MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Discriminative Subspace Pooling for Action Recognition

Wang, J.; Cherian, A.

TR2018-141 September 26, 2018

Abstract

Adversarial perturbations are noise-like patterns that can subtly change the data, while failing an otherwise accurate classifier. In this paper, we propose to use such perturbations for improving the robustness of video representations. To this end, given a well-trained deep-model for per-frame video recognition, we first generate adversarial noise adapted to this model. Using the original data features from the full video sequence and their perturbed counterparts, as two separate bags, we develop a binary classification problem that learns a set of discriminative hyperplanes - as a subspace - that will separate the two bags from each other. This subspace is then used as a descriptor for the video, dubbed discriminative subspace pooling. As the perturbed features belong to data classes that are likely to be confused with the original features, the discriminative subspace will characterize parts of the feature space that are more representative of the original data, and thus may provide robust video representations. To learn such descriptors, we formulate a subspace learning objective on the Stiefel manifold and resort to Riemannian optimization methods for solving it efficiently. We provide experiments on several video datasets and demonstrate state-of-the-art results.

Workshop on Perceptual Organization in Computer Vision as part of the European Conference on Computer Vision (ECCV)

Discriminative Subspace Pooling for Action Recognition

Jue Wang[1⋆] and Anoop Cherian[2]

[1]Data61/CSIRO, ANU, Canberra [2]MERL Cambridge, MA
jue.wang@anu.edu.au cherian@merl.com

Abstract. Adversarial perturbations are noise-like patterns that can subtly change the data, while failing an otherwise accurate classifier. In this paper, we propose to use such perturbations for improving the robustness of video representations. To this end, given a well-trained deep model for per-frame video recognition, we first generate adversarial noise adapted to this model. Using the original data features from the full video sequence and their perturbed counterparts, as two separate bags, we develop a binary classification problem that learns a set of discriminative hyperplanes – as a subspace – that will separate the two bags from each other. This subspace is then used as a descriptor for the video, dubbed *discriminative subspace pooling*. As the perturbed features belong to data classes that are likely to be confused with the original features, the discriminative subspace will characterize parts of the feature space that are more representative of the original data, and thus may provide robust video representations. To learn such descriptors, we formulate a subspace learning objective on the Stiefel manifold and resort to Riemannian optimization methods for solving it efficiently. We provide experiments on several video datasets and demonstrate state-of-the-art results.

1 Introduction

Deep learning has enabled significant advancements in several areas of computer vision; however, the sub-area of video-based recognition continues to be elusive. In this paper, we present a novel pooling framework for learning video representations for action recognition. A robust video representation is one that can avoid the action classifier from using features (or feature dimensions) that are sensitive to data perturbations. One way to learn such representations is to explicitly find out which features are vulnerable and avoid them. The recent advances in adversarial learning has allowed us to generate such perturbations, one popular model is the universal adversarial perturbations proposed in [9, 8]. While, these prior works have used these noise-like image patterns for fooling a well-trained classifier, we propose to use such patterns on videos, and for finding data parts that are sensitive to mis-classifications; and use these patterns as a means to robust representation learning.

Assuming we have adversarial patterns generated, we make two bags, one consisting of the original video features, while the other one consisting of features

⋆ Work done while interning at MERL.

perturbed by noise. Next, we learn a discriminative hyperplane that separates the bags in a max-margin framework. Such a hyperplane, which in our case is produced by a primal support vector machine (SVM), finds decision boundaries that could well-separate the bags; the resulting hyperplane could be a vector which is a weighted combination of the data points (support vectors) in the bags. Given that the data features are non-linear, and given that a kernelized SVM might not scale well with sequence lengths, we propose to instead use multiple hyperplanes for the classification task, by stacking several such hyperplanes into a column matrix. We propose to use this matrix as our data representation for the video sequence. For generalizability of our representation, we assume the hyperplanes are columns of an orthogonal frame (a tall matrix with orthogonal columns), and propose a non-linear Riemannian optimization on the Stiefel manifold (which is the mathematical manifold of such orthogonal frames) for representation learning. Our overall pipeline is illustrated in Figure 1.

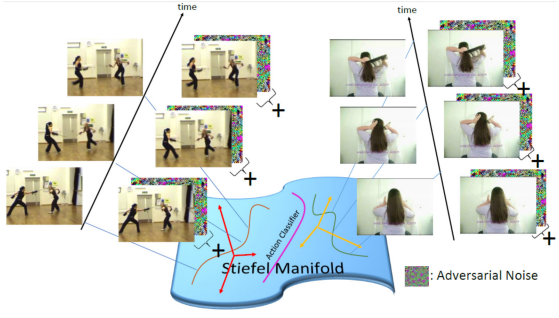


Fig. 1. An illustration of our discriminative subspace pooling with adversarial noise.

2 Related work

With the success of deep learning methods, feeding video data as RGB frames, optical flow subsequences, RGB differences, or 3D skeleton data directly into CNNs is preferred. One successful such approach is the two-stream model (and its variants) [11, 6, 5] that use video segments (of a few frames) to train deep models, the predictions from the segments are fused via average pooling to generate a video level prediction. The above architectures are usually trained for improving the classification accuracy, however, do not consider the robustness of their internal representations – accounting for which may improve their generalizability to unseen test data. To this end, we explore the vulnerable factors in a model (via generating adversarial perturbations [8]), and learn representations that are resilient to such factors in a network-agnostic manner. Our main inspiration comes from the recent work of Moosavi et al. [8] that show the existence of quasi-imperceptible image perturbations that can fool a well-trained CNN model. They provide a systematic procedure to learn such perturbations in an image-agnostic way. In Xie et al. [14], such perturbations are used to improve

the robustness of an object detection system. Similar ideas have been explored in [9]. While these schemes share similar motivation as ours, the problem setup and formulations are entirely different. Our contribution is inspired by the recent work of Wang et al [13] that proposes using decision boundaries of a support vector machine classifier that separates data features from independently sampled noise. In this paper, we argue that using data dependent adversarial noise is significantly more powerful in learning useful representations.

3 Proposed Method

In this section, we detail our main approach. Let us assume $X = \langle x_1, x_2, ..., x_n \rangle$ be a sequence of video features, where $x_i \in \mathbb{R}^d$ represents the feature from the i-th frame. The feature representation x_i could be the outputs from intermediate layers of a CNN. As alluded to in the introduction, our key idea is the following. We look forward to an effective representation of X that is (i) compact, (ii) preserves characteristics that are beneficial for the downstream task (such as video dynamics), and (iii) efficient to compute. Recent methods such as generalized rank pooling [3] have similar motivations and propose a formulation that learns compact temporal descriptors that are closer to the original data in ℓ_2 norm. However, such a reconstructive objective may also capture noise, thus leading to sub-optimal performance. Instead, we take a different approach. Specifically, we assume to have access to some noise features $Z = \{z_1, z_2, ..., z_m\}$, each $z_i \in \mathbb{R}^d$. Let us call X the positive bag, with a label $y = +1$ and Z the negative bag with label $y = -1$. Our main goal is to find a discriminative hyperplane that separates the two bags; these hyperplanes can then be used as the representation for the bags. A problem in this context is how much coverage does the hyperplanes have to represent a majority of the data points in the positive bag (on their robust dimensions). We could achieve this by expecting the classification accuracy of this binary problem to be very high, which happens when overfitting the hyperplanes to a majority of the data points (assuming the features are non-linear and the hyperplanes are linear, that is limited representational capacity). There are two important problems in this context: (i) how to find the noise patterns, and (ii) finding a good robust representation, which are addressed below.

3.1 Finding Noise Patterns

As alluded to above, having good noise distributions that help us identify the vulnerable parts of the feature space is important for our scheme to perform well. To this end, we resort to the recent idea of universal adversarial perturbations (UAP) [8]. This scheme is dataset-agnostic and provides a systematic and mathematically grounded formulation for generating adversarial noise that when added to the original features is highly-likely to mis-classify a pre-trained classifier. Precisely, suppose $\mathcal{X}$ denotes our dataset, let h be a CNN trained on $\mathcal{X}$ such that $h(x)$ for $x \in \mathcal{X}$ is a class label predicted by h. Universal perturbations are noise vectors ϵ found by solving the following objective:

$$\min_{\epsilon} \|\epsilon\| \text{ s.t. } h(x + \epsilon) \neq h(x), \forall x \in \mathcal{X}, \tag{1}$$

where $\|\epsilon\|$ is a suitable normalization on ϵ such that its magnitude remains small, and thus will not change x significantly. In [8], it is argued that this

norm-bound restricts the optimization problem in (1) to look for the minimal perturbation ϵ that will move the data points towards the class boundaries; i.e., selecting features that are most vulnerable – which is precisely the type of noise we need in our representation learning framework.

3.2 Discriminative Subspace Pooling

Once a "challenging" noise distribution is chosen, the next step is to use a subspace of discriminative directions (as against a single one as in [13]) for separating the two bags such that frame-level feature x_i in the video is classified by at least one of the hyperplanes to the correct class label. Such a scheme can be looked upon as an approximation to a non-linear decision boundary by a set of linear ones, each one separating portions of the data. Mathematically, suppose $W \in \mathbb{R}^{d\times p}$ is a matrix with each hyperplane as its columns, then we seek to optimize:

$$\min_{W,\xi} \Omega(W) + \sum_{\theta\in X\cup Z} \left[\max\left(0, 1 - \max\left(\mathbf{y}(\theta) \odot \mathbf{W}^\top\theta\right) - \xi_\theta\right) + C\xi_\theta\right], \quad (2)$$

where $\mathbf{y}$ is a vector with the label y repeated p times along its rows. The quantity Ω is a suitable regularization for discriminative subspace W, of which one possibility is to use $\Omega(W) = W^\top W = \mathbf{I}_p$, in which case W spans a p dimensional subspace of $\mathbb{R}^d$. The operator $\odot$ is the element-wise multiplication and the quantity $\max(\mathbf{y}(\theta) \odot \mathbf{W}^\top\theta)$ captures the maximum value of the element-wise multiplication, signifying that if at least one hyperplane classifies the input feature θ correctly, then the hinge-loss will be zero.

There is a further consideration to make given that we are working with videos, and that the features that we use are temporally-ordered. To include this criteria in our representation learning, we include additional ordering constraints on the matrix W and introduce our complete **order-constrained discriminative subspace pooling optimization** as:

$$\min_{\substack{W^\top W=\mathbf{I}_p,\\ \xi,\zeta\geq 0}} \sum_{\theta\in X\cup Z}\left[\max\left(0, 1 - \max\left(\mathbf{y}(\theta) \odot \mathbf{W}^\top\theta\right) - \xi_\theta\right)\right] + C_1\sum_{\theta\in X\cup Z}\xi_\theta + C_2\sum_{i<j}\zeta_{ij}, \quad (3)$$

$$\left\|W^\top x_i\right\|^2 + 1 \leq \left\|W^\top x_j\right\|^2 + \zeta_{ij}, \quad i < j, \forall (i,j) \in \mathcal{T} \quad (4)$$

where (4) captures the temporal order, while $\mathcal{T}$ is the set of all frames in a contiguous action cycle (i.e., we include only one action cycle in case of a repeated action, such as *clapping*).

As is clear, the above optimization problem is non-linear and looks cumbersome. However, the orthogonality constraint on W puts it on the so-called Stiefel manifold, which is a non-linear manifold, but having efficient optimization schemes available that could help solve this objective efficiently. We use one popular such optimization scheme, dubbed Riemannian conjugate gradient descent (RCG) that takes gradients in conjugate directions on the tangent spaces of this manifold [3, 1]. The important component to derive the Riemannian gradient for

this descent is to compute the derivatives of the objective in (3), which is:

$$\min_{W\in\mathcal{S}(d,p)} g(W) := \sum_{\theta\in X\cup Z} \left[\max\left(0, 1-\max\left(\mathbf{y}(\theta)\odot \mathbf{W}^\top\theta\right)-\xi_\theta\right)\right] + \frac{1}{n(n-1)}\sum_{i<j}\max(0, 1+\left\|W^\top x_i\right\|^2-\left\|W^\top x_j\right\|^2-\zeta_{ij}), \quad (5)$$

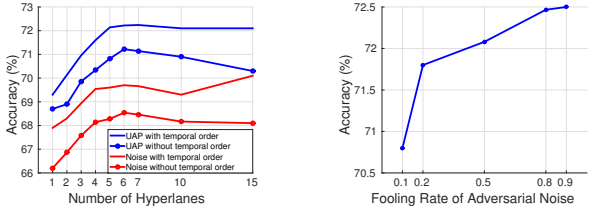
$$\frac{\partial g}{\partial W} = \sum_{\theta\in X\cup Z} A(W;\theta,y(\theta)) + \frac{1}{n(n-1)}\sum_{i<j} B(W;x_i,x_j), \text{where} \quad (6)$$

$$A(W;\theta,y(\theta)) = \begin{cases} 0, & \text{if } \max(y(\theta)\odot W^\top\theta-\xi_\theta)\geq 1 \\ -\left[\mathbf{0}_{d\times r-1}\ y(\theta)\theta\ \mathbf{0}_{d\times p-r}\right], & r=\arg\max_q y(\theta)\odot W_q^\top\theta, \text{ else} \end{cases} \quad (7)$$

$$B(W;x_i,x_j) = \begin{cases} 0, & \text{if } \left\|W^\top x_j\right\|^2 \geq 1+\left\|W^\top x_i\right\|^2-\zeta_{ij} \\ 2(x_i x_i^\top - x_j x_j^\top)W, & \text{else.} \end{cases} \quad (8)$$

In the definition of $A(W)$, we use $W_q^\top$ to denote the q-th column of W. To reduce clutter in the derivations, we have avoided including the terms using $\mathcal{T}$. Assuming the matrices of the form xx^T can be computed offline, on careful scrutiny we see that the cost of gradient computations on each data pair is only $O(d^2p)$ for $B(W)$ and $O(dp)$ for the discriminative part $A(W)$. If we include temporal segmentation with k segments, the complexity for $B(W)$ is $O(d^2p/k)$. Once the directions W are computed, we use an action classifier using a multi-layer neural network or a non-linear SVM using an exponential projection metric kernel. Note that our scheme can be **learned end-to-end in a CNN**, however this will need some knowledge on the implicit function theorem and bi-level optimization [7], which we skip in this paper due to the lack of space.

4 Experiments and Conclusions



(a) (b)

Fig. 2. Analysis of the hyper parameters. All experiments use ResNet-152 features on HMDB-51 split-1 with a fooling rate of 0.8 in (a) and 6 hyperplanes in (b).

We demonstrate the performance of our discriminative subspace pooling (DSP) on three benchmarks: HMDB-51 using two-stream ResNet-152 features from [11], and NTU-RGBD for 3D skeleton based action recognition by using temporal residual network from [10], and (iii) YUP++ dynamic video texture understanding using an Inception-ResNet-v2 model. In Figure 2(a) and 2(b), we analyze the influence of DSP hyperparameters, (i) the number of hyperplanes,

(ii) the importance of the temporal order, (iii) and the fooling rate for UAP. It is clear that: 1. UAP shows significant benefit compared with random noise; 2. temporal ranking constraint will help the discriminative subspace capture the video dynamics. 3. 6 subspaces and fooling rate of 0.8 could result in better performance. Finally, we achieve the state-of-the-art performance across three datasets in the Table 1, including in comparisons on the recent Inflated-3D models (pre-trained on the large Kinetics dataset) [2].

HMDB-51	
Method	Accuracy
TS I3D [2]	80.9%
ST-ResNet [4]	66.4%
ST-ResNet+IDT [4]	70.3%
STM Network [5]	68.9%
STM Network+IDT [5]	72.2%
GRP [3]	70.9%
SVMP [13]	71.0%
Ours(TS ResNet)	**72.4%**
Ours(TS ResNet+IDT)	**74.3%**
Ours(TS I3D)	**81.5%**

NTU-RGBD		
Method	Cross-Subject	Cross-View
SVMP [13]	78.5%	86.4%
GRP [3]	76.0%	85.1%
Res-TCN [12]	74.3%	83.1%
Ours	**81.6%**	**88.7%**
YUP++		
Method	Stationary	Moving
TRN [6]	92.4%	81.5%
SVMP [13]	92.5%	83.1%
GRP [3]	92.9%	83.6%
Ours	**95.1%**	**88.3%**

Table 1. Comparisons to the state-of-the-art on each dataset following their respective official evaluation protocols. 'TS' refers to 'Two-Stream'.

Conclusions: To conclude, in this paper we investigated the problem of representation learning for video sequences by using adversarial perturbations, which affect vulnerable parts of the features. We propose a discriminative classifier, in a max-margin setup, via learning a set of hyperplanes as a subspace, that could separate our synthetic noise from data, which demonstrates state-of-the-art performance on the benchmarks.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press (2009)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (July 2017)
3. Cherian, A., Fernando, B., Harandi, M., Gould, S.: Generalized rank pooling for activity recognition. In: CVPR (2017)
4. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: NIPS (2016)
5. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: CVPR (2017)
6. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Temporal residual networks for dynamic scene recognition. In: CVPR (2017)
7. Gould, Stephen et al., j.y.: On differentiating parameterized argmin and argmax problems with application to bi-level optimization
8. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations (2017)
9. Oh, S.J., Fritz, M., Schiele, B.: Adversarial image perturbation for privacy protection–a game theory perspective. In: ICCV (2017)
10. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+ D: A large scale dataset for 3d human activity analysis. In: CVPR (2016)

11. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
12. Soo Kim, T., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: CVPR Workshops (2017)
13. Wang, J., Cherian, A., Porikli, F., Gould, S.: Video representation learning using discriminative pooling. In: CVPR (2018)
14. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)