Article

Not peer-reviewed version

Exploring the Collaboration Between Vision Models and LLMs for Enhanced Image Classification

Bhavya Rupani *, Dmitry Ignatov, Radu Timofte

Posted Date: 15 December 2025

doi: 10.20944/preprints202512.1276.v1

Keywords: vision models; large language models; image classification; multimodal integration; CIFAR-10; CIFAR-100; vision-language collaboration; model performance benchmarks; deep learning; multimodal learning; computer vision; neural network fusion; visual reasoning; semantic alignment

 doi:10.20944/preprints202512.1276.v1

Article

Exploring the Collaboration Between Vision Models and LLMs for Enhanced Image Classification

Bhavya Rupani *, Dmitry Ignatov and Radu Timofte

Computer Vision Lab, CAIDAS, University of Würzburg, Germany
* Correspondence: bhavyarupani@gmail.com

Abstract

This paper defines a task that utilizes vision and language models to improve benchmarks through analysis of CIFAR-10 and CIFAR-100 datasets. The work divides its operations into image categorization followed by visual description production. The task utilizes BEiT and Swin models as state-of-the-art application-specific components for both parts of this research. We selected the current best image classification checkpoints available in the market which delivered **99.00%** accuracy on CIFAR-10 and **92.01%** on CIFAR-100. For dense contextually rich text output we used BLIP. The expert models performed well on their target responsibilities using minimal noisy data. The BART model achieved new state-of-the-art accuracies when used as a text classifier to compare synthesized descriptions and reached **99.73%** accuracy on CIFAR-10 while attaining **98.38%** accuracy on CIFAR-100. This paper demonstrates how our integrated vision and language decomposition-hierarchical model surpasses all existing state-of-the-art results on these common benchmark classifications. The full framework, along with the classified images and generated datasets, is available at https://github.com/bhavyarupani/LLM-Img-Classification.

Keywords: vision models; large language models; image classification; multimodal integration; CIFAR-10; CIFAR-100; vision-language collaboration; model performance benchmarks; deep learning; multimodal learning; computer vision; neural network fusion; visual reasoning; semantic alignment

1. Introduction

The two fields of image description and classification have improved a lot in the past few years thanks to leaps made in computer vision, natural language processing, and the continuous improvement of large language models (LLMs) [1–3]. Initial approaches include Deep CNNs models like AlexNet [4], VGG [5] and ResNet [6] necessary for image classification. However, the models such as Vision Transformers (ViT) [7] turned the domain by capturing information properly. EfficientNet [8], BEiT [9], Swin Transformer [10] as well as EVA-02 [11] are the modern architectures that have improved the CIFAR-10 [12] and CIFAR-100 [12] and ImageNet-1K [13] by providing scalable and accurate classification solutions [14].

In the field of image captioning, the recent models such as CLIP [15], BLIP [16] and GIT [17] have also shown remarkable improvement over the previous models by creating the connection between vision and language tasks, which in turn produces more meaningful captions. Even so, some issues still remain in generating high-quality and semantically precise captions of visually rich datasets such as ImageNet-1K. New approaches invented nowadays, such as BLIP and ViT-GPT2 [18] have strong and reliable visual encoders combined with the latest generation of language models, which allows better understanding of scenes and adding more semantic depth. These advances provide further support for the future of end-to-end image classification and captioning.

1.1. Novelty of the Approach

The apraoch distinguishes itself by combining structured classification outputs with improved caption generation to produce text-based classification through BART. Our system unifies classification

 doi:10.20944/preprints202512.1276.v1

methods with captioning procedures to generate improved datasets, which lead to better accuracy levels. Through an integration of BEiT and Swin followed by BLIP text-caption generation and BART for final text-based classification followed by an evaluation of semantic correspondences between the BLIP and classification outputs, we can produce better image cclassification pipeline. Through this mandatory enrichment process our data becomes clearer which improves both reliability level and overall performance of multimodal learning systems.

1.2. Datasets Used

Our experiments leverage two benchmark datasets: **CIFAR-10** and **CIFAR-100**. The CIFAR-10 dataset presents 60000 colored images which are 32*32 pixels in size and split among ten categories thus achieving fundamental image classification goals [19]. Classifying CIFAR-100 requires higher difficulty because it contains a total of 100 classes within identical 32*32 image dimensions [20].

1.3. Importance of Robust Captioning and Classification Models

Expert models must obtain high accuracy levels for trustworthy image captioning and classification which serves vital applications that need precise medical diagnosis [21] and operates autonomous vehicles [22] and surveillance systems [23]. The three models BLIP, CLIP, and BART use transformers [24] and advanced attention mechanisms [7] for processing and combining vision with text information. These systems demonstrate advanced properties of adaptability and extensibility which combined with their strong ability to decrease noise and errors within complex systems that combine multiple data modes. The working methods of semantic embedding alongside context enhancement presented in this research direction help to enhance the reliability of these models as reported in [15].

1.4. Research Objectives

The study investigates how to improve results by using existing classification and captioning procedures. The research produces context-rich diverse BLIP captions for accuracy enhancement followed by an evaluation of semantic correspondences between the BLIP outputs and classification models for efficiency improvement modeling. The research conducts its evaluation of BLIP-based pipeline performance by applying standard classification models to both CIFAR-10 and CIFAR-100 benchmarks.

1.5. Contribution

The proposed research method expands dataset content through the use of available classification and captioning systems. The integration of BEiT with Swin and BLIP models on CIFAR-10 and CIFAR-100 proves to produce results according to our analysis. The approach generates high-accuracy captions of better quality because of contextual inputs that produces clearer results suitable for real-world vision-language application integration. The principles from this work provide direction for future research which focuses on uniting classification and captioning approaches to create improved image classification results [25].

2. Previous Works

This part examines research relevant to image classification and image captioning and dataset augmentation strategies [26] that are relevant to this work.

2.1. State-of-the-Art Models for Image Classification

The initial substantial breakthroughs in image classification occurred using convolutional neural networks (CNN) [4,5,14]. The introduction of attention-based architectures came after models were already built without the ability to recognize global interconnectivity between components. Vision Transformers (ViT) [7] changed the world of classification through self-attention functionality that further enhanced large-scale image database performance. The EfficientNet [8] compound scaling

 doi:10.20944/preprints202512.1276.v1

technique enhanced accuracy while reducing processing requirements through its approach to enhance network depth and resolution growth and width.

2.2. State-of-the-Art Models for Image Captioning

The development of better image captioning models stems directly from enhanced vision-language pretraining techniques. The CNN-RNN model structures from traditional methods experienced challenges with contextual accuracy according to references [27,28]. Transformers have brought about more complex captioning models since the introduction of UNITER [29] and ViLT [30].

The BLIP [31] improves caption generation through query-based pretraining to boost the alignment of image. GIT [17] creates contextually steady captions along with ViT-GPT2 [18] that connects Vision Transformers to autoregressive language models for permitting syntactically and semantically coherent descriptions. The research efforts of SimVLM and Flamingo [32] brought various enhancements to vision-language architecture models which resulted in better results for captioning and VQA tasks.

2.3. Integrating Classification and Captioning for Dataset Enrichment

Researchers have established that the combination of classification and captioning models represents a vital method to enhance dataset augmentation for multimodal learning. Former studies on multimodal learning established language and visual representation integration as an approach to enhance model generalization [33,34]. VLP models [25,35] show transformers create enhanced captions which maintain classification precision.

Synthetic caption creation methods improve data collection effectiveness. Web crawled datasets at scale have proven useful for pretraining purposes which strengthens the model [36,37]. The FuseCap system enhances complex application support through its ability to extend captions using multiple data types [38,39]. The authors of RS-LLaVA [40] describe that accurate dataset curation in remote sensing applications requires combining classification with captioning.

Pretraining strategies have developed to help expand the size of available datasets. Through Pyramid-based approaches [29,41] image models achieve better semantic understanding which enables them to analyze delicate details. The development of multimodal learning has improved vision-language models through two major automated pretraining methods: Flamingo [32] and CoCa. The recent advancements require improved datasets through combination of classification and captioning to produce comprehensive training material for subsequent applications.

2.4. Relevance to Current Work

Research progress in this field incorporates Swin and BEiT along with BLIP captioning model for advancement of classification results. BLIP results in precise captions that preserve their classification accuracy values. The proposed models have undergone evaluation to show their ability in improving multimodal datasets as proven in current research on dataset enrichment.

This approach solves past research difficulties through implementing contemporary enhancements to CIFAR-10 and CIFAR-100 systems.The incorporation of classification together with captioning models creates a new standard in vision-language operations that boosts performance of multimodal systems and dataset quality.

3. Proposed Methodology

Our proposed system improves multimodal learning by uniting the latest version (SOTA) of picture classification technology with captioning models. A sophisticated image classifier performs first classification before the process begins as shown in Figure 1. The structured classification output from the first stage goes into a subsequent unconditional caption generator to produce verbose descriptions from images.

 doi:10.20944/preprints202512.1276.v1

A final text-based classifier utilizes structured classification together with resulting captions to deliver the ultimate classification outcome that enhances accuracy performance and interpretability metrics.

All information related to conditional caption generation method can be found in the supplementary section.

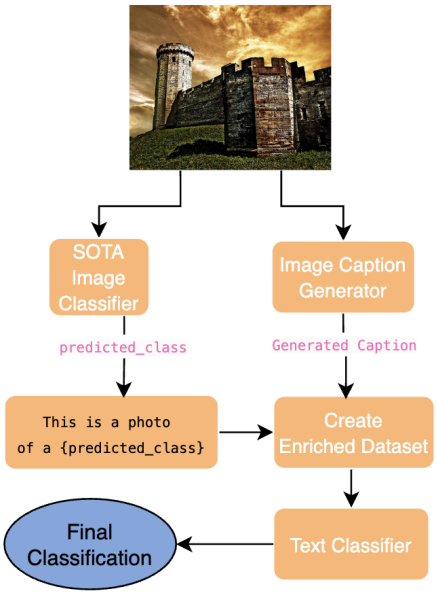

Figure 1. The proposed pipeline unites an image classification model along with a captioning.

4. Model Evaluation and Selection

Fine-tuning of models is performed using the AI Linux docker image abrainone/ai-linux[1] on NVIDIA GeForce RTX 3090/4090 GPUs of the CVL Kubernetes cluster at the University of Würzburg. The evaluation phase utilized multiple image classification and captioning models that operated on CIFAR-10 and CIFAR-100 datasets [12]. We wanted to find out which models created the most accurate results for integrating in our pipeline and classification. The ablation study analysis contains crucial findings that helped select the best models for each dataset according to the results presented.

4.1. Ablation Study and Model Selection

The development of an optimal classification pipeline required us to perform a state-of-the-art evaluation of classification and captioning models. The analysis evaluated model performance regarding their ability to create structured descriptions and captions for integration processes in our pipeline.

4.1.1. Image Classification Models

The initial part of our methodology applies contemporary approaches from image classification. Our research evaluated the performance of BEiT, ViT [42] and CLIP models together with the Swin Transformer model on both **CIFAR-10** and **CIFAR-100**. Our studies based on accuracy results in Table 1 match state-of-the-art performances although recent model checkpoints were scarce.

[1] AI Linux: https://hub.docker.com/r/abrainone/ai-linux

 doi:10.20944/preprints202512.1276.v1

Table 1. The experimental results indicate that BEiT with fine-tuning demonstrates excellent outcomes on CIFAR-10 yet Swin-Base performs best on CIFAR-100.

Dataset	Model	Accuracy (%)
CIFAR-10	ViT-B/16 [43]	97.88
	CLIP-ViT-Base-Patch32 [43]	97.59
	Fine-tuned BEiT (Ours)	**99.00**
CIFAR-100	ViT-B/16 [43]	91.48
	CLIP-ViT-Base-Patch32 [43]	88.37
	Swin-Base [43]	**92.01**

Due to the limited availability of state-of-the-art CIFAR-10 classification model checkpoints, we resorted to fine-tuning the BEiT model. Table 2 details the complete hyperparameter configuration employed during the fine-tuning process.

Table 2. Hyperparameter configuration for BEiT fine-tuning on CIFAR-10 for performance-enhancing.

Training Parameter	Value
Batch Size	8 (Gradient Accumulation: 4 steps)
Learning Rate	5×10^{-5} (Phase 1), 1×10^{-5} (Phase 2)
Optimizer	AdamW
Loss Function	Cross-Entropy Loss
Epochs	3 (Phase 1), 2 (Phase 2)
Scheduler	Linear Decay
Mixed Precision Training	Enabled (FP16)

4.1.2. Image Captioning Models

The research utilized BLIP [16] along with the GIT [17] and ViT-GPT2 [18] models to extract caption responses from images within each dataset collection. During processing of CIFAR-10 and CIFAR-100 low-resolution image samples all models faced difficulties creating precise capturing sentences.

Table 3. BLIP, GIT and ViT-GPT2 generated the following captions from images belonging to CIFAR-10 and CIFAR-100.

Dataset	BLIP Base Caption	GIT Base Caption	ViT-GPT2 Caption
CIFAR-10	A small white dog with a green collar	A white flower	A small bird standing on top of a dirt ground
CIFAR-100	A man riding a horse in a field	Red nose on cow	A cow standing on top of a lush green field

An examination was carried out to determine the influence of captions classification accuracy using BART and DeBERTa. Table 4 reveals that captions from BLIP offered maximum accuracy.

 doi:10.20944/preprints202512.1276.v1

Table 4. Evaluation of caption-based classification on CIFAR-10 and CIFAR-100. The system produced captions by BLIP resulted in better accuracy levels throughout the classification process.

Dataset	Captioner	Text Classifier	Accuracy (%)
CIFAR-10	BLIP	BART	**89.48**
	GIT	BART	74.65
	ViT-GPT2	BART	74.79
	BLIP	DeBERTa	87.94
	GIT	DeBERTa	71.98
	ViT-GPT2	DeBERTa	73.62
CIFAR-100	BLIP	BART	**49.98**
	GIT	BART	29.87
	ViT-GPT2	BART	26.07
	BLIP	DeBERTa	49.81
	GIT	DeBERTa	31.45
	ViT-GPT2	DeBERTa	26.59

4.2. Key Observations

During evaluation for image classification, the fine-tuning of BEiT produced the best CIFAR-10 accuracy results and Swin for CIFAR-100. During the evaluation for Image captioning part, BLIP produced captions that contained higher semantic value while providing better contextual match than GIT and ViT-GPT2.

4.3. Final Model Selection

The chosen models for pipeline deployment in Table 7 include BEiT for CIFAR-10 along with Swin Transformer for CIFAR-100. BLIP functions across both datasets for captioning tasks while BART achieves caption-based classification.

5. Enriched Dataset Generation and Statistics

This section outlines process flows for enriched dataset development in CIFAR-10 and CIFAR-100. State-of-the-art image classification operates alongside unregulated caption generation to create detailed descriptive text from structured labels that are applied to images. The enrichment technique adds more data to the dataset which improves performance levels in subsequent applications.

5.1. Workflow Overview

Data preprocessing of raw images leads to application of sophisticated classification algorithms which then create structured labels. The unconditional caption generator functions independently to use the BLIP model for producing long captions. The enriched datasets undergo analysis through procedures that yield presentation results available in Table 5. The workflow shows in Figure 3 produces improved datasets through its integrated structure. Readers can access complete information about conditional caption generation in the supplementary material.

5.2. Generated Dataset Examples

Images from CIFAR-10 and CIFAR-100 datasets serve to demonstrate the output generation as presented in Figure 2. Table 5 demonstrates such datasets which were produced through the application of BEiT, Swin alongside BLIP models. The produced text outputs from each model stand out for their ability to create detailed descriptions through individual usage.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 December 2025 doi:10.20944/preprints202512.1276.v1

(a) CIFAR-10 **(b)** CIFAR-100

Figure 2. Different captioning models were evaluated these samples to demonstrate their performance across datasets CIFAR-10 and CIFAR-100 with different resolutions and complexities in Table 3.

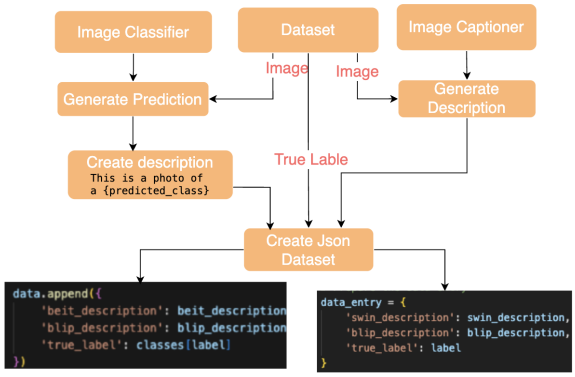

Figure 3. The process to integrate classification models BEiT and Swin with BLIP to create CIFAR-10 and CIFAR-100 JSON-formatted datasets.

Table 5. BEiT, Swin and EVA-02 and BLIP models which dataset for the CIFAR-10, CIFAR-100 and ImageNet.

Dataset	Model	Generated Text	True Label
CIFAR-10	BEiT	This is a photo of a frog.	Frog
	BLIP	A small white dog with a green collar	
CIFAR-100	Swin	This is a photo of an horse.	Horse
	BLIP	A man riding a horse in a field	

5.3. Statistics of Enriched Datasets

A statistical evaluation of the enriched datasets was performed to measure the richness of the generated content. Mean word counts and standard deviations were calculated for each model and dataset. Table 6 provides a detailed summary of these statistics.

Table 6. Statistics show BEiT and Swin with BLIP model produce descriptions with various word counts expressed through Mean - σ values (Mean - Standard Deviation) in the dataset.

Dataset	Model	Mean ± σ
CIFAR-10 (60,000 Images)	BEiT	7.00 - 0.00
	BLIP	11.15 - 0.83
	Combined	18.15 - 0.83
CIFAR-100 (60,000 Images)	Swin	7.00 - 0.00
	BLIP	11.13 - 1.33
	Combined	18.13 - 1.33

 doi:10.20944/preprints202512.1276.v1

6. Fine-Tuning Pipeline Models on Enriched Descriptions

6.1. Overview of Fine-Tuning Pipelines

The fine-tuning pipelines unite classification models using BLIP for captioning and BART for classification. We use individual dataset-specific pipeline configurations, which are shown in Table 7.

Table 7. Selected pipelines for for CIFAR-10 and CIFAR-100, combined with BLIP-generated captions to final classification through BART.

Dataset	Pipeline
CIFAR-10	BEiT + BLIP + BART
CIFAR-100	Swin Transformer + BLIP + BART

6.2. Pipeline for CIFAR-10 and CIFAR-100

A processing pipeline integrates image classifier models with BLIP image captioner and BART text classifier to classify images of CIFAR-10 and CIFAR-100.

The **BEiT** model executes structured classification assessments with **CIFAR-10** along with the **BLIP** system which generates elaborate captions. The text descriptions developed in previous operations allow BART to carry out exact image classification while Swin architecture creates specific descriptions by utilizing its attention structure to handle complex CIFAR-100 data. The captions generated by BLIP system creates contextually rich description so BART receives expanded contextual details that boost its classification accuracy.

6.3. Data Processing Pipeline

Our data processing flow joins text descriptions produced by image classifiers with those from captioners to process them further according to Figure 4. A modern classifier applies structured captions to images during the initial step. An unconditional captioning model simultaneously generates detailed descriptions from visual features only. The semantic similarity operations validate the outputs against each other in order to ensure proper alignment with classifier and captioner descriptions.

Through the fusion process the system produces an extensive structured input which maintains the analytical aspects of the classifier and maintains the contextual details from the captioner. Semantic enhancement of the final data allows it to serve as input for better performance in subsequent vision-language applications.

For details on the conditional data processing workflow, where caption generation is guided by classifier outputs, please refer to the supplementary material.

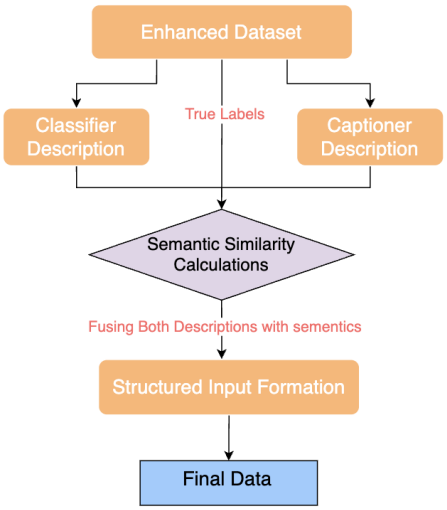

Figure 4. Workflow for processing captions from enriched dataset which was created

 doi:10.20944/preprints202512.1276.v1

6.4. Hyperparameter Configuration & Fine-Tuning

Model performance optimization demanded the selection of special hyperparameters during the fine-tuning process. The essential training hyperparameter values used on CIFAR-10 and CIFAR-100 appear in Table 8.

For details on the hyperparameter configuration used for BART fine-tuning feeding conditional data, where caption generation is guided by classifier outputs, please refer to the supplementary material.

Table 8. Important hyperparameters for BART training across datasets.

Hyperparameter	CIFAR-10	CIFAR-100
Epochs	18	11
Batch Size	32	32
Warmup Steps	500	500
Learning Rate	3e-6	3e-6
Weight Decay	0.005	0.005
LR Scheduler	cosine	cosine
Loss Function	MSE	MSE

The **BART** model receives fine-tuning during for adaptation to our particular classification problem. Through fine-tuning the model maintains its initial language processing knowledge alongside target information acquisition from the new enriched dataset. The model parameter update through iterative optimization allows for better representation creation which enhances its classification abilities.

The approach delivers additional semantic information about captions that boosts the model's contextual capabilities as well as its capability to link text descriptions to image categories. BART adjusts its underlying weight values throughout training to decrease loss which results in performance improvements.

6.5. Performance Comparison with State-of-the-Art

A comparison of CIFAR-10 and CIFAR-100 test results from the fine-tuned BART model against other SOTA models are in Table 9.

Table 9. Comparison of evaluation metrics for CIFAR-10 and CIFAR-100 between SOTA and our approach.

Metric	CIFAR-10			CIFAR-100		
	ViT-H/14	Used: BEiT	Ours	EffNet-l2	Used: Swin	Ours
Accuracy	99.50%	99.00%	**99.73%**	96.08%	92.01%	**98.39%**
Precision	—	—	**99.46%**	—	—	**97.11%**
Recall	—	—	**99.73%**	—	—	**98.39%**
F1 Score	—	—	**99.59%**	—	—	**97.65%**

7. Conclusion and Future Work

The experimental results validate the successful integration of VL models to achieve image classification. We evaluate the leading results and obstructing elements of our methodology and suggest new research opportunities together with practical deployment possibilities.

7.1. Conclusion

The analysis proved its superiority through evaluation of every tested dataset. Richer data sources enabled the model to reach an optimum performance level of **99.73%** when used for CIFAR-10 tasks. Detailed descriptions enable the system to reach an accuracy of **98.22%** in processing complex data

 doi:10.20944/preprints202512.1276.v1

provided in the CIFAR-100 dataset. The combination of structured labels with detailed captions produces enhanced classification accuracy based on the obtained results.

7.2. Challenges and Limitations

Results of the study are promising, yet multiple obstacles persist as they signal upcoming improvement opportunities.

Rich captions by captioner Issues Create better enriched datasets which havew more context-richtext rich captions where captioner alone can provide high accuracy.

Generalization Issues The difficulty of the pipelines to deal with unusual situations restricts their practical implementation.

Low-Resolution Image Challenges BLIP exhibits limited captioning capability for developing detailed descriptions of images at low resolution.

Contextual Limitations The difficulty for machine-generated captions to distinguish objects with similar appearances degrades the performance of classification and captioning tasks.

High Computational Costs The training process of large-scale multimodal models demands high computational power that creates a challenge for wide-scale deployment.

Bias in Captions The descriptions that are generated from the system tend to display training data biases some times which presents practical deployment challenges due to ethical reasons.

7.3. Future Work

The research creates multiple prospects to enhance methodology along with expanding its areas of application for future scholarly works.

Improving Captioning for Low-Resolution Images Developing suitable methods to generate detailed accurate captions for low-resolution datasets.

Exploring Advanced Captioning Models The overall system performance could benefit from adding state-of-the-art captioning models other then BLIP.

Scaling to Complex Datasets The methodology should scale up to larger commercial-grade datasets from domains like industry to evaluate the operational performance of this proposed methodology in practical settings.

Optimizing Computational Efficiency Future research needs to develop methods that decrease the training requirements of large-scale multi-modal models for better accessibility by environments under limited resource constraints.

Exploring New Applications The enriched datasets and multimodal pipelines should undergo evaluation tests for VQA and medical diagnostic applications in addition to image retrieval to advance their practical utility range.

Scaling to Larger and More Complex Datasets Apply methodology to broader and more advanced datasets together with industrial domain-specific industrial datasets to assess scalability and generalization potential.

Real-Time Deployments Efficient optimization of pipelines should focus on real-time applications which require critical domains such as autonomous driving due to their need for immediate and dependable decisions.

7.4. Final Remarks

The combination of classification models with caption models and further classifying them with text classifiers using semantic similarities creates exceptionally strong capabilities to classify images according to the presented research.

References

1. Kochnev, R.; Goodarzi, A.T.; Bentyn, Z.A.; Ignatov, D.; Timofte, R. Optuna vs Code Llama: Are LLMs a New Paradigm for Hyperparameter Tuning? *arXiv preprint arXiv:2504.06006* **2025**.

 doi:10.20944/preprints202512.1276.v1

2. Gado, M.; Taliee, T.; Memon, M.D.; Ignatov, D.; Timofte, R. VIST-GPT: Ushering in the Era of Visual Storytelling with LLMs? *arXiv preprint arXiv:2504.19267* **2025**.
3. Kochnev, R.; et al. NNGPT: Neural Network Model Generation. *arXiv preprint* **2025**.
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. https://doi.org/10.1145/3065386.
5. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), 2015.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, 2020.
8. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning; Chaudhuri, K.; Salakhutdinov, R., Eds. PMLR, 09–15 Jun 2019, Vol. 97, *Proceedings of Machine Learning Research*, pp. 6105–6114.
9. Bao, H.; Dong, L.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), 2022.
10. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 10012–10022.
11. Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
12. Krizhevsky, A.; Hinton, G.; et al. Learning multiple layers of features from tiny images.(2009), 2009.
13. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 248–255.
14. Goodarzi, A.T.; Kochnev, R.; Khalid, W.; Qin, F.; Uzun, T.A.; Dhameliya, Y.S.; Kathiriya, Y.K.; Bentyn, Z.A.; Ignatov, D.; Timofte, R. LEMUR Neural Network Dataset: Towards Seamless AutoML. *arXiv preprint arXiv:2504.10552* **2025**.
15. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning; Meila, M.; Zhang, T., Eds. PMLR, 18–24 Jul 2021, Vol. 139, *Proceedings of Machine Learning Research*, pp. 8748–8763.
16. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning; Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; Sabato, S., Eds. PMLR, 17–23 Jul 2022, Vol. 162, *Proceedings of Machine Learning Research*, pp. 12888–12900.
17. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research* **2022**.
18. Vasireddy, I.; HimaBindu, G.; B., R. Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing. In Proceedings of the International Journal of Innovative Research in Engineering and Management, 2023, Vol. 10, pp. 55–59.
19. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images (CIFAR-10). Technical report, University of Toronto, 2009. CIFAR-10 dataset.
20. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images (CIFAR-100). Technical report, University of Toronto, 2009. CIFAR-100 dataset.
21. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118.
22. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 3354–3361.
23. Akhtar, M.J.; Mahum, R.; Butt, F.S.; Amin, R.; El-Sherbeeny, A.M.; Lee, S.M.; Shaikh, S. A Robust Framework for Object Detection in a Traffic Surveillance System. *Electronics* **2022**, *11*. https://doi.org/10.3390/electronics11213425.

 doi:10.20944/preprints202512.1276.v1

24. Need, A.I.A.Y. Ashish Vaswani, Noam Shazeer, NikiParmar, JakobUszkoreit, Llion Jones, Aidan N. *Gomez, Lukasz Kaiser, IlliaPolosukhin* **2017**.
25. Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; Huang, J. Vision-Language Pre-Training With Triple Contrastive Learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 15671–15680.
26. Aboudeshish, N.; Ignatov, D.; Timofte, R. AUGMENTGEST: CAN RANDOM DATA CROPPING AUGMENTATION BOOST GESTURE RECOGNITION PERFORMANCE? *arXiv preprint* **2025**.
27. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
28. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.
29. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal Image-TExt Representation Learning. In Proceedings of the Computer Vision – ECCV 2020; Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J.M., Eds., Cham, 2020; pp. 104–120.
30. Kim, W.; Son, B.; Kim, I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning; Meila, M.; Zhang, T., Eds. PMLR, 18–24 Jul 2021, Vol. 139, *Proceedings of Machine Learning Research*, pp. 5583–5594.
31. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning; Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; Sabato, S., Eds. PMLR, 17–23 Jul 2022, Vol. 162, *Proceedings of Machine Learning Research*, pp. 12888–12900.
32. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: a Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds. Curran Associates, Inc., 2022, Vol. 35, pp. 23716–23736.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.
34. Hu, R.; Singh, A. UniT: Multimodal Multitask Learning With a Unified Transformer. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 1439–1449.
35. Wang, W.; Bao, H.; Dong, L.; Wei, F. VLP: Vision-Language Pre-Training of Fusion Transformers. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning, 2021.
36. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Gurevych, I.; Miyao, Y., Eds., Melbourne, Australia, 2018; pp. 2556–2565. https://doi.org/10.18653/v1/P18-1238.
37. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 3558–3568.
38. Rotstein, N.; Bensaïd, D.; Brody, S.; Ganz, R.; Kimmel, R. FuseCap: Leveraging Large Language Models for Enriched Fused Image Captions. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2024, pp. 5689–5700.
39. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In Proceedings of the Computer Vision – ECCV 2016; Leibe, B.; Matas, J.; Sebe, N.; Welling, M., Eds., Cham, 2016; pp. 382–398.
40. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Ricci, R.; Melgani, F. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery. *Remote Sensing* **2024**, *16*, 1477.
41. Huang, P.Y.; Hu, R.; Schwing, A.; Murphy, K. Understanding and Improving the Masked Language Modeling Objective for Vision-and-Language Pretraining. In Proceedings of the Advances in Neural Information Processing Systems, 2019.

 doi:10.20944/preprints202512.1276.v1

42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* **2020**.
43. Face, H. Hugging Face: Advancing Natural Language Processing and Machine Learning, 2019.