KTH VETENSKAP OCH KONST

ROYAL INSTITUTE OF TECHNOLOGY

From Generic to Specific Deep Representations for Visual Recognition

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, Stefan Carlsson

KTH (Royal Institute of Technology), Computer Vision and Active Perception Lab. (CVAP), Stockholm, Sweden

CVPR 2015 BOSTON JUNE 7-12

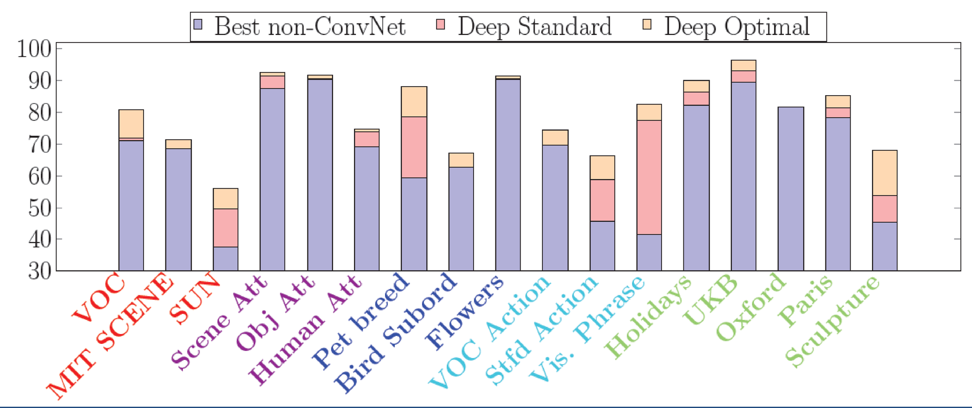
DeepVision: Deep Learning in Computer Vision 2015

Goal

- Understanding factors for transferability of a Generic ConvNet representation to different target tasks
- Analyze the correlation of the transferability factors and source to target tasks distance

Motivation

- Numerous computer vision tasks are affected by deep learning
  - Object detection
  - Scene recognition
  - Pose Estimation
  - Semantic Segmentation
- Better ConvNet representation often beats more complicated reasoning/modeling
  - *e.g.* Deeper networks rep. + SVM often outperforms shallow networks representation + complicated model



Contributions

- We propose a set of pre-training and post-training factors when transferring a generic ConvNet representation
- Categorically organize the different computer vision tasks
- Extensively study the proposed factors and their correlation with distance of source to target tasks
- state of the art performance on various (=16) recognition tasks

Transferability Factors

- We divide the transferability factors into two groups
- We call the decisions involved before learning the generic ConvNet representation on the source task, ***learning factors***
- We further identify factors which are relevant after optimizing the ConvNet on the source task: ***post-learning factors***

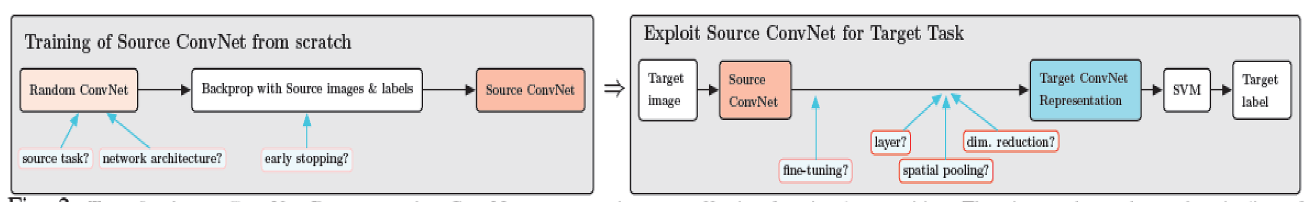


Fig. 2: **Transferring a ConvNet Representation** ConvNet representations are effective for visual recognition. The picture above shows the pipeline of transferring a source ConvNet representation to a target task of interest. We define several factors which control the transferability of such representations to different tasks (questions with blue arrow). These factors come into play at different stages of transfer. Optimizing these factors is crucial if one wants to

Learning Factors

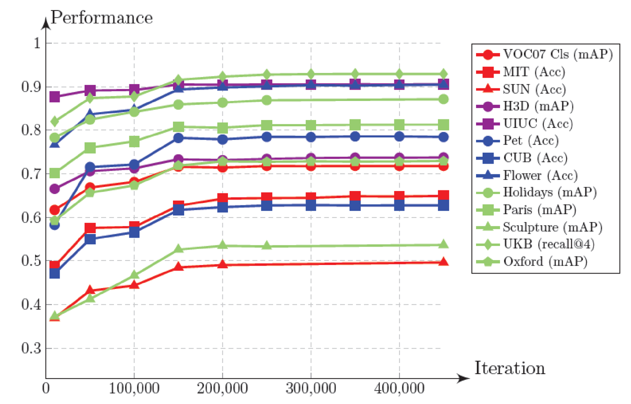
Source Task

	Image Classification			Attribute Detection		Fine-grained Recognition			Compositional	Instance Retrieval		
Source task	VOC07	MIT	SUN	H3D	UIUC	Pet	CUB	Flower	Stanf. Act40	Oxf.	Scul.	UKB
ImageNet	71.6	64.9	49.6	73.8	**90.4**	**78.4**	**62.7**	**90.5**	58.9	71.2	52.0	93.0
Places	68.5	69.3	55.7	68.0	88.8	49.9	42.2	82.4	53.0	70.0	44.2	88.7
Hybrid	72.7	69.6	56.0	72.6	90.2	72.4	58.3	89.4	58.2	**72.3**	52.3	92.2
Concat	**73.8**	**70.8**	**56.2**	**74.2**	**90.4**	75.6	60.3	90.2	**59.6**	72.1	**54.0**	**93.2**

TABLE V: **Source Task:** Results on all tasks using representations optimized for different source tasks. ImageNet is the representation used for all experiments of this paper. Places is a new ConvNet trained on 3.5M images labeled with scene categories [46]. Hybrid is a model proposed by Zhou et al. [46] which combines the ImageNet and Places datasets and train a single network for the combination. Concat indicates results of concatenating the feature obtained from ImageNet ConvNet and Places ConvNet for each input image. All results are for first fully connected layer (FC6).

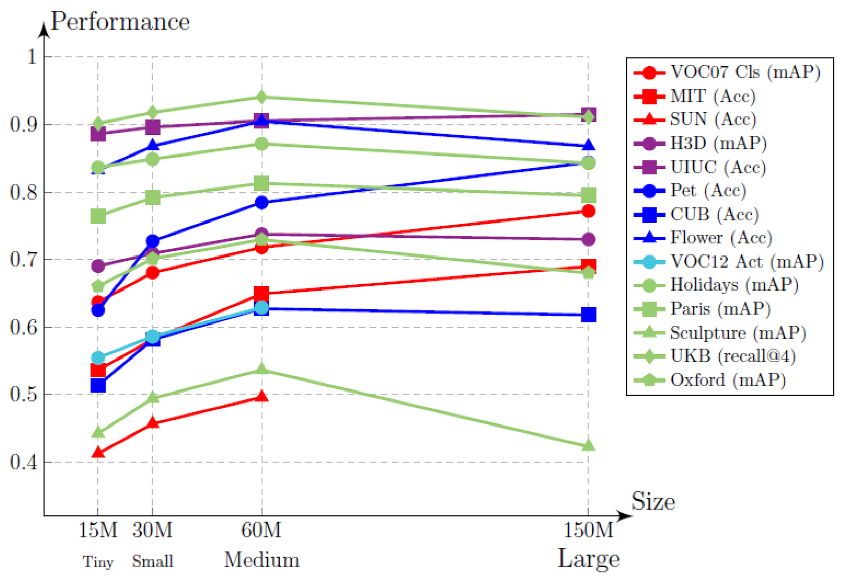
- How does changing the source task affect the transferability?
- Concatenation of features or Hybrid ConvNets?
- Cases where the source task is completely irrelevant

Early Stopping



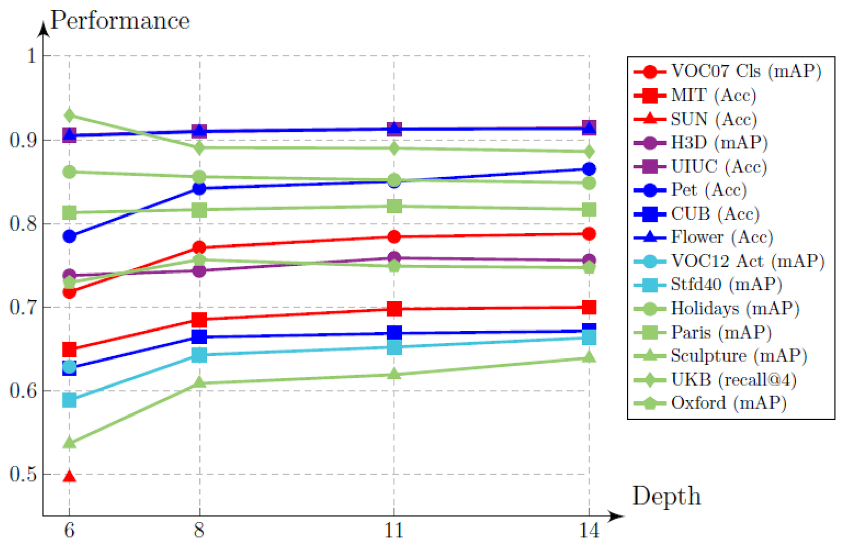
- Early stopping not a good idea! Except…

Network Width



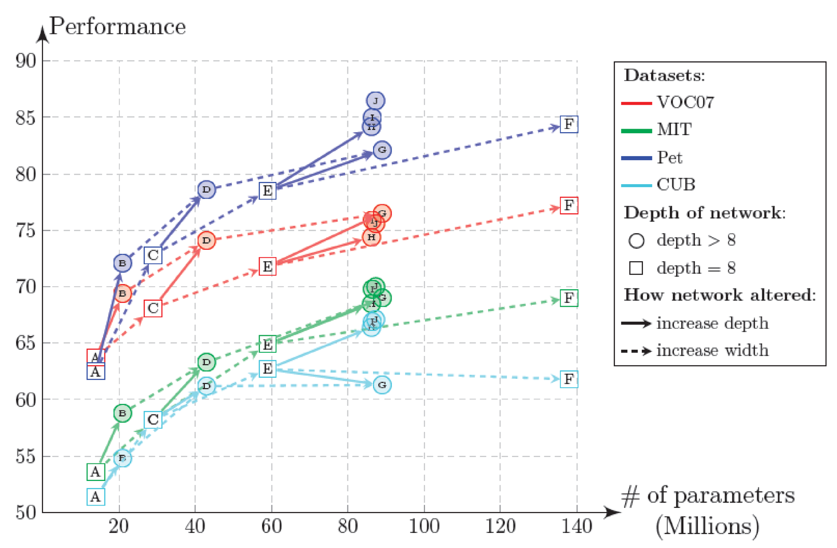
- Wide networks are good for source task and target tasks close to source

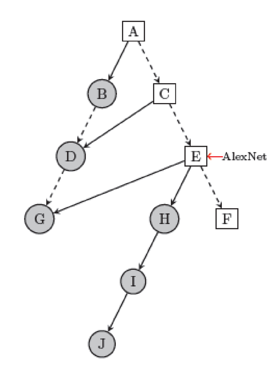
Network Depth



- Deep networks are almost always good, retrieval is an exception

Width v.s. Depth





- Better transferred performance to number of parameters ratio for deep networks

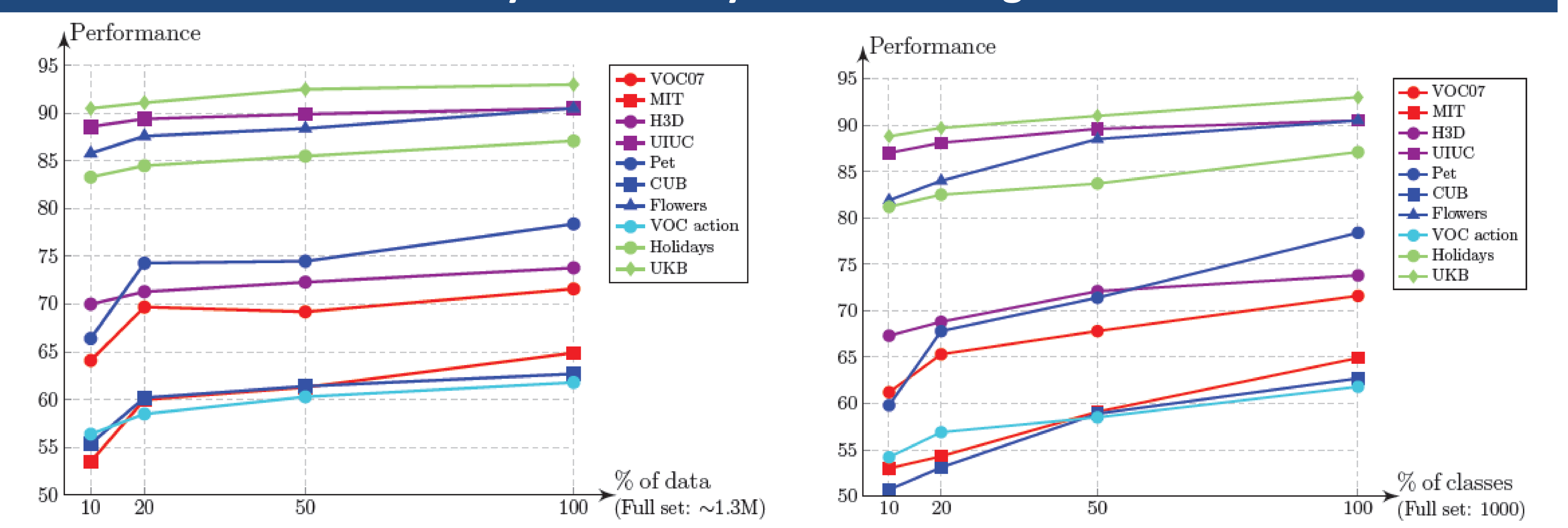
Additional Data

	Classification		Attribute		Fine-grained		Retrieval	
Dataset	VOC07	MIT	H3D	UIUC	Pet	Flower	Oxf.	Scul.
SUN	57.8	62.6	45.0	86.3	45.0	75.9	64.5	39.2
Places	**68.5**	**69.3**	**49.9**	**88.8**	**49.9**	**82.4**	**70.0**	**44.2**

TABLE VII: **Additional data (source task):** Results with the ConvNet representation optimized for different amount of training data. First row shows the results when the network is trained on scene recognition dataset of SUN397 [39] dataset with 100K images. The second row corresponds to the network trained on Places dataset [46] with 2.5M images annotated with similar categories. In all cases the ConvNet trained on Places dataset outperforms the one trained on SUN.

- Additional data helps the performance of the transferred feature, even if the target task is far from the source task

Diversity v.s. Density of the Training Data



(a) **Increasing Density:** We trained 3 "Medium" networks using different number of images per ILSVRC 2012 class. We decreased the number of images for each class to 10%, 20% and 50% of its original set separately. The images are sampled randomly. Heavier regularization is applied for networks with smaller data size.

(b) **Increasing Diversity:** We trained 3 "Medium" networks using different number classes from ImageNet. We decreased the number of classes to 100, 200 and 500 but kept the images within each class the same as ILSVRC 2012. The classes are sampled randomly. Heavier regularization is applied for networks with smaller data size.

- Even using only 100K subset of ImageNet, one can learn pretty amazing features, it is better to increase diversity of the training data as opposed to density!

Range of Tasks

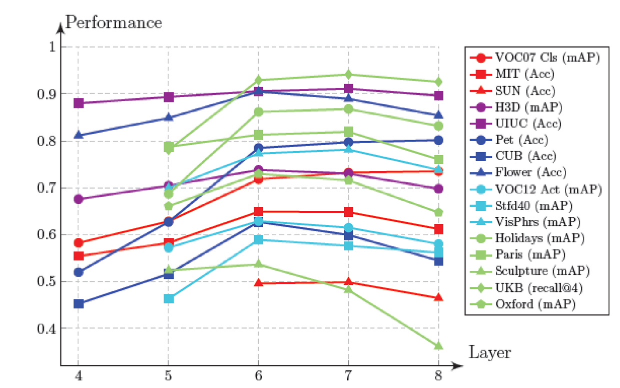
- Distance of source-target task can be analyzed from different viewpoints and can become ambigious. We take the following parameters into consideration:
- Target classes being super-category or sub-category of the source task
- Leaning based or metric based tasks
- Explanatory classes
- Image acquisition
- …

Increasing distance from ImageNet

Image Classification	Attribute Detection	Fine-grained Recognition	Compositional	Instance Retrieval
PASCAL VOC Object [9]	H3D human attributes [6]	Cat&Dog breeds [25]	VOC Human Action [9]	Holiday scenes [16]
MIT 67 Indoor Scenes [29]	Object attributes [10]	Bird subordinate [38]	Stanford 40 Actions [41]	Paris buildings [27]
SUN 397 Scene [40]	SUN scene attributes [26]	102 Flowers [23]	Visual Phrases [30]	Sculptures [4]

Post-Learning Factors

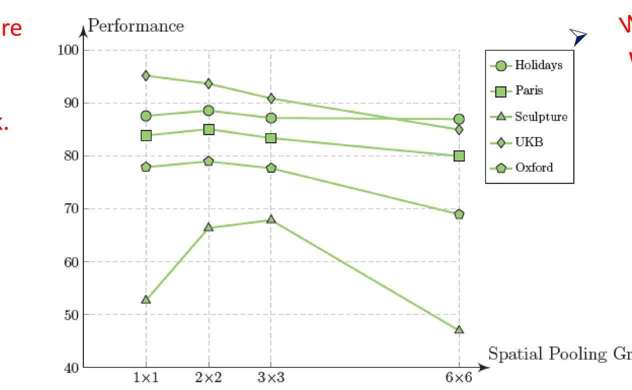
Network Layer



- Earlier layers are more suitable for target tasks further away from the source task.

Fig. 7: **Representation Layer:** Efficacy of representations extracted from AlexNet's different layers for different visual recognition tasks. A distinct pattern can be observed: the further the task moves from object image classification, the earlier layers are more effective. For instance, layer 8 works best for PASCAL VOC image classification which is very similar to ImageNet while the best performance for all retrieval tasks is at layer 6.

Spatial Pooling

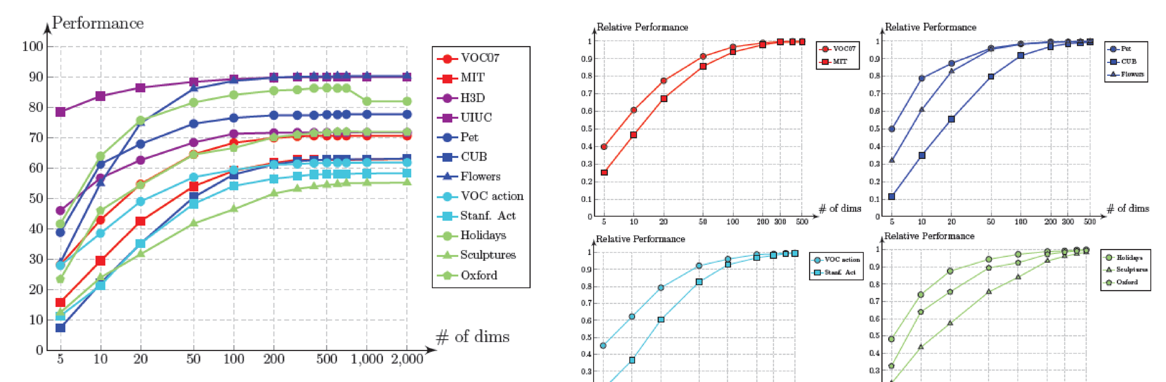


While the representation taken from convolutional layers are large and ineffective, pooled convNet features prove useful.

- Different pooling windows for different tasks based on the importance of fine details

Fig. 9: **Spatial Pooling:** In order to obtain meaningful results for retrieval with representations from convolutional layers we applied spatial pooling of different sizes for different tasks. Objects of more complex structures such as sculptures and buildings need more spatial resolution for optimal performance.

Dimensionality Reduction



(a) **Effective Dimensionality:** It can be observed that almost all tasks reach their maximum performance at an dimensionality of below 500 indicating a low (class-conditional) effective dimensionality of ConvNet representations. The accuracy of all tasks for dimensions under 50 are surprisingly high. Thus, he fact that these transformations are obtained using a linear transform supports capability of ConvNet in generalization by disentangling underlying generating factors.

(b) **Saturation:** As we move further from the original task, performance saturation happens with slightly more dimensions. For more clarity, the performance values of each task are divided by the maximum value for these plots. Horizontal axis is in log scale.

- Effective dimensionality of a ConvNet representation for various target task is between 200 to 500, closer tasks have slightly lower effective dimensionality

Fine Tuning

Representation	MIT	CUB	Flower
Medium FC7	65.9	62.9	90.4
Medium FT	**66.3**	**66.4**	**91.4**

TABLE VI: **Fine-tuning:** The first row shows the original ConvNet results. The second row shows the results when we fine-tune the ConvNet toward the target task and specialize the learnt representation. Fine-tuning is consistently effective. The proportional improvement is higher for the more distant tasks from ImageNet.

Representation	bird	cat	dog
ConvNet [13]	38.5	51.4	46.0
ConvNet-FT VOC [13]	50.0	60.7	56.1
ConvNet-FT VOC+CUB+Pet	**51.3**	**63.0**	**57.2**

TABLE VIII: **Additional data (fine-tuning):** The table presents the mAP accuracy of a sliding window detector based on different ConvNet representations for 3 object classes from VOC 2007. ImageNet contains more than 100,000 dog images and Pascal VOC has 510 dog instances. For the representation in the second row, image patches extracted from the VOC training set are used to fine-tune the ConvNet representation [13]. It results in a big jump in performance. But including cat, dog and bird images from the Oxford Pet and Caltech bird datasets boosts the performance even further.

Final Results Table

	Image Classification			Attribute Detection			Fine-grained Recognition			Compositional			Instance Retrieval				
	VOC07	MIT	SUN	SunAtt	UIUC	H3D	Pet	CUB	Flower	VOCa.	Act40	Phrase	Holid.	UKB	Oxf.	Paris	Scul.
non-ConvNet	[34] 71.1	[22] 68.5	[39] 37.5	[26] 87.5	[37] 90.2	[44] 69.1	[25] 59.2	[12] 62.7	[18] 90.2	[24] 69.6	[41] 45.7	[30] 41.5	[36] 82.2	[45] 89.4	[36] **81.7**	[36] 78.2	[4] 45.4
Deep Standard	71.8	64.9	49.6	91.4	90.6	73.8	78.5	62.8	90.5	69.2	58.9	77.3	86.2	93.0	73.0	81.3	53.7
Deep Optimized[4]	**80.7**	**71.3**	**56.0**	**92.5**	**91.5**	**74.6**	**88.1**	**67.1**	**91.3**	**74.3**	**66.4**	**82.3**	**90.0**	**96.3**	79.0	**85.1**	**67.9**
Err. Reduction	32%	18%	13%	13%	10%	4%	45%	12%	8%	17%	18%	22%	28%	47%	22%	20%	31%
Source Task	ImgNet	Hybrid	Hybrid	Hybrid	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	ImgNet	Hybrid	ImgNet	ImgNet	ImgNet	ImgNet
Network Width	Medium	Medium	Medium	Medium	Large	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Network Depth	16	8	8	8	8	16	16	16	16	16	16	16	8	8	16	16	16
Rep. Layer	last	last	last	last	2nd last	2nd last	2nd last	3rd last	3rd last	3rd last	3rd last	3rd last	4th last	4th last	4th last	4th last	4th last
PCA	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
Pooling	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	1×1	1×1	2×2	2×2	3×3
Deep Optimized MS	**80.7**	**71.3**	**56.0**	**92.5**	**91.5**	**74.6**	**88.1**	**67.1**	**91.3**	**74.3**	**66.4**	**82.3**	**90.0**	**96.3**	79.0	**85.1**	**67.9**
Deep Optimized MS+ML	**80.7**	**71.3**	**56.0**	**92.5**	**91.5**	**74.6**	**88.1**	**67.1**	**91.3**	**74.3**	**66.4**	**82.3**	**90.0**	**96.3**	79.0	**85.1**	**67.9**

TABLE IX: **Final Results:** Final results of the deep representation with optimized factors along with a linear SVM compared to the non-ConvNet state of the art. In the bottom half of the table the factors used for each task are noted. We achieve up to a **50% reduction of error** by optimizing transferability factors. Relative error reductions refer to how much of the remaining error (from Deep Standard) is decreased. "Deep Standard" is the common choice of parameters - a Medium sized network of depth 8 trained on ImageNet with representation taken from layer 6 (FC6).

- These factors are important! Taking different factors into account, we achieved up to 40% reduction of classification error.

[1] Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for visual recognition CVPR DeepVision Workshop (2014) [2] Sharif Razavian, A., Sullivan, J., Maki, A., Carlsson, S.: A Baseline for Visual Instance Retrieval with Deep ConvNets. ICLR Workshop (2014)

*** We gratefully acknowledge NVIDIA's help with the donation of multiple GPUs which made this work possible ***