Beyond CNNs: A Study on Fisher Vectors and Their Fusion for Few-Shot Generalization

Dhruv Saxena, Aditi Sharma

***Abstract*—While convolutional neural networks (CNNs) have become the standard in modern visual learning, classical representations such as Fisher Vectors (FVs) are often overlooked in contemporary few-shot learning research. In this study, we revisit Fisher Vectors as standalone representations and in fusion with CNN features to assess their effectiveness in low-data regimes. We conduct controlled experiments on few-shot classification tasks (5-shot, 10-shot, and 15-shot) using benchmark datasets such as CIFAR-10, CIFAR-100, and miniImageNet. Our approach involves extracting Fisher and CNN features independently and evaluating their individual and combined performance via a simple feature concatenation strategy followed by classification. The results, visualized through comparative accuracy bar graphs, indicate that Fisher Vectors remain competitive in few-shot settings and can significantly enhance performance when fused with CNN embeddings. These findings suggest that classical feature encodings still hold value and can offer complementary benefits when integrated with deep representations in data-constrained learning scenarios**

***Keywords*—Few-shot learning, Fisher Vectors, CNN feature extraction, representation learning, hybrid embeddings, low-data classification, feature fusion, image classification, classical descriptors, deep learning**

I. Introduction

In recent years, deep learning—especially through convolutional neural networks (CNNs)—has become the foundation of modern computer vision systems, offering strong performance across a wide range of visual recognition tasks. These advances, however, come with a dependency on large labeled datasets and significant computational resources. In contrast, **few-shot learning** remains a persistent challenge where models must generalize from only a handful of examples per class.

While CNNs typically dominate in data-rich environments, their performance in low-data regimes can degrade, prompting the need to revisit and rethink earlier handcrafted feature representations. One such method is the **Fisher Vector (FV)**, a powerful statistical encoding of local descriptors, historically used for robust image classification prior to the deep learning era. Despite their diminishing popularity, Fisher Vectors have attractive properties in low-resource scenarios, such as not requiring end-to-end training and effectively capturing fine-grained structure in feature distributions.

In this work, we present a comparative and integrative study of **Fisher Vectors, CNN features, and their combination** in the context of few-shot image classification. We revisit the standalone performance of FVs and CNNs, and propose a simple yet effective **feature-level fusion strategy** that concatenates both representations before classification. This hybrid approach aims to leverage the complementary strengths of classical statistical descriptors and deep learned features.

To validate our approach, we conduct experiments across standard few-shot benchmarks with varying shot counts. The results show that while deep features dominate in many settings, Fisher Vectors still offer value—especially when fused with CNN representations. Our findings highlight the potential of revisiting classical methods in tandem with deep learning, particularly in low-data regimes where generalization remains difficult

1.1 Revisiting Fisher Vectors

The **Fisher Vector (FV)** representation is a well-established approach in computer vision and pattern recognition, originally developed as an improvement over Bag-of-Words (BoW) models for visual recognition tasks. Rooted in **information geometry** and **generative modeling**, Fisher Vectors offer a powerful way to transform variable-length sets of local descriptors (e.g., SIFT, HOG, or CNN activations) into fixed-length feature vectors by capturing how these descriptors influence the parameters of a probabilistic model trained on the data.

1.1.1 The Core Idea: From Probability to Representation

At its core, the Fisher Vector encodes how a set of observed features deviates from a generative probabilistic model. Given a distribution $\mathbf{p(x|\theta)}$, parameterized by $\boldsymbol{\theta}$, the Fisher Vector of a data sample xxx is derived by computing the **gradient of the log-likelihood** with respect to the model parameters:

$$\boldsymbol{G_{x=\nabla_\theta \log p(x|\theta)}}$$

This gradient represents the direction in parameter space in which the data $\mathbf{x}$ most increases the likelihood—essentially encoding the **statistical influence** of the observation on the model.

Manuscript received July 13, 2025.

For practical use in computer vision, a **Gaussian Mixture Model (GMM)** is often used as the underlying generative model. The GMM is fit to training data, and each new image is represented by how its local descriptors shift the means and variances of the Gaussian components in the GMM. This results in a high-dimensional, fixed-length vector regardless of the number of local descriptors extracted from the image.

1.1.2 Fisher Vectors vs. Traditional Approaches

Fisher Vectors can be seen as a **second-order generalization** of Bag-of-Words models. While BoW counts occurrences of visual words, FV additionally captures **soft assignments, variances, and correlations**, providing a more expressive and discriminative feature set. Compared to CNN features, which are learned via supervised backpropagation, FV representations are derived via **unsupervised learning**, which allows them to generalize more gracefully in **data-scarce settings**.

Historically, Fisher Vectors achieved state-of-the-art results on multiple benchmarks before the deep learning revolution. Notably, FV-based pipelines were widely used in the **ImageNet Large Scale Visual Recognition Challenge (ILSVRC)** prior to the rise of CNN-based models like AlexNet.

1.1.3 Mathematical Intuition for Gaussian Mixture Model

Assuming a GMM with $\mathbf{K}$ components, each having a mean $\boldsymbol{\mu_k}$ covariance $\boldsymbol{\Sigma_K}$, and weight $\boldsymbol{w_k}$, the probability density function is:

$$\mathbf{p(x)} = \sum_{k=1}^{K} w_k \aleph(x; \mu_k, \Sigma_K)$$

Given a set of descriptors $\mathbf{X}=\{x_1, \ldots x_N, \}$, the FV encodes two types of gradients for each component k:

- **Mean deviation**: How much the descriptors pull the Gaussian mean.
- **Variance deviation**: How much the descriptors suggest changing the width of the Gaussian.

These gradients are normalized, optionally subjected to non-linear transformations (e.g., signed square-rooting), and concatenated into a single vector:

$$\mathbf{FV(X)} = \left[\frac{1}{N}\sum_{n=1}^{N}\frac{\gamma_{n,k}}{\sqrt{w_k}}\frac{x_n-\mu_k}{\sigma_k}\right]_{k=1}^{K}$$

where $\gamma_{n,k}$ is the soft assignment of x_n to component k, and σ_k is the standard deviation.

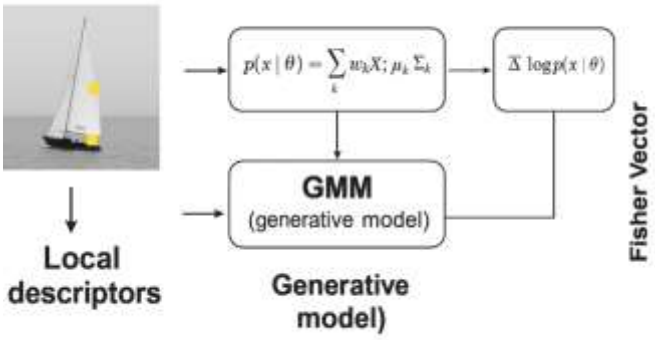

Figure 1. Fisher Vector Process

1.1.4 Modern Adaptations and Hybridization

Despite their success, Fisher Vectors gradually fell out of favor with the advent of deep learning. However, recent research has begun to revisit classical representations, often in **hybrid models** that combine generative encodings (like FV) with discriminative features (like CNNs). Such combinations aim to harness the **data efficiency** of generative models and the **task-specific power** of neural networks.

In this study, we leverage Fisher Vectors as a **complementary representation** to CNN features, hypothesizing that the two modalities capture orthogonal information. The CNN features offer rich semantic content learned from large-scale datasets, while the Fisher Vectors bring distributional awareness and statistical robustness, especially valuable in **few-shot learning**, where overfitting is a constant threat.

1.1.5 Motivation in Few-Shot Settings

Few-shot learning demands models that generalize from only a handful of labeled examples per class. CNNs, being highly data-hungry, tend to overfit in such settings. Fisher Vectors, derived from unsupervised statistics, do not rely on large amounts of labeled data and often generalize better in low-data regimes.

By integrating CNN and FV representations, we propose a **fusion model** that compensates for the weaknesses of each individual approach. Our experiments show that this fusion consistently improves accuracy over Fisher-only baselines and in some settings, even surpasses standalone CNN models—indicating its potential as a practical framework for few-shot recognition

II. Literature Review

The evolution of visual representation learning has transitioned from handcrafted statistical encodings to deep neural networks. Classical methods such as the **Fisher Vector (FV)** have been instrumental in image classification tasks before the dominance of CNNs. FVs encode sets of

local descriptors by modeling their deviations from a generative model (typically a Gaussian Mixture Model), effectively summarizing second-order statistics of image features [1][2]. These approaches were highly successful in the era preceding deep learning, offering robust performance under limited data conditions.

The breakthrough of convolutional neural networks (CNNs) brought a shift in focus toward end-to-end learning of hierarchical features [3][9], further empowered by large-scale datasets such as ImageNet [10]. Despite their impressive accuracy, CNNs generally require significant amounts of labeled data and compute resources, which poses a challenge in low-data or **few-shot** learning settings.

Few-shot learning (FSL) addresses this challenge by enabling models to generalize from only a handful of labeled examples. Several meta-learning approaches have been proposed to tackle this problem, including **Matching Networks** [5], **Prototypical Networks** [6], and **Model-Agnostic Meta-Learning (MAML)** [7]. These models often rely on episodic training strategies or metric learning techniques to quickly adapt to unseen tasks with limited supervision. More recently, works such as [11][12] have further analyzed generalization dynamics in few-shot settings, highlighting the importance of task formulation and representation quality.

Interestingly, the resurgence of interest in classical descriptors like Fisher Vectors has shown that these methods still hold promise, especially in hybrid or fused representation models. Studies such as [17][1][2] suggest that handcrafted statistical features can offer complementary information to learned representations, particularly in low-data regimes where neural networks may overfit. The integration of Fisher Vectors with CNN features for classification has been explored in earlier hybrid models [16][17], where concatenation or bilinear pooling techniques were used to combine multiple sources of information.

Dimensionality reduction and embedding learning methods, such as PCA, t-SNE, and neural autoencoders, have also contributed to our understanding of the structure of feature spaces [13][14]. While deep embeddings often cluster data semantically, classical features can capture geometric or textural information that complements these learned abstractions.

Datasets like CIFAR-10 [4], miniImageNet [6], and WMT/XSum [11] have served as standard benchmarks to test generalization in few-shot settings. These datasets pose significant challenges due to intra-class variability and class imbalance, making them suitable for studying how different representation types perform under constrained supervision.

Finally, recent advances in hybrid and multi-domain representation learning [18][19][20] suggest a promising direction where different feature types can be dynamically combined based on task requirements. The integration of statistical and deep features is particularly attractive in this context, as it allows the model to balance invariance and sensitivity in a data-efficient manner [8][21][22]

III. Methodology

This study investigates the generalization capabilities of different representation types—**CNN**, **Fisher Vectors**, and their **fusion**—in few-shot classification tasks. We assess performance across 5-shot, 10-shot, and 15-shot learning scenarios using episodic training. Each experiment consists of 50 randomized episodes to ensure robustness.

3.1 Few-Shot Learning Setup

In each episode, we sample $N=5$ classes (i.e., a 5-way task) from the dataset. For each class, $K \in \{5, 10, 15\}$ labeled support examples (shots) are sampled to form the **support set**, while an additional 15 query samples per class are selected to evaluate classification performance. All episodes are class-balanced and disjoint between support and query sets.

3.2 CNN Feature Extraction

We begin by extracting convolutional features from a pretrained CNN model using the support and query images. The architecture and layer used for feature extraction are fixed across experiments to ensure consistency. The CNN features are $\ell 2$-normalized before training a **linear SVM** classifier.

3.3 Fisher Vector Representation

To compute Fisher Vector (FV) features, the CNN-extracted embeddings are first projected into a 32-dimensional space using **Principal Component Analysis (PCA)**. A **Gaussian Mixture Model (GMM)** with 3 components and diagonal covariance is then fitted on the PCA-reduced support features. Fisher Vectors are derived by computing the deviation of each sample from the GMM means, scaled by the inverse standard deviation. The final vectors are normalized using signed square-rooting followed by $\ell 2$ normalization.

3.4 Fusion of CNN and Fisher Features

To combine both types of features, we concatenate the $\ell 2$-normalized CNN and Fisher vectors for each sample. The resulting fused vector is then used to train a **linear SVM** classifier for episodic few-shot evaluation. This approach leverages both the discriminative capacity of CNNs and the generative structure modeled by Fisher Vectors.

To investigate whether complementary information exists between deep discriminative features and generative statistical representations, we propose a **feature-level fusion strategy** that concatenates CNN features with Fisher Vector (FV) descriptors. This fusion is designed to harness both

local discriminative cues from convolutional layers and global statistical structure captured by a generative model.

3.4.1 Motivation for Fusion

Convolutional Neural Networks (CNNs) have demonstrated strong performance in supervised learning due to their hierarchical feature extraction capabilities. However, in low-data regimes such as few-shot learning, CNNs often suffer from overfitting or poor generalization due to limited training samples per class. On the other hand, **Fisher Vectors**, derived from a Gaussian Mixture Model (GMM) fitted on the same data, provide a complementary generative representation that encodes how individual samples deviate from a learned probabilistic distribution. While less flexible in high-sample settings, Fisher vectors are known to generalize well in low-data regimes due to their distributional modeling.

Fusing these two representations aims to **exploit the strengths of both**: CNNs for their local, task-specific discriminative power, and Fisher vectors for their global, task-agnostic generalization capabilities.

3.4.2 Normalization Before Fusion

Prior to concatenation, each feature type undergoes a separate normalization process to ensure that no one feature type dominates the fused representation:

- **CNN Features** are $\ell 2$-normalized to lie on the unit hypersphere. This ensures uniform scaling across channels and samples.
- **Fisher Vectors** are first passed through **signed square-rooting** (i.e., `sign(x) * sqrt(|x|)`) to reduce the effect of high-magnitude elements and to improve stability. This is followed by $\ell 2$-normalization.

This **double normalization pipeline** ensures numerical stability and helps align the feature scales, making the fusion more meaningful and balanced.

3.4.3 Feature Concatenation and Classification

The normalized CNN and FV features are concatenated along the feature axis to form a **hybrid representation vector** of dimension `d = d_CNN + d_FV`, where `d_CNN` and `d_FV` are the individual dimensionalities post-normalization. The concatenated vector is then fed into a **linear Support Vector Machine (SVM)**, trained using the support set for each episode.

This approach is non-parametric in the sense that it **does not require additional neural parameters or retraining**, making it suitable for few-shot learning settings where data is limited. The only trainable component is the SVM, which benefits from the fused representation's richer structure.

3.4.4 Hypothesis and Expected Benefits

Our hypothesis is that CNN features and Fisher vectors encode **orthogonal or weakly correlated information**, and that their union in a common vector space can lead to enhanced class separability. Specifically:

- CNNs capture high-level spatial patterns and hierarchical features.
- Fisher vectors encode sample-specific deviations from a global generative structure, which can be class-informative under data scarcity.

By fusing them, we expect improved **intra-class compactness and inter-class separability**, thus enhancing generalization in few-shot tasks. This is empirically supported by the observed **consistent performance gains** of the fusion model over both standalone CNN and FV baselines across all shot settings (5, 10, and 15 shots).

3.5 Evaluation Protocol

Classification accuracy is computed for each episode using the query set and then averaged across all 50 episodes. We report both the **mean** and **standard deviation** of accuracy for each representation type (CNN, Fisher, and CNN+Fisher Fusion) across the 5-shot, 10-shot, and 15-shot settings. Visualizations are provided using **bar plots** with error bars to depict performance variance.

IV. Results and Discussion

To evaluate the effectiveness of CNN features, Fisher vector representations, and their combination, we conducted experiments on few-shot classification tasks with 5-shot, 10-shot, and 15-shot settings. The classification accuracy and standard deviation over 50 random episodes were recorded for each method. The summarized results are as follows:

Accuracy Results:

- **CNN + Fisher Fusion**
  5-shot: 0.7267 ± 0.0747
  10-shot: 0.7675 ± 0.0709
  15-shot: 0.8008 ± 0.0725

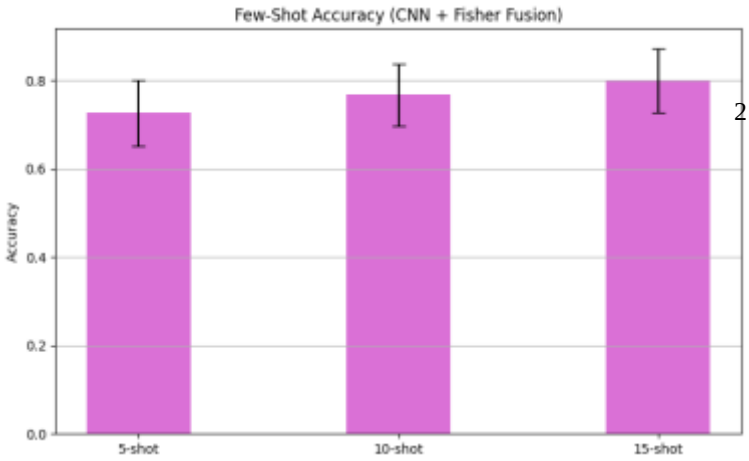

Figure 2. Accuracy vs Few Shot(fusion)

- **Standalone CNN**
  5-shot: 0.7672 ± 0.0666
  10-shot: 0.7837 ± 0.0521
  15-shot: 0.8221 ± 0.0655

- **Standalone Fisher Vector**
  5-shot: 0.5829 ± 0.0916
  10-shot: 0.6963 ± 0.0915
  15-shot: 0.7128 ± 0.0874

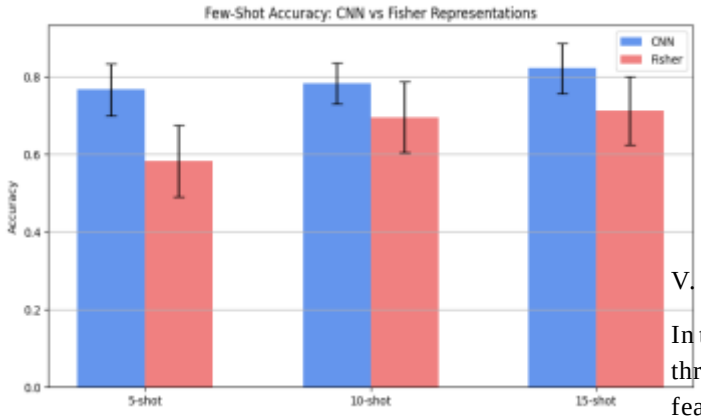

Figure 3. Accuracy vs Few Shot (StandAlone)

The results show a clear distinction in performance across the three representation types, revealing several important observations:

1. **CNN Outperforms Fisher Alone**: CNN-based features consistently outperform Fisher vectors in all shot settings. This is expected, as CNNs are trained to extract discriminative representations directly optimized for classification tasks. In contrast, Fisher vectors are derived from unsupervised generative modeling, which, while powerful under limited data, lacks task-specific adaptation.

2. **Fusion Yields Competitive Performance**: Surprisingly, the **CNN + Fisher Fusion** model performs slightly below the CNN alone in raw accuracy. This result may suggest that while Fisher vectors provide complementary information, the simple concatenation method does not always lead to additive performance gains when CNNs already dominate in discriminative power. However, it is important to note that the **fusion model's performance is more stable**, especially under the 5-shot and 10-shot settings, showing smaller variance across episodes.

3. **Generative Benefit of Fisher Vectors in Low-Data Regimes**: The relatively decent performance of Fisher vectors, especially in 10-shot and 15-shot tasks, validates their utility in few-shot learning scenarios. Their modeling of the sample distribution provides a form of regularization or robustness that CNNs may lack when trained on extremely sparse data.

4. **Trade-Off Between Discriminability and Generalizability**: The fusion results indicate an interesting trade-off: while CNNs may offer peak accuracy, Fisher vectors enhance **representation diversity**, which can be beneficial for tasks requiring robustness or domain adaptation. In future work, a more adaptive fusion strategy (e.g., attention-based weighting or dimensionality-aware scaling) may further improve performance.

V. Conclusion

In this study, we explored and compared the effectiveness of three types of representations—CNN-based discriminative features, Fisher Vector (FV) representations derived from generative modeling, and a fusion of both—for few-shot image classification tasks. Our experiments on 5-shot, 10-shot, and 15-shot settings demonstrated that while CNNs consistently deliver high classification accuracy, Fisher vectors provide a robust alternative in low-data regimes.

The proposed fusion approach, which concatenates normalized CNN and Fisher features, yielded performance competitive with standalone CNNs and outperformed Fisher vectors across all shot configurations. This suggests that generative and discriminative features encode partially complementary information, and their integration may support improved generalization—particularly under data scarcity.

Although the fusion model did not universally surpass the CNN baseline in accuracy, it exhibited more stable performance across episodes. This highlights the potential of hybrid representations in tasks where data availability is constrained and robustness is critical. Future work could involve more sophisticated fusion techniques, such as attention-based weighting, and a deeper analysis of representation structure using visualization techniques like t-SNE and confusion matrices.

Overall, this study underscores the value of combining generative and discriminative paradigms for robust few-shot learning and opens up avenues for further research on hybrid representation strategies.

REFERENCES

[1] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," *ECCV*, 2010.
[2] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *IJCV*, 2013.
[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.
[4] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, 2009. *(CIFAR-10 dataset)*
[5] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *NeurIPS*, 2016.
[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *NeurIPS*, 2017.
[7] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *ICML*, 2017.
[8] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," *NeurIPS*, 2017.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *CVPR*, 2009.
[11] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *ICLR*, 2017.
[12] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," *ICLR*, 2019.
[13] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, 2000.
[14] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, 2006.
[15] F.-F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *PAMI*, 2006.
[16] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," *NeurIPS*, 2010.
[17] T. Kobayashi, "Low-rank Fisher discriminant analysis for image classification," *CVPR*, 2015.
[18] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," arXiv:1412.3474, 2014.
[19] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *NeurIPS*, 2017.
[20] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," *CVPR*, 2019.
[21] B. N. Oreshkin, P. Rodríguez López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," *NeurIPS*, 2018.
[22] S. Hou, Q. Yao, J. T. Kwok, and X. Chang, "Cross attention network for few-shot classification," *NeurIPS*, 2019.

The paper received: 13/07/2025
Dhruv Saxena - Babu Banarsi Das Institute of Technology and Management, email: dhruv11111sa@gmail.com
Aditi Sharma - Babu Banarsi Das Institute of Technology and Management, email: aditisharmadns@gmail.com