Comparison of Fine-tuning and Extension Strategies for Deep Convolutional Neural Networks

Nikiforos Pittaras[1], Foteini Markatopoulou[1,2], Vasileios Mezaris[1], and Ioannis Patras[2]

[1]Information Technologies Institute / Centre for Research and Technology Hellas

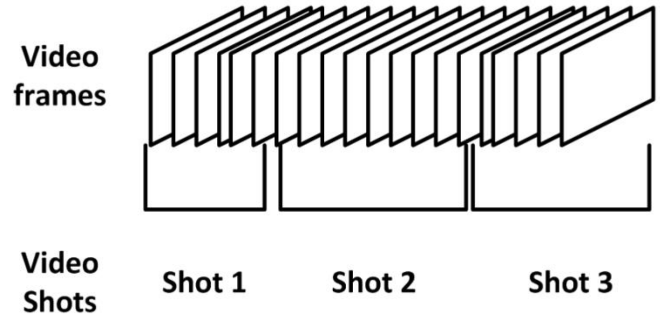
[2]Queen Mary University of London

Problem

Pool of Concepts

Video Concept Detection

Video shot segmentation



Video shot sampling and annotation

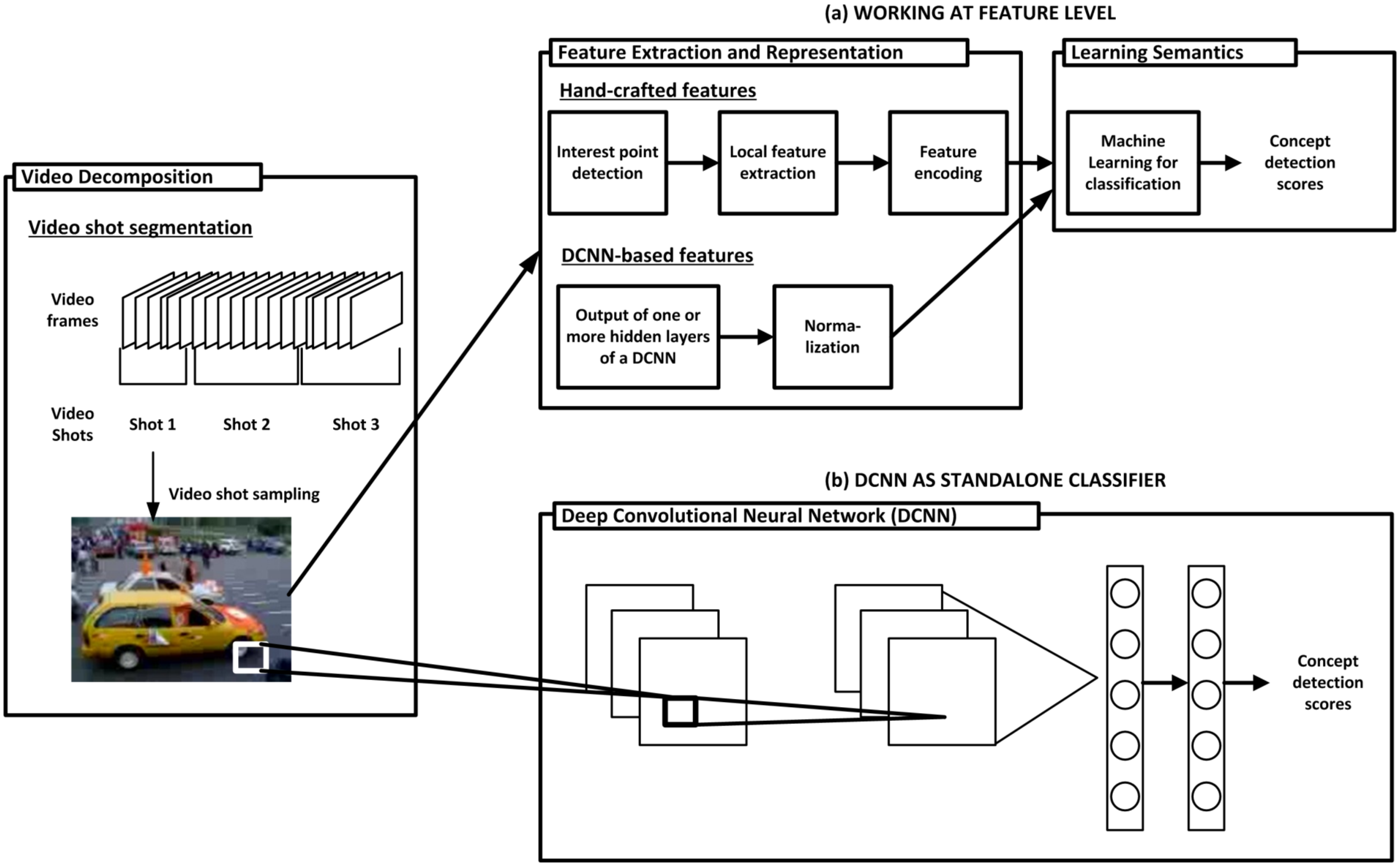
Car, 0.95
Road, 0.98
People, 0.8
....
Indoor, 0.01
Dog, 0.1

Image Concept Detection

Bicycle, 0.95
Road, 0.98
People, 0.8
....
Car, 0.01
Indoor, 0.1

Information Technologies Institute
Queen Mary University of London
InVID

Typical solution

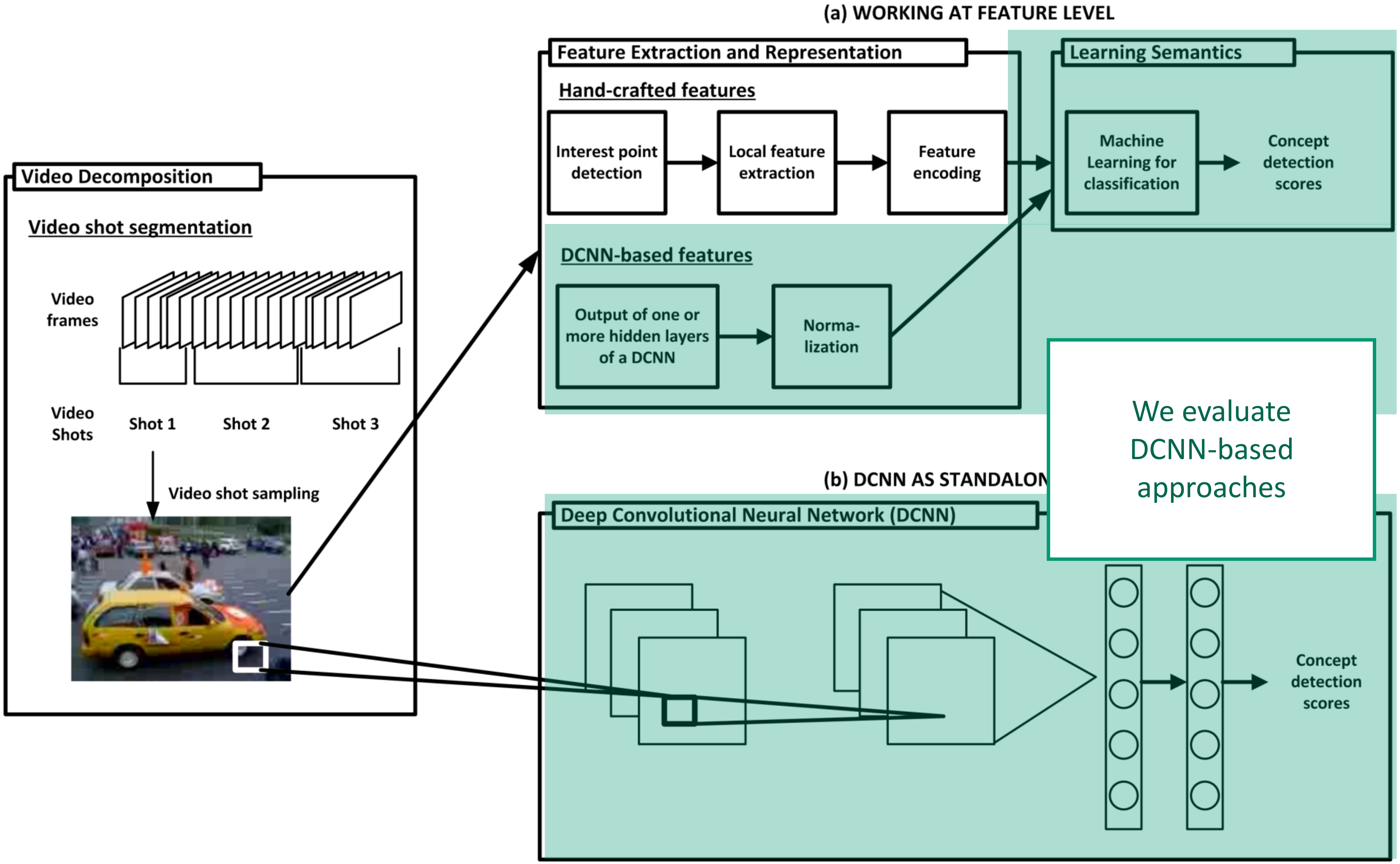


Information Technologies Institute

Queen Mary University of London

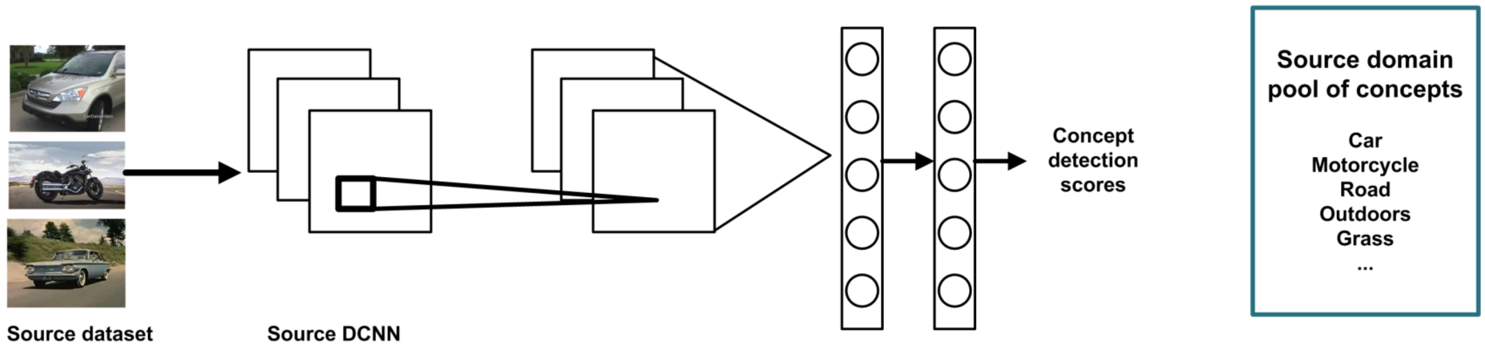
InVID

Typical solution

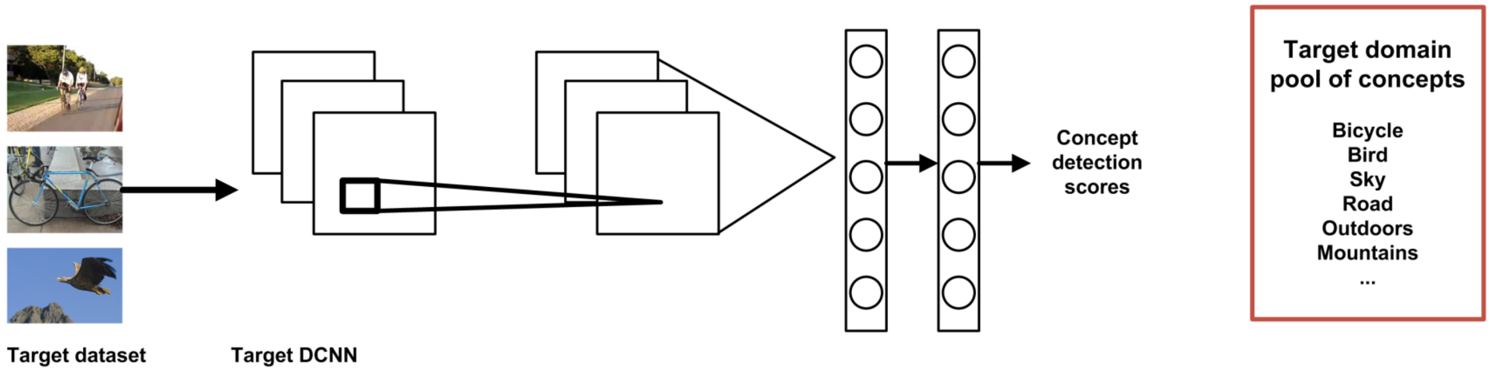


Transfer learning

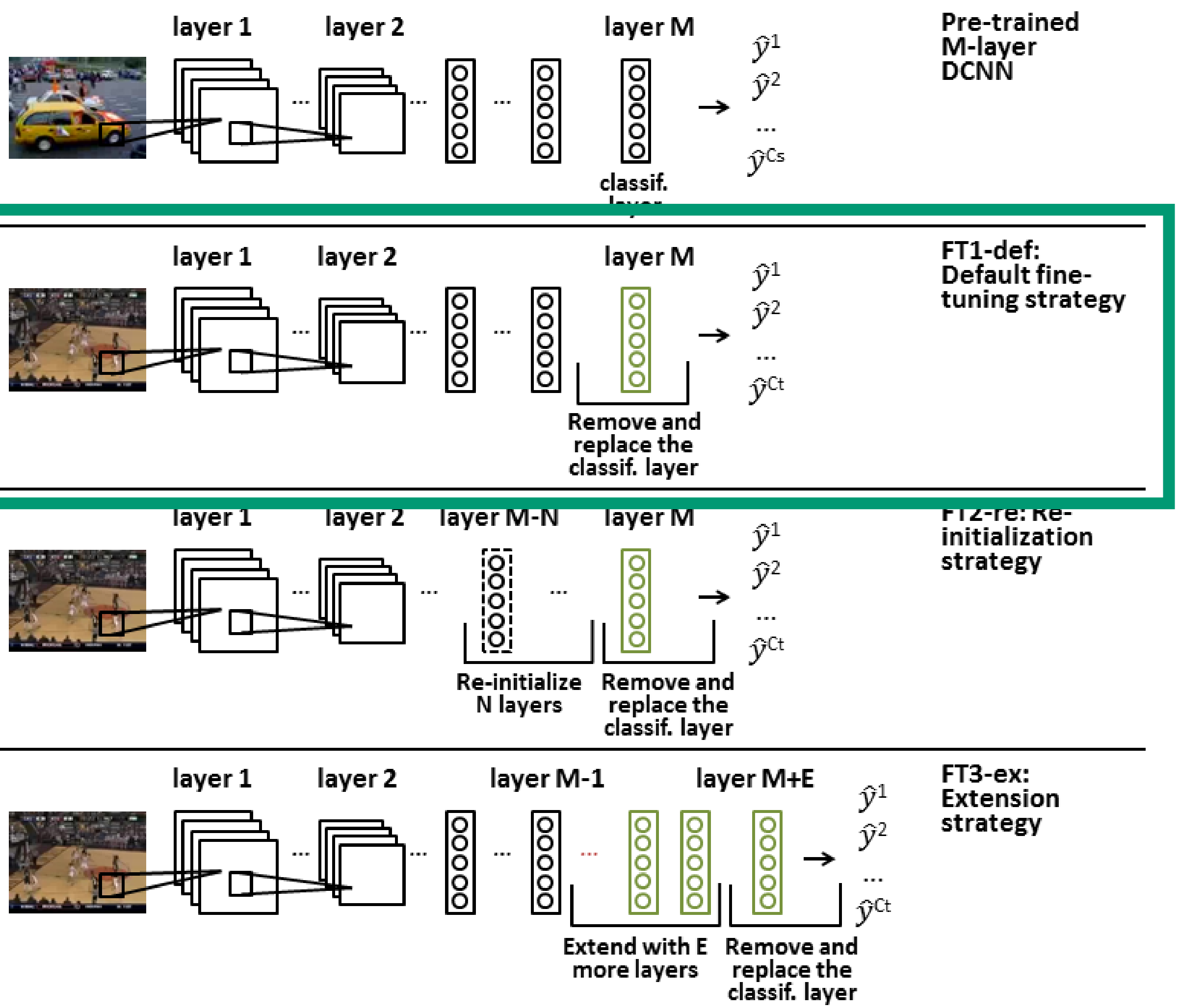
Source domain DCNN



Target domain DCNN

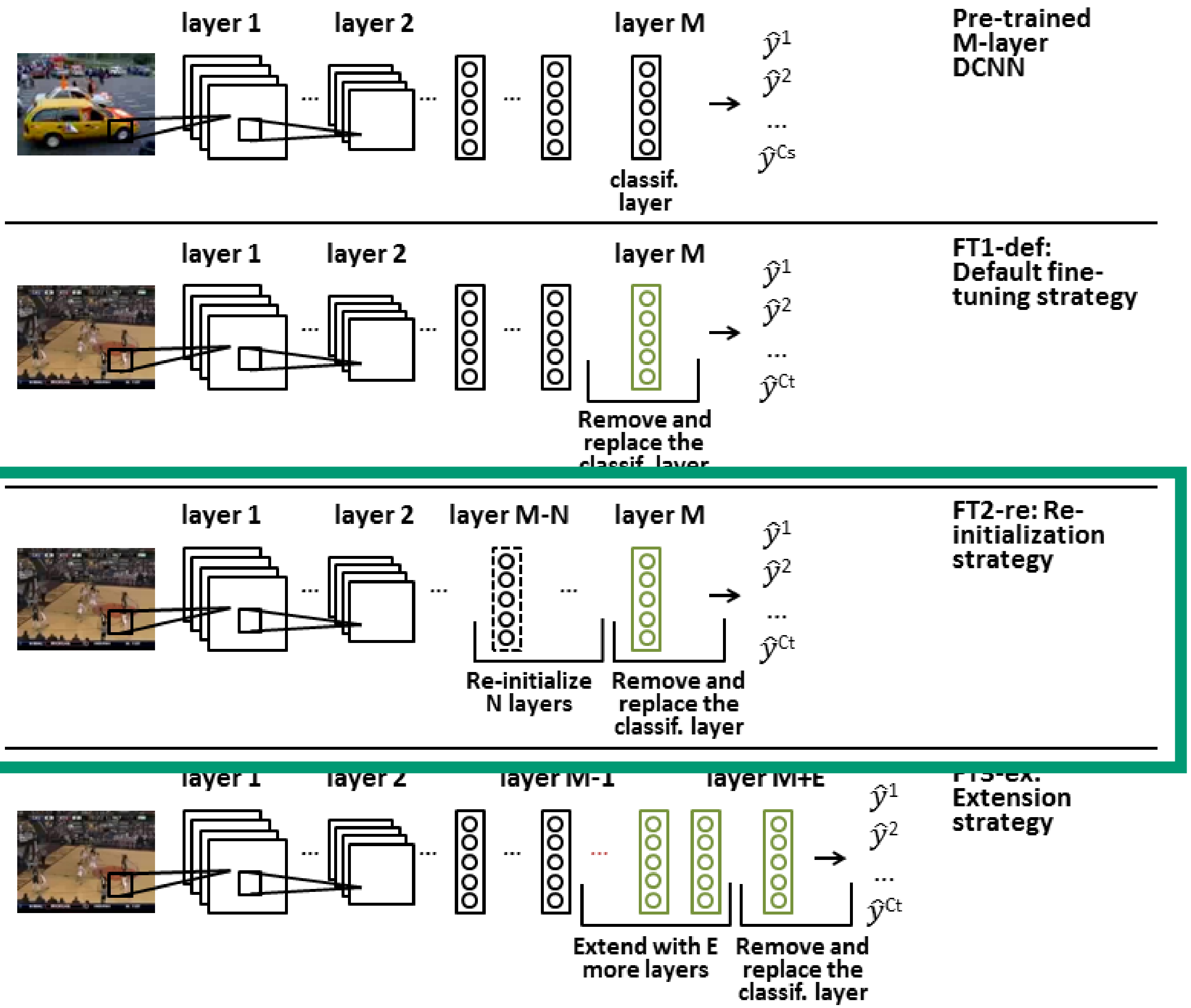


Literature review: Fine-tuning strategies



- Replacing the classification layer with a new output layer [2,5,18]
  - The new layer is learned from scratch
  - All the other layers are fine-tuned

Literature review: Fine-tuning strategies



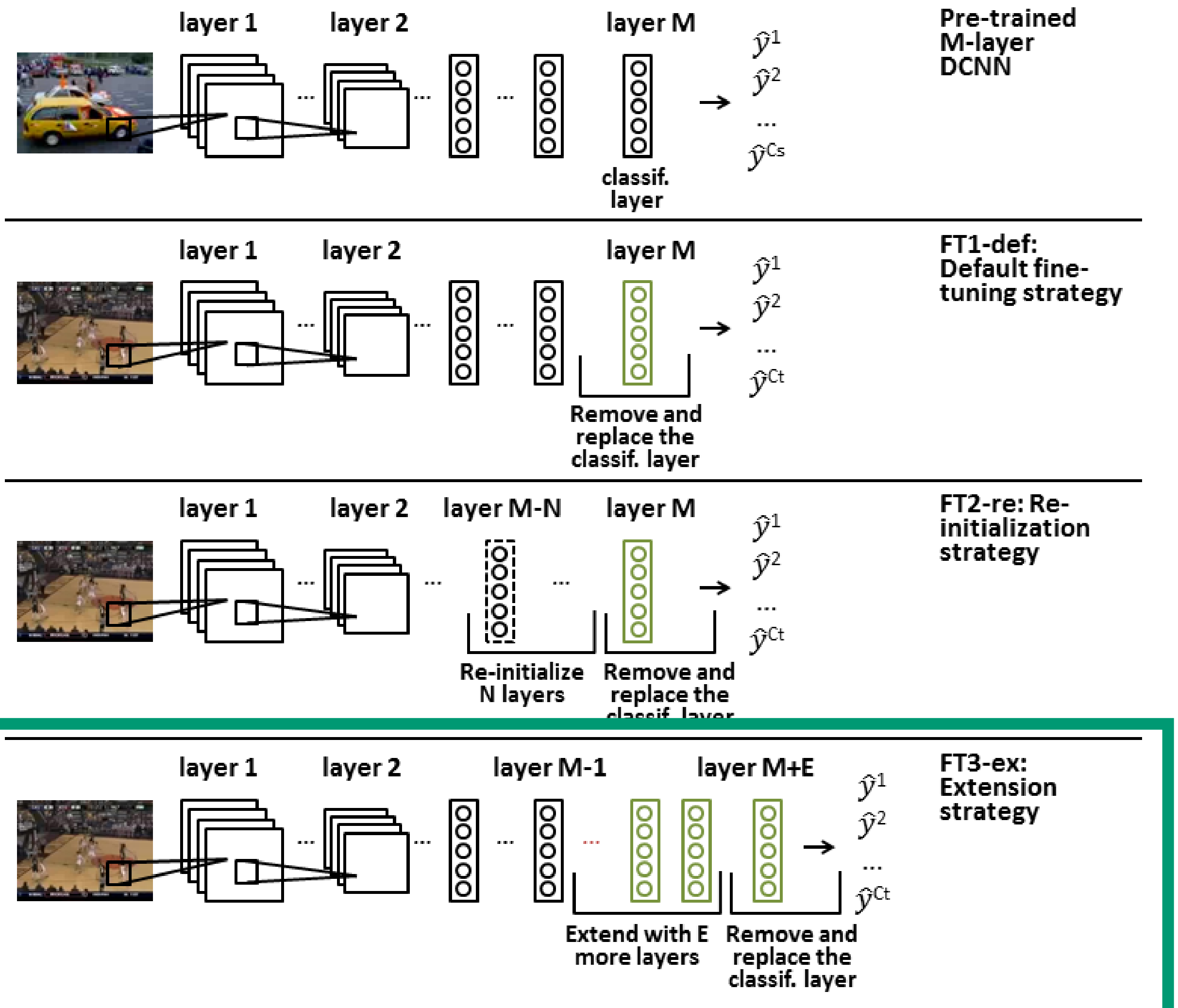
- Re-initializing the last N layers [1,9,18]
  - The last N layers are learned from scratch
  - The first M-N layers are fine-tuned with a low learning rate [18] (or could remain frozen [1,9])

Information Technologies Institute

Queen Mary University of London

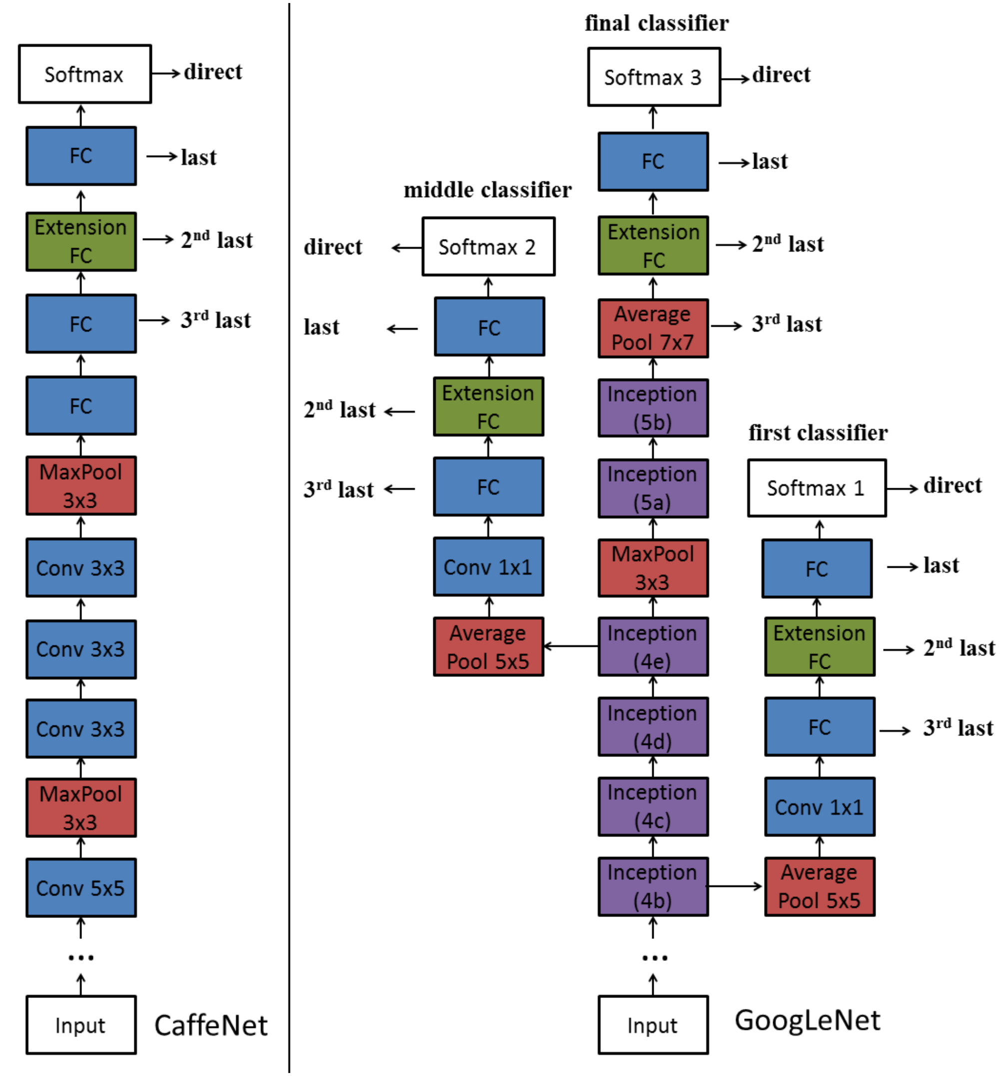
InVID

Literature review: Fine-tuning strategies

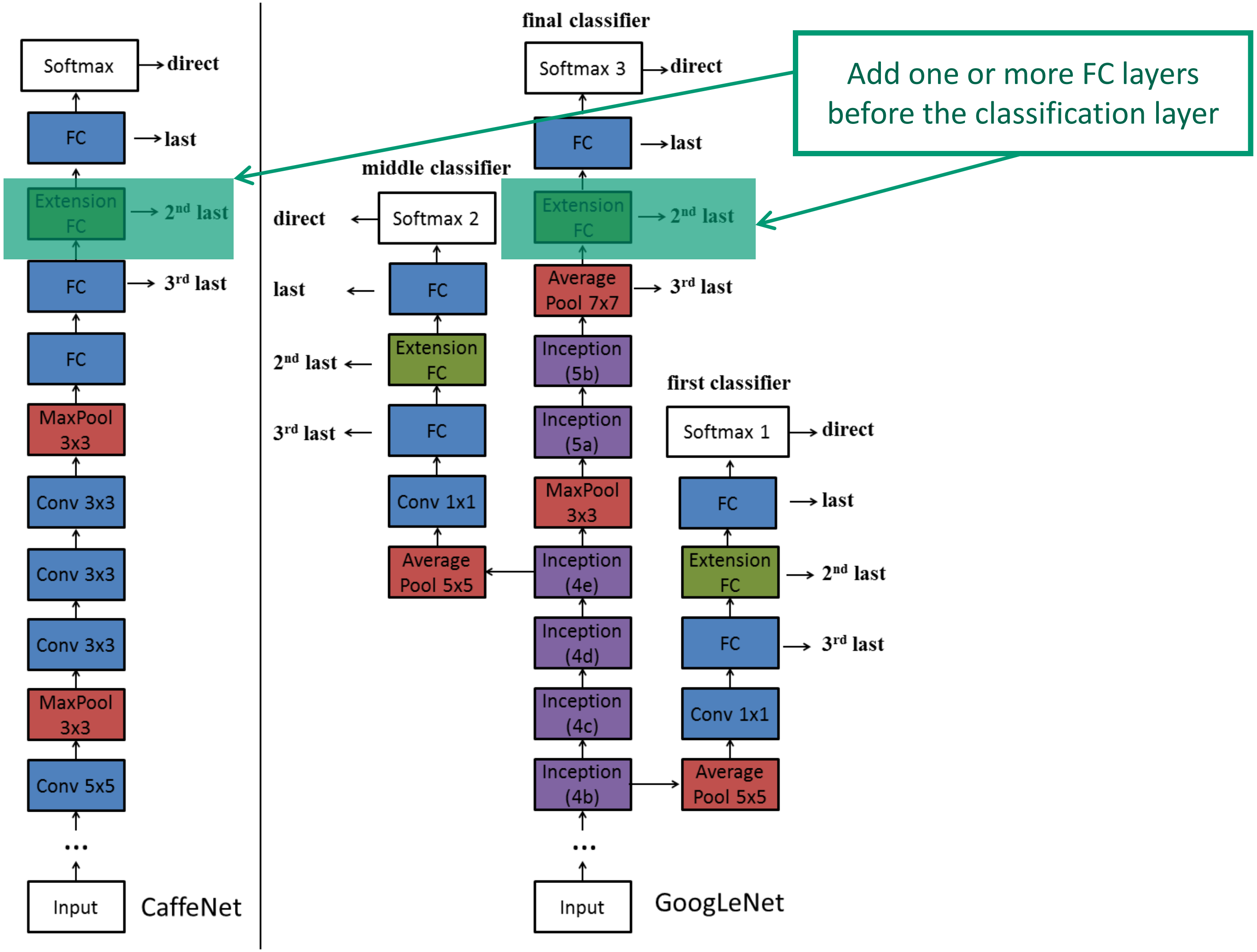


- Extending the network by one or more fully connected layers [1,8,9,15]

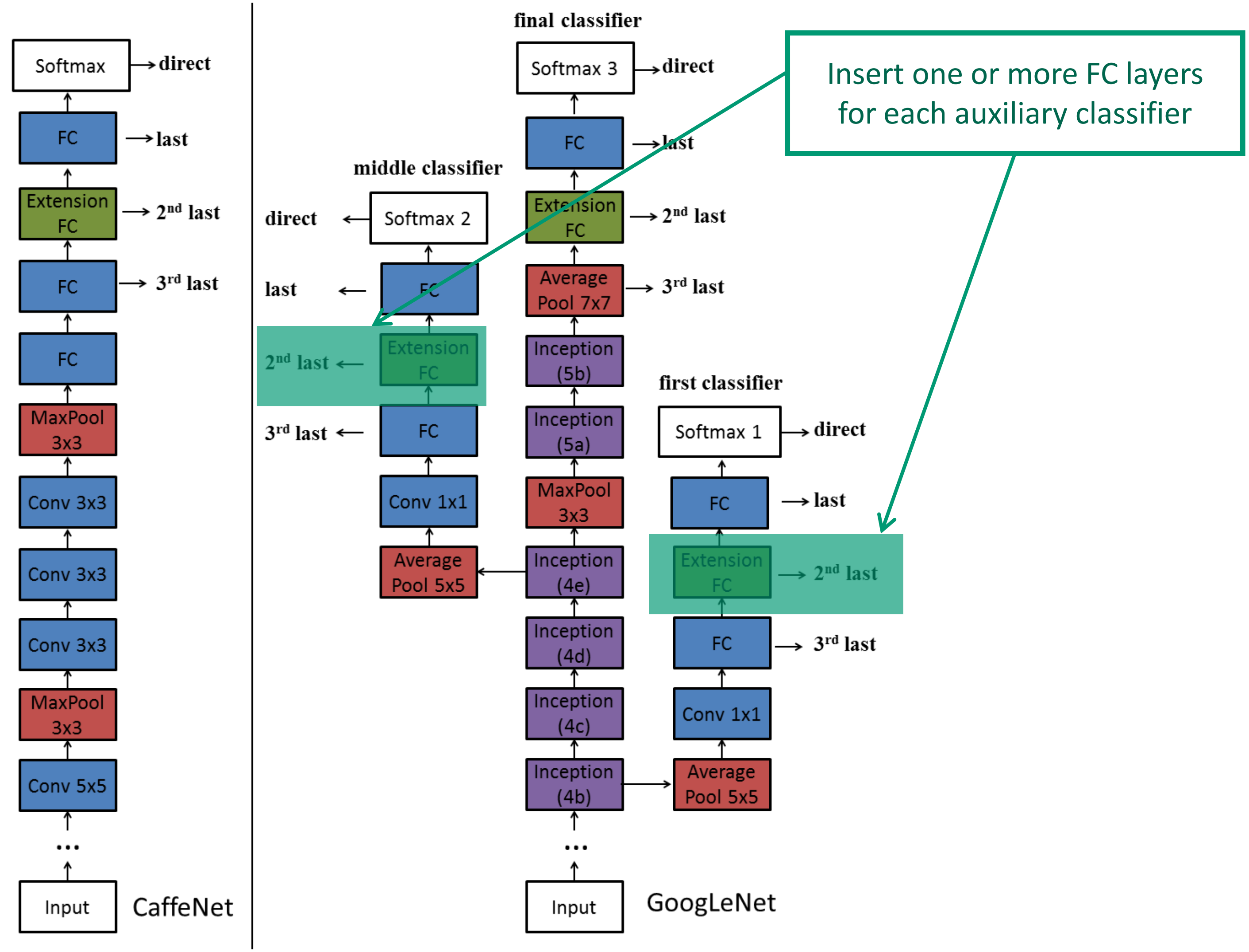
FT3-ex: extension strategy



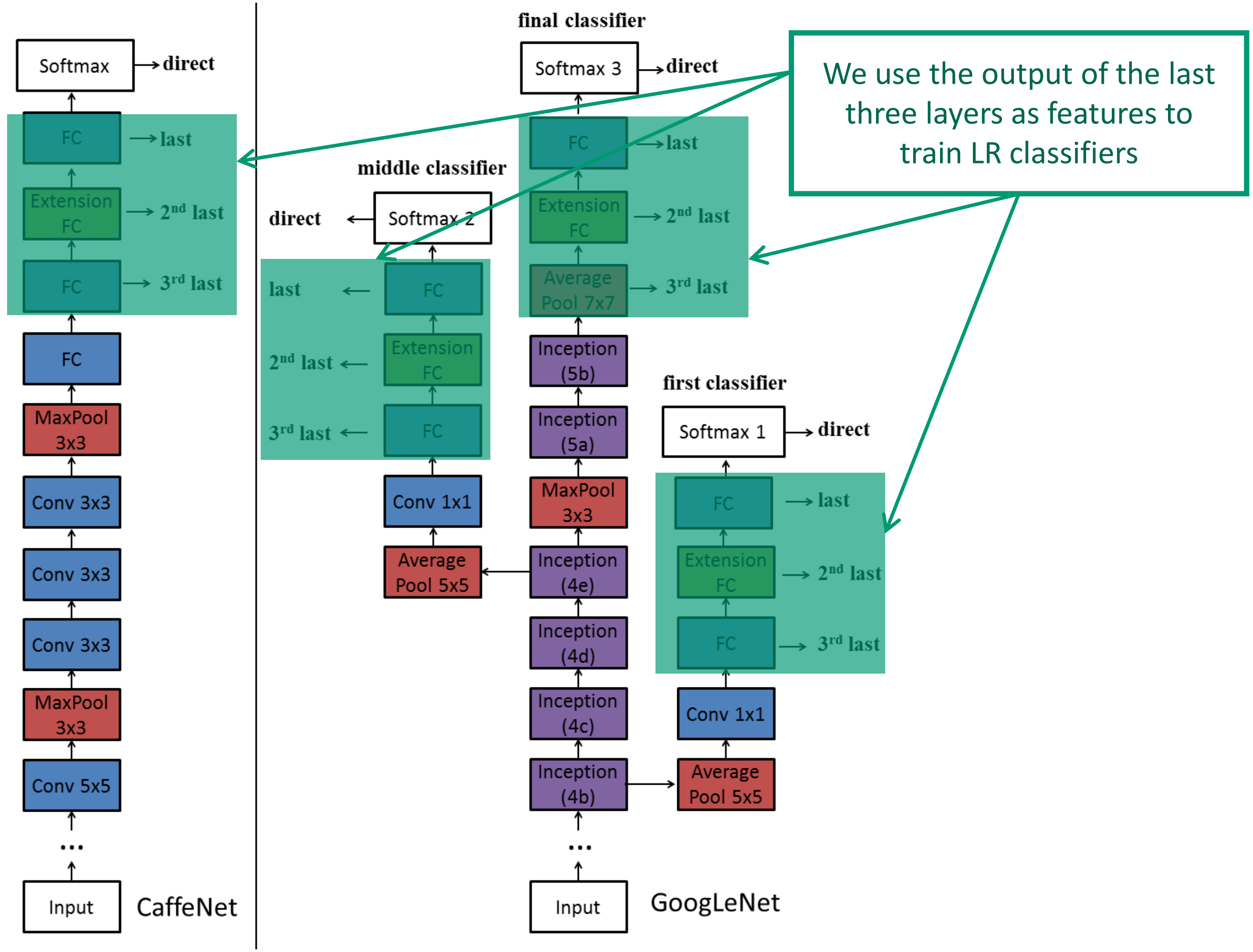
FT3-ex: extension strategy



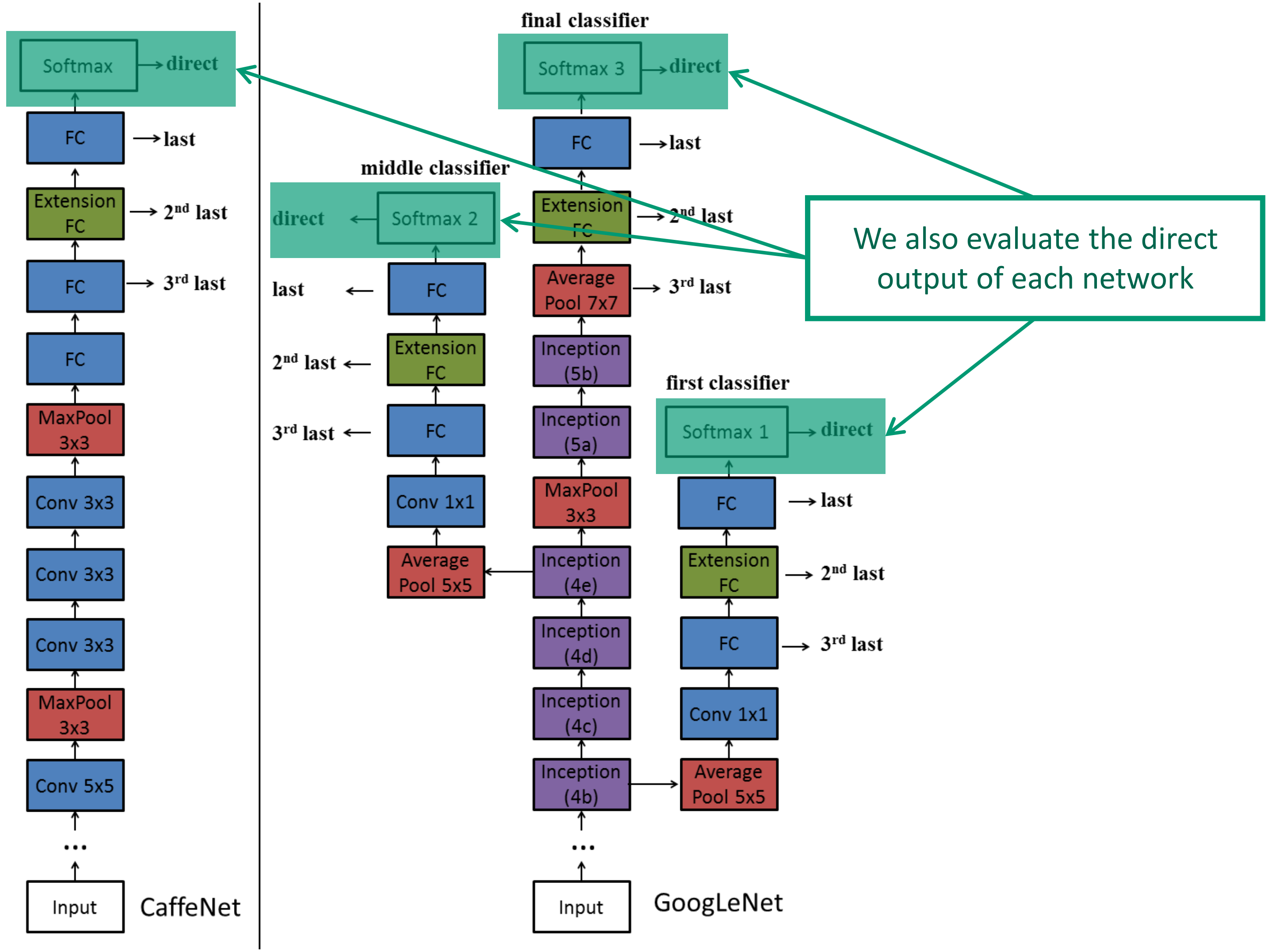
FT3-ex: extension strategy



FT3-ex: extension strategy



FT3-ex: extension strategy



Information Technologies Institute

Queen Mary University of London

InVID

Evaluation setup

Dataset: TRECVID SIN 2013

- 800 and 200 hours of internet archive videos for training and testing
- One keyframe per video shot
- Evaluated concepts: 38, Evaluation measure: MXinfAP

Dataset: PASCAL VOC 2012

- 5717 training, 5823 validation and 10991 test images
- Evaluation on the validation set instead of the original test set
- Evaluated concepts: 20, Evaluation measure: MAP

We fine-tuned 3 pre-trained ImageNet DCNNs:

- CaffeNet-1k, trained on 1000 ImageNet categories
- GoogLeNet-1k, trained on the same 1000 ImageNet categories
- GoogLeNet-5k, trained using 5055 ImageNet categories

Evaluation setup

For each pair of utilized network and fine-tuning strategy we evaluate:

- The direct output of the network
- Logistic regression (LR) classifiers trained on DCNN-based features
  - One LR classifier per concept trained using the output of one layer
  - The late-fused output (arithmetic mean) of LR classifiers trained using the last three layers

We also evaluate the two auxiliary classifiers of the GoogLeNet-based networks

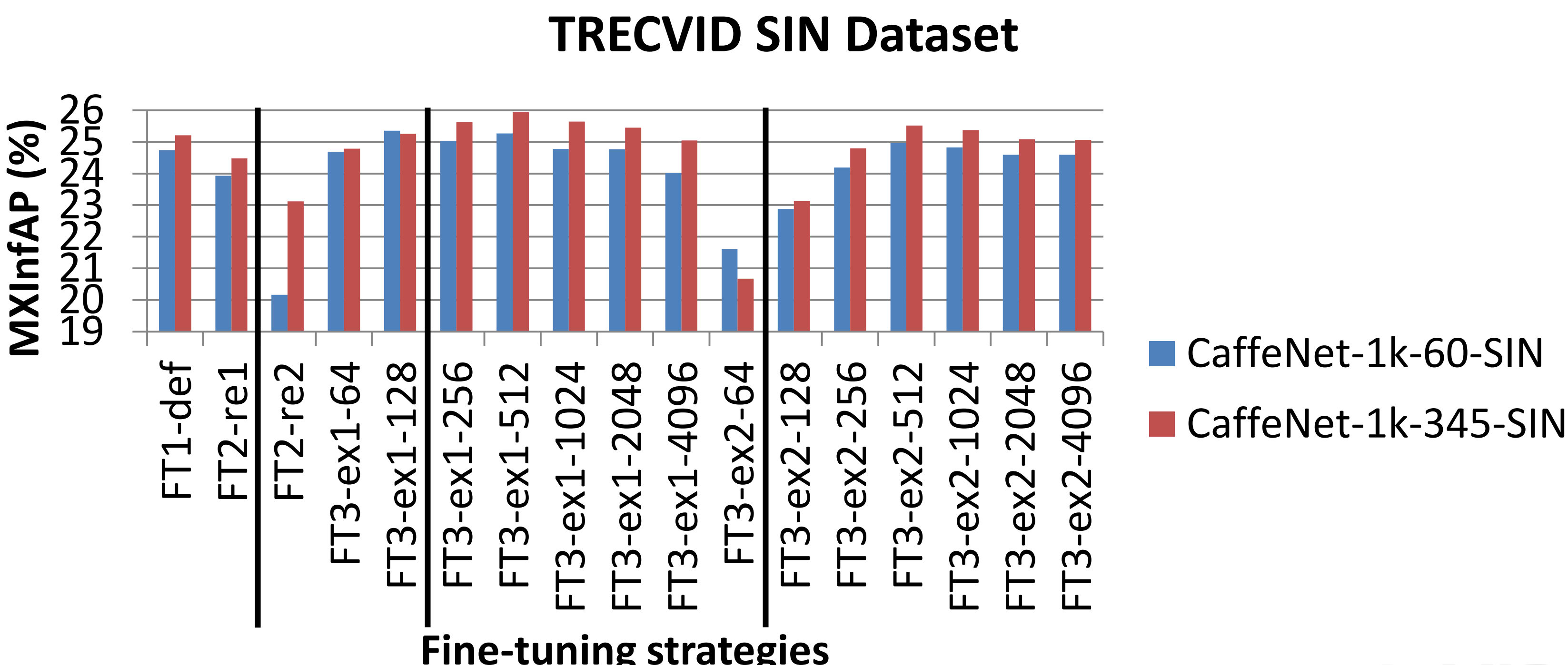
Preliminary experiments for parameter selection

- Classification accuracy for the **FT1-def strategy** and **CaffeNet-1k-60-SIN**
  - *k*: the learning rate multiplier of the pre-trained layers
  - *e*: the number of training epochs
  - The best accuracy per *e* is underlined; the globally best accuracy is **bold and underlined**
- Improved accuracy for:
  - Smaller learning rate values for the pre-trained layers
  - Higher values for the training epochs

k/e	0.25	0.5	1	2	4	8	16	32
0.050	0.348	0.362	0.402	0.417	0.437	0.434	0.451	0.462
0.075	0.341	0.349	0.388	0.412	0.438	0.453	0.462	0.462
0.100	0.346	0.354	0.388	0.420	0.434	0.455	0.463	**0.470**
0.250	0.328	0.361	0.397	0.421	0.430	0.450	0.455	0.468
0.500	0.306	0.354	0.388	0.415	0.439	0.447	0.451	0.444
0.750	0.284	0.349	0.381	0.410	0.431	0.443	0.448	0.448
1.000	0.257	0.321	0.367	0.390	0.430	0.442	0.450	0.436

Experimental results – # target concepts

- **Goal**: Assess impact of number of target concepts
- **Experiment**: We fine-tune the network for either 60 or all 345 concepts; we evaluate the same 38⊆60 concepts
- **Conclusion**: Fine-tuning a network for more concepts improves concept detection accuracy



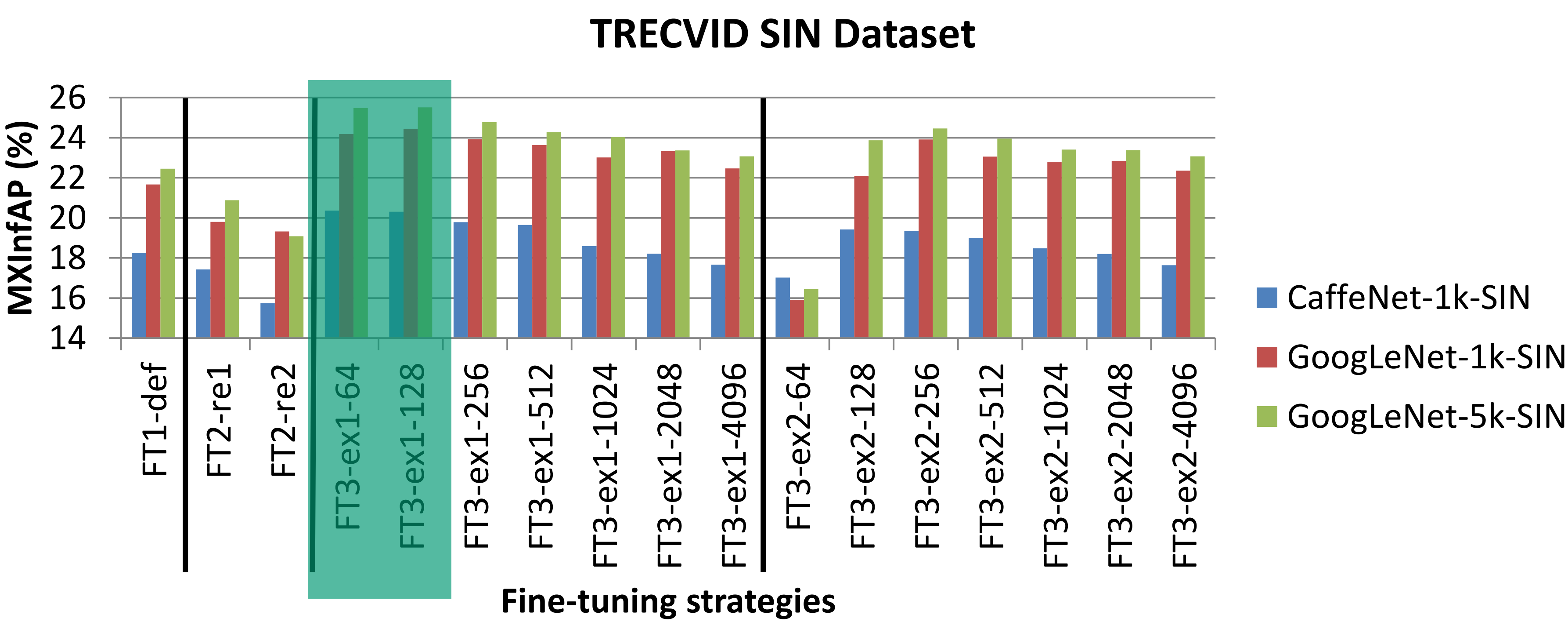
Information Technologies Institute

Queen Mary University of London

InVID

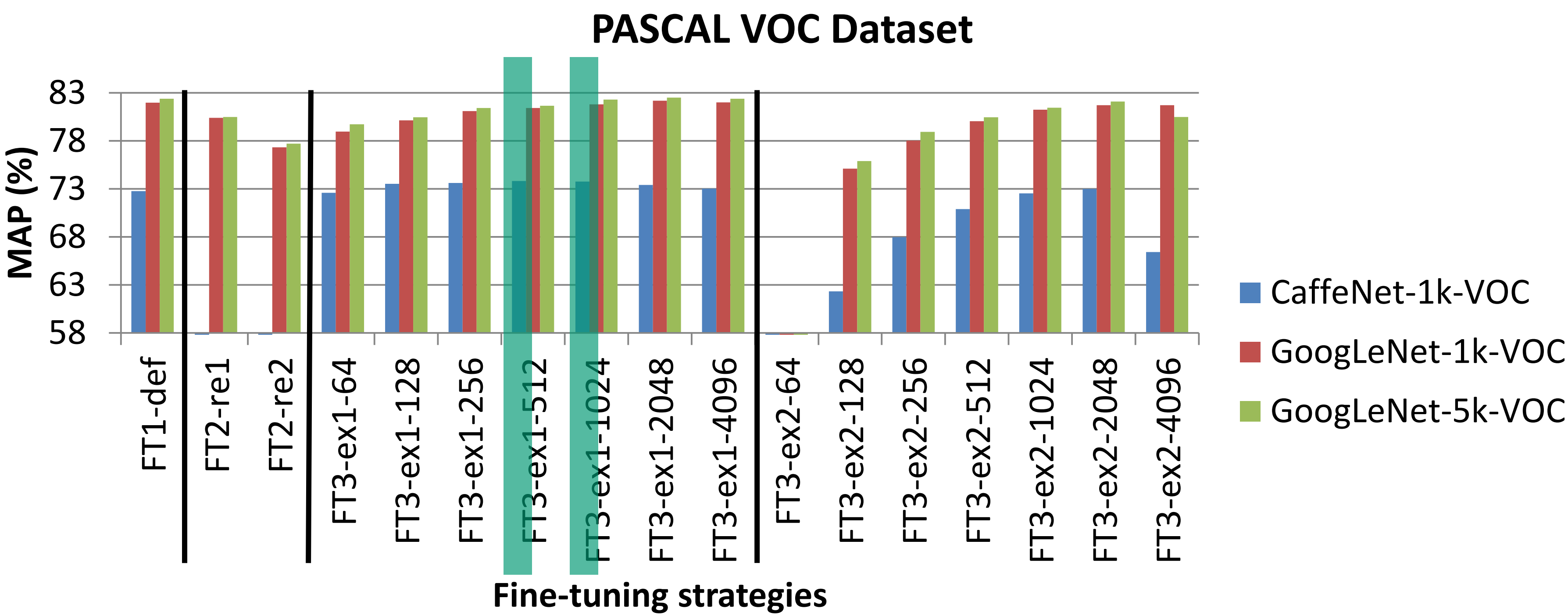
Experimental results – direct output

- **Goal**: Assess DCNNs as standalone classifiers (direct output)
- **Experiment**: Fine-tuning on the TRECVID SIN dataset
- **Conclusion**: FT3-ex1-64 and FT3-ex1-128 constitute the top-two methods irrespective of the employed DCNN



Experimental results – direct output

- **Goal**: Assess DCNNs as standalone classifiers (direct output)
- **Experiment**: Fine-tuning on the PASCAL VOC dataset
- **Conclusion**: FT3-ex1-512 and FT3-ex1-1024 the best performing strategies **for the CaffeNet network**



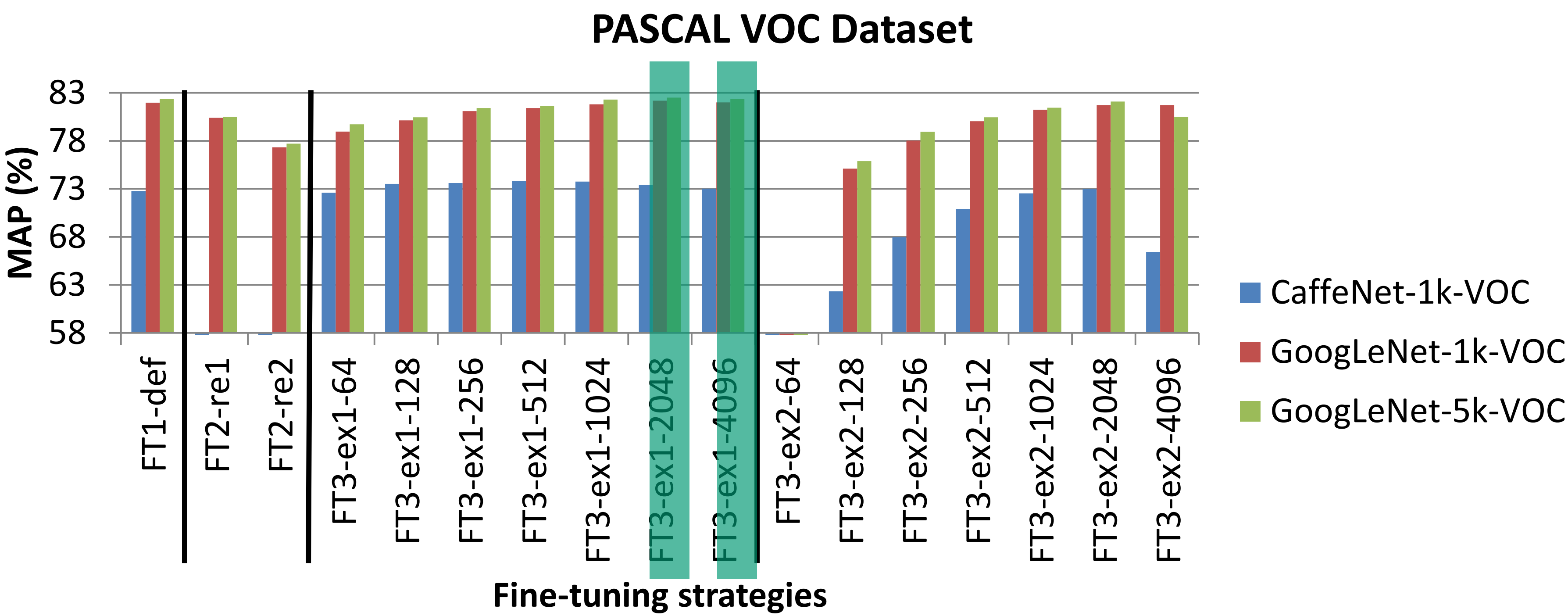
Information Technologies Institute

Queen Mary University of London

InVID

Experimental results – direct output

- **Goal**: Assess DCNNs as standalone classifiers (direct output)
- **Experiment**: Fine-tuning on the PASCAL VOC dataset
- **Conclusion**: FT3-ex1-2048 and FT3-ex1-4096 the top-two methods **for the GoogLeNet-based networks**



Information Technologies Institute

Queen Mary University of London

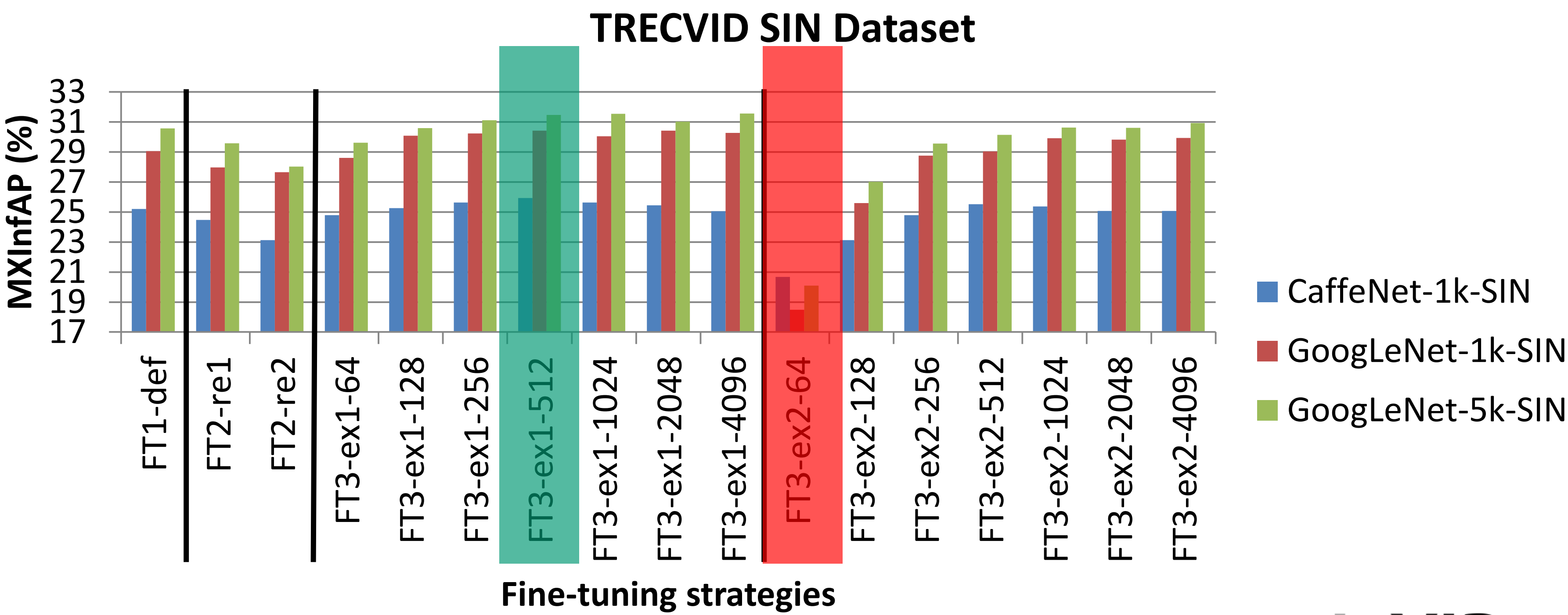
InVID

Experimental results – direct output

- **Main conclusions**: FT3-ex strategy with one extension layer is always the best solution
- The optimal dimension of the extension layer depends on the dataset and the network architecture

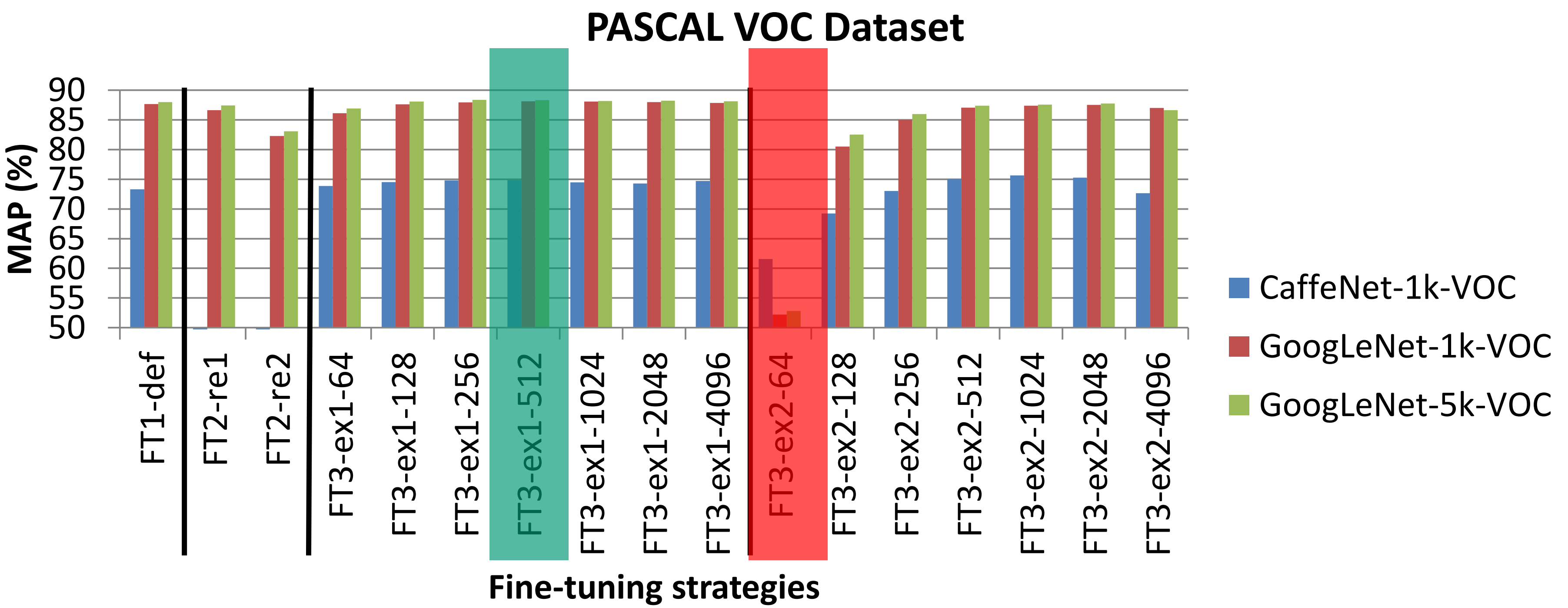
Experimental results – DCNN features

- **Goal**: Assess DCNNs as feature generators (DCNN-based features)
- **Experiment**: LR concept detectors trained on the output of the last 3 layers and fused in terms of arithmetic-mean
- **Conclusion**: FT3- ex1-512 in the top-five methods; FT3-ex2-64 is always among the five worst fine-tuning methods



Experimental results – DCNN features

- The same conclusions hold for the PASCAL VOC Dataset



Experimental results – DCNN features

- **Main conclusions**: FT3-ex strategy almost always outperforms the other two fine-tuning strategies
  - FT3-ex1-512 is in the top-five methods
- **Additional conclusions**: drawn from results presented in the paper
  - Features extracted from the top layers are more accurate than layers positioned lower in the network; the optimal layer varies, depending on the target domain dataset
  - Better to combine features extracted from many layers
  - The presented results correspond to the fused output of the last 3 layers

Conclusions

- Extension strategy almost always outperforms all the other strategies
  - Increase the depth with one fully-connected layer
  - Fine-tune the rest of the layers
- DCNN-based features significantly outperform the direct output
  - In a few cases the direct output works comparably well
  - Choose based on the application that the DCNN will be used; e.g., real time applications' time and memory limitations
  - Better to combine features extracted from many layers

References

[1] Campos, V., Salvador, A., Giro-i Nieto, X., Jou, B.: Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction. In: 1st International Workshop on Affect and Sentiment in Multimedia (ASM 2015), pp. 57–62. ACM, Brisbane (2015)

[2] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference (2014)

[5.] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR 2014) (2014)

[8] Markatopoulou, F., et al.: ITI-CERTH participation in TRECVID 2015. In: TRECVID 2015 Workshop. NIST, Gaithersburg (2015)

[9] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR 2014) (2014)

[15] Snoek, C., Fontijne, D., van de Sande, K.E., Stokman, H., et al.: Qualcomm Research and University of Amsterdam at TRECVID 2015: recognizing concepts, objects, and events in video. In: TRECVID 2015 Workshop. NIST, Gaithersburg (2015)

[18] Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? CoRR abs/1411.1792 (2014)

Thank you for your attention!
Questions?

More information and contact:
Dr. Vasileios Mezaris
bmezaris@iti.gr
http://www.iti.gr/~bmezaris

Information Technologies Institute

Queen Mary University of London

InVID