The Multiverse Loss for Robust Transfer Learning

Etai Littwin
Department of Electrical Engineering
Tel-Aviv University

Lior Wolf
The Blavatnik School of Computer Science
Tel Aviv University

Abstract

Deep learning techniques are renowned for supporting effective transfer learning. However, as we demonstrate, the transferred representations support only a few modes of separation and much of its dimensionality is unutilized. In this work, we suggest to learn, in the source domain, multiple orthogonal classifiers. We prove that this leads to a reduced rank representation, which, however, supports more discriminative directions. Interestingly, the softmax probabilities produced by the multiple classifiers are likely to be identical. Experimental results, on CIFAR-100 and LFW, further demonstrate the effectiveness of our method.

1. Introduction

One of the hallmarks of the recent success of deep learning methods in computer vision is the ability to learn effective representations in one domain and apply these on another domain [15, 7, 6, 39, 22, 34, 16]. The source and target domains might differ in the underlying probability distribution, the imaging modality, and, often, in the task performed. A striking example is image captioning [12], in which image representations trained on ImageNet [4] are transferred along with word embeddings trained on Wikipedia and other corpora [21] in order to solve a seemingly complex task of describing images with sentences.

Another task where transfer learning has been shown to be effective is face recognition. In this task, face representations are trained on large datasets collected from social networks or search engines. The representations are trained to solve the multiclass classification problem using a cross entropy loss and are then transferred to a different domain, e.g., the celebrity images of the LFW dataset [11]. Moreover, the task changes post-transfer to face verification (same/not-same).

An effective algorithm for face verification based on engineered or learned representations is the Joint Bayesian (JB) method [2]. JB, similarly to other Bayesian methods, such as Linear Discriminant Analysis (LDA), is based on the interplay between the within class covariance matrix S_w and the between class covariance matrix S_b. We prove (Thm. 5) that JB fails to be discriminative whenever LDA fails, i.e., when the Fisher ratio (Eq. 20 below) is low.

When empirically observing the spectrum of Fisher ratios associated with the transferred representations, we noticed that only a handful of the generalized eigenvectors of S_b and S_w present large eigenvalues. The other directions are therefore non-discriminative and the representation can be considered flat.

To amend this situation, we propose to employ, in the source domain, a generalization of the cross entropy loss. In this generalization, multiple sets of classifiers are learned, such that the group of classifiers for each class is orthogonal. Each set of classifiers is trained using a separate cross entropy loss, and gives rise to its own set of probabilities.

When performing such training a few non-trivial properties emerge: (i) For each training sample, the vector of probabilities obtained is identical across the classifier sets; (ii) The dimensionality of the representation is reduced and (iii) The Fisher Spectrum displays multiple directions with high Fisher scores. In a series of theorems, we expose how the new loss leads to these properties.

Finally, we demonstrate experimentally the both the effectiveness of our method and the consequences of the emerging properties. For example, using a single network, we obtain 2nd best results for a single network on LFW [11]. This is achieved using a a training set that is a few orders of magnitude smaller than that of the leading literature network [23], and using a very compact representation of only 51 dimensions.

2. Related work

Compound losses for training deep neural networks that are created by combining multiple losses are now commonplace. In the very deep GoogLeNet network [30] multiple cross entropy losses are distributed at different intermediate layers of the deep network in order to help avoid vanishing gradients. In contrast, our work supports multiple cross entropy losses at the top layer and for different reasons.

In many other cases, multiple losses are used in order to support multiple tasks by the same network. For example,

in object detection and object segmentation, the location of the object is recovered jointly with the associated detection probability [20, 8, 31]. This is in contrast to our case, where the same loss is used multiple times in order to improve the performance of one task.

In our work, we create multiple losses by constructing multiple top classification layers on top of a shared network representation. Each classification layer has one output neuron per class. The weights from the representation to this neuron are the classifier weights for this specific classifier. In order to enforce multiplicity among the classifiers of the same class, we add an orthogonality constraint, which is enforced either in the representation space or in a Fisher spectrum aligned space. A number of ways to encourage diversity in a classifier ensemble by enforcing orthogonality have been studied in the machine learning literature [1] and in computer vision [17]. However, note that in our case orthogonality does not lead to diversity since all classifiers end up presenting the same set of probabilities.

A prominent example of the success of transfer learning can be seen in the task of face recognition. Starting with the work of Taigman et al. [32], a neural network has been employed for extracting representations from face images that are shown to outperform humans. Sun et al. [28, 25, 29, 26] further improve the state-of-art by extracting features from multiple face patches, incorporating architectures into the domain of deep face recognition that are inspired by recent architectures that are used for object recognition [24], and most relevant to our work, combining, during training, both classification and contrastive loss. Another recent work [23] further improves the training criterion by using a triplet cost to increase the discriminability between identities. The idea presented here, of combining multiple copies of the same loss, was not pursued in previous works.

The deep face networks mentioned above, are all trained on large scale proprietary datasets, which are not publicly available. Yi et al. [41] built a publicly available dataset by mining images from the internet. Furthermore, they demonstrated the quality of the data collected by training a state-of-the-art network on it. Their network architecture is similar to that of the VGG model [24]. JB is used to effectively enhance performance. In our work, we use the same architecture suggested in [41] as the basis of our face recognition experiments. We also employ JB to learn similarities for faces and other objects. A recent paper using JB outside the domain of face recognition is [40].

3. Preliminaries and notation

The notations used in this work are summarized in Tab. 1. n training samples, indexed by $i = 1 \dots n$ are represented, using a network of any depth as "neural code" vectors of length d, $D_{d\times n} = [d_1 \dots d_n]$. Each sample is associated with a label $y_i \in [1 \dots c]$.

Symbol	
c	Number of classes, typically indexed by j.
n	Number of data points, typically indexed by i.
y	$n \times 1$ vector of labels. Each label is y_i.
d	Dimensionality of the representation vectors.
D	$d \times n$ features matrix.
d_i	A column of D, the representation of sample i.
F	$d \times c$ classifier matrix of weights.
f_i	A column of F; A normal to the separating hyperplane of class i.
b	$c \times 1$ vector of parameters (biases).
L_i	Loss function value evaluated for data point i.
L	The aggregated loss function: $\sum_{i=1}^{n} L_i$.
F^*, b^*	(any) global minimizers of $L(D, y)$.
$L^*(D, y)$	The minimum value of L given D, y, i.e L(F^*, b^*, D, y).
K	The linear kernel matrix of the data: $DD^\top$.
$\mathbf{1}_c$	An all 1 vector of length c: $[1, 1 \dots 1]_{c\times 1}$.
$p_i(j)$	Vector of probabilities associated with d_i.
S_b (S_w)	The between (within) class covariance matrix.

Table 1. Summary of notations.

Classification is performed by projecting the representations d_i by a $d \times c$ classifier matrix $F = [f_1 \dots f_c]$ and adding biases $b \in \mathbb{R}^c$. Softmax probabilities are obtained as $p_i(y_i) = \frac{e^{d_i^\top f_{y_i} + b_{y_i}}}{\sum_{j=1}^{k} e^{d_i^\top f_j + b_j}}$.

The training loss of a single example is the negative-log likelihood and is a function of the classifier parameters F,b, the representation D, and the labels y: $L_i(F, b, D, y) = -\log p_i(y_i)$. The aggregated cross entropy loss is $L(F, b, D, y) = \sum_{i=1}^{n} L_i(F, b, d_i, y_i)$.

The loss function L is a convex function of F, b [3]. F and b do not define the mapping from sampled d_i to probability vectors p_i in a unique way, and there are multiple minimizers for L as the following lemma shows.

Lemma 1. *The minimizers F^*, b^* of L are not unique, and it holds that for any vector $v \in \mathbb{R}^c$ and scalar s, the solutions $F^* + v\mathbf{1}_c^\top$, $b^* + s\mathbf{1}_c$ are also minimizers of L.*

Proof. This proof and all other omitted proofs are provided in the supplementary material. □

In this work, we study the compound loss that is obtained as $\sum_{r=1}^{m} L(F^r, b^r, D, y)$ for m different sets of classifiers F^r, b^r. More specifically, let the set of classifier parameters be $F^1 = \left[f_1^1 \dots f_c^1\right]$,$b^1$, ...$F^m = [f_1^m \dots f_c^m]$,$b^m$, we enforce orthogonality for each class. This is done either in the conventional way: $\forall jrs \;\; f_j^{r\top} f_j^s = 0$, or in the domain of the within class covariance matrix $\forall jrs \;\; f_j^{r\top} S_w f_j^s = 0$. We call the second type of orthogonality "S_w-orthogonality".

The S_w orthogonality is directly related to our goal of improving the number of distinct discriminative directions, as captured by the Fisher ratios. This is explored in Sec. 5. It resembles, other methods that down-regulate the contribution of the directions in the vector space that account for much of the within class covariance, such as WCCN [9].

In practice, this orthogonality is enforced by adding loss terms of the form $\lambda|f_j^{r\top} f_j^s|$ or $\lambda|f_j^{r\top} S_w f_j^s|$. The value of λ used throughout our experiments is 0.005, which is, for comparison, 10 times larger than the weight decay used during training. This value is high enough to ensure solutions that are very close to orthogonality (normalized dot products lower than 10^{-3}) in all of our experiments. Higher weights might hinder an effective exploration of the parameter space during optimization.

For the S_w orthogonality, S_w depends on the representation and is estimated for each train mini-batch separately. In all experiments, a mini-batch of 200 samples was used. While the values of S_w change between mini-batches, we found the estimations to be reliable.

Since multiple copies of the same loss are used, we term our loss "the multiverse loss". The choice of term is further motivated by the property, discussed below, that all copies are different (due to orthogonality) but provide the same probabilistic outcome.

4. Properties of the learned representation

When employing the multiverse loss $\sum_{r=1}^{m} L(F^r, b^r, D, y)$ for training the neural network, under either orthogonality constraint, the learned representation displays a few desirable properties. The first property is that for every two classifiers F^r,b^r and F^s,b^s the parameters are intimately related. The nature of this link depends on the rank of D. For a full rank D, the solutions are highly constrained, which can be seen as a very restrictive form of regularization. This leads to a lower rank representation, where orthogonal solutions are linked by rank-1 modifications.

We will be using the following Lemma in order to prove Thm 1.

Lemma 2. *Let $K = \sum_{i=1}^{n} d_i d_i^\top$ be a full rank $d \times d$ matrix, i.e., it is PD and not just PSD, then for all vector $q \in \mathbb{R}^n$ such that $\forall i \quad q_i > 0$, the matrix $\hat{K} = \sum_{i=1}^{n} q_i d_i d_i^\top$ is also full rank.*

Proof. For every vector $v \in \mathbb{R}^d$, $v^\top \hat{K} v \geq (min_i q_i) v^\top K v > 0$. □

The following theorem links any two optimal solutions in the case in which D is full rank. Note that the orthogonality constraint is not assumed.

Theorem 1. *Assume the minimal loss $L^*(D, y)$ is obtained at two solutions F^1, b^1 and F^2, b^2. If $rank(D) = d$, then there exists some vector $v \in \mathbb{R}^c$ and some scalar s such that $F^1 - F^2 = v\mathbf{1}_c^\top$ and $b^1 - b^2 = s\mathbf{1}_c$.*

Proof. For simplicity we prove the case where $b^1 = b^2 = \mathbf{0}$, the case where $b^1, b^2 \neq \mathbf{0}$ is similar. Let $\Psi = [\psi_1, \psi_2, \ldots, \psi_c] = F^2 - F^1$, and let ψ denote the concatenation of the column vectors ψ_j into a single column vector. Given that F^1, F^2 achieve minimal loss, from convexity it must hold that:

$$\psi^T \nabla^2 L(D, y)\Big|_{F^1} \psi = \psi^T \frac{\partial L(D, y)^2}{\partial F \partial F}\Big|_{F^1} \psi = 0 \qquad (1)$$

where $\nabla^2 L^*(D, y)$ is the hessian of the loss. We will show that in order for ψ to lie in its kernel it must hold that $\psi_1 = \psi_2 \ldots = \psi_c$. Recall that $p_i(j)$is the vector of softmax probabilities associated with d_i.

$$\frac{\partial}{\partial F_{ju}} L(D, y) = -\sum_{i=1}^{n} d_{iu} p_i(j) - \sum_{i, y_i = u} d_{iu} \qquad (2)$$

$$\frac{\partial^2}{\partial F_{ju} F_{j'v}} L(D, y) = -\sum_{i=1}^{n} d_{iu} d_{iv} p_i(j)(\delta_{j=j'}(1 - p_i(j)) - \delta_{j \neq j'} p_i(j')) \qquad (3)$$

Therefore:

$$\psi^\top \frac{\partial^2}{\partial F \partial F} L(D, y)\psi = \sum_{j=1}^{c} \psi_j^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j)(1 - p_j(u))\psi_j - \sum_{j=1}^{c} \sum_{j' \neq j} \psi_j^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j) p_{j'}(v) \psi_{j'} \qquad (4)$$

Since $(1 - p_i(j)) = \sum_{j' \neq j} p_i(j')$, the first term of Eq. 4 can be written as follows:

$$\begin{aligned} &\sum_{j=1}^{c} \psi_j^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j)(1 - p_i(j))\psi_j \\ &= \sum_{j=1}^{c} \sum_{j' \neq j} \psi_j^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j) p_i(j') \psi_j \\ &= \sum_{j=1}^{c} \sum_{j'=j+1}^{c} [\psi_j^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j) p_i(j') \psi_j \\ &\qquad + \psi_{j'}^\top \sum_{i=1}^{n} d_i d_i^\top p_i(j) p_i(j') \psi_{j'}] \end{aligned} \qquad (5)$$

Similar manipulation can be done with the second term of

Eq. 4:

$$-\sum_{j=1}^{c}\sum_{j'\neq j}\psi_j^\top\sum_{i=1}^{n}d_i d_i^\top p_i(j)p_i(j')\psi_{j'} = -\sum_{j=1}^{c}\sum_{j'=j+1}^{c}2\psi_j^\top\sum_{i=1}^{n}d_i d_i^\top p_i(j)p_i(j')\psi_{j'} \quad (6)$$

Adding the two term we get:

$$\psi^\top\frac{\partial^2}{\partial F\partial F}L(D,y)\Big|_{F^1}\psi = \sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\psi_j-\psi_{j'})^\top\sum_{i=1}^{n}d_i d_i^\top p_i(j)p_i(j')(\psi_j-\psi_{j'}) \quad (7)$$

Since $\forall i,j \quad p_i(j) > 0$ and since $rank(D)$ is full, $\sum_{i=1}^{n} d_i d_i^\top p_i(j)p_i(j')$ is PD. Eq 7 is therefor the sum of positive values, and can only vanish if and only if $\psi_j = \psi_{j'}$ for all j, j'. □

In our method, we require that the multiple solutions found F^1,F^2 (possibly more) lead to orthogonal (or S_w-orthogonal) separating hyperplanes for each class. The theorem below shows that unless D is degenerate, this requirement leads to either an increase of the total loss, or to a very specific and limiting type of regularization on F^1. Such a stringent regularization would hinder effective learning. For convenience, we state and prove Thm. 2, 3, 4 for the case of conventional orthogonality. The analog theorems for S_w-orthogonality are stated in the same way, and proven similarly, after applying the transformation $S_w^{\frac{1}{2}}$.

Theorem 2. *Assume that $rank(D) = d$, that $d < c$, and that the minimal loss $L^*(D,y)$ is obtained at a solution F^1, b^1. If there exists a second minimizer F^2, b^2 such that for all $j \in [1...c]$ the orthogonality constraint $f_j^1 \perp f_j^2$ holds, then F^1 admits to a stringent second order constraint.*

The situation described in Theorem 2 is even worse for more than two sets of orthogonal weights on top of the representation D. The solution in the case of m orthogonal sets would be restricted to lie on the intersection of $\binom{m}{2}$ hyperellipses.

The crux of Theorem 2 is the full rank property of D. As the theorems below show, if D has $m-1$ low singular values, we can construct solutions with m orthogonal sets of weights that present loss that is only slightly higher than $mL^*(D,y)$.

Specifically, let $\lambda_1, \lambda_2, ..., \lambda_d$ denote the (all non-negative) eigenvalues of the kernel matrix $K = DD^\top$, ordered from largest to smallest. We can bound the loss based on the last eigenvalues.

Theorem 3. *There exist sets of weights $F^1 = \left[f_1^1, f_2^1, ..., f_c^1\right], b^1, F^2 = \left[f_1^2, f_2^2, ..., f_c^2\right], b^2$ which are orthogonal as follows $\forall j \quad f_j^1 \perp f_j^2$, for which the joint loss:*

$$J(F^1,b^1,F^2,b^2,D,y) = L(F^1,b^1,D,y)+L(F^2,b^2,D,y) \quad (8)$$

is bounded by

$$2L^*(D,y) \leq J(F^1,b^1,F^2,b^2,D,y) \leq 2L^*(D,y)+A\lambda_d \quad (9)$$

where A is a bounded parameter.

Proof. We prove the theorem by constructing such a solution. Let v be the eigenvector of K corresponding to the smallest eigenvalue λ_d. We consider the solution $F^1 = F^*$,$b^1 = b^2 = b^*$, $F^2 = F^1 + v\alpha^\top$, for some vector $\alpha_j = -\frac{||f_j^1||^2}{v^\top f_j^1}$.

From the construction, it is clear that $L(F^1,b^1,D,y) = L^*(D,y)$ and that the orthogonality constraints $(f_j^1 + \alpha_j v)^\top f_j^1 = 0$ hold for all j.

Let $\Psi = [\psi_1, \psi_2, \ldots, \psi_c] = F^2 - F^1$, and let ψ denote the concatination of the column vectors ψ_j into a single column vector. The expansion of $L(F^1+\Psi, b^1)$ into a multivariate taylor series is as follows:

$$L(F^1+\Psi,b^1) = L(F^1,b^1)+(\vec{\nabla}\cdot\psi)L(D,y)\Big|_{F^1,b^1}+R(\psi). \quad (10)$$

Where $R(\psi)$ represents the remainder term, and can be written in the Lagrange form [13] as follows:

$$R(\psi) = \frac{1}{2}(\vec{\nabla}\cdot\psi)^2 L(D,y)\Big|_{\rho,b^1} = \frac{\psi^\top}{2}\frac{\partial^2}{\partial F\partial F}L(D,y)\Big|_{\rho,b^1}\psi \quad (11)$$

where the derivatives are evaluated at some point ρ, b^1 such that $||\rho - F^1||_F \leq ||\Psi - F^1||_F$. The first order terms in Eq. 10 vanishes due to the optimality of F^1, b^1. Therefore:

$$L(F^1+\Psi,b^1) = L^*(D,y)+\frac{1}{2}\psi^\top\frac{\partial^2}{\partial F\partial F}L(D,y)\Big|_{\rho,b^1}\psi \quad (12)$$

Using Eq. 7 we can form a bound on the remainder term that does not depend on ρ:

$$\begin{aligned} &L(F^1+\Psi,b^1) = L^*(D,y) \\ &+\frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\psi_j-\psi_{j'})^\top\sum_{i=1}^{n}d_i d_i^\top p_i(j)p_i(j')(\psi_j-\psi_{j'}) \\ &\leq L^*(D,y)+\frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\psi_j-\psi_{j'})^\top K(\psi_j-\psi_{j'}) \end{aligned} \quad (13)$$

Since $\psi_j = \alpha_j v$ we get:

$$L(F^1+\Psi, B^1) \leq L^*(D,y)+\frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\alpha_j-\alpha_{j'})^2 v^T K v$$
$$= L^*(D,y) + \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\alpha_j-\alpha_{j'})^2\lambda_d \quad (14)$$

Denoting $A = \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\alpha_j-\alpha_{j'})^2$ we have:

$$J(F^1, B^2, F^1, B^2, D, y) \leq L^*(D,y) + A\lambda_d \quad (15)$$

□

Thm. 3 can be generalized to the case of m cross entropy losses as follows.

Theorem 4. *There exist a set of weights* $F^1 = \left[f_1^1, f_2^1, ..., f_C^1\right], b^1, F^2 = \left[f_1^2, f_2^2, ..., f_C^2\right], b^2 ... F^m = \left[f_1^m, f_2^m, ..., f_C^m\right], b^m$ *which are orthogonal* $\forall jrs\ f_j^r \perp f_j^s$ *for which the joint loss:*

$$J(F^1, b^1 ... F^m, b^m, D, y) = \sum_{r=1}^{m} L(F^r, b^r, D, y) \quad (16)$$

is bounded by:

$$mL^*(D,y) \leq J(F^1, b^1 ... F^m, b^m, D, y)$$
$$\leq mL^*(D,y) + \sum_{l=1}^{m-1} A_l \lambda_{d-j+1} \quad (17)$$

where $[A_1 \dots A_{m-1}]$ *are bounded parameters.*

5. Fisher spectrum properties

We next tie the outcome of the multiverse minimization to the Fisher scores used in LDA classification, which served as motivation to our approach. The Fisher spectrum $\gamma_1 \dots \gamma_d$ is obtained by solving the generalized eigenproblem $S_b v = \gamma S_w v$, where S_b and S_w may be approximated by the between class and within class covariance matrices:

$$S_b = \frac{1}{n}\sum_{j=1}^{c} n_j(\mu-\mu_j)(\mu-\mu_j)^\top \quad (18)$$

$$S_w = \frac{1}{n}\sum_{j=1}^{c}\sum_{i\in I_j}(d_i-\mu_j)(d_i-\mu_j)^\top \quad (19)$$

where $\mu = \frac{\sum_{i=1}^{n} d_i}{n}$ is the mean of all data points, and $\mu_j = \frac{\sum_{i\in I_j} d_i}{n_j}$ is the mean of class j. S_b and S_w are the same matrices used in LDA.

The Fisher ratio is defined for any vector v as:

$$\sigma(v, S_b, S_w) = \frac{v^T S_b v}{v^T S_w v} \quad (20)$$

In the JB formulation, an instance of a class member is influenced by two factors, its class identity and interclass variation. Each class member d_i is modeled as the sum of two Gaussian variables: $d_i = \mu_{y_i} + \epsilon$, where μ_{y_i} is the mean of class y_i, and ϵ represents the intraclass variation. The two terms are modeled as multivariate Gaussians $N(\mathbf{0}, S_b)$, $N(\mathbf{0}, S_w)$.

Given the above multivariate Gaussian distribution for d_i, the joint distribution $(d_i, d_{i'})$ is also a zero mean multivariate Gaussian. Let H represent the hypothesis that d_i and $d_{i'}$ belong to the same class, and I represent the hypothesis that they belong to different classes. Under the JB formulation, the covariance matrix of the probability distributions $P(d_i, d_{i'}|H)$ and $P(d_i, d_{i'}|I)$ can be derived:

$$\Sigma_H = \begin{pmatrix} S_b+S_w & S_b \\ S_b & S_b+S_w \end{pmatrix}, \Sigma_I = \begin{pmatrix} S_b+S_w & \mathbf{0} \\ \mathbf{0} & S_b+S_w \end{pmatrix} \quad (21)$$

Let $\hat{d} = ((d_i-\mu)^\top, (d_{i'}-\mu)^\top)^\top$. The log probabilities of the two hypotheses are given, up to a const, by $\hat{d}^\top \Sigma_H^{-1} \hat{d}$ and $\hat{d}^\top \Sigma_I^{-1} \hat{d}$. The following theorem links the Fisher spectrum to the success of the JB method.

Theorem 5. *Given data representation* D*, mean* μ *and labels* y*, for any centered data point* $\hat{d}_i = d_i - \mu$*, we denote* $d_i' = (S_b+S_w)^{-1}\hat{d}_i$*. Given two centered data points* $\hat{d}_1, \hat{d}_2$ *such that the fisher ratios* $\sigma(d_1', S_b, S_w), \sigma(d_2', S_b, S_w) < T$*, it holds that:*

$$1-2T \leq \frac{\log P(d_1, d_2|H) + \eta_1}{\log P(d_1, d_2|I) + \eta_2} \leq 1+6T \quad (22)$$

Where $\eta 1, \eta 2$ *are fixed constants.*

Theorem 5 indicates that in the directions of low Fisher ratio the JB method cannot distinguish between the two competing hypotheses and determine whether the two samples d_i and $d_{i'}$ belong to the same class.

We observed during experiments performed on a number of datasets, that training of a CNN using a single cross entropy loss produces a representation that has a rapidly decreasing Fisher spectrum, and is highly discriminative in only a few directions. Reducing the representation dimension, i.e., using a bottleneck technique helps in reducing the total number of dimensions but does not seem to increase the number of discriminative dimensions. We next show that by optimizing for multiple orthogonal solutions, we promote more directions that have high Fisher scores.

Since the hyperplanes f_i^r learned during optimization are discriminative, we can expect most of these to have high Fisher ratios. The multiplicity created by the multiverse loss, leads to multiple orthogonal hyperplanes. Since the probabilities produced by the matching hyperplanes are identical, it is likely that all matching hyperplanes f_j^r, and f_j^s have similar Fisher ratios. The theorem below shows that adding more S_w-orthogonal classifiers with high Fisher ratios increases the $L1$ norm of the Fisher spectrum.

Theorem 6. *Let $f^1 ... f^m$ be a set of m classifiers that are S_w-orthogonal for data D and labels y, and let $\gamma = [\gamma_1 ... \gamma_d]$ denote the Fisher spectrum. Given that $\forall 1 \leq r \leq m$, for some value θ, $\sigma(f^r, S_b, S_w) \geq \theta$, it holds that $\sum_{k=1}^{d} \gamma_k \geq \sqrt{m}\theta$.*

In Thm. 6 we used the S_w orthogonality of the solutions to guarantee the result, however it is not a necessary condition. From our experiment we noticed an improved Fisher spectrum when both S_w and the standard orthogonality condition were used.

6. Experiments

In order to evaluate the effect of using the multiverse loss on performance, we have conducted experiments on two widely used datasets: CIFAR-100 and LFW. While the CIFAR-100 experiments are performed using a new transfer learning protocol, the LFW experiments provide a direct empirical comparison to a large body of previous work.

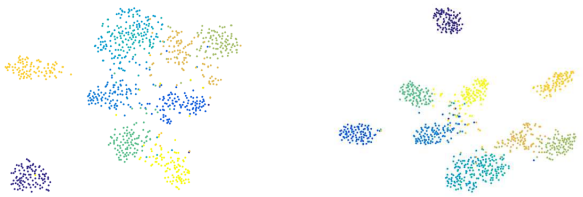
6.1. Network architecture

In our experiments, we employ three network architectures. For the CIFAR-100 experiments, we use the architecture of network in network [18]; for the face recognition experiments, we use an architecture similar to the scratch architecture [41] for most of our experiments (Denoted by $N1$). We also use a higher capacity network similar to [24] for further evaluation (Denoted by $N2$). The networks were trained from scratch at each experiment, using the MatConvNet framework [37].

All networks are fully convolutional, and we added a hidden layer on top of the $N1$ network to apply our method on top of a vector of activations. This modification is not strictly needed and was made for implementation convenience. This top layer was used as the representation. The architectures used are fully described in the supplementary material.

6.2. Results

The CIFAR-100 [14] contains 50,000 32×32 color images, split between 100 categories. The images were extracted from the tiny image collection [35]. Throughout our experiments, the first 90 classes (class ids 0 to 89) are used as the source domain, and the last 10 as the target domain.



(a) (b)

Figure 1. 2D embedding (TSNE [36]) of the representation of (a) conventional cross entropy and (b) multiverse (M5). The 10 target classes of the CIFAR-100 experiment are shown.

Our experiments compare six architectures: a baseline with one cross entropy loss ("M1"); four multiverse architectures with 2–5 such losses ("M2–M5"); and an ensemble of five networks with a single cross entropy loss each. The last method was added to demonstrate that our method's benefit is greater than that of combining multiple networks. Note, however, that when compounding losses, the overall network architecture resembles that of a single network and is almost as efficient to train and deploy as the baseline network. One can easily create ensembles of networks with multiverse losses, as we do for the LFW benchmark.

We report the methods' performance in multiple ways. The validation error reports the error rate obtained, in the source domain of 90 classes, on the 10% of the data reserved for this purpose. In the target domain, two metrics are used: same/not-same accuracy using either the cosine distance or the JB method. Note that the cosine distance is unsupervised, and that we train the JB on the validation set of the source domain. Hence, no training was done in the target domain. For the same/not-same evaluation, 3000 matching and 3000 non-matching pairs were randomly sampled from the 10 classes of the target domain.

As can be seen in Tab. 2, the multiverse method outperform the baseline and the ensemble methods on the target domain, in each of the accuracy metrics. It is also evident that adding more cross entropy losses improves performance. The preferable separation between the classes is also depicted visually in Fig. 1, where the 2D embedding of the baseline (M1) representation is compared to that obtained using the M5 multiverse method. For the purpose of this visualization, the TSNE [36] embedding method is used.

As mentioned above, for the face recognition experiments, we use the scratch model [41]. The networks are trained on the CASIA dataset [41]; LFW dataset [11] is used as the target domain.

Models are evaluated in the source domain by measuring the classification accuracy on the CASIA dataset, which we split to 90% training and 10% validation. For the target domain, the LFW benchmark in the unrestricted mode [10] is used (we do not use person ID from LFW, but do use the

Domain	Source	Target (transfer)	
Metric	Val error	Cosine	JB
M1	0.340	0.789	0.800
M2	0.340	0.791	0.804
M2 (S_w-orthogonal)	0.344	0.798	0.803
M3	0.345	0.801	0.812
M3 (S_w-orthogonal)	0.346	0.799	0.811
M4	0.351	0.807	0.82
M4 (S_w-orthogonal)	0.353	0.808	0.823
M5	0.360	0.812	0.833
M5 (S_w-orthogonal)	0.362	0.811	0.831
M6	0.369	0.816	0.838
M6 (S_w-orthogonal)	0.371	0.816	0.834
M7	0.375	0.815	0.831
M7 (S_w-orthogonal)	0.377	0.816	0.830
Ensemble of 5 times M1	NA	0.803	0.82

Table 2. CIFAR-100 Results. Multiverse networks of multiplicity 1–7 are shown, for both types of orthogonality. Also shown is the result obtained by an ensemble of 5 conventional networks. The numbers indicate either the validation error or the same/not-same accuracy in the target domain.

IDs of the CASIA dataset). The LFW results are mean and Standard Error estimated over the fixed ten cross-validation splits. JB is either trained on the CASIA validation split or on the LFW dataset itself in a cross validation manner.

In the LFW experiments, we performed the M1 (baseline), M3, and M5 experiments multiple times, in order to show the stability of the results and to support ensembles. The S_w-orthogonality multiverse method, which is slower to train, was not tested on LFW by the submission date. As can be seen in Tab. 3, the multiverse loss outperforms, in the target domain, the baseline method and also outperforms the ensemble of multiple baseline networks. This is true for the cosine similarity, as well as for the two JB experiments. Interestingly, multiverse does not show an advantage in the source domain (this does not weaken our claims).

In face recognition, the effect of the training dataset sometimes overshadows that of the method. We, therefore, employed a proprietary 800k images 3rd party dataset, which does not intersect the identities of the LFW dataset. In comparison to CASIA's 500k images, the 3rd party dataset is slightly larger and contains fewer tagging mistakes. As can be seen in Tab. 3, this leads to an improvement in performance.

The results we obtained are compared in Tab. 4 to the state of the art as reported on the LFW webpage on the date of the submission. Our results, which use a fairly simple fully convolutional architecture, achieves the highest ranking for a single network outperforming all results, except one result [23], which was obtained using 200 million images.

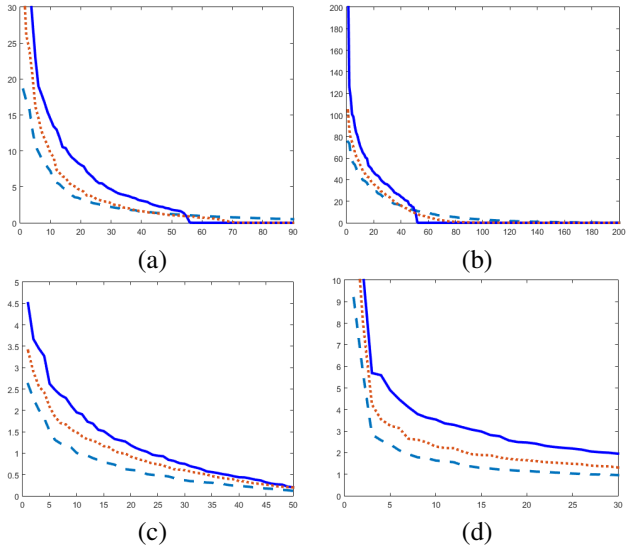


Figure 2. Various spectrums obtained in the target domain. Each plot shows singular values (first row) or generalized eigenvalues (second row), sorted separately for each of three methods. Solid blue is the result obtained for M5. The dotted red line is the M3 result, and the baseline M1 is shown as dashed green. (a-b) For CIFAR-100 employing conventional orthogonality, CIFAR-100 S_w orthogonality, and LFW, respectively, the singular values of the kernel matrix $K = DD^\top$ are shown. The multiverse loss leads to higher values until it drops to zero earlier than the conventional spectrum. (c-d) The Fisher spectrum, i.e., the generalized eigenvalues of S_w and S_b for the same datasets:(c) CIFAR conventional orthogonality and (d) LFW. As a result of applying the multiverse method, there is an increase in the magnitude of the eigenvalues..

In addition to performance, we also examined the effect of the multiverse loss on the properties of the representation. Fig. 2 demonstrate the singular values of the data representation in the transfer domain on (a) CIFAR-100 using conventional orthogonality and (b) LFW. As can be seen, the multiverse network (M5) has larger singular values. However, these drop to zero abruptly whereas the spectrum of the baseline representation continues to decay gradually. As a result, the representation of our method is of a lower dimension, and is more balanced among the dimensions. Fig. 2 (c) and (d) show the generalized eigenvalues of S_b and S_w in the target domain. As can be seen, the multiverse method promotes larger Fisher ratios.

The sharp drop in the data dimensionality that is promoted by the multiverse method leads to very compact representations. The dimensionality of our best single network (M5, 3rd party dataset), is only 51 (Fig. 4(b)). This is a very compact representation, which is much lower than any other state of the art network.

Domain		Source	Target (transfer)		
Metric	Network	Val error	Cosine	JB on source	JB on LFW splits
CASIA trained M1	$N1$	0.07	0.962 ± 0.0032	0.966 ± 0.0022	0.970 ± 0.0016
CASIA trained M1 (2)	$N1$	0.07	0.962 ± 0.0021	0.966 ± 0.0019	0.971 ± 0.0022
CASIA trained M1 (3)	$N1$	0.07	0.961 ± 0.0022	0.966 ± 0.0013	0.971 ± 0.0015
Ensemble of 3 CASIA M1	$N1$		0.968 ± 0.0019	0.972 ± 0.0021	0.975 ± 0.0025
CASIA trained M2	$N1$	0.08	0.970 ± 0.0021	0.974 ± 0.0017	0.976 ± 0.0016
CASIA trained M3	$N1$	0.11	0.972 ± 0.0012	0.977 ± 0.0015	0.980 ± 0.0034
CASIA trained M3 (2)	$N1$	0.11	0.971 ± 0.0031	0.977 ± 0.0028	0.979 ± 0.0027
CASIA trained M5 (1)	$N1$	0.12	0.973 ± 0.0011	0.978 ± 0.0014	0.981 ± 0.0019
CASIA trained M5 (2)	$N1$	0.12	0.972 ± 0.0015	0.977 ± 0.0019	0.980 ± 0.0031
3rd party DB, M5	$N2$	0.12	0.982 ± 0.0034	0.986 ± 0.0031	0.988 ± 0.0035

Table 3. Face recognition results. Shown are the validation error on CASIA, and transfer results on LFW. The cosine similarity as well as learned JB similarities are shown. The JB was either trained on CASIA or on the LFW training splits. The LFW results confirm with the unrestricted mode and report mean and Standard Error of the accuracy obtained for the ten cross validation splits. Network architecture N1 is the scratch architecture [41]; N2: a VGG style network [24].

Method	Single network	Ensemble result	#nets	Training dataset
M5	0.9814 ± 0.0019	–		CASIA [41]
M5, 3rd party DB	0.9883 ± 0.0035	0.9905 + 0.0027	2	proprietary 800k images
DeepFace [32]	0.9700 ± 0.0087	0.9735 ± 0.0025	7	proprietary, 4M images
DeepID [28]	–	0.9745 ± 0.0026	25	proprietary,160k
Original scratch [41]	0.9773 ± 0.0031	–	1	CASIA [41]
Web-Scale Training [33]	0.9800	0.9843	4	proprietary, 500M images
MSU TR [38]	0.9745 ± 0.0099	0.9823 ± 0.0068	7	CASIA [41]
MMDFR [5]	0.9843 ± 0.0020	0.9902 ± 0.0019	8	CASIA [41]
DeepID2 [25]	0.9633	0.9915 ± 0.0013	25	proprietary,160k
DeepID2+ [29]	0.9870	0.9947 ± 0.0012	25	proprietary,290k
FaceNet [23]	0.9967 ± 0.0015	0.9963 ± 0.0009	8	proprietary, 200M
FR+FCN [43](*)	–	0.9645 ± 0.0025	5	CelebFaces [27], 88k
betaface.com(*)	–	0.9808 ± 0.0016	NA	NA
Uni-Ubi(*)	–	0.9900 ± 0.0032	NA	NA
Face++ [42](*)	–	0.9950 ± 0.0036	4	proprietary, 5M face images
DeepID3 [26](*)	–	0.9953 ± 0.0010	25	proprietary,300k
Tencent-BestImage(*)	–	0.9965 ± 0.0025	20	proprietary, 1M face images
Baidu [19](*)	–	0.9977 ± 0.0006	10	proprietary, 1.2M face images
AuthenMetric(*)	–	0.9977 ± 0.0009	25	proprietary, 500k face images

Table 4. Comparison to state of the art results on LFW. We present the best result for a single network, with the exception of FaceNet, which was trained on a dataset which is a hundred times larger than ours. A star (*) indicates commercial systems whose claimed results were not peer reviewed.

7. Conclusions

This work presented the emergence of surprising and desirable properties of the representation layer of a deep neural network when learning multiple orthogonal solutions. The practical implications of our work are far reaching since the suggested method is easy to incorporate into almost any architecture.

Acknowledgments

This research is supported by a 2015 IBM Faculty Award. We are thankful for RealFace for providing the 3rd party DB mentioned in Sec. 6.

References

[1] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. 2

[2] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conf. Computer Vision*, 2012. 1

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[5] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *CoRR*, abs/1509.00244, 2015. 8

[6] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611, 2006. 1

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 1

[8] R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 2

[9] A. O. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for svm-based speaker recognition. In *Proc. of ICSLP*, page 14711474, 2006. 3

[10] G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. UM-CS-2014-003, 2014. 6

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1, 6

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[13] M. Kline. *Calculus: an intuitive and physical approach*. Courier Corporation, 1998. 4

[14] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Master's thesis, 2009. 6

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[16] I. Kuzborskij, F. Orabona, and B. Caputo. From n to n+ 1: Multiclass transfer incremental learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3358–3365. IEEE, 2013. 1

[17] N. Levy and L. Wolf. Minimal correlation classification. In *Computer Vision–ECCV 2012*, pages 29–42. Springer, 2012. 2

[18] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations (ICLR)*, 2013. 6

[19] J. Liu, Y. Deng, T. Bai, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *CoRR*, abs/1506.07310, 2015. 8

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR (to appear)*, Nov. 2015. 2

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013. 1

[22] F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla, and G. Sandini. Model adaptation with least-squares SVM for adaptive hand prosthetics. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 2897–2903. IEEE, 2009. 1

[23] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. June 2015. 1, 2, 7, 8

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 2, 6, 8

[25] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc., 2014. 2, 8

[26] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. 2, 8

[27] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496, Dec 2013. 8

[28] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2, 8

[29] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 8

[30] 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 2015*, 2015. 1

[31] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *CoRR*, abs/1412.1441, 2014. 2

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1701–1708, Washington, DC, USA, 2014. IEEE Computer Society. 2, 8

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 8

[34] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model

knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3081–3088. IEEE, 2010. 1

[35] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970, Nov 2008. 6

[36] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

[37] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015. 6

[38] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *CoRR*, abs/1507.07242, 2015. 8

[39] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197. ACM, 2007. 1

[40] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. June 2015. 2

[41] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 2, 6, 8

[42] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *CoRR*, abs/1501.04690, 2015. 8

[43] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *CoRR*, abs/1404.3543, 2014. 8