exploration but is crucial for the comprehensive analysis of figures in scientific papers. Models like LLaMA, which offer advanced capabilities in language understanding, provide a robust framework for interpreting captions and labels that accompany images, thereby enriching the classification process [8].

Ensemble Methods in Machine Learning: Ensemble methods, which combine multiple models to improve the accuracy of predictions, have been effectively used in various fields, from natural language processing to computer vision [2]. These approaches are particularly useful in scenarios where different types of data (such as images and text) need to be integrated to make informed decisions.

Gaps in Current Research: Despite these advancements, few studies have explored the synergistic use of cutting-edge models across different modalities (text and image) specifically for the classification of figures in scientific articles. Most existing datasets do not provide the granular labels necessary for training such sophisticated models, and the lack of comprehensive, labeled datasets remains a significant barrier. The work by [5] on the ACL-Fig dataset provides a significant step toward addressing this issue by presenting a dataset tailored for the classification of scientific figures. This dataset features manually annotated labels, a traditional but labor-intensive approach. We aim to replace human annotation with our novel technique of combining multiple models and leveraging GPT-4.0 as a judge to refine and verify labels automatically, which we believe will significantly enhance efficiency and scalability in dataset creation and maintenance [5]

Our work addresses these gaps by proposing a novel ensemble model that leverages both state-of-the-art image transformers and language models. This approach not only enhances the accuracy of figure classification but also paves the way for a more nuanced understanding of scientific content.

3. DataSet

The CS-FIG dataset forms a comprehensive foundation for our research, facilitating the development and rigorous evaluation of our ensemble model tailored for the classification of figures in computer science articles. With 37,000 figures extracted from approximately 4,000 CS scientific papers published from 2015 onwards, our dataset is positioned to significantly advance the field of automated figure classification due to its scale and diversity.

Data Collection: Employing the PDFFigures 2.0 tool, we meticulously extracted a rich assortment of figures from the ArXiv repository. This assortment includes various types of figures such as line charts, bar graphs, scatter plots, and architectural diagrams. Each figure is carefully paired with its corresponding caption and textual descriptions extracted directly from the articles. This meticulous pairing ensures that each figure is contextualized with the relevant explanatory text, which is crucial for the subsequent stages of classification and analysis.

Data Labeling: Our data labeling process began with the ACL-FIG dataset, which includes a smaller but crucial subset of labeled figures. We used this dataset to train two advanced image classification models: Vision Transformer (ViT) and Swin Transformer. These models were initially trained and validated on the ACL-FIG dataset to ensure they could effectively classify scientific figures, particularly architectural diagrams. The classification report from the ViT model, as shown in the **attached figure**, demonstrates strong performance on the ACL-FIG test split, highlighting the model's robustness with high precision, recall, and F1-scores across various classes. Following this validation, we then applied these trained models to our larger dataset of 37,000 figures extracted from CS scientific papers. This application involved using the models to infer labels for the figures in our dataset, leveraging their trained capabilities to generate preliminary labels across a broad array of figure types. The subset of these inferred labels were then refined through a combination of automated techniques such as GPT as judge [10] and different subset for manual review, ensuring a high level of accuracy and relevance for subsequent uses in training and validation processes.

We are still working on leveraging Llama model for classification of images by only using captions and labels

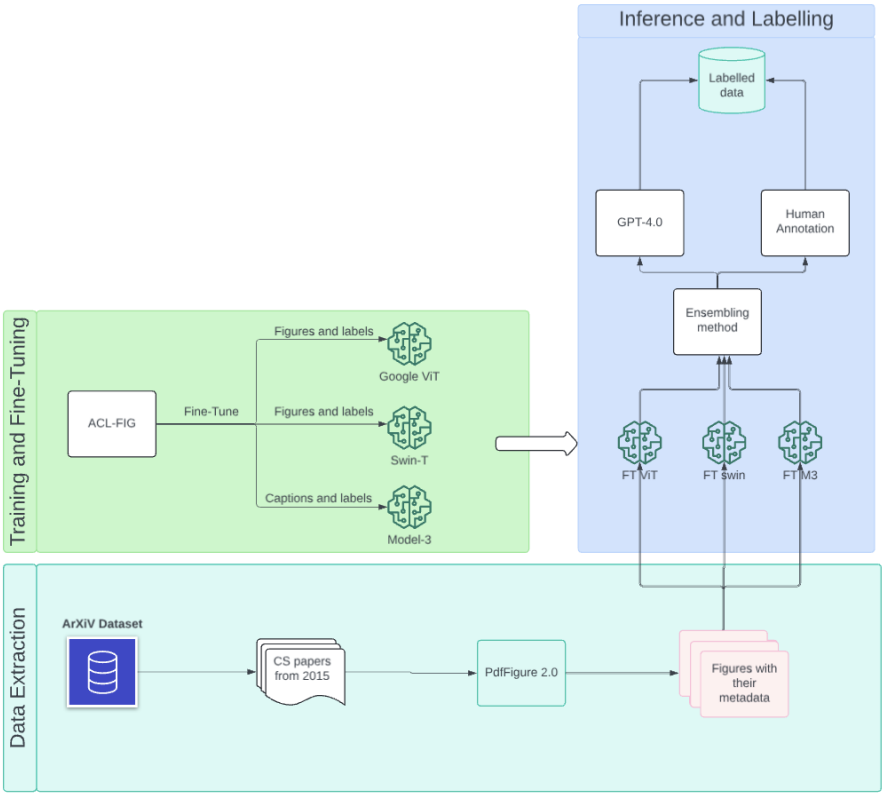

Fig. 1. Architecture of the CS-FIG dataset processing workflow.

from the ACL-FIG [5] dataset.This method of leveraging pre-trained models on a well-curated subset of data (ACL-FIG) and extending their application to a larger, more diverse dataset (CS-FIG) exemplifies a scalable and efficient approach to data labeling in machine learning, particularly in domains where labeled data are scarce. We will take the best performing model for each category based on their weighted F1 score and decide via ensembling methods and then later the output label of ensembling method will directly sent to GPT-4 for judgement which acts as another layer of fool proof and precision labeling and subset of the predicted data will be sent to manual labelling.

Preprocessing: Standardization is key in preparing the dataset for effective model training. Each figure undergoes a series of preprocessing steps designed to ensure uniformity and optimize model performance. This includes resizing images to a consistent dimension, normalizing color channels to reduce model bias due to color variations, ensuring that the input format is optimally structured for processing by our Image based models. We have preprocessed captions and finetune LLama(In-Progress) for text based classification.

Dataset Structure and Utilization: The dataset(ACL-FIG) is meticulously structured into training, validation and test sets, with the training set comprising the majority of the figures. This structure is designed to ensure that the models are not only trained on a comprehensive set of data but are also rigorously evaluated on an independent set to confirm their robustness and generalizability. The structure and utilization of the dataset are illustrated in Figure hich outlines the entire workflow from data collection to model application.

Implications for Research and Development: The creation of the CS-FIG dataset marks a significant advancement in the field of scientific document analysis. By providing an extensively labeled and well-structured dataset, we not only facilitate the development of more precise classification models but also pave the way for future research into automated analysis techniques. This dataset is expected to be a valuable resource for researchers and practitioners alike, fostering further innovations in the automated processing of scientific figures.

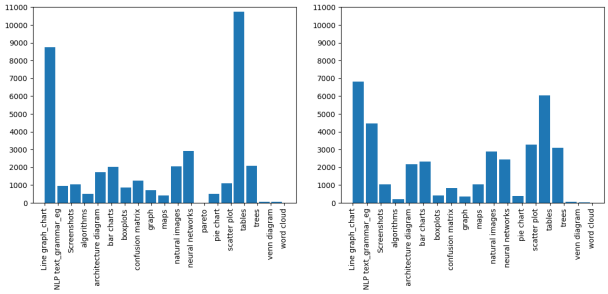

Fig. 2. Left Histogram of Google ViT, Right Histogram of Swin-Window7 of all 19 categories on CS-FIG dataset.

4. Results

The deployment of the trained Vision Transformer (ViT) and Swin Transformer models on our expansive CS-FIG dataset has delivered substantial insights into the capabilities and limitations of contemporary image classification technologies within the scientific figure classification domain.

Model Performance:

Vision Transformer (ViT): The ViT model demonstrated exemplary precision, recall, and F1-scores on the test split of the ACL-FIG dataset, which served as an initial validation before broader application. Specifically, the model achieved an average precision of 0.92, recall of 0.93, and F1-score of 0.92, showcasing its ability to effectively process and classify a wide array of figure types.

	precision	recall	f1-score	support
0	0.95	1.00	0.98	20
1	0.78	0.90	0.84	20
2	0.95	1.00	0.98	20
3	0.84	0.80	0.82	20
4	0.95	0.86	0.90	21
5	0.91	1.00	0.95	21
6	0.67	0.57	0.62	7
7	1.00	1.00	1.00	23
8	0.92	1.00	0.96	22
9	1.00	0.90	0.95	10
10	0.96	1.00	0.98	25
11	0.87	0.95	0.91	21
12	0.00	0.00	0.00	1
13	0.95	1.00	0.98	20
14	1.00	1.00	1.00	13
15	1.00	0.85	0.92	20
16	0.93	0.93	0.93	44
17	0.00	0.00	0.00	2
18	1.00	0.60	0.75	5
accuracy			0.93	335
macro avg	0.83	0.81	0.81	335
weighted avg	0.92	0.93	0.92	335

Fig. 3. Classification report of the Vision Transformer (ViT) model on the ACL-FIG test split, demonstrating high effectiveness in figure classification.

Swin Transformer: Similarly, the Swin Transformer model reflected robust performance, with precision, recall, and F1-score each at 0.92 across the test dataset. This consistent performance across metrics highlights the model's

strengths in handling complex visual data with high accuracy.

	precision	recall	f1-score	support
0	0.91	1.00	0.95	20
1	0.82	0.90	0.86	20
2	0.95	0.95	0.95	20
3	0.77	0.85	0.81	20
4	0.94	0.81	0.87	21
5	0.95	0.95	0.95	21
6	0.75	0.43	0.55	7
7	0.96	1.00	0.98	23
8	0.92	1.00	0.96	22
9	1.00	0.90	0.95	10
10	0.96	1.00	0.98	25
11	0.72	1.00	0.84	21
12	1.00	1.00	1.00	1
13	1.00	1.00	1.00	20
14	1.00	1.00	1.00	13
15	1.00	0.75	0.86	20
16	0.97	0.89	0.93	44
17	1.00	0.50	0.67	2
18	1.00	0.80	0.89	5
accuracy			0.92	335
macro avg	0.93	0.88	0.89	335
weighted avg	0.92	0.92	0.91	335

Fig. 4. Classification report of the Swin Transformer model, showcasing its robust performance across different categories.

Comparison with Existing Methods: Leveraging these advanced machine learning models significantly enhances the efficiency and accuracy of data labeling, especially compared to traditional methods that rely extensively on manual annotation. This advantage is crucial for managing large-scale datasets like the CS-FIG, which contains tens of thousands of figures, making manual annotation impractical.

Innovations in Data Labeling: Incorporating GPT-4.0 to refine and verify the model outputs introduces an innovative, scalable solution for data labeling. This approach utilizes the generative capabilities of GPT-4.0 to enhance label accuracy, demonstrating a significant improvement over traditional methods and setting new standards for efficiency in scientific research.

Challenges Encountered: As it can be seen in the Fig 2 the results of the CS-FIG dataset. There are some imbalance classification on few categories. Those imbalances are addressed by our ensembling and GPT-4 as judging method.

Implications for Future Research: The successful application of ViT and Swin Transformer models in this study not only underscores the effectiveness of using state-of-the-art machine learning techniques for scientific figure classification but also opens avenues for further research into automated document analysis. The methodologies developed here have the potential to be extended to other domains, broadening the impact and applicability of automated image and text analysis in various fields.

5. Conclusion

This study has presented a novel approach to the classification of scientific figures within computer science articles using state-of-the-art machine learning models, namely the Vision Transformer (ViT) and Swin Transformer, coupled with the innovative use of GPT-4.0 for label refinement. Our work not only addresses the challenges posed by the scarcity of labeled data in scientific figure classification but also demonstrates a significant enhancement in efficiency and scalability compared to traditional manual annotation methods.

The CS-FIG dataset, meticulously curated and labeled through our advanced ensemble methodology, has enabled the detailed evaluation of these models, revealing their strong performance across a diverse array of figure types. The success of these models in accurately classifying complex images confirms the viability of applying advanced AI techniques to automate aspects of academic literature review and analysis.

Key findings include:

High Classification Accuracy: Both the ViT and Swin Transformer models achieved impressive metrics on precision, recall, and F1-scores, underscoring their capability to effectively parse and understand diverse scientific figures.

Efficiency in Labeling: By integrating GPT-4.0 into our workflow, we have significantly reduced the time and labor required for data annotation, setting a new benchmark for the processing of large datasets.

Challenges and Adaptations: While the models performed excellently on standard datasets, challenges arose with figures exhibiting complex structures or ambiguous content, indicating areas for future improvement in algorithm adaptation and training. Looking forward, the potential for expanding this research is vast. Future work could explore the integration of additional modalities of data, such as full-text analysis or deeper semantic understanding of content, to further improve classification accuracy. Additionally, extending our methods to other domains of scientific literature or even to different fields entirely could provide broader benefits to the academic and research communities.

In conclusion, the methodologies developed and implemented in this study are not only a step forward for the field of document analysis but also serve as a foundation for future innovations in automated data processing. By continuing to enhance these models and expand their applicability, we can look forward to a new era of efficiency in academic research and beyond.

References

1. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
2. Zabit Hameed, Sofia Zahia, Begonya Garcia-Zapirain, José Javier Aguirre, and Ana María Vanegas. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, 20(16), 2020.
3. Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.
4. KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.
5. Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. Acl-fig: A dataset for scientific figure classification, 2023.
6. Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. *CoRR*, abs/2111.09883, 2021.
7. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021.
8. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
9. Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.
10. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.