Characterizing Visual Representations within Convolutional Neural Networks: Toward a Quantitative Approach

Chuan-Yung Tsai CHUANYUNGTSAI@FAS.HARVARD.EDU
Center for Brain Science, Harvard University, Cambridge, MA 02138 USA

David D. Cox DAVIDCOX@FAS.HARVARD.EDU
Center for Brain Science, Harvard University, Cambridge, MA 02138 USA

Abstract

Even when parameters of a deep neural network are fully known, it is still not always clear how and why a given network "works," and how neurons in the network contribute to overall network performance. In this paper, we propose a suite of tools for visualizing and characterizing deep networks that aims to reveal their key representation properties. We present preliminary results with a large collection of randomly generated networks to explore how representation properties relate to network depth and performance, and we find that these properties can explain a substantial fraction of variation in overall network performance.

1. Introduction

Deep convolutional neural networks have increasingly become a core tool in computer vision in recent years, delivering groundbreaking performance across a wide range of challenging datasets. Their success is commonly explained by the notion that deeper networks have greater "representation power," even though what exactly this means in practice is not always well-defined.

In this work, we focused on developing empirical tools to quantify the properties of representations within convolutional neural networks, and allow statistical comparison of these properties. Related work is briefly summarized next. (Erhan et al., 2010; Le et al., 2010; 2012; Zeiler & Fergus, 2014; Simonyan et al., 2014; Yosinski et al., 2015) showed that deep neurons are most responsive to complex patterns, including object-part- and even whole-object-resembling patterns, which are also called their optimal stimuli. (Mahendran & Vedaldi, 2015; Wang et al., 2015) generalized such analysis onto groups of neurons. Besides the optimal stimulus, (Goodfellow et al., 2009; Zeiler & Fergus, 2014; Fawzi & Frossard, 2015) also studied the invariance and selectivity of neurons using parametric deformations of images, like translation, rotation, scaling, *etc.*, while (Berkes & Wiskott, 2006; Le et al., 2010) used Hessian information and (Bakry et al., 2016) used kernel analysis. (Lenc & Vedaldi, 2015) also studied more advanced properties like equivariance and equivalence.

Here, we describe a suite of quantitative tools that allow us to compare representation properties across different layers and different networks using statistical tests. With statistical testing, not only descriptions about their differences can be made more rigorously, minor differences that are harder to tell can also emerge through such an approach. As a first test of our methods, we focused on algorithms from (Cox & Pinto, 2011) where a large number of shallow and minimally deep networks were randomly generated and characterized, as in this simple setup, representation properties and differences that had been observed before can be easily verified.

Interestingly, our approach not only statistically confirmed some known properties, but also identified at least one key difference which has not been reported before—instead of invariance, selectivity is actually the most significant difference between shallow *vs.* deep representations. When used to compare deep representations against their performance on image recognition (face pair matching), our methods explained 71% of the variance of the performance using our representation measures (*i.e.* which factors make a network perform well) with strong statistical confidence. Invertibility, in this case, turned out to be the most important property.

The rest of this paper is organized as follows. All proposed methods are detailed in Sec. 2. Experimental setup, *viz.* the networks and dataset used in this work, is briefly described in Sec. 3. Results for comparing representations using our

Presented at the *Workshop on Visualization for Deep Learning*, the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016.

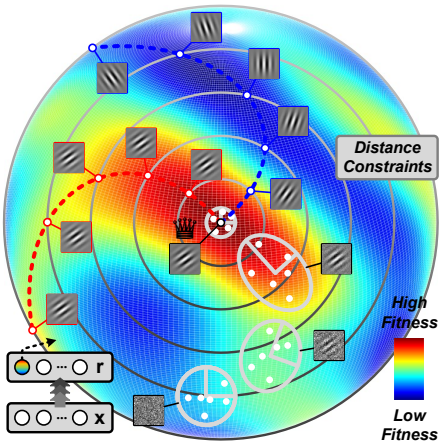

Figure 1. Overview of methods. On the constant-energy spherical constraint, the optimal stimulus (1) of a representation is first iteratively searched (Hansen & Ostermeier, 2001) and quantified using (4) and (5). Then, the invariant (2) and selective stimuli (3) at distances 0.1π to 0.5π away from the optimal stimulus, forming the invariance and selectivity paths (dashed red and blue curves), are searched as well and quantified using (6), (7), (8) and (9). All quantitative measures are then compared using permutation tests (Sec. 2.3).

methods are in Sec. 4. Implications of our major findings, their potential applications and other future directions are addressed in Sec. 4 and 5. More results can be found in the longer draft version (Tsai & Cox, 2015) of this paper too.

2. Methods

Given a network f, our methods extensively utilize both first-order (*i.e.* optimal stimulus related) and second-order (*i.e.* invariance and selectivity related) characterization for its visual representations (partly generalized and improved from previous papers). Throughout this paper, $\mathbf{r} = f(\mathbf{x})$ is called a scalar representation (*i.e.* $\mathbf{r} \in \mathbb{R}$) when considering single neurons, and a vector representation (*i.e.* $\mathbf{r} \in \mathbb{R}^M$) when considering multiple neurons. We also call f a fitness function and its value fitness. Figure 1 depicts an overview of our methods.

2.1. Representation Search

The optimal stimuli for both scalar and vector representations are defined in (1) subject to $\|\mathbf{x}\| = E$, where E is set to the task-related image's average energy. When considering the vector representation, it is equivalent to maximizing the response of an "auxiliary neuron" tuned to certain reference stimulus $\tilde{\mathbf{x}}$ (*e.g.* a natural image).[1]

$$\hat{\mathbf{x}} = \begin{cases} \arg\max_{\mathbf{x}} f(\mathbf{x}) & \text{if } f(\mathbf{x}) \in \mathbb{R} \\ \arg\max_{\mathbf{x}} e^{-\|f(\mathbf{x}) - f(\tilde{\mathbf{x}})\|} & \text{if } f(\mathbf{x}) \in \mathbb{R}^M \end{cases} \tag{1}$$

$$\mathbf{x}_\delta^+ = \begin{cases} \arg\max_{\mathbf{x}_\delta} f(\mathbf{x}_\delta) & \text{if } f(\mathbf{x}) \in \mathbb{R} \\ \arg\max_{\mathbf{x}_\delta} e^{-\|f(\mathbf{x}_\delta) - f(\hat{\mathbf{x}})\|} & \text{if } f(\mathbf{x}) \in \mathbb{R}^M \end{cases} \tag{2}$$

$$\mathbf{x}_\delta^- = \begin{cases} \arg\min_{\mathbf{x}_\delta} f(\mathbf{x}_\delta) & \text{if } f(\mathbf{x}) \in \mathbb{R} \\ \arg\min_{\mathbf{x}_\delta} e^{-\|f(\mathbf{x}_\delta) - f(\hat{\mathbf{x}})\|} & \text{if } f(\mathbf{x}) \in \mathbb{R}^M \end{cases} \tag{3}$$

With respect to the optimal stimulus $\hat{\mathbf{x}}$, the invariant and selective stimuli are defined in (2) and (3) respectively, where $0 < \delta \leq \frac{\pi}{2}$, subject to $\|\mathbf{x}_\delta\| = E$ and $\langle \mathbf{x}_\delta, \hat{\mathbf{x}} \rangle = E^2 \cos(\delta)$. The distance constraint, while being simple and linear, enforces exploration of the fitness landscape, which is one of the main differences compared to (Erhan et al., 2010). The invariance path $\{\mathbf{x}_\delta^+\}$ and selectivity path $\{\mathbf{x}_\delta^-\}$ are then searched through multiple runs of maximization and minimization on discretized $\delta \in \{0.1\pi, 0.2\pi, 0.3\pi, 0.4\pi, 0.5\pi\}$ as the distance constraints shown in Fig. 1, where each run is initialized with the result from the previous run (and the 0.1π run directly with optimal stimulus $\hat{\mathbf{x}}$) to increase the path continuity and searching speed. This method is more generic than using parametric deformations, and more efficient than using the Hessian. Similarly, it can be performed with respect to certain reference stimulus $\tilde{\mathbf{x}}$ (especially for vector representations) where the invariance and selectivity of the auxiliary neuron are characterized equivalently.

2.2. Representation Quantification

Complexity: It is often observed that deeper neurons hold more complex optimal stimuli, as extensively visualized in (Zeiler & Fergus, 2014; Yosinski et al., 2015). However, most previous work only provided qualitative results. With the $\|\mathbf{x}\| = E$ constraint, we can directly define its spectral complexity measure as the L_1 norm of the Fourier power spectrum of the optimal stimulus, *i.e.*

$$\text{Complexity} = \|\mathcal{F}(\hat{\mathbf{x}})\|_1 . \tag{4}$$

Since all stimuli have the same $\|\mathcal{F}(\hat{\mathbf{x}})\|$ (L_2 norm), higher L_1 norm also implies higher non-sparsity (*i.e.* complexity), as simple stimuli (*e.g.* sine gratings or Gabor filters) usually have sparse Fourier power spectrum.

Invertibility: As demonstrated by previous papers, to understand the meaning of a representation $\mathbf{r}$ formed by the network f, one can simply perform the inversion $f^{-1}(\mathbf{r})$,

[1] Although a vector representation's optimal stimulus can also be defined closer to a scalar representation's, *e.g.* as the stimulus maximizing the sum of all responses, in practice we found such a definition usually led to much less informative results.

e.g. using (1), to visualize the properties of this representation. To further quantify this process, we perform multiple inversions of a representation and define

$$\text{Invertibility} = \frac{1}{n}\sum_{i=1}^{n} \text{SSIM}\left(\tilde{\mathbf{x}}, \hat{\mathbf{x}}_i\right) \tag{5}$$

as the average structural similarity (Wang et al., 2004) between the reference stimulus $\tilde{\mathbf{x}}$ and n inverted stimuli from random initialization, where $n = 10$ in our experiments.

Invariance & selectivity: The major advantage of searching the invariance and selectivity paths is that we can subsequently define a neuron's normalized invariance and selectivity easily. First we define the normalized fitness $\hat{f}(\mathbf{x})$ as $f(\mathbf{x})/f(\tilde{\mathbf{x}})$. Then, as δ increases on the fitness *vs.* distance diagram (Jones & Forrest, 1995), a neuron with high invariance should have its invariance curve $\hat{f}(\mathbf{x}_\delta^+)$ staying close to 1, and a neuron with high selectivity should have its selectivity curve $\hat{f}(\mathbf{x}_\delta^-)$ dropping fast (see Fig. 4). Knowing the fact that a single inner-product neuron (the simplest form of neural network) has its invariance and selectivity curves $\hat{f}(\mathbf{x}_\delta^+) = \hat{f}(\mathbf{x}_\delta^-) = \cos(\delta)$,[2] we can thus easily define the baseline curve (*i.e.* zero invariance and selectivity curve) as $\cos(\delta)$, and

$$\text{Invariance} = \int_0^{0.5\pi} \left|\cos^{-1}(\hat{f}(\mathbf{x}_\delta^+)) - \delta\right| \mathrm{d}\delta, \tag{6}$$

$$\text{Selectivity} = \int_0^{0.5\pi} \left|\cos^{-1}(\hat{f}(\mathbf{x}_\delta^-)) - \delta\right| \mathrm{d}\delta, \tag{7}$$

being the normalized areas sandwiched between the invariance and selectivity curves and the baseline curve.

Capacity: While the invariance (6) and selectivity (7) defined above only characterize best solutions of $\{\mathbf{x}_\delta^+\}$ and $\{\mathbf{x}_\delta^-\}$, we are also interested in how diverse different solutions can be, particularly for $\mathbf{x}_\delta^+$, as it indicates the dimensionality of the "invariance subspace" (*i.e.* the central high fitness region as in Fig. 1), or the capacity of the representation. To this goal, we perform multiple runs of invariant stimulus searches, and measure

$$\text{Capacity} = \left\|\left[\mathbf{x}_{\delta,1}^+, \ldots, \mathbf{x}_{\delta,n}^+\right]\right\|_* \tag{8}$$

as the nuclear norm of the concatenation of n search results, where $n = 20$ and $\delta = 0.1\pi$ in our experiments.

Alignment against natural images: Another very important quality regarding invariance and selectivity is how well they may in reality benefit visual recognition. To quantify this, we first perform PCA on task-related natural images (*e.g.* face images in our experiments) to obtain its PC vector space $\mathbf{V}$ (*i.e.* eigenfaces in our case), and with respect to a reference stimulus $\tilde{\mathbf{x}}$ measure

$$\text{Alignment} = \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{V}\mathbf{x}_{\delta,i}\right\|_1, \tag{9}$$

the average sparsity of $n = 20$ invariant and selective stimuli at $\delta = 0.1\pi$ represented in the PC vector space. This estimates how likely these invariant and selective directions are pointing onto other task-related image as well, in which case they should benefit the task more.

2.3. Statistical Testing

Comparing two representations: Given 2 distributions of measures A and B, we use permutation tests to determine the significance of their difference (*e.g.* Bhattacharyya distance) as follows. First compute the original distance $d = \text{dist}(A, B)$. Then without replacement resample A' and B' from $A \cup B$ while keeping their sizes (*i.e.* $|A'| = |A|$ and $|B'| = |B|$) and compute the new distance d'. Repeat this process for a larger number of times (*e.g.* 10^6) and calculate the probability p of $d' > d$. A small p suggests the null hypothesis that A and B are actually coming from the same distribution and thus d is not significant, can be rejected. Notation-wise, *, **, and *** mean $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

Comparing representations against their performance: Given a sequence of distributions of measures $A_1, \ldots, A_n$ and their performance numbers $P_1, \ldots, P_n$, we again use permutation tests to determine the significance of their correlation (*e.g.* Spearman's rank correlation). First compute the original correlation $\rho = \text{corr}\left(\{\mathbb{E}[A_i]\}, \{P_i\}\right)$ ordinarily. Then resample A_i' from $\bigcup_i A_i$ similarly and compute the new ρ'. Repeat this process to calculate the probability p of $\rho' > \rho$. A small p rejects the null hypothesis that the sequence of distributions are actually the same and thus their means do not have a significant correlation ρ against their performance. One may use this to compare representation measures against their depths as well.

3. Experimental Setup

In our experiments, 100 shallow (*i.e.* one `conv` layer) networks and 100 deep (*i.e.* two `conv` layers) networks with 11×11 and 21×21 receptive field sizes respectively were randomly generated following (Cox & Pinto, 2011). `conv` and `pool` sizes were chosen from $\{3, 5, 7, 9\}$ and `relu` was applied right after `conv`. For shallow networks, 32 `conv1` filters were adopted. For deep networks, number of `conv1` filters were randomly chosen from $\{8, 16, 32, 64\}$ and 32 `conv2` filters were used. `pool` style can be *average*, *squared*, or *max-like*. Shallow and deep representa-

[2] Given $\|\mathbf{x}\| = E$, a single inner-product neuron $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$ has optimal stimulus $\hat{\mathbf{x}} = E\|\mathbf{w}\|^{-1}\mathbf{w}$ and all its invariance and selectivity curves $\hat{f}(\mathbf{x}_\delta) = f(\mathbf{x}_\delta)/f(\hat{\mathbf{x}}) = \cos(\delta)$.

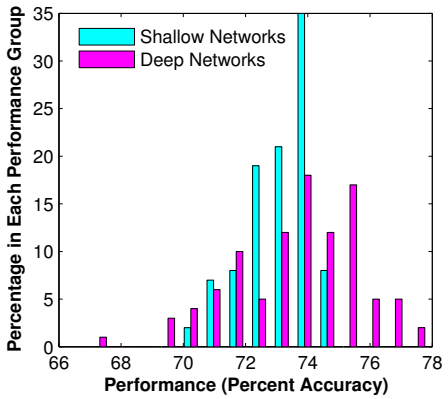

Figure 2. Performance of shallow and deep networks on LFW-a face pair matching. We tried to address the following 2 questions using our methods: (1) why do deep networks perform better than shallow networks? and (2) among all networks of the same depth, why do certain networks perform better than others?

tions correspond to `pool1` and `pool2` neurons.[3] Accuracies of linear SVMs reading out from these 100 shallow and deep representations against the LFW-a dataset (Huang et al., 2007; Wolf et al., 2011) are summarized in Fig. 2.

Scalar representations, *i.e.* totally 3,200 shallow and 3,200 deep individual neurons, were measured, when comparing shallow and deep representations (see Sec. 4.1), where all the optimal, invariant and selective stimulus searches were performed twice, and the better numerical result were kept.

Vector representations, on the other hand, were used when comparing all the 100 deep networks against their performance. We randomly picked 16 reference stimuli (*i.e.* face images), from which more runs of searches were performed (as described in Sec. 2.2) to acquire a sufficient number of measures for statistical comparison (see Sec. 4.2).

4. Results

4.1. Shallow *vs.* Deep Representations

Figure 3 demonstrates search results of shallow *versus* deep representations. First, the optimal stimuli of shallow neurons are visually simpler than those of deep neurons. Second, in shallow neurons, invariance paths are mostly phase changes and selectivity paths are leading toward meaningless noises (all at the same falloff rate). However, in deep neurons, both types of paths consist of sophisticated shape deformations. Finally, the invariance subspaces of shallow

[3]These representations, though being randomly generated and relatively simple, in fact performed competitively well on LFW-a (among neural network based algorithms without using outside data) and served as an efficient testbed, particularly since relative, instead of absolute performance, is of our primary interest.

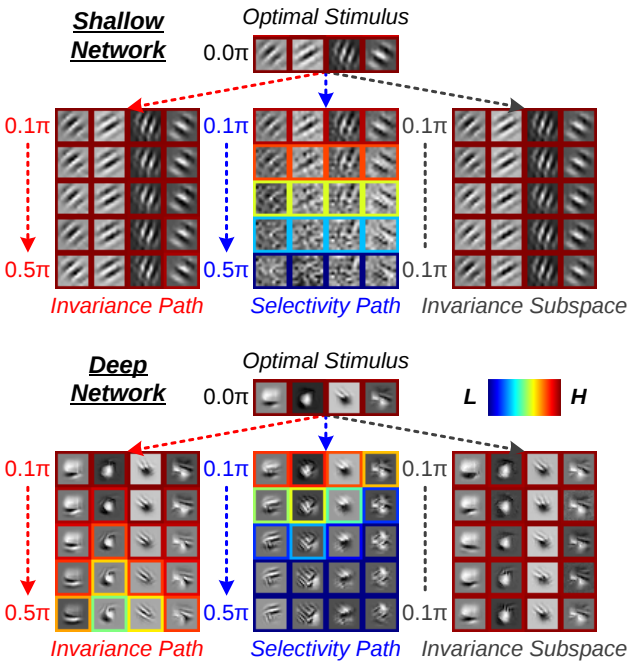

Figure 3. Visualization of shallow and deep representations. Optimal stimuli, invariance paths, selectivity paths, and invariance subspaces of 4 neurons randomly selected from the best performing shallow and best performing deep networks are shown respectively. Color of the boarder of an image indicates the fitness (here, response of the neuron) elicited by the image.

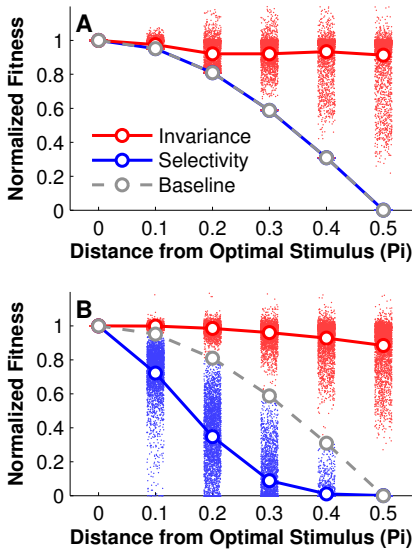

Figure 4. Fitness-distance diagrams of shallow and deep representations. Invariance and selectivity curves of all 3,200 shallow (panel A) and 3,200 deep neurons (panel B) are shown for comparison.

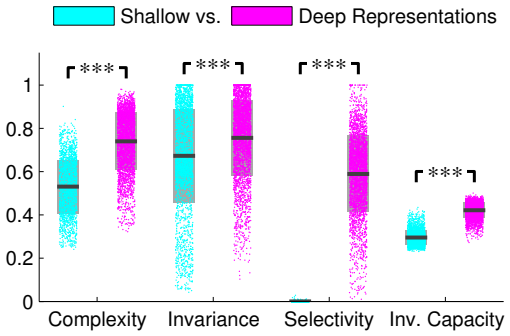

Figure 5. Differences between shallow and deep representations. Bhattacharyya distances and their significance levels between the shallow and deep representations under the 4 measures from left to right are 0.39^{***}, 0.07^{***}, 6.42^{***} and 1.27^{***} respectively. Dark gray lines indicate the means, and light gray boxes indicate the ranges of one standard deviations.

neurons are also visually simpler than those of deep neurons.

When further plotting the fitness-distance diagram (Fig. 4), we observed that, although shallow neurons can have good invariance, they surprisingly have *zero selectivity* (*i.e.* blue curve drops as slow as the baseline), while deep neurons show both good invariance and selectivity.[4] We also tested manually rotated optimal stimuli of shallow neurons, since they are generally most selective to orientation changes (of *e.g.* Gabor filters). However, the resultant fitnesses still did not drop faster then the numerical solutions like illustrated in Fig. 4. This suggests our numerical searches likely had found the steepest selectivity curves for shallow neurons—the cosine falloff.

Comparisons between the shallow and deep representations are summarized in Fig. 5. In addition to the quantities of differences, the significance levels are also reported. With statistical testing, we can confirm deep representations are indeed quantitatively and significantly better than shallow representations. Otherwise, such claim is hardly supported, particularly since their distributions of complexity and invariance measures are actually very close.

Selectivity (as the most significantly different measure) can strongly benefit visual recognition given similar invariance, because subtle visual differences can induce more neuronal response changes when the gap between the invariance and selectivity curves at a small δ is larger (*i.e.* stronger selectivity).

[4] A small fraction of invariance curves actually went over 1 as more optimization (2) were run with the increasing δ, simply due to the non-convex nature of these networks. This however did not cause any noticeable difference in our results as shown in Fig. 5.

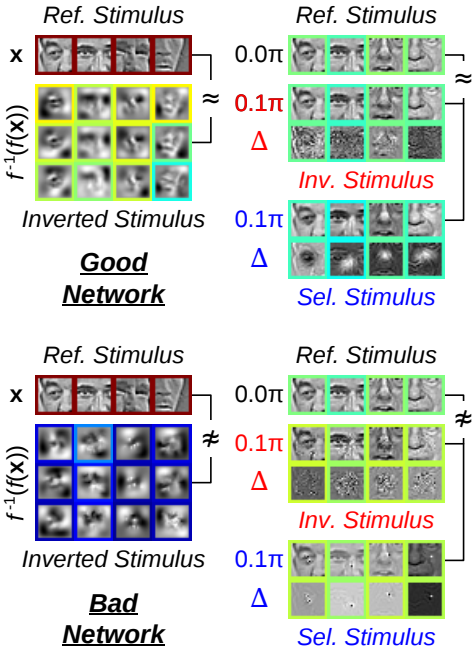

Figure 6. Visualization of good and bad representations. Examples of good and bad invertibility, invariance and selectivity alignments, from deep representations, are shown respectively. Color of the boarder of an image indicates the corresponding measure value.

4.2. Good *vs.* Bad Representations

Figure 6 demonstrates deep representations with good and bad measures. A representation with good invertibility simply implies it can be inverted with results visually closer to the reference stimulus, compared to a representation with bad invertibility. Similarly, a representation with good invariance (or selectivity) alignment against natural images, implies its most invariant (or selective) directions, starting from a reference stimulus, point to other more natural looking images. One may observe such differences more easily in the directions Δ, where a good representation has more natural (*e.g.* structural deformations, lighting changes, *etc.*) Δ and a bad representation has noisy Δ.

Correlations between all the 100 deep networks' representation measures and performance are summarized in Fig. 7, where multiple correlation analysis (*i.e.* best linear combination) is used to integrate multiple representations.[5] Overall, the proposed representation measures can explain 71%

[5] The remaining measures included in the multiple correlation analysis but not individually plotted in Fig. 7 are the invariance, selectivity, invariance capacity, and optimal stimulus' explanation power, and their correlations and significance levels are -0.31^{**}, -0.24^{*}, -0.39^{***} and 0.60^{***} respectively. Details can be found in (Tsai & Cox, 2015).

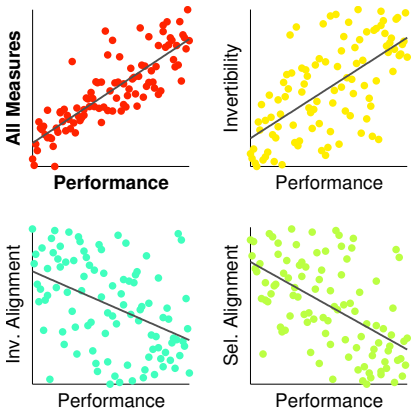

Figure 7. Correlations between deep representations' measures and performance. Spearman's rank correlations and their significance levels of all measures combined and the 3 individual measures *versus* the performance are 0.84^{***}, 0.64^{***}, -0.44^{***} and -0.56^{***} respectively.

of the variance (ρ^2, where $\rho = 0.84$) of the networks' performance with strong statistical significance.

Invertibility is noticeably the strongest measure in predicting the network performance, despite that those networks, like in most previous work, were not trained to perform an inversion task. This can be seen as an evidence supporting the view that deep networks, although being discriminative models, actually perform like generative models implicitly (Patel et al., 2015; Arora et al., 2016), in which case better "invertibility" naturally corresponds to better performance.

5. Discussion

In this work, we verified the effectiveness of our methods in characterizing visual representations from randomly generated networks performing face pair matching tasks. One of the future directions is to apply our methods onto more recent and deeper networks trained under various recognition tasks and see if similar known or other unknown properties can be identified. Both of these cases can not only enhance our understandings toward deeper representations, but also help us in improving their performance potentially. For instance, better network performance may be achieved (Reed et al., 2015) by enforcing invertibility—the most important property identified in this work. This suggests regularizing crucial representation properties to improve their measures (thus the performance potentially) can be a valuable future direction as well.

References

Arora, Sanjeev, Liang, Yingyu, and Ma, Tengyu. Why are deep nets reversible: A simple theory, with implications for training. In *ICLR Workshop*, 2016.

Bakry, Amr, Elhoseiny, Mohamed, El-Gaaly, Tarek, and Elgammal, Ahmed. Digging deep into the layers of CNNs: In search of how CNNs achieve view invariance. In *ICLR*, 2016.

Berkes, Pietro and Wiskott, Laurenz. On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural Computation*, 2006.

Cox, David and Pinto, Nicolas. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *IEEE FG*, 2011.

Erhan, Dumitru, Courville, Aaron, and Bengio, Yoshua. Understanding representations learned in deep architectures. Technical report, Université de Montréal, 2010.

Fawzi, Alhussein and Frossard, Pascal. Manitest: Are classifiers really invariant? In *BMVC*, 2015.

Goodfellow, Ian, Lee, Honglak, Le, Quoc, Saxe, Andrew, and Ng, Andrew. Measuring invariances in deep networks. In *NIPS*, 2009.

Hansen, Nikolaus and Ostermeier, Andreas. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 2001.

Huang, Gary, Ramesh, Manu, Berg, Tamara, and Learned-Miller, Erik. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

Jones, Terry and Forrest, Stephanie. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *International Conference on Genetic Algorithms*, 1995.

Le, Quoc, Ngiam, Jiquan, Chen, Zhenghao, Chia, Daniel, Koh, Pang, and Ng, Andrew. Tiled convolutional neural networks. In *NIPS*, 2010.

Le, Quoc, Ranzato, Marc'Aurelio, Monga, Rajat, Devin, Matthieu, Chen, Kai, Corrado, Greg, Dean, Jeff, and Ng, Andrew. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

Lenc, Karel and Vedaldi, Andrea. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, 2015.

Mahendran, Aravindh and Vedaldi, Andrea. Understanding deep image representations by inverting them. In *CVPR*, 2015.

Patel, Ankit, Nguyen, Tan, and Baraniuk, Richard. A probabilistic theory of deep learning. *CoRR*, abs/1504.00641, 2015.

Reed, Scott, Lee, Honglak, Anguelov, Dragomir, Szegedy, Christian, Erhan, Dumitru, and Rabinovich, Andrew. Training deep neural networks on noisy labels with bootstrapping. In *ICLR Workshop*, 2015.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICML Workshop*, 2014.

Tsai, Chuan-Yung and Cox, David. Measuring and understanding sensory representations within deep networks using a numerical optimization framework. *CoRR*, abs/1502.04972, 2015.

Wang, Jianyu, Zhang, Zhishuai, Premachandran, Vittal, and Yuille, Alan. Discovering internal representations from object-CNNs using population encoding. *CoRR*, abs/1511.06855, 2015.

Wang, Zhou, Bovik, Alan, Sheikh, Hamid, and Simoncelli, Eero. Image quality assessment: from error visibility to structural similarity. *IEEE T-IP*, 2004.

Wolf, Lior, Hassner, Tal, and Taigman, Yaniv. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE T-PAMI*, 2011.

Yosinski, Jason, Clune, Jeff, Nguyen, Anh, Fuchs, Thomas, and Lipson, Hod. Understanding neural networks through deep visualization. In *ICML Workshop*, 2015.

Zeiler, Matthew and Fergus, Rob. Visualizing and understanding convolutional networks. In *ECCV*, 2014.