Paper #1256: A Feature Embedding Strategy for High-Level CNN Representations from Multiple ConvNets

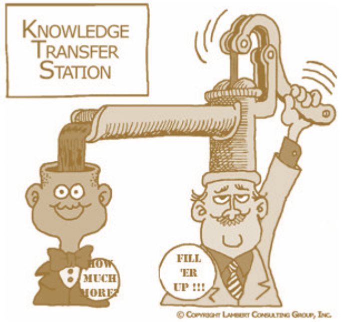
A Transfer learning approach

T. Akilan, Q.M.J. Wu
Department of Electrical and Computer Engineering
University of Windsor
Canada

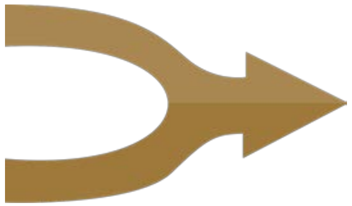
W. Jiang
Department of Control Science and Engineering
Zhejiang University
China

OUTLINE

Transfer learning

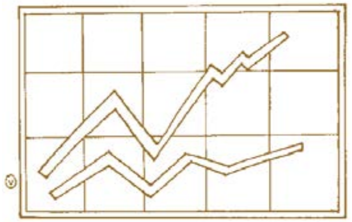


Fusion in computer vision



The proposed approach

Experimental Results, Discussion, Conclusion



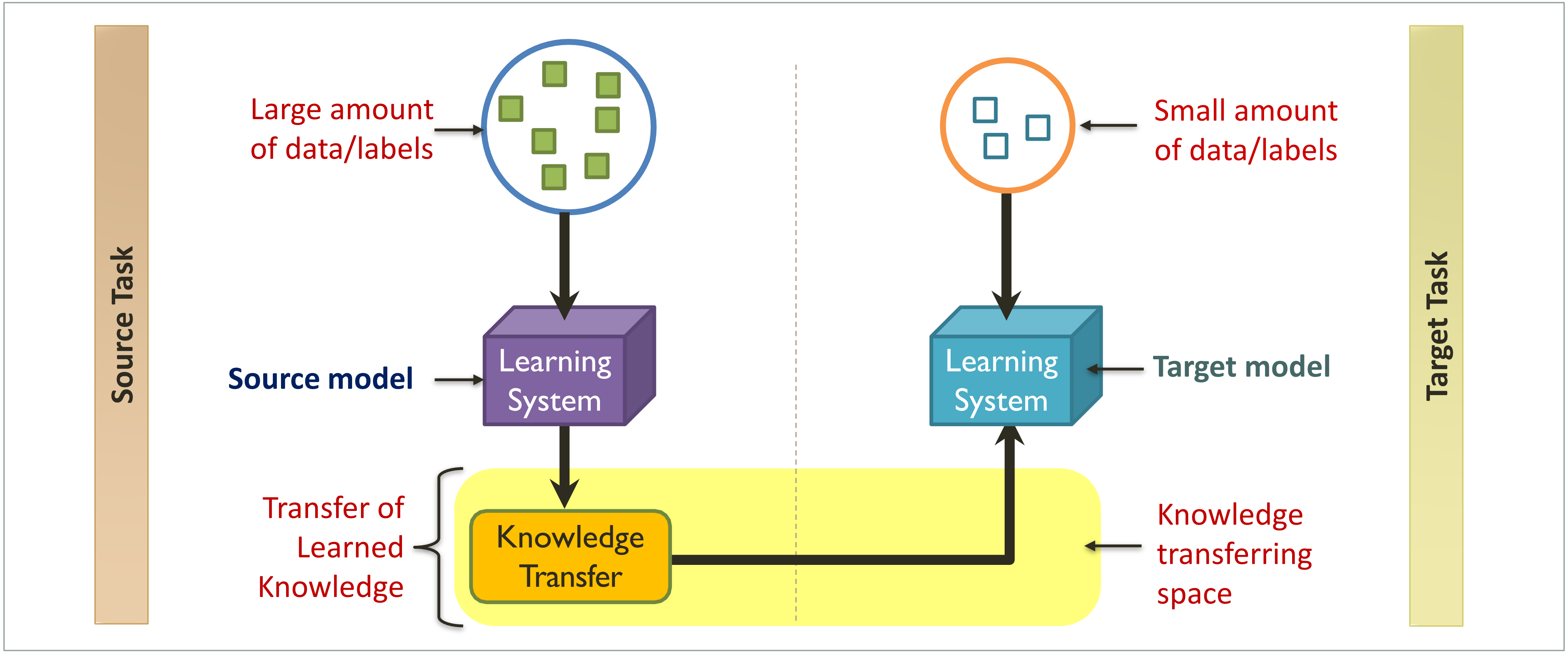
TRANSFER LEARNING AS PART OF DEEP LEARNING

* http://www.destination-innovation.com/

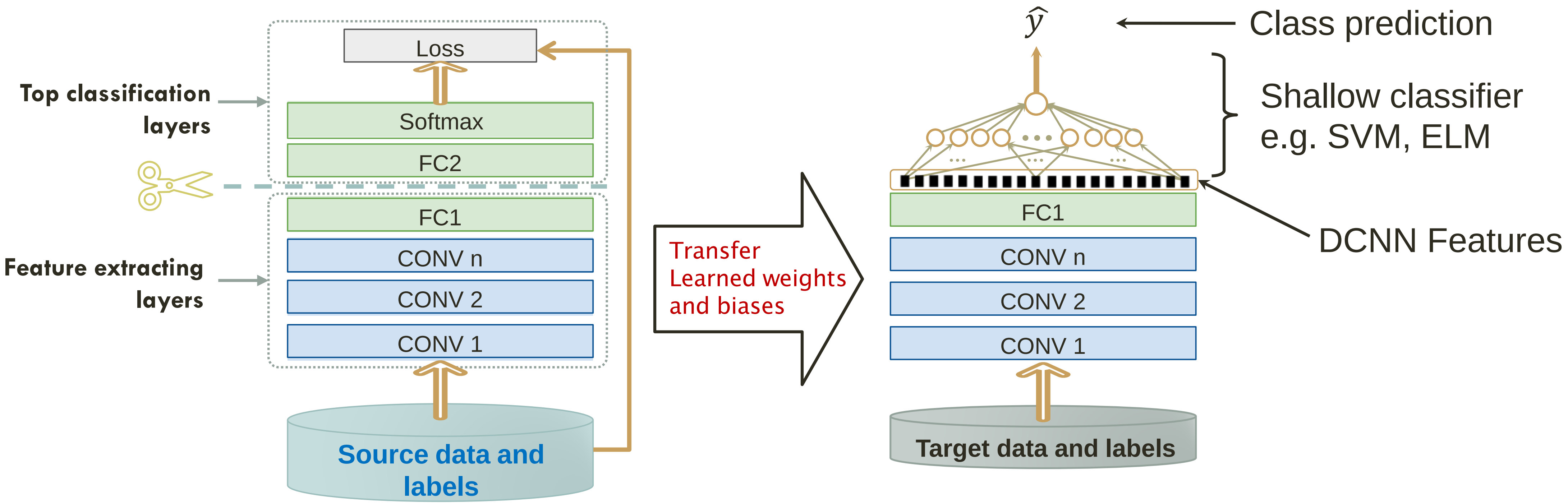
- Deep learning (DL) has become a driving force of the current revolution in computing.
- DL is the cornerstone everything from Self-Driving cars to language translation even generated art.
- Transfer learning is a technique to take the burden off from training deep neural networks (DNN).

- Transfer learning has been coexisted in Machine Learning (ML), Artificial Intelligence (AI), and Neural Network (NN).
- It is termed as knowledge transfer, meta learning, inductive transfer, parameter transfer, life-long learning, or context-sensitive learning [1].
- Transfer learning has the ability to extend what has been learned in one context to new contexts.

GENERAL CONTEXT OF TRANSFER LEARNING



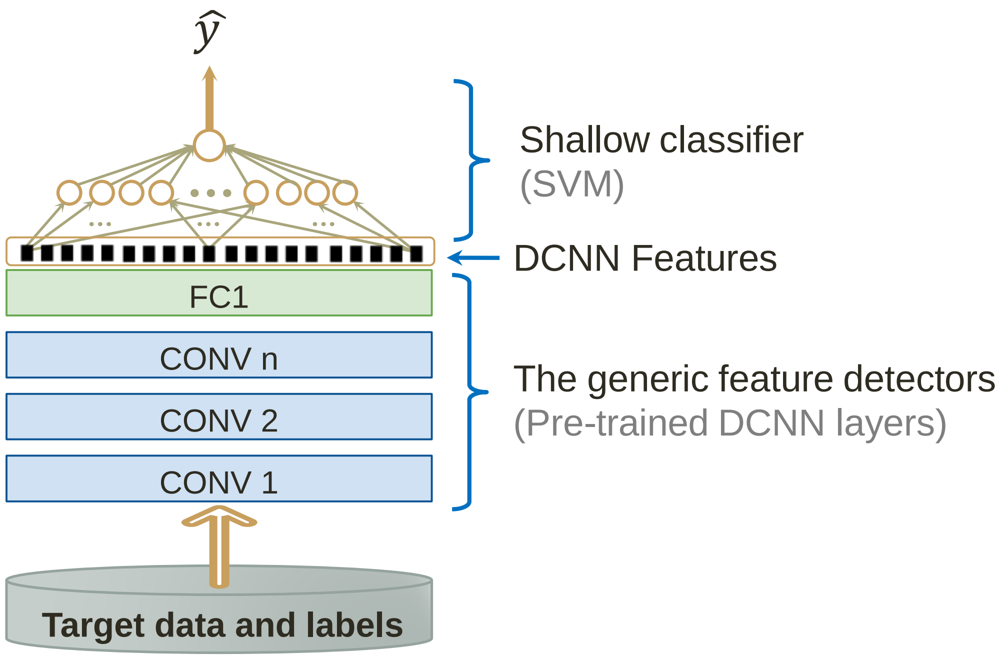
TRANSFER LEARNING: USING "OFF-THE-SHELF" DCNN



Idea:

- Use outputs of one or more layers of a Deep Convolutional Neural Network (DCNN) trained on a different task as **generic feature detectors**.
- Train a new shallow model on these features.

TRANSFER LEARNING: USING "OFF-THE-SHELF" DCNN



- Works surprisingly well in practice!
- Surpassed or on par with state-of-the-art in several tasks

Method	mean Accuracy
HSV [27]	43.0
SIFT internal [27]	55.1
SIFT boundary [27]	32.0
HOG [27]	49.6
HSV+SIFTi+SIFTb+HOG(MKL) [27]	72.8
BOW(4000) [14]	65.5
SPM(4000) [14]	67.4
FLH(100) [14]	72.7
BiCos seg [7]	79.4
Dense HOG+Coding+Pooling[2] w/o seg	76.7
Seg+Dense HOG+Coding+Pooling[2]	80.7
CNN-SVM w/o seg	74.7
CNNaug-SVM w/o seg	**86.8**

Sample results from [3]: Oxford 102 flowers dataset.

[3] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, CVPRW '14.

FUSION OF MULTI-MODEL CNN FEATURES

If a single modality CNN features provide improvement in classification accuracy, **the natural question arises asking:**

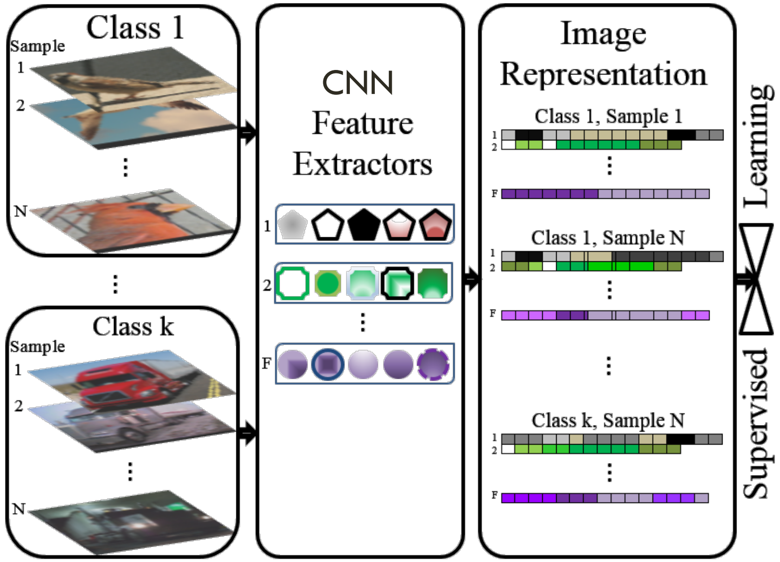
- How about a **fusion** of **multi-modality CNN features**?
- Would they have clues that **complement** each other?

FUSION OF MULTIMODAL CNN FEATURES

- **Fusion of multiple features and/or ensemble of classifiers** are efficient techniques to achieve better results for applications like classification and recognition [4].
- Multi-modal learning has been shown to improve learned representations in the unsupervised setting [16] and when used as a-priori known unrelated tasks [17].

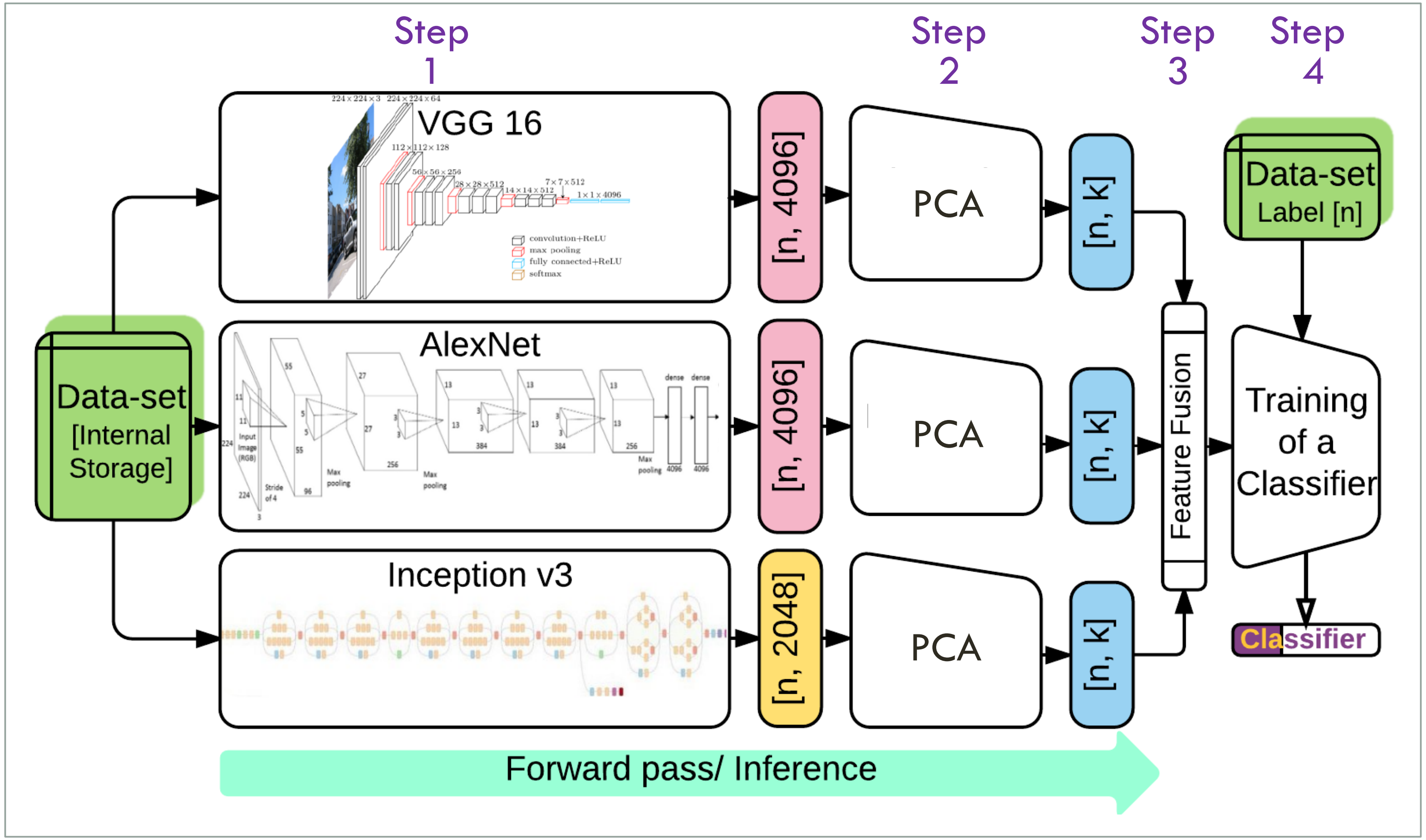
The rationale:

- All our **knowledge is based upon experience.**
- What we call **inferential knowledge**, in which we go **from less general to the more general.**
- General nature of the human brain, which is able to learn many different tasks (**benefit from transfer learning**).



General Overview of a Multi-modal Feature Representation

FUSION OF MULTIMODAL CNN FEATURES



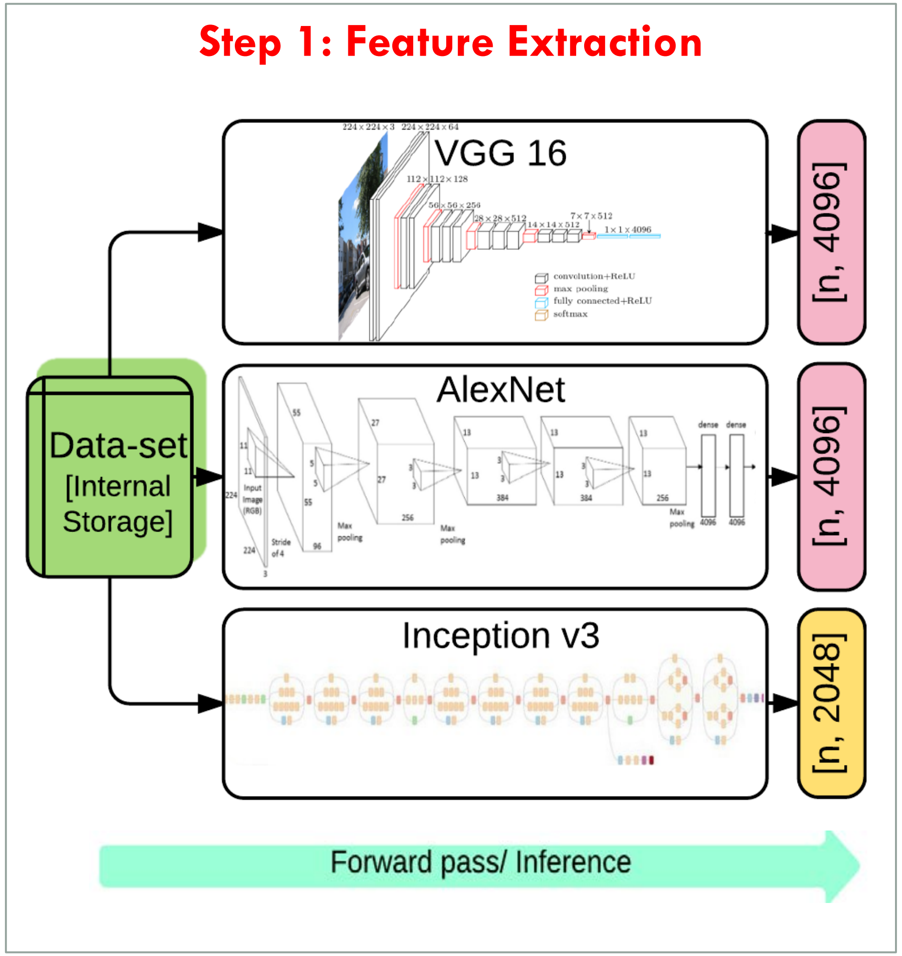
Step 1: Feature Extraction
Using multi-CNN models

Step 2: Feature Transformation
Using feature-energy and PCA

Step 3: Feature Fusion
Using arithmetic operators and pooling

Step 4: Classifier Training
Using a multi-class SVM

FUSION OF MULTIMODAL CNN FEATURES

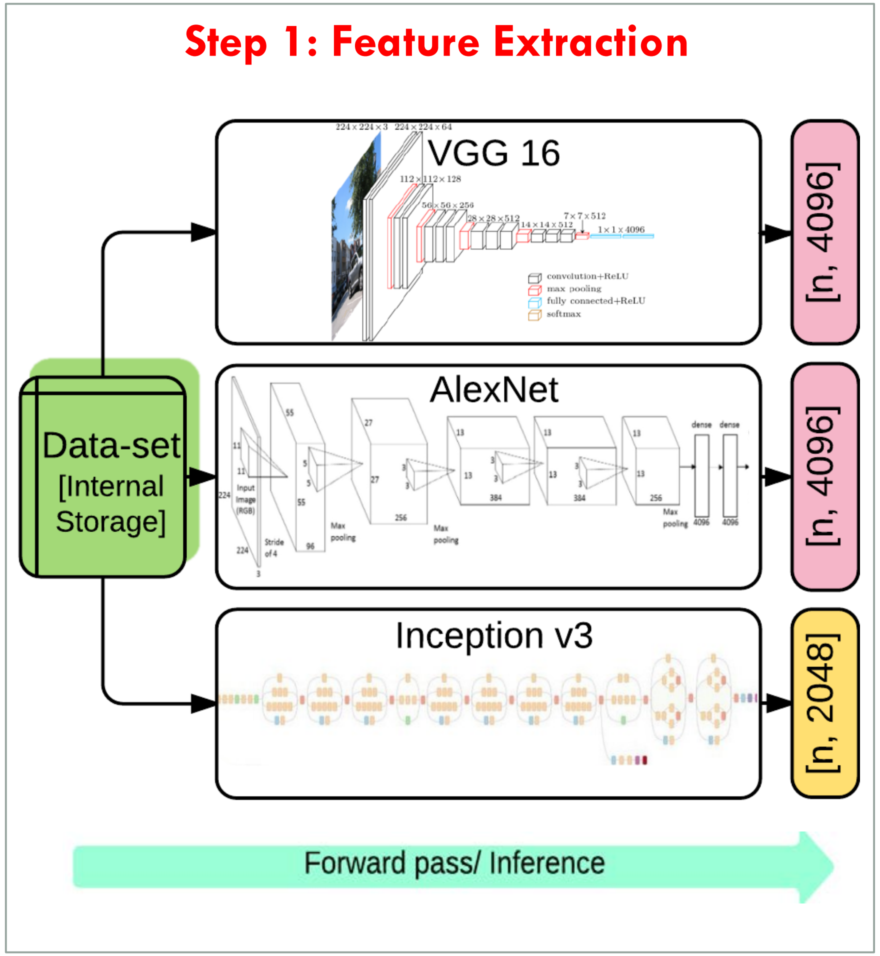


Model	Top-1 Accuracy	Top-5 Accuracy	Notes
Inception-V3	78.0%	93.9%	48 layers, $P \simeq 5M$
VGG-16	71.5%	89.8%	16+ layers, $P \simeq 180M$
AlexNet	63.3%	84.7%	5+ layers, $P \simeq 60M$

- **AlexNet** [17] is the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012.
- **VGG-16** [18] is the winner of ILSVRC-2014 on localization task and runner-up on classification task.
- **Inception-v3** [19] is an advanced version of the winner of ILSVRC-2014 classification task, the GoogLeNet.

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

FUSION OF MULTIMODAL CNN FEATURES



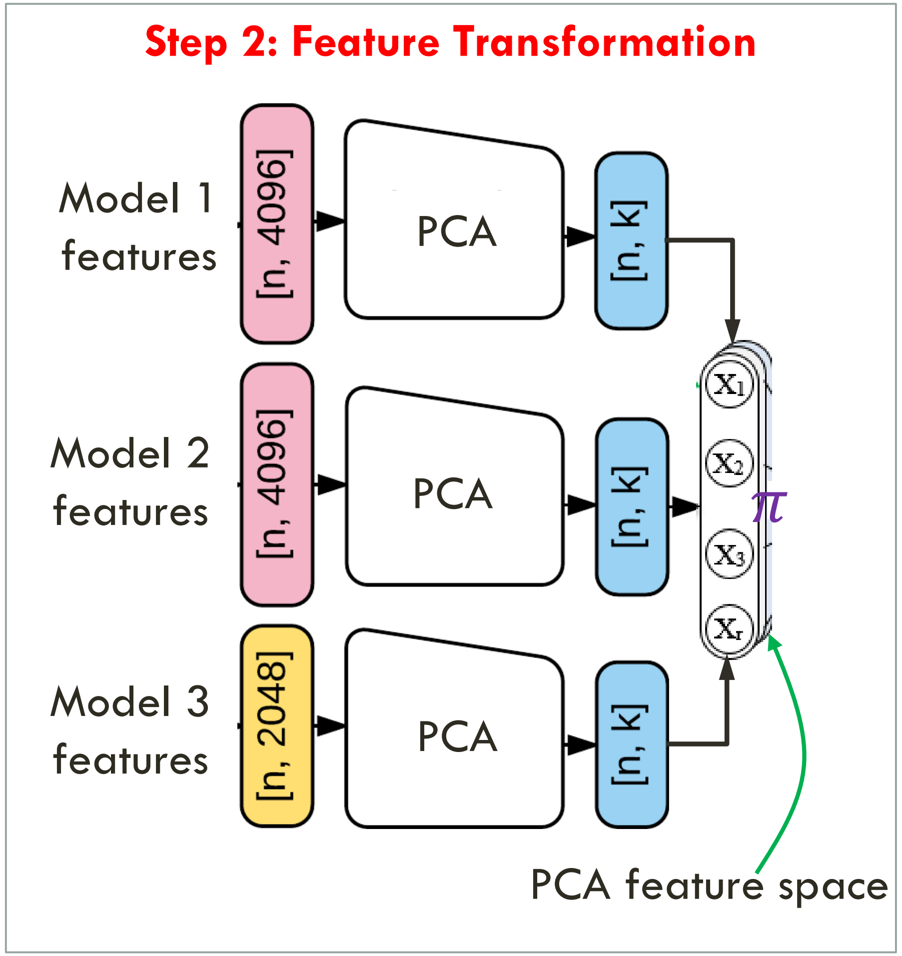
Model	Top-1 Accuracy	Top-5 Accuracy	Notes
Inception-V3	78.0%	93.9%	48 layers, $P \simeq 5M$
VGG-16	71.5%	89.8%	16+ layers, $P \simeq 180M$
AlexNet	63.3%	84.7%	5+ layers, $P \simeq 60M$

DCNN architecture	Total no. of layers	ϱ name	GPU time (s)
AlexNet	8	FC7	0.004
VGG16	16	FC2	0.018
Inception-v3	48	pool_3:0	0.023

The Average Time Taken by AlexNet, VGG16, and Inception-v3 as Feature Extractors.

- The per sample feature extraction computational complexity is taken as average time taken for an image over 10,000 test samples of CIFAR10 using NVIDIA GeForce GTX 1060 6 GB and Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz.

FUSION OF MULTIMODAL CNN FEATURES

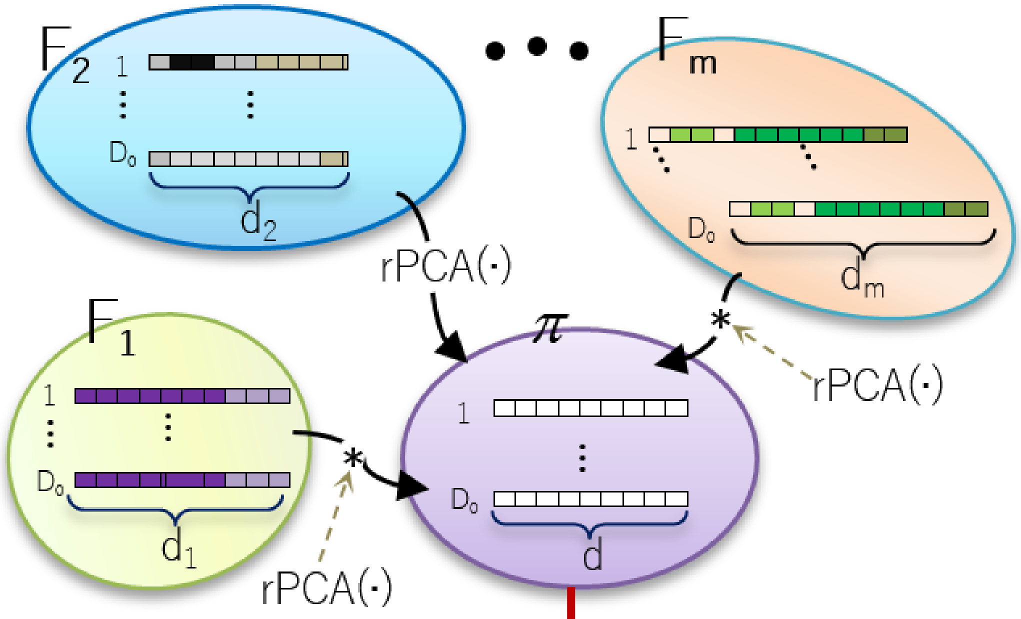


- The **high-level feature** representation is **generated through similar operations** like convolution, spatial sampling, and non-linear rectification.
- Thus, it is an **effective way** to take advantage of **PCA** for dimensionality **reduction** and data **transformation**.
- Then, the individual generalized features are **normalized based on their energy levels** (i.e. the area under the curve of feature F_i denoted as E_{F_i}) as given by $F_i' = \Omega_i \cdot F_i$, where the weight Ω_i is computed as,

$$\Omega_i = \frac{1}{E_{F_i}} \times \sum_{i=1}^{m} \frac{1}{E_{F_i}}.$$

FUSION OF MULTIMODAL CNN FEATURES

Step 3: Feature Generalization with Fusion



π – PCA space

$f^* : \hat{Y}_1, \hat{Y}_2, \hat{Y}_3 \longmapsto Y^*$

- Five different fusion rules are applied to form a generalized feature vector.

Concatenation: $\mathrm{FFV} = (F'_1, \cdots, F'_m)$ **(R1)**

Product: $\mathrm{FFV} = \prod_{i=1}^{m} F'_i$ **(R2)**

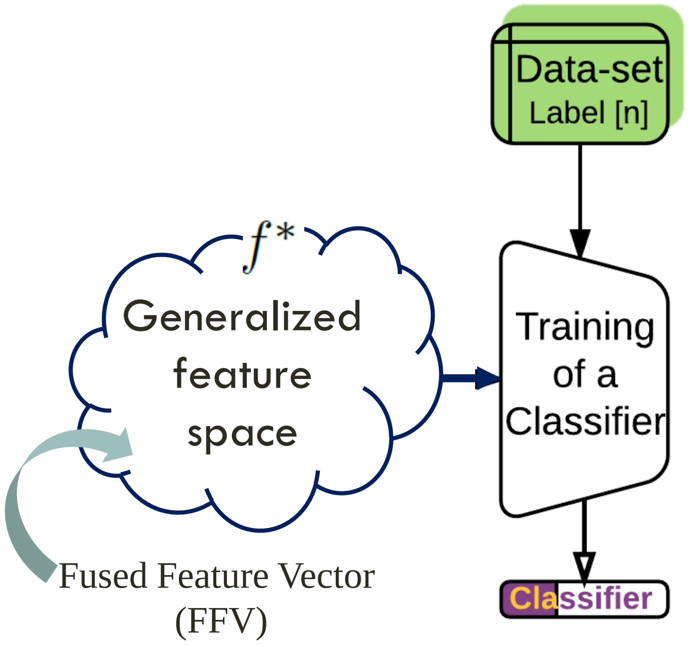
Summation: $\mathrm{FFV} = \sum_{i=1}^{m} F'_i$ **(R3)**

Average: $\mathrm{FFV} = \mathrm{mean}({\mathrm{F}'_1}^{\mathrm{T}}, \cdots, {\mathrm{F}'_m}^{\mathrm{T}})$ **(R4)**

Max: $\mathrm{FFV} = \max({F'_1}^{T}, \cdots, {F'_m}^{T})$ **(R5)**

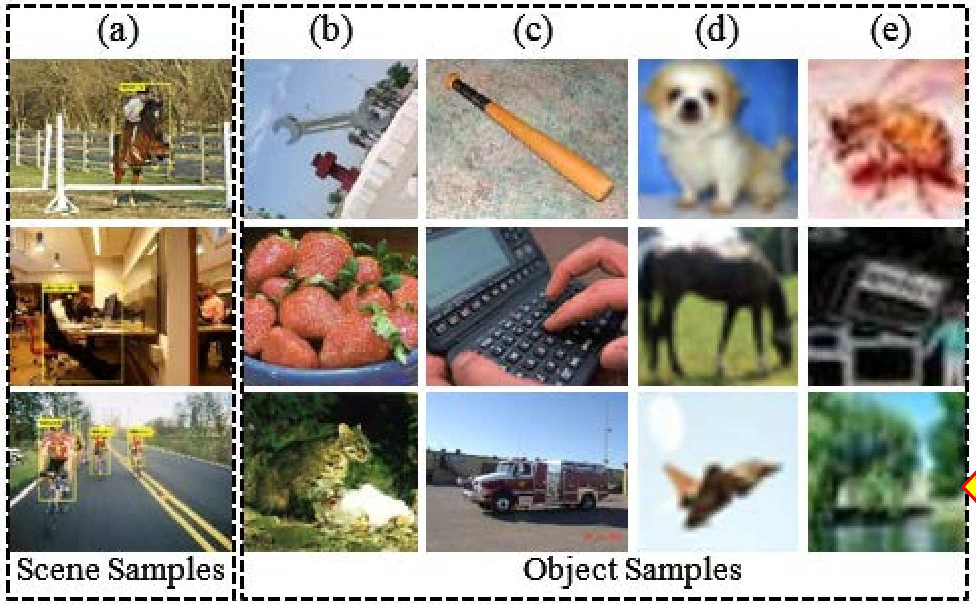
FUSION OF MULTIMODAL CNN FEATURES

Step 4: Shallow Classifier Training



- A multi-class SVM is trained on the fused feature vectors (FFV) to achieve a **multi-class** linear **classifier** C based on **one-versus-rest** (OVR) training procedure.
- In this work, the Scikit-learn Python multi-class linear SVM using **Crammer and Singer's** strategy with **L2 penalty** and **squared hinge loss** is employed.
- The learning rate of this network is set to $\lambda = 2^{-13}$.

EXPERIMENTAL RESULTS



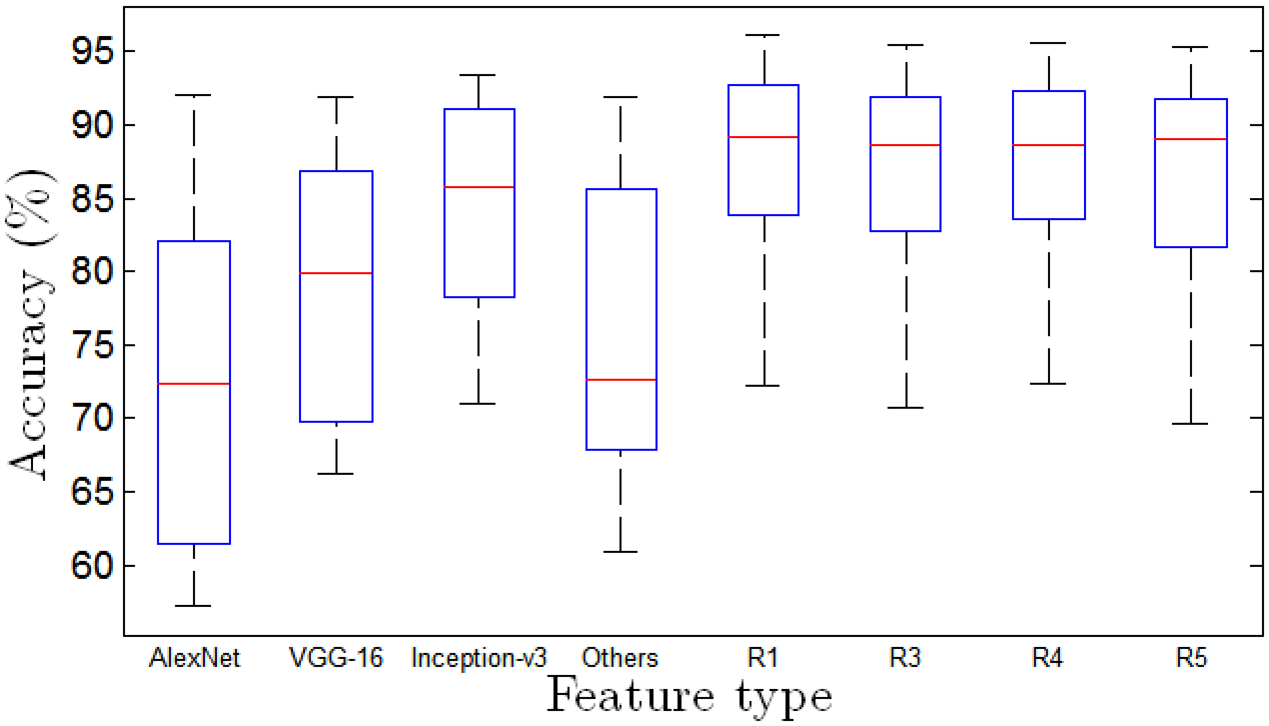
Data-set	No. of classes	Train. samples	Test samples	Ref.
CIFAR10	10	50,000	10,000	[17]
CIFAR100	100	50,000	10,000	[17]
Caltech101	101	6,076	2,601	[20]
Caltech256	256	21,363	9,146	[18]
Pascal VOC	10	4,588	4,569	[19]

(a). Pascal VOC 2012 (riding horse, using computer, ridding bike), (b). Caltech101 (wrench, strawberry, wild cat), (c). Caltech256 (baseball bat, calculator, firetruck), (d). CIFAR10 (dog, horse, airplane), (e). CIFAR100 (insects, household furniture, large natural outdoor scenes)

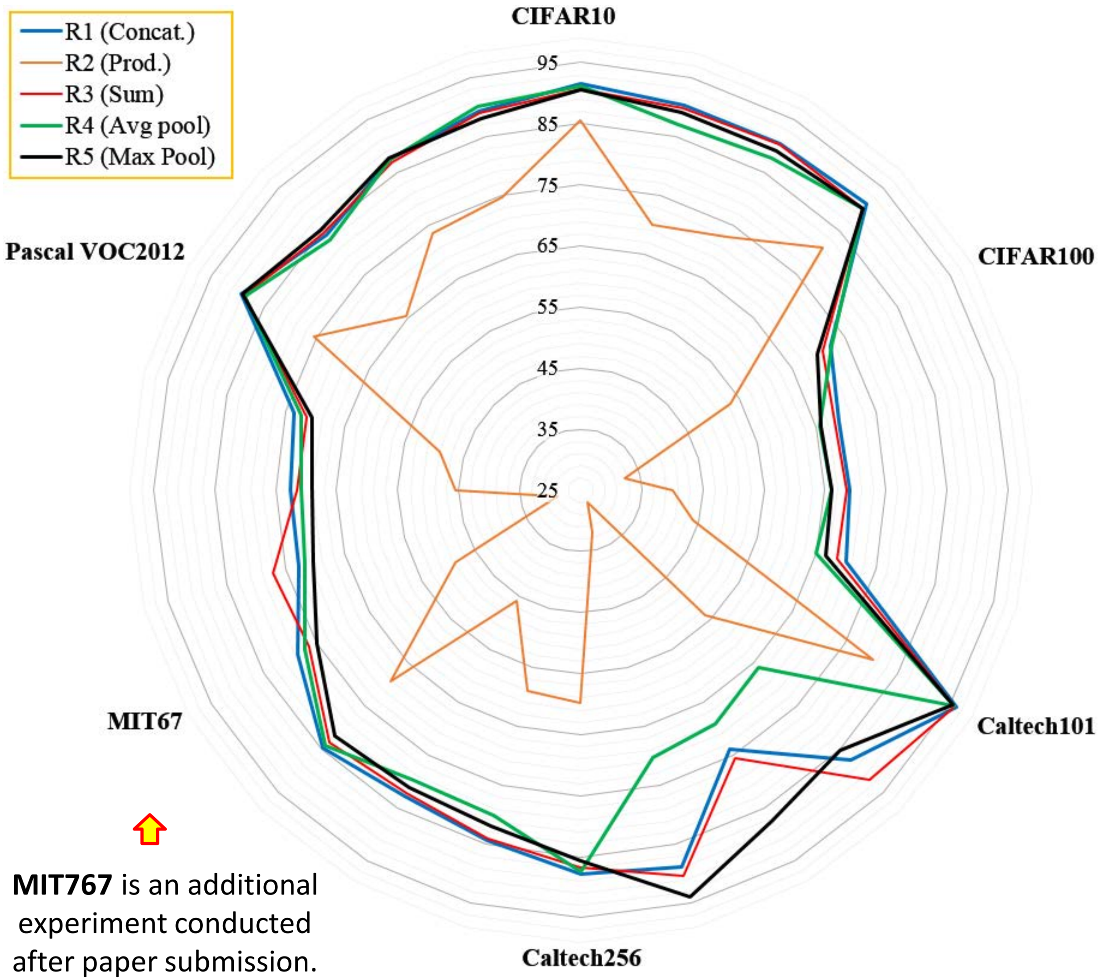
Comparisons of the proposed feature embedding with other methods on five datasets (top-1 accuracy in %).

Task	Data-set	Embedded Feature Space					AlexNet	VGG-16	Ince.-v3	Other methods
		R1(2)	R2(3)	R3(4)	R4(5)	R5(6)				
Object classification	CIFAR10	91.60	85.50	90.70	91.20	90.60	78.40	85.20	90.40	**91.87** [21], 85.02 [22], 74.50 [23]
	CIFAR100	72.20	53.20	70.70	72.40	69.70	57.20	66.20	71.00	**72.60** [24], 66.64 [21]
	Caltech101	**96.10**	80.40	95.50	95.60	95.30	92.00	91.90	93.40	83.60 [2], 82.10 [5], 76.10 [6]
	Caltech256	**87.80**	59.90	86.80	87.40	85.70	72.40	79.90	85.70	60.97 [7], 50.80 [5]
Scene classification	Pascal VOC	**89.20**	75.40	88.60	88.60	89.10	62.90	71.00	80.70	70.20 [13], 69.84 OXFORD [19]

EXPERIMENTAL RESULTS

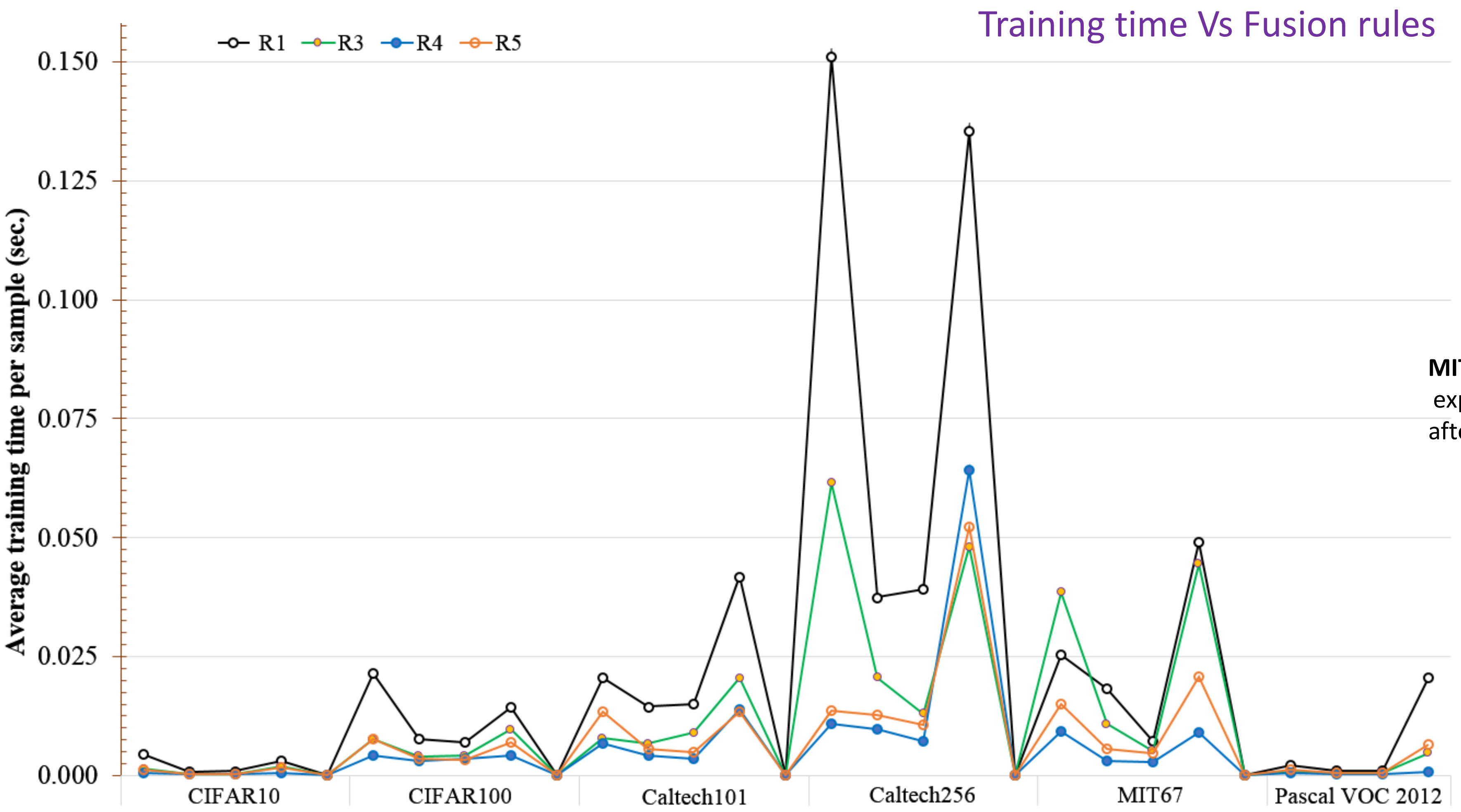


Note that, for visualization easiness results of product-based fusion rule (R2) is omitted, since it is evident from the tabulated numerical results and the radar plot that its performance is the poorest among all the fusion techniques.



MIT767 is an additional experiment conducted after paper submission.

EXPERIMENTAL RESULTS



MIT767 is an additional experiment conducted after paper submission.

CONCLUSION

- Re-emphasize that the **high-level features from DCNN provide abstract** information about objects/scenes and such features **are superior to the state-of-the-art low-level** local features.
- Taking advantage of **complementary cues of multiple DCNN creates** more **generalized feature space** that is somewhat appearance invariant and more discriminative of intra-class variations.
- **Fusion of multiple deep ConvNet architecture's high-level features enhances** the classification **accuracy** than **a single modality** and produces **very competitive results to fully trained DCNN** and fusion of hand-crafted features as well.
- Features **from distinct neural architectures yet posses complementary cues** that can be integrated later to accurately classify visual objects or scenes.
- In the future, it would be interesting to consider such feature fusion for video content analysis (VCA), semantic segmentation, and medical image classification.

REFERENCES

[1] T. Akilan, Q. M. J. Wu, Y. Yang, and A. Safaei, “Fusion of transfer learning features and its application in image classification,” in IEEE 30th Canadian Conf. Electri. Compu. Engg. (CCECE), pp. 1–5, 2017.

[2] D.-C. Park, “Multiple feature-based classifier and its application to image classification,” IEEE Inter. Conf. on Data Mining WS, pp. 65–71, 2010.

[3] J. Kwon and K. M. Lee, “Visual tracking decomposition,” in CVPR, pp. 1269–1276, 2010.

[4] T.-B. Fernando, E. Fromont, D. Muselet, and M. Sebban, “Discriminative feature fusion for image classification,” ICPR, pp. 3434–3441, 2012.

[5] P.-V. Gehler and S. Nowozin, “On feature combination for multiclass object classification,” in ICCV, pp. 221–228, Sept. 2009 .

[6] F.-S. Khan, J. van de Weijer, and M. Vanrell, “Modulating shape features by color attention for object recognition,” IJCV, vol. 98, pp. 49–64, 2012.

[7] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos, “Scene classification with semantic fisher vectors,” in CVPR, pp. 2974–2983, June 2015.

[8] S. Guo, W. Huang, L. Wang, and Y. Qiao, “Locally supervised deep hybrid model for scene recognition,” IEEE TIP, vol. 26, pp. 808–820, Feb 2017.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in CVPR, pp. 1–9, June 2015.

[10] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik,“Simultaneous detection and segmentation,” ECCV,pp. 297–312, 2014.

[11] A. Thangarajah, Q. M. J. Wu, and J. Huo, “A unified threshold updating strategy for multivariate gaussian mixture based moving object detection,” in Inter. Conf. HPCS., pp. 570–574, July 2016.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” IJCV., vol. 115, no. 3, pp. 211–252, 2015.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in Proc. CVPR, pp. 1717–1724, 2014.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, vol. abs/1409.1556, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in In Advan. Neu. Info. Process. Sys., pp. 1097–1105, 2012. 2

[16] Y. Luo, D. Tao, Y.Wen, K. Ramamohanarao, and C. Xu, “Tensor canonical correlation analysis for multi-view dimension reduction,” IEEE Trans.Knowle. and Data Engg., vol. 27, no. 11, 2015.

[17] A. Krizhevsky, “Learning multiple layers of features from tiny images,” 2009.

[18] G. Griffin, A. Holub, and P. Perona, “The caltech-256: Caltech technical report,” vol. 7694, 2007.

[19] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” IJCV, vol. 111, no. 1, pp. 98–136, 2015.

[20] L. Fei-Fei, L.-R. Fergus, and P. Perona, “Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories,” in CVPR, pp. 178–178, 2004.

[21] M. Sun, T.-X. Han, L. M.-C. Xu, X., and K. Ahmad Khodayari-Rostamabad, “Latent model ensemble with auto-localization,” in Proc. ICPR, 2016.

[22] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” Advan. Neu.l Informa. Process. Sys., vol. 25, pp. 2951– 2959, 2012.

[23] K. Yu and T. Zhang, “Improved local coordinate coding using local tangents,” in ICML-10, pp. 1215–1222, Omnipress, 2010.

[24] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams, “Scalable bayesian optimization using deep neural networks,” in JMLR WS and Conf. Proc., pp. 2171–2180, 2015.

[25] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in Advan. Neu. Info. Process. Sys., vol. 27, pp. 487–495, 2014.

THE END

Thank you very much for your attention