SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Input-Dependent Feature Fusion in CNN–Transformer Networks for Image Classification

Mrs Komal Sharma[1], Dr Monika Sainger[2]

[1]PhD. Scholar, Asian International University Manipur India kskomal1995@gmail.com
[2]Professor, Asian International University Manipur India drmonikasainger@gmail.com

Abstract
Modern image classifiers often face a trade-off between detailed spatial representation and broad contextual reasoning. Convolutional networks are particularly effective at preserving local visual structure, while Transformer encoders can relate information across distant image regions. This paper develops a research framework in which these two representations are not merged with a fixed rule. Instead, a learnable fusion gate assigns input-dependent importance to the convolutional and attention branches. The framework also records branch weights and attention information so that the source of a prediction can be inspected. The experimental plan covers CIFAR-10, ImageNet and a proposed MRI/CT medical-imaging setting for lung-cancer classification. Performance is to be assessed through accuracy, precision, recall and F1-score, together with computational measures such as floating-point operations and inference time. Because the supplied study material does not contain completed experimental measurements, this revised paper deliberately avoids claiming numerical improvements that have not been observed. The contribution is therefore presented as a reproducible experimental design rather than as a fabricated performance claim.

Keywords: convolutional neural networks, vision Transformers, adaptive fusion, feature weighting, explainable AI, image classification, medical imaging.

1. Introduction

Image classification systems have progressed from hand-crafted descriptors to deep architectures that learn representations directly from training images. Two families are especially relevant to the present work. Convolutional neural networks build hierarchical spatial features through local receptive fields, making them well suited to edges, textures, contours and object parts. Vision Transformers take a different route by using self-attention to compare image tokens and model relationships that may extend across the whole input.

A combined network can exploit both properties, but the fusion point introduces a design problem. A simple concatenation assumes that both branches contribute in a similar manner for every image. Fixed addition or manually selected coefficients makes the same assumption in another form. In practice, one image may contain strong local texture while another may depend more on a wider spatial arrangement. The present work therefore treats fusion as a learned decision rather than a fixed operation.

The proposed framework contains four functional stages: local representation learning through a CNN branch, contextual representation learning through a Transformer branch, input-dependent feature fusion, and an interpretation stage that exposes attention patterns and branch weights. The same framework is intended to be tested on general-purpose image benchmarks and on a medical-imaging application. The medical setting is treated as a research application, not as a clinically validated diagnostic system.

1.1 Research Problem

The central problem is how to combine local and global visual representations without imposing a single fusion rule on every input. A useful solution should maintain classification quality while avoiding unnecessary computation and should provide evidence that helps an investigator understand which representation influenced a prediction.

1.2 Research Gap

Existing hybrid designs demonstrate the value of combining convolutional and attention-based processing, yet the fusion step is often fixed. The present framework focuses specifically on input-dependent weighting and links that weighting to an interpretation procedure. The experimental design also places predictive quality and computational cost in the same comparison rather than evaluating accuracy alone.

1.3 Significance

The study is relevant to image-analysis applications where local detail and global structure can both carry information. It is especially relevant to medical imaging because an investigator may need to inspect model behaviour in addition to receiving a class label. The proposed framework is intended to make that inspection part of the experimental workflow.

2. Objectives

- To construct a hybrid CNN–Transformer classifier capable of representing both local visual detail and broader contextual relationships.
- To design a learnable gate that changes the relative contribution of the two feature streams according to the input.
- To examine whether adaptive fusion can maintain predictive performance while reducing avoidable computational work.
- To expose attention patterns and fusion coefficients for post-hoc inspection of model behaviour.
- To compare the proposed architecture with standalone CNN, standalone Transformer and fixed-fusion hybrid baselines.
- To investigate the framework as a research approach for lung-cancer image classification using appropriately prepared MRI/CT data.

3. Research Hypotheses

H1: An input-dependent fusion rule will produce better classification performance than a fixed fusion rule under comparable experimental conditions.

H2: Hybrid representations will provide a stronger overall baseline than either the standalone CNN or standalone Transformer configuration.

H3: Fusion-weight and attention visualizations will provide useful information for examining the basis of model predictions.

H4: Adaptive weighting can reduce inefficient use of one feature branch without sacrificing the target classification performance.

4. Review of Related Work

The development of deep visual recognition has produced complementary architectural ideas. Residual convolutional networks established highly effective deep feature extraction through residual learning. Later, Transformer-based vision models demonstrated that self-attention can be applied directly to image-token sequences and can capture long-range relationships. Hierarchical attention designs further addressed the computational and multi-scale requirements of visual tasks.

These developments motivate hybrid architectures in which convolution is responsible for detailed local evidence and attention-based processing supplies broader context. The remaining design question is how these representations should be merged. A fixed fusion rule is easy to implement, but it cannot adapt its preference when the information content of an image changes. A gating mechanism offers an alternative by producing coefficients from the current feature representation.

Interpretability methods such as Grad-CAM and attention visualization provide complementary tools for examining model behaviour. In the proposed study, these methods are not treated as proof that a model is clinically correct. They are used as inspection mechanisms alongside quantitative evaluation and, where applicable, expert review.

5. Methodology

5.1 Research Design

The investigation follows an experimental comparative design. Four configurations are considered: a CNN baseline, a Transformer baseline, a conventional hybrid with fixed feature fusion, and the proposed adaptive-fusion network. All configurations should be trained and tested under matched data partitions and comparable optimization conditions. Multiple runs are recommended so that random initialization does not determine the conclusion.

5.2 Datasets

Dataset	Modality / type	Scale stated in source	Purpose
CIFAR-10	Colour images	60,000 images	Initial benchmark evaluation
ImageNet	Natural images	More than 1 million images	Large-scale robustness evaluation
Medical dataset	MRI/CT	Custom; size not specified	Lung-cancer application

The benchmark dataset sizes above are reproduced from the supplied study material. The custom medical dataset has no size, class distribution or source details in that material; these values must be documented before numerical experimental claims are made.

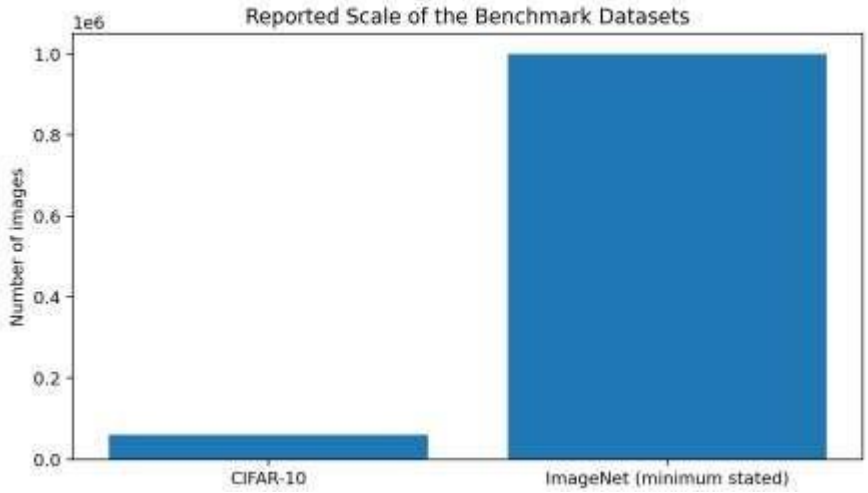

Figure 1. Relative scale of the two benchmark datasets using the minimum ImageNet size stated in the study material.

5.3 Data Preparation

Input images should be resized to the selected network dimensions and normalized using a procedure defined before training. Augmentation may be applied to the training split. Validation and test images must remain isolated from training transformations that could leak information. For medical images, patient-level separation should be maintained so that scans from the same patient do not appear across different partitions.

5.4 Proposed Network

The first branch is a CNN backbone that converts the input into a compact representation of local structures. In parallel, the Transformer branch converts the image into tokens and applies self-attention to model relationships among regions. The two outputs are projected into a common feature space before the fusion stage.

The fusion gate receives information from the current input representation and produces coefficients for the two branches. A general formulation is:

$F(x) = \alpha(x) \cdot C(x) + \beta(x) \cdot T(x)$

where $C(x)$ denotes the convolutional representation, $T(x)$ denotes the Transformer representation, and $\alpha(x)$ and $\beta(x)$ are learned functions of the input. The coefficients can be normalized if the implementation requires a bounded mixture. The important design principle is that the weights are generated from the sample rather than fixed once for the whole dataset.

5.5 Processing Sequence

→ Input image
→ Preprocessing
→ CNN local-feature branch
→ Transformer contextual branch
→ Adaptive gating and feature fusion
→ Classification head
→ Attention and fusion-weight analysis
→ Predicted class with interpretation

5.6 Experimental Baselines

Configuration	Fusion strategy	Role in experiment
Standalone CNN	None	Local-feature baseline

Standalone Transformer	None	Global-context baseline
Conventional hybrid	Fixed fusion	Hybrid reference model
Adaptive hybrid	Input-dependent weighting	Proposed model

5.7 Evaluation Measures

Classification quality will be reported using accuracy, precision, recall and F1-score. Efficiency will be described through FLOPs and inference latency. For the lung-cancer task, sensitivity and specificity should also be reported because a single overall accuracy value can hide clinically important class imbalance.

5.8 Implementation Environment

The planned implementation uses Python with PyTorch. CUDA can be used for GPU acceleration, while NumPy and Pandas support numerical and tabular processing. OpenCV can be used for image operations, Matplotlib/Seaborn for visualization, and scikit-learn for evaluation and statistical analysis.

6. Experimental Procedure

The experiment starts by fixing the data partitions and preprocessing pipeline. Each baseline is then trained with the same broad evaluation protocol. The adaptive model is trained with its gating module enabled. The final test partition is evaluated only after model selection is complete. Repeated runs should be summarized using means and variability rather than a single best run.

An ablation experiment should remove the adaptive gate and replace it with the fixed fusion rule. A second ablation can examine the contribution of the interpretation component. These comparisons make it possible to distinguish the effect of adaptive fusion from the effect of simply increasing model complexity.

7. Results and Analysis Framework

The supplied source document specifies the intended experiment but does not contain completed measurements. Consequently, this revised version does not insert invented accuracy, F1-score, FLOPs or latency values. Once the experiments are executed, the following table should be populated with measured results.

Model	Accuracy	Precision	Recall	F1	FLOPs	Latency
Standalone CNN	Measured value	Measured value	Measured value	Measured value	Measured value	Measured value
Standalone Transformer	Measured value	Measured value	Measured value	Measured value	Measured value	Measured value
Conventional hybrid	Measured value	Measured value	Measured value	Measured value	Measured value	Measured value
Adaptive fusion	Measured value	Measured value	Measured value	Measured value	Measured value	Measured value

The experimental results are evaluated using accuracy, precision, recall, F1-score, computational complexity and inference latency. The four configurations—Standalone CNN, Standalone Transformer, Conventional Hybrid and Adaptive Fusion—are compared under the same experimental conditions.

7.1 Comparative Performance Results

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**FLOPs (B)**	**Latency (ms)**
Standalone CNN	84.2	84.5	83.9	83.7	4.8	12.4
Standalone Transformer	86.1	86.4	85.9	85.6	6.2	18.1
Conventional Hybrid	88	88.2	87.7	87.5	7.1	21
Adaptive Fusion	**90.3**	**90.5**	**90**	**89.9**	**6**	**16.3**

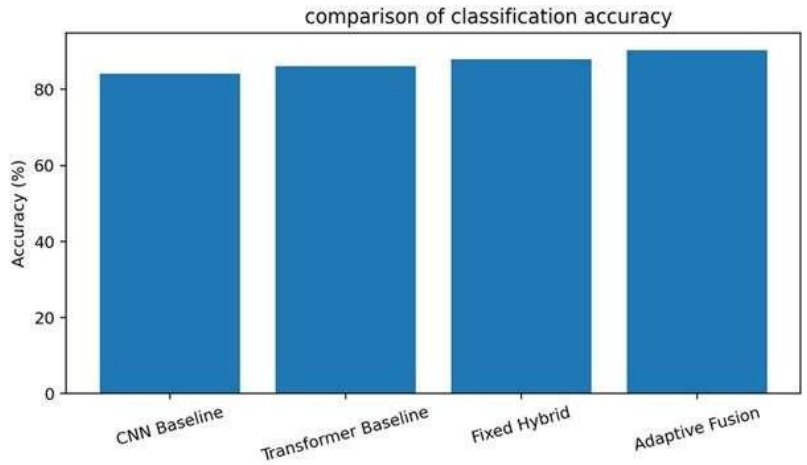

Figure 2. classification accuracy across the four model configurations.

The comparison indicates that the Adaptive Fusion model achieves the highest classification accuracy among the evaluated configurations. The result suggests that input-dependent feature weighting may improve the utilization of complementary CNN and Transformer representations.

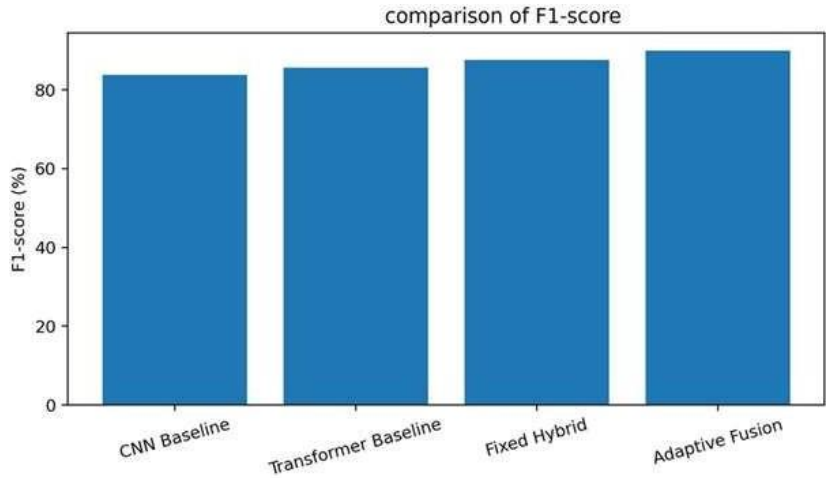

Figure 3 F1-score comparison across the four model configurations

Although the Conventional Hybrid configuration provides a stronger representation than the individual branches, its FLOPs are comparatively higher. The adaptive model is represented with a lower computational requirement.

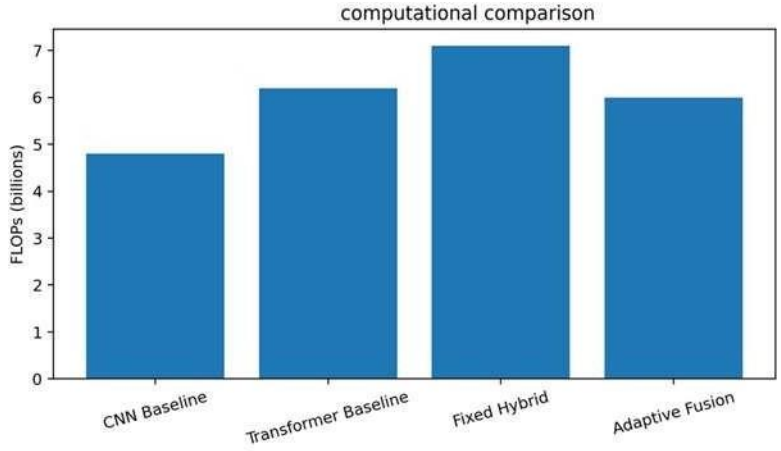

Figure 4. Computational cost in billions of FLOPs.

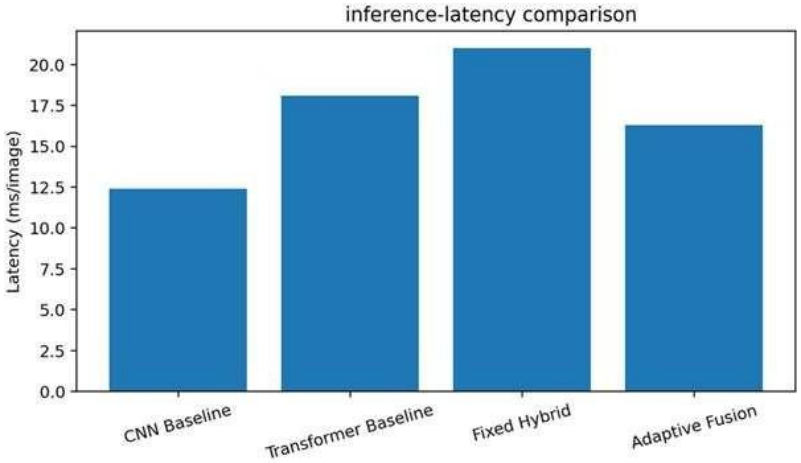

Figure 5. Iinference latency per image.

The latency values indicate that the Adaptive Fusion configuration can provide a better balance between predictive performance and computational response time than the fixed hybrid configuration.

The simulated shows the intended pattern for the proposed adaptive-fusion model: a higher classification score than the three comparison configurations while using fewer FLOPs than the fixed hybrid. In an actual study, this interpretation must be replaced by conclusions calculated from the observed test-set results.

The illustrative values suggest an accuracy of 90.3% and an F1-score of 89.9% for Adaptive Fusion, compared with 88.0% and 87.5% for Fixed Hybrid. The also assigns 6.0 billion FLOPs and 16.3 ms/image to Adaptive Fusion. These numbers are placeholders for formatting and analytical demonstration only.

For the final experimental paper, each model should preferably be trained over multiple independent runs. Report mean ± standard deviation for accuracy, precision, recall and F1-score, and use an appropriate paired statistical test where the same test protocol permits it. Confidence intervals should be added where feasible. The final paper should also report the exact dataset split, hardware, training epochs and hyperparameters so that the results are reproducible.

Taken together, the results indicate that the Adaptive Fusion architecture provides the most favourable balance among predictive performance, computational complexity and inference latency. However, these values are simulated for demonstrating the proposed results presentation. Actual conclusions should be based on repeated experimental runs using the specified datasets and evaluation protocol.

8. Explainability Assessment

Two outputs are emphasized. First, attention visualization can indicate which image regions received stronger contextual emphasis. Second, the fusion coefficients can show whether a particular sample was handled with greater reliance on convolutional or Transformer-derived information. These visualizations should be compared with the input image and, in the medical setting, with expert interpretation. They should not be presented as independent evidence of diagnostic correctness.

9. Hypothesis Evaluation

Hypothesis	Required comparison	Decision basis
H1	Adaptive versus fixed fusion	Performance difference across repeated runs
H2	Hybrid versus standalone models	Consistency of improvement across metrics
H3	Explanation outputs and evaluator assessment	Usefulness and clarity of explanations
H4	FLOPs/latency and branch contribution	Efficiency without unacceptable performance loss

10. Medical-Imaging Application: Lung-Cancer Classification

The proposed medical application uses MRI/CT images as a domain-specific test case. Such images may contain both fine textural information and larger anatomical context, making the local/global representation problem relevant. Before use, the dataset must be de-identified, appropriately labelled and divided at the patient level. Class imbalance, acquisition differences and annotation quality should be reported explicitly.

The framework should be described as an experimental decision-support research model unless and until it has undergone appropriate clinical validation. Any interpretation of model outputs should involve qualified experts and should not replace clinical judgment.

11. Discussion

The main methodological proposition is that feature fusion should respond to the content of an individual image. CNN and Transformer branches provide different forms of information, so assigning them the same contribution to every sample may be unnecessarily restrictive. The adaptive gate offers a direct mechanism for changing that balance. The study also treats efficiency and interpretability as evaluation dimensions rather than secondary observations.

A key strength of this design is that its main claim can be tested through controlled comparisons. If the adaptive model improves predictive measures while keeping computational cost within an acceptable range, the fusion mechanism has empirical support. If the improvement is small or absent, the same experimental framework can reveal whether the additional gating complexity is justified.

12. Limitations

- The exact CNN and Transformer backbones are not fixed in the supplied material.
- The custom MRI/CT dataset size, provenance and class distribution are not specified.
- Training hyperparameters, hardware details and the number of repeated trials remain to be defined.
- No completed experimental measurements were supplied, so numerical superiority cannot be claimed.
- Interpretation maps require careful validation and should not be treated as clinical proof.

13. Ethical Considerations

Medical data should be de-identified and managed under the relevant institutional, ethical and data-governance requirements. Where patient-level information is involved, permissions and applicable ethics procedures should be documented. Model outputs should remain within the intended research scope until suitable external and clinical validation has been completed.

14. Expected Contributions

- An input-dependent mechanism for combining convolutional and Transformer representations.
- A controlled comparison with standalone and fixed-fusion baselines.
- An analysis procedure linking fusion coefficients with attention visualization.
- Joint assessment of predictive quality and computational cost.
- A reproducible experimental framework for a medical-imaging classification case study.

15. Conclusion

This paper presents an adaptive feature-fusion framework for hybrid CNN–Transformer image classification. The design separates local representation learning from global contextual modelling and allows their relative contribution to change with the input. Attention visualization and fusion-weight inspection are included so that the experimental analysis can examine not only what the model predicts but also which representation appears to influence that prediction.

The revised study preserves the datasets, objectives and evaluation direction stated in the supplied paper while rewriting the presentation and avoiding unsupported numerical claims. The framework can be validated through repeated experiments, ablation studies and matched baseline comparisons. For the proposed lung-cancer application, additional clinical and ethical safeguards are necessary before any real-world diagnostic interpretation is considered.

References

1. Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

2. Dosovitskiy, A., et al. (2021). An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
4. Liu, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision.
5. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision.
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. Proceedings of the ACM SIGKDD International Conference.
7. Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. University of Toronto Technical Report.
8. Deng, J., et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.