Do Textual Descriptions Help Action Recognition?

Matteo Bruni, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo
Media Integration and Communication Center, Università degli Studi di Firenze
{name.surname}@unifi.it

ABSTRACT

We present a novel method to improve action recognition by leveraging a set of captioned videos. By learning linear projections to map videos and text onto a common space, our approach shows that improved results on unseen videos can be obtained. We also propose a novel structure preserving loss that further ameliorates the quality of the projections. We tested our method on the challenging, realistic, Hollywood2 action recognition dataset where a considerable gain in performance is obtained. We show that the gain is proportional to the number of training samples used to learn the projections.

CCS Concepts

•Information systems → Video search; •Computing methodologies → Activity recognition and understanding;

Keywords

Action recognition, Multi-modal learning

1. INTRODUCTION

Imagine listening to someone describing a movie clip. When watching the actual footage afterwards, there will be many divergences from what one may have pictured. Many of the missing visual details from the description will likely appear different in pictures, having been filled in by one's brain. That is because text and videos are very much different domains.

A sentence describing a video clip will inevitably not illustrate every detail of a scene, usually focusing on the relevant semantic entities. A video, instead, carries plenty of information. In particular, the semantic of a video is still hardly represented through visual features due to the well known *semantic gap* [20]. Features are designed to capture appearance and motion, with invariance to lighting, camera movements and viewpoint. They are not usually geared to represent meaning in the form of objects and actions, unless some learning is involved. Consequently, both representations have natural limitation on the information they may deliver. Yet two video sequences with similar descriptions should share semantic properties.

We believe that the semantics of a text associated with a video can be exploited to improve the visual features. This approach has been successfully applied to image annotation [1]. Textual descriptions of videos have some interesting properties. First they contain a high level representation that is hard to obtain from the bare visual data. Second they have a strong sequential structure that in some cases, visual feature representation struggle to maintain.

In this paper we try to give an answer to the question: "Do textual descriptions help action recognition?" To satisfy this curiosity we devise a learning scheme to incorporate the textual information into the video representation. We require this process to be done once in advance, so that textual descriptions are not needed when attempting to predict actions in a new video.

Our learning scheme has a first step, creating a connection between visual and textual features, in the form of a feature transformation. Finally this transformation is applied to visual feature vectors on which we want to learn human action classifiers. When testing new videos only the transformation and the classifier are needed.

The main contributions of this paper are:

- A method exploiting paired visual and textual data to learn a common space, in which text semantic improves visual features, leading to better action recognition performance.
- A novel structure preserving loss, that is able to avoid excessive distortion of the relationships in the initial feature spaces.

To the best of our knowledge, this is the first work that incorporates textual information into the video representation for the task of action recognition. Experiments on the challenging Hollywood2 dataset show several benefits of our approach, obtaining state of the art performance. Even more, we show that performance increases with the amount of training data and that such benefit is observable even with relatively few samples.

2. PREVIOUS WORK

There are few contributions on cross-modal representation for video and there is no previous work on action recognition exploiting joint multi-modal representations. For this

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967301

reason we briefly review works pertaining cross-modal space learning and action recognition and highlight the few intersections in these two lines of research.

2.1 Multimodal Joint Spaces

There is little prior art on video-text embedding. Das *et al.* use latent topics modeling and a tripartite template graph to map visual features to words and finally to generate a textual description [2]. Xu *et al.* jointly model language and visual space to improve retrieval [29]. Zhu *et al.* [31] proposed to use a context-aware CNN to combine video and text for the task of aligning books and movies.

Additionally, embedding visual and textual features into a common space has been done for other tasks.

For the task of image annotation, Ballan *et al.* [1] proposed to use Kernel Canonical Correlation Analysis to fuse labels with image features into a semantic space. Considering that the top k ranked list of annotations are usually transferred when annotating a novel image, in [28] a novel embedding built with a rank loss is proposed. They show that optimizing the precision at k in a joint word-image embedding is useful. A deep visual semantic embedding model is proposed in [4] to train visual models exploiting both labeled image data and semantic information gathered from unannotated text. Socher *et al.* [21] proposed to employ a neural network to map the training images to their respective word vectors, in order to perform zero-shot learning on novel image categories.

For the task of captioning, in [22] the authors proposed to use Recurrent Neural Networks (RNN) on dependency trees, together with a learned multimodal representation for describing images with sentences. In [8], another approach based on RNN for describing images is proposed. They perform alignment of textual descriptions with an embedding that is learned on training pairs of images and descriptions.

Differently from all these works, we address the task of action recognition in video. We propose a novel loss to learn a structure preserving common space in which action classification is easier, thanks to the incorporation of knowledge from the textual space.

2.2 Action Recognition

Action recognition has received significant attention in the past. The majority of the works follow a standard pipeline where features are first extracted from the video and then used as inputs to a classifier.

Since videos provide a huge amount of information through motion, many works focused on the development of features such as the improved Dense Trajectories (iDT) [26] and the recent Convolutional Neural Networks (CNN) [12, 23].

Improved Dense Trajectories exploit optical flow to perform feature tracking and extract consistent local descriptors. Optical flow is estimated compensating camera movements, registering subsequent frames with a transformation [26]. They are considered state-of-the-art of video handcrafted features.

After local feature extraction, generic action recognition methods typically apply a pooling strategy to obtain a global feature of a video. As in the image domain, Fisher Vectors [17] obtain the best performance [26].

In addition to handcrafted features, CNNs have been recently found to be very good at learning meaningful features [3, 23]. In [23], 3D convolutions are used to extend an AlexNet CNN [12] to videos, with the aim of obtaining learned features from its activations. They show performance comparable to iDT features on datasets obtained from YouTube.

Classification is usually performed using SVM. Only recently few works began to employ CNN to train end-to-end classifiers [9, 19].

3. FUSING TEXT AND VIDEO

In this section we introduce our embedding model which learns, via linear projections, a novel representation for videos.

3.1 Common Space

Given a pair made of a video $x \in \mathbb{R}^D$ and a sentence $y \in \mathbb{R}^V$, we define $W_v \in \mathbb{R}^{K \times D}$ and $W_t \in \mathbb{R}^{K \times V}$ respectively as the video and sentence projection matrices, where K is the dimensionality of the common space. The pair embedded representation is:

$$u = W_v \cdot x \qquad v = W_t \cdot y \tag{1}$$

Similarly as [8, 31], the idea is that paired elements should be located very near in the projected space while unpaired ones should be more distant. In particular, we require that the cosine similarity $\cos(x, y) = x \cdot y / (||x|| \cdot ||y||)$ between two paired elements (u_i, v_i) versus any other unpaired elements should be greater than $\alpha \in (0, 1)$. To this end, we learn matrices W_v and W_t solving an unconstrained problem, by minimizing the following contrastive ranking loss:

$$\begin{aligned} \mathcal{L}_{rank} = \; & \sum_i \sum_k \max\{0, \alpha - \cos(u_i, v_i) + \cos(u_i, v_k)\} + \\ & \sum_i \sum_k \max\{0, \alpha - \cos(v_i, u_i) + \cos(v_i, u_k)\} \end{aligned} \tag{2}$$

where v_k and u_k are contrastive terms, respectively incorrect videos and sentences that should not be associated to u_i and v_i.

3.2 Structure Preserving Space

One shortcoming of the pairwise ranking-loss defined in Equation 2 is that no information of the original manifolds of visual and textual features is preserved in the common space. For instance, similar videos with close visual features can be mapped far apart in the common space or, conversely, dissimilar videos may end up close together.

We address this shortcoming by defining additional constraints that, in contrast to Equation 2, preserve the original similarities in the common space. Thus, they act as a regularizer that induces some structure of the original manifolds into the common space.

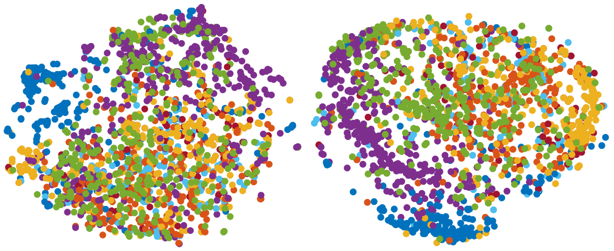
We consider the constraint:

$$|\cos(x_i, x_k) - \cos(u_i, u_k)| < \beta_v \tag{3}$$

which enforces solutions that keep the similarity between two videos, before and after the projection, within a margin β_v. Similarly, we formulate the following constraint for sentences:

$$|\cos(y_i, y_k) - \cos(v_i, v_k)| < \beta_s \tag{4}$$

We add these constraints to the problem defined by the minimization of loss in Equation 2. In order to maintain an unconstrained problem, we relax the constraints defined in



(a) Video Space (b) Embedding Space

Figure 1: Effect of our projection, using t-SNE visualization on Hollywood2 dataset. Each color corresponds to a different label.

Equation 3 and Equation 4 into a structure preserving loss. For videos, we define:

$$\begin{aligned}\Omega = & \sum_i \sum_k \max\{0, -\beta_v + \cos(x_i, x_k) - \cos(u_i, u_k)\} + \\ & \sum_i \sum_k \max\{0, -\beta_v - \cos(x_i, x_k) + \cos(u_i, u_k)\} \quad (5)\end{aligned}$$

and for sentences, we define:

$$\begin{aligned}\Theta = & \sum_i \sum_k \max\{0, -\beta_s + \cos(y_i, y_k) - \cos(v_i, v_k)\} + \\ & \sum_i \sum_k \max\{0, -\beta_s - \cos(y_i, y_k) + \cos(v_i, v_k)\} \quad (6)\end{aligned}$$

We add both terms to the ranking loss in Equation 2 and define our novel structure preserving loss:

$$\mathcal{L}_{struct} = \mathcal{L}_{rank} + \Omega + \Theta \qquad (7)$$

We show in Figure 1 a t-SNE [25] visualization of the original video features and their projection in the common space. It can be seen that a structure emerges in the embedding space. Compared to the original features, video sharing the same labels (i.e. the same colors) are put closer. This suggests that a classification algorithm may be able to find a better hypothesis.

Our embedding procedure uses a vector representation for videos and descriptions. In the following we detail the visual and textual features we have used in our experiments.

3.3 Video Representations

We explore two different video representations: learned spatio-temporal features and handcrafted features. The former is based on recent developments in deep learning, specifically we used the temporal extension of convolutional neural network obtained by performing 3D convolutions on video volumes namely C3D, proposed by Tran [23]. The latter is the well known improved dense trajectory (iDT) features proposed by Wang *et al.* [27].

For each video we compute C3D features on subsequences of 16 frames using the network pre-trained on Sports-1M [9]. Specifically, we obtain the activations of the sixth network layer which is Fully Connected (FC6). To obtain a global video representation, we use average-pooling of the FC6 over the whole sequence, ending with a 4096-dimensional vector.

We use all iDT descriptors, namely HOG, HOF, MBH$_x$, MBH$_y$, MBH$_{xy}$ and trajectories coordinates, encoding them with Fisher Vectors [17]. First we learn a PCA projection on 200k randomly selected descriptors for each local feature. We retain the first 80 components of each feature (except for trajectories which is compressed to 20). Then we concatenate the space-time coordinates of the central trajectory point, normalized in $[-1, 1]$, to the PCA compressed features. This adds spatial context to each local feature.

We estimate the GMM using the same subset on which we have learned PCA. We consider 256 Gaussians as previous work [26]. This results in a high dimensional representation that can be prone to overfitting. Thus, we further compress this representation to 4,096 dimensions with PCA, the same of C3D representation.

3.4 Textual Representation

We employ the Skip-Thought representation [11] which learns a distributed sentence representation using a neural model similar in the spirit to Word2Vec [15]. Skip-Thought are learned through an encoder-decoder architecture. A sentence s_i is fed to the encoder one word w_i^t at a time to produce a hidden state h_i representing the entire sentence. Then, the decoder has to output the previous s_{i-1} and the next s_{i+1} sentence of the corresponding text, exploiting the encoder output h_i. Thus, the model is learned optimizing the sum of the log-probabilities for the forward and backward sentences conditioned on the encoder representation.

Given sentences s_i, s_{i-1} and s_{i+1}, the log loss to be optimized is:

$$\sum_t \log p(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t \log p(w_{i-1}^t | w_{i-1}^{<t}, h_i) \qquad (8)$$

where notation $w^{<t}$ indicates all words before w_t, in the same sentence. We use the combine-skip model that has a hidden state h_i dimensionality of 4,800.

3.5 Learning Details

We initialize matrices weights, with random uniform distribution, $w_{ij} \propto \mathcal{U}\left(-\sqrt{(2/n_{in})}, \sqrt{(2/n_{in})}\right)$, where n_{in} is the input dimensionality of each domain. We set the projection space dimensionality K to 1000. Parameters α and β were both set to 0.2.

We optimize all losses using ADAM [10]. Considering all the contrastive pairs is unfeasible on large datasets, thus we only select a set of random pairs at each iteration.

We set the batch size to 100 and randomly sample 50 contrastive elements for each batch element. We run no more than 20 epochs and keep the model yielding the highest action classification mAP. We measure the mAP using 5-fold cross-validation on Hollywood2 training set.

After computing textual and visual features, we learn an embedding space for each type of feature. Afterwards, we apply the relative projections defined in Equation 1 to obtain new visual features. We learn a one-vs-rest linear SVM for each action class. For iDT we use late fusion, i.e. we sum SVM scores of each single-feature classifier.

4. EXPERIMENTS

4.1 Datasets

Since no public dataset with textual descriptions and action classes is publicly available, we consider one dataset to learn the embedding and one to perform the actual action recognition.

(a) iDT

Approach	mAP
Wang [26] *et al.*	65.7
Jain [6] *et al.*	62.5
Zhu [30] *et al.*	61.4
Mathe [14] *et al.*	61.0
Jiang [7] *et al.*	59.5
Gaidon [5] *et al.*	54.0
iDT+Fisher	56.4
iDT+Fisher+PCA	58.4
ours	66.2
ours + structure	**67.4**

(b) CNN

Approach	mAP
Max pooled LSTM [18]	43.2
Soft attention model [18]	43.9
C3D	44.7
ours	48.1
ours + structure	**48.7**

Table 1: Comparison with state-of-the art on Hollywood2.

The MPII Video Description Dataset [16] contains 68,337 short video sequences with associated textual descriptions, gathered from 94 HD movies. The particularity of this dataset is the availability of aligned Audio Descriptions. They are textual descriptions of video sequences that, read by a professional narrator, make movies accessible to visually impaired people. We use this dataset to learn the embedding.

The Hollywood2 [13] dataset contains 1,707 video snippets gathered from 69 HD movies labeled with 12 different actions. It is a difficult dataset since actions have high variation in appearance, context and motion. We used the clean training dataset with the provided train and test split. We use this dataset to measure the actual action recognition performance.

Both datasets are built from movies. We keep the datasets independent by removing from MPII all clips extracted from movies that are used in both datasets. We are left with 57,613 clips.

4.2 Results

We compare classification performance on Hollywood2 of baseline features and their projected counterparts.

First, we report in Table 1 (a) the performance of our approach using iDT features, compared to the state-of-the-art. Our baseline is similar to the approach of Wang *et al.* in [26]. However, for the sake of simplicity we do not use Spatio-Temporal Pyramids in conjunction with Spatial Fisher Vector. We just rely on the concatenation of local feature coordinates to each descriptor to inject contextual information. Thus, our implementation reaches 56.4 vs 65.7 of [26].

Applying PCA to the baseline does not compromise performance in classification, actually slightly improving the mAP. This behavior has been reported previously in image retrieval tasks [24]. Our projection, learned with the structure preserving loss, obtains an improvement of more than 9 mAP points with respect to the FV+PCA baseline.

In Table 1 (b) we report classification results using the C3D descriptor, compared with other CNN based approaches. Unfortunately convolutional representations do not perform very well on this challenging dataset. Our baseline is slightly superior to the Soft attention model by Sharma *et al.* [18]. Nonetheless, using our re-projected features improves mAP by 4 points. The structure preserving loss adds a slight improvement.

For both visual features we obtain a substantial improvement by learning a novel representation from video-text pairs. Moreover our structure preserving loss helps in not corrupting the video similarity in the starting space consistently improving the performance on both features.

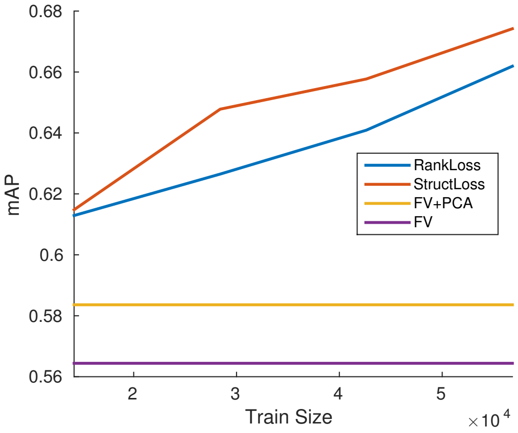


Figure 2: MAP on Hollywod2 varying the training set size compared to baseline.

We also analyze how the size of embedding training set affects the action classification performance. Increasing the set of captioned videos should increment the quality of the common space. We are also interested to understand if the learning method saturates. We perform this analysis using the late fusion of iDT descriptors, that performed better than C3D.

Interestingly, using 25% of the training set, we already get a sensible improvement in classification performance. Even more intriguing is the fact that the quality of the features steadily increases until the whole training set is used, with no observable saturation, as can be seen in Figure 2. We believe that further increasing the amount of captioned videos may lead to even better projections.

5. CONCLUSIONS

Our intuition that textual description can improve visual features for action recognition in video proved correct. The proposed method can easily leverage large sets of videos paired with a description to improve action recognition. Projections in structure preserving common space, can be profitably learned even with a smaller data set and performance improvement, in our experiments, does not show saturation.

Acknowledgments. This work is partially supported by the "Social Museum and Smart Tourism" project (CTN01_00034_231545).

6. REFERENCES

[1] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *Proc. of ICMR*. ACM, 2014.

[2] P. Das, C. Xu, R. Doell, and J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proc. of the CVPR*, pages 2634–2641, 2013.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proc. of NIPS*.

[5] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. *International Journal of Computer Vision*, pages 1–20, 2013.

[6] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *Proc. of CVPR*, 2013.

[7] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *Proc. of ECCV*. 2012.

[8] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*, 2015.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. of CVPR*, 2014.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Proc. of NIPS*, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[13] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. of CVPR*, 2009.

[14] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Proc. of ECCV*. 2012.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[16] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proc. of CVPR*, 2015.

[17] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[18] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *NIPS Workshop on Time Series*, 2015.

[19] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[20] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[21] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Proc. of NIPS*, 2013.

[22] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *CoRR*, abs/1412.0767, 2014.

[24] T. Uricchio, M. Bertini, L. Seidenari, and A. Del Bimbo. Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In *Proc. of ICCV Workshops*, 2015.

[25] L. van Der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[26] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, pages 1–20, 2015.

[27] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. *CoRR*, abs/1505.04868, 2015.

[28] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[29] R. Xu, C. Xiong, W. Chen, and J. J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *In Proc. of AAAI*. Citeseer, 2015.

[30] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action recognition with actons. In *Proc. of ICCV*, 2013.

[31] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.