ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

VIDEO CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS

Boga Vaishnavi [1], *Gopari Gouthami [2], Gopari Prasanna [3], B. Rahul kumar [4]

[1]UG Student VignanS Institute of Management and Technology for Women, Ghatkesar, Telangana -501301

[2]Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

[3, 4] UG Student, St. Martin's Engineering College, Secunderabad, Telangana – 500100

*Corresponding Author

Email: ggouthamiit@smec.ac.in

Abstract:

Convolutional Neural Networks (CNNs) have been established as a powerful class of models for image recognition problems. Encouraged by these results, we pro-vide an extensive empirical evaluation of CNNs on large- scale video classification using a new dataset of 1 million YouTube videos belonging to 487 classes. We study multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatial-temporal information and suggest a multi resolution, for related architecture as a promising way of speeding up the training. Our best spatio temporal networks display significant performance improvements compared to strong feature-based baselines (55.3% to 63.9%), but only a surprisingly mod- est improvement compared to single-frame models (59.3% to60.9%).We further study the generalization performance of our best model by retraining the top layers on the UCF- 101 Action Recognition dataset and observe significant performance improvements compared to theUCF-101 baseline model (63.3% up from 43.9%).

1. Introduction

Images and videos have become ubiquitous on the internet, which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization. Recently, Convolutional Neural Networks (CNNs) [15] have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval [11, 3,2, 20,9, 18].The key enabling factors behind these results were techniques for scaling up the networks to tens of millions of parameters and massive labeled datasets that can support the learning process. Under these conditions, CNNs have been shown to learn powerful and interpretableimage features [28].Encouraged by positive results in do- main of images, we study the performance of CNNs in large-scale video classification, where the networks have access to not only the appearance information present in single, static images, but also their complex temporal evolu- tion. There are several challenges to extending and applying CNNs in this setting.

2. Related Work

The standard approach to video classification [26, 16, 21, 17] involves three major stages: First, local visual fea- tures that describe a region of the video are extracted ei- ther densely[25]orates parseset of interest points[12,8]. Next, the features get combined into a fixed-sized video- level description. One popular approach is to quantize all features using a learned k-means dictionary and accumulate the visual words over the duration of the video into histograms of varying spatio-temporal positions and extents [13].Lastly, a classifier (such as an SVM) is trained on the resulting "bag of words" representation to distinguish among the visual classes of interest.

Convolutional Neural Networks[15]area biologically- inspired class of deep learning models that replace all three stages with a single neural network that is trained end to end from raw pixel values to classifier outputs. The spa- tial structure of images is explicitly taken advantage of for regularization through restricted connectivity between layers(local filters), parameter sharing(convolutions) and special local in variance-building neurons(max pooling).Thus, these architectures effectively shift the required engineering from feature design and accumulation strategies to de-sign of the network connectivity structure and hyper param- eter choices.

3.Models

Unlike images which can be cropped and rescaled to a fixed size, videos vary widely in temporal extent and can- not be easily processed with a fixed-sized architecture. In this work we treat every video as a bag of short, fixed-sized clips. Since each clip contains several contiguous frames in time, we can extend the connectivity of the network in time dimension to learn spatio-temporal features. There are multiple options for the precise details of the extended connectivity and we describe three broad connectivity pattern categories(EarlyFusion, Late Fusion and Slow Fusion)be- low. Afterwards, we describe a multi resolution architecture for addressing the computational efficiency.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

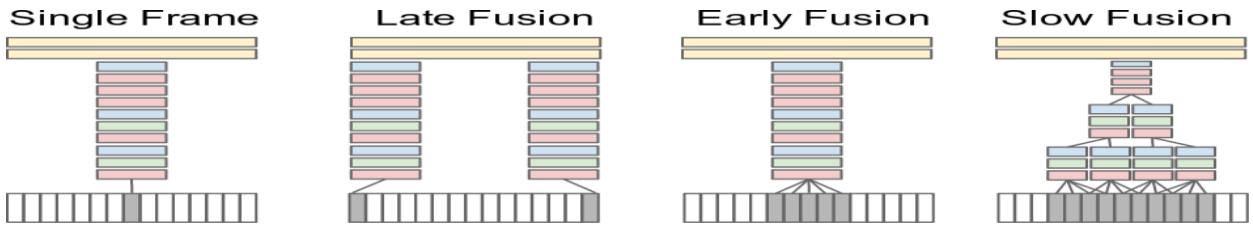

Figure1: Explored approaches for fusing information over temporal dimension through the network.
Red, green and blue boxes indicate convolutional, normalization and pooling layers respectively. In the Slow Fusion model, the depicted columns share parameters.

Time Information Fusion in CNNs

We investigate several approaches to fusing information across temporal domain (Figure 1): the fusion can be done early in the network by modifying the first layer convolutional filters to extend in time, or it can be done late by placing two separate single-frame networks some distance in time apart and fusing their outputs later in the processing. We first describe a base line single frame CNN and then discuss its extensions in time according to different types of fusion.

Multi resolution CNNs

Since CNNs normally take on order so weeks to train on large-scale datasets even on the fastest available GPUs, the runtime performance is a critical component to our ability to experiment with different architecture and hyper parameter settings. This motivates approaches for speeding up the models while still retaining their performance. There are multiple front soothes endeavors, including improvements in hardware, weight quantization schemes, better optimization algorithms and initialization strategies, but in this work we focus on changes in the architecture that enable faster running times without sacrificing performance.

One approach to speeding up the networks is to reduce the number of layers and neurons in each layer, but simi- lar to [28] we found that this consistently lowers the performance. Instead of reducing the size of the network, we conducted further experiments on training with images of lower resolution. However, while this improved the running time of the network, the high-frequency detail in the images proved critical to achieving good accuracy.

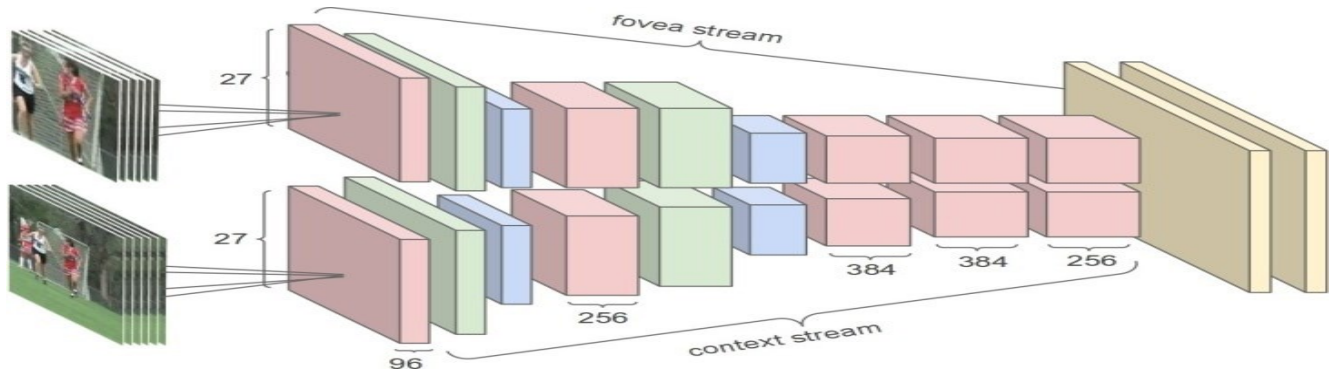

Figure 2: Multiresolution CNN architecture.

4.Learning

Data augmentation and preprocessing.

Following[11],we take advantage of data augmentation to reduce the effects of overfitting. Before presenting an example to a network, we preprocess all images by first cropping to center region, resizing them to 200 ×200 pixels, randomly sampling a 170×170 region, and finally randomly flipping the images horizontally with 50% probability. These preprocessing steps are applied consistently to all frames that are part of the same clip. A sala step of preprocessing. We subtract a constant value of 117 from raw pixel values, which is the approximate value of the mean of all pixels in our

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

images.

5.Results

We first present results on our Sports-1M dataset and qualitatively analyze the learned features and network predictions. We then describe our transfer learning experiments on UCF-101.

Experiments on Sports-1M

Dataset. The Sports-1M dataset consists of 1 million YouTube videos annotated with 487 classes. The classes are arranged in a manually curated tax on omy that contains internal nodes such as *Aquatic Sports, Team Sports, Winter Sports, Ball Sports, Combat Sports, Sports with Animals,* and generally becomes fine-grained by the leaf level. For example, our data set contains 6 different types of bowling, 7 different types of American football and 23 types of billiards.

Training. We trained our models over a period of one month, with models processing approximately 5 clips per second for full-frame networks and up to 20 clips per second for multi-resolution networks on assign le model replica. The rate of 5 clips per second is roughly 20 times slower than what one could expect from a high-end GPU, but, we expect or each comparable speeds overall given that we use 10-50 model replicas. We further estimate the size of our dataset of sampled frames to be on the order of 50 million examples and that our networks have each seen approximately 500 million examples throughout the training period in total.

Video-level predictions. To produce predictions for an entirevideowerandomlysample20clipsandpresenteach clip individually to the network. Every clip is propagated throughthenetwork4times(with different crops and flips).

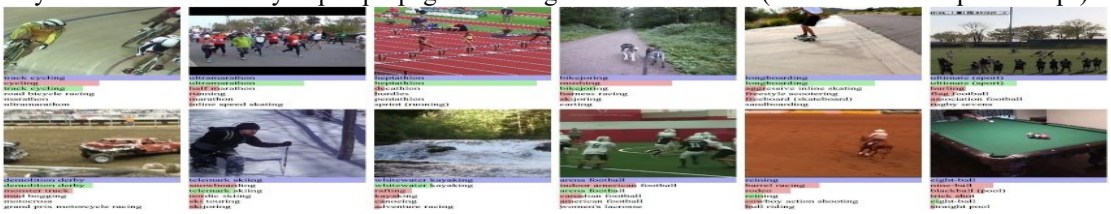

Figure 3:Predictions on Sports-1M test data.

Blue (first row) indicates ground truth label and the bars below show model predictions sorted in decreasing confidence. Green and red distinguish correct and incorrect predictions, respectively.

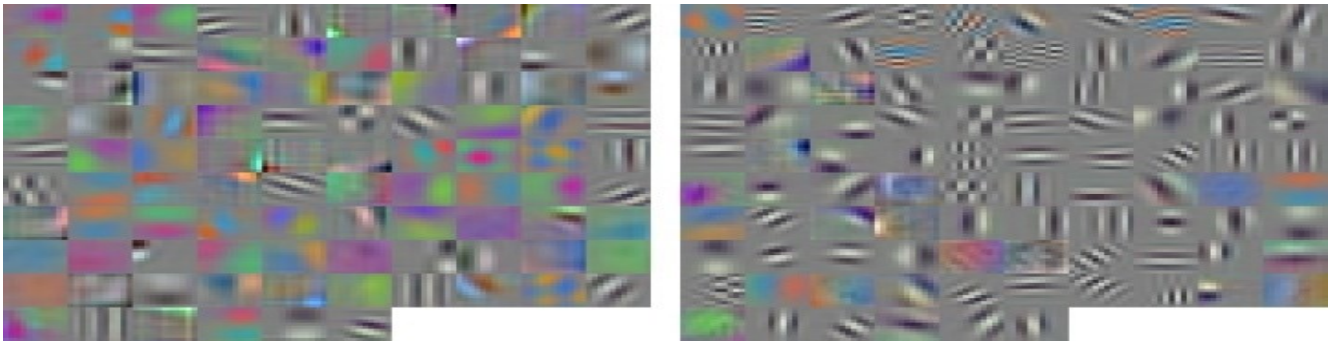

Figure 4. Filters learned on first layer of a multiresolution network.

Left: context stream, Right: fovea stream. No- table the fovea stream learns gray scale, high-frequency features while the context stream model slower frequencies and colors. GIF soft moving video features can be found on Number of nodes in all layers.

Quantitative results. The results for the Sports-1M dataset test set, which consists of 200,000 videos and 4,000,000 clips, are summarized in Table 1.As can be seen from the table, our networks consistently and significantly out, perform the

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

Model	ClipHit@ 1	VideoHit@ 1	VideoHit@ 5
Feature Histograms +Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multi res	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average(Single +Early +Late+ Slow)	41.4	63.9	82.4

feature-based baseline. feature-based approach densely over the produces predictions video-level feature networks only see 20 individual ally. seem to learn well

We emphasize that the computes visual words duration of the video and based on the entire vector, while our randomly sampled clips Moreover, our networks despite significant label noise the training videos are subject to incorrect annotations and even the correctly-labeled videos often contain a large amount of artifacts such as text,effects, cuts, and logos, none of which we attempted to filter out explicitly.

Contributions of motion. We conduct further experiments to understand the differences between the single- frame network and networks,that have access to motion information. We choose the Slow Fusion network as a representative motion-aware network because it performs best. We compute and compare the per-class average precision for all Sports classes and highlight the ones that exhibit largest differences. Manually inspecting some of the associated clips(Figure4),we qualitatively observe that the motion-aware network clearly benefits from motion in- formation in some cases, but these seem to be relatively uncommon .On the other hand, balancing the improvements from access to motion information, we observe that motion aware networks are more likely to under,perform when there is camera motion present. We hypothesize that the CNNs struggle to learn complete in variance across all possible angles and speeds of camera translation and zoom.

Table 1: Classes for which a (motion-aware) Slow Fusion CNN performs better than the single-frame CNN (left)and vice versa (right), as measured by difference in per-class average precision

5.Conclusions

Our results indicate that while the performance is not particularly sensitive to the architectural details of the connectivity in time, a Slow Fusion model consistently per- forms better than ,the early and late fusion alternatives. Surprisingly, we find that a single-frame model already displays very strong performance, suggesting that local motion cues may not be critically important, even for a dynamic dataset such as Sports. We also identified mixed-resolution architectures that consist of a low-resolution context and a high- resolution fovea stream as an effective way of speeding up CNNs without sacrificing accuracy.

Our transfer learning experiments on UCF-101 suggest that the learned features are generic and generalize other video classification tasks. In particular, we achieved the highest transfer learning performance by retraining the top 3 layers of the network.

In future work we hope to incorporate broader categories in the dataset to obtain more powerful and generic features, investigate approaches that explicitly reason about camera motion, and explore recurrent neural networks as a more powerful technique for combining clip-level predictions into global video-level predictions.

References

[1] M. Baccouche, F. Malamet , C Wolf, C. Garcia, and A.Baskar .Sequential deep learning for human action recognition. In *Human Behavior Understanding*, pages29–39.Springer,2011.2,3

[2] D.Ciresan, A.Giusti, J. Schmidhuber,etal .Deep neural net- works segment neuronal membranes in electron microscopy images. In *NIPS*, 2012. 1

[3] L. N. Clement Farabet, Camille Couprie and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8), 2013.1, 2

[4] C. Couprie, C. Farabet, L. Najman, and Y. LeCun.Indoor semantic segmentation using depth information. *Internatinal Conference on Learning Representation*, 2013.2

[5] N.Dalaland B.Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, 2005.5

[6] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le,M.Z.Mao,M.Ranzato,A.Senior,P.Tucker,K.Yang, and A. Y. Ng. Large scale distributed deep networks . In *NIPS*, 2012. 44

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei- Fei. ~~Imagenet~~Image net: A large-scale hierarchical image database. In *CVPR*, 2009. ~~2~~2

[8] P.Dolla´r V.Rabaud,G.Cottrell,andS.Belongie.Behav- iorrecognitionviasparsespatio-temporalfeatures.In*InternationalWorkshoponVisualSurveillanceandPerformance Evaluation of Tracking and Surveillance*, 2005.2, 5

[9] R.Girshick, J.Donahue, T.Darrell, andJ.Malik.Richfea- ture hierarchies for accurate object detection and semantic segmentation.In *CVPR*, 2014.1, 2

[10] S.Ji,W.Xu,M.Yang,andK.Yu.3Dconvolutionalneural networks for human action recognition.*PAMI*, 35(1):221– 231, 2013.2, 3

[11] A.Krizhevsky, I.Sutskever, andG.Hinton.Imagenetclas- sificationwithdeepconvolutionalneuralnetworks.In*NIPS*, 2012.1, 2, 3, 4

[12] I.Laptev. Onspace-timeinterestpoints. *IJCV*,64(2-3):107– 123, 2005. 2

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies.In *CVPR*, 2008.2

[14] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng.Learn- inghierarchicalinvariantspatio-temporalfeaturesforaction recognition with independent subspace analysis.In *CVPR*, 2011.2

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.Gradient- based learning applied to document recognition.*Proceed-ings of the IEEE*, 86(11):2278–2324, 1998.1, 2

[16] J. Liu, J. Luo, and M. Shah.Recognizing realistic actions from videos “in the wild”.In *CVPR*, 2009.2