Convolutional Neural Networks with Generalized Attentional Pooling for Action Recognition

Yunfeng Wang*, Wengang Zhou†, Qilin Zhang‡ and Houqiang Li§
*†§*Department of Electronic Engineering and Information Science, University of Science and Technology of China,* Hefei, China
‡*Highly Automated Driving, HERE Technologies*, Chicago, Illinois, USA
Email: *wangyf11@mail.ustc.edu.cn, †zhwg@ustc.edu.cn, ‡samqzhang@gmail.com, §lihq@ustc.edu.cn

***Abstract*—Inspired by the recent advance in attentional pooling techniques in image classification and action recognition tasks, we propose the Generalized Attentional Pooling (GAP) based Convolutional Neural Network (CNN) algorithm for action recognition in still images. The proposed GAP-CNN can be formulated as a new approximation of the second-order/bilinear pooling techniques widely used in fine-grained image classification. Unlike the existing rank-1 approximation, a generalized factoring (with non-linear functions) is introduced to exploit the intrinsic structural information of the sample covariance matrices of convolutional layer outputs. Without requiring preprocessing steps such as object (*e.g.*, human body) bounding boxes detection, the proposed GAP-CNN automatically focuses on the most informative part in still images. With the additional guidance of keypoints of human pose, the proposed GAP-CNN algorithm achieves the state-of-the-art action recognition accuracy on the large-scale MPII still image dataset.**

***Index Terms*—Action Recognition, Generalized Attentional Pooling, Convolutional Neural Network**

I. Introduction

Human action recognition is a fundamental and well explored research area in computer vision, due to its widespread applications in human-computer interaction, surveillance and game control. Traditional methods are based on handcrafted features, such as dense trajectory [1], object detection [2] or context mining [3] in image content. Recent convolutional neural networks (CNNs) based approaches have achieved impressive performance in action recognition with both still images and videos. Among them, multi-stream CNN methods such as "Two-Stream" [4] and its derivatives [5], [6] are among the top performers on the UCF101 [7] and HMDB51 [8] video action recognition datasets. Currently, the ResNet-101 based attentional pooling method [9] keeps the record of highest action recognition accuracy in still images with the MPII [10] dataset.

Previously in action recognition in still images, it is the norm to feed entire images to a CNN for classification. Later the hard attention concept is introduced with fine-grained features around the human bounding boxes or human pose keypoints, and such features are subsequently fed to CNNs for action classification [3]. Despite their performance advantages over the standard full-image based CNNs, the hard attention based CNNs suffer from significantly higher computational complexity due to the extra human bounding boxes detection step. Worse still, the required manual labeling of such bounding boxes in training data is prohibitively time-consuming and potentially expensive.

Pooling layer is an indispensable component of a modern CNN. Popular pooling algorithms include mean pooling and max pooling, both of which are first-order pooling (pooling operates on the feature map/matrix itself). Alternatively, the second-order pooling (pooling operates on the sample covariance matrix of the feature map/matrix) is advocated in [11], especially in applications such as semantic segmentation and fine-grained image classification. In [9], an evolved variant of the second-order pooling is proposed, with low-rank approximation and reformulation as attentional pooling. However, it assumes a rank-1 approximation of the weight matrix, which is arguably too restrictive and could potentially lead to performance penalties.

Inspired by [9], we propose a generalized factoring scheme (with additional non-linear functions) of the weight matrix, to exploit the intrinsic structural information of the sample covariance matrices of convolutional layer outputs. With the proposed factoring scheme, the weights matrix of a pooling layer is approximated by a top-down vector, a bottom-up vector and multiple bottom-up matrices. Parameters such as the optimal number of bottom-up matrices are empirically determined via cross validation. By incorporating extra supervision in the form of human pose keypoints, our proposed Generalized Attentional Pooling (GAP) based CNN+Pose (GAP-CNN+Pose) method can achieve even better results than the original attentional pooling [9] on the large-scale MPII still image action recognition dataset, indicating that GAP is complementary to hard attention.

The primary contribution of this paper is a new, generalized factoring/approximation to the weight matrix in the second order pooling layer of a CNN, with the action recognition application in a large-scale MPII still image dataset.

II. Related Work

Visual recognition has been widely studied in recent years, with both still image datasets and video datasets [1], [3], [12]–

This work was supported in part to Dr. Houqiang Li by 973 Program under contract No. 2015CB351803 and NSFC under contract No. 61390514, and in part to Dr. Wengang Zhou by NSFC under contract No. 61472378 and No. 61632019, the Fundamental Research Funds for the Central Universities, and Young Elite Scientists Sponsorship Program By CAST (2016QNRC001).

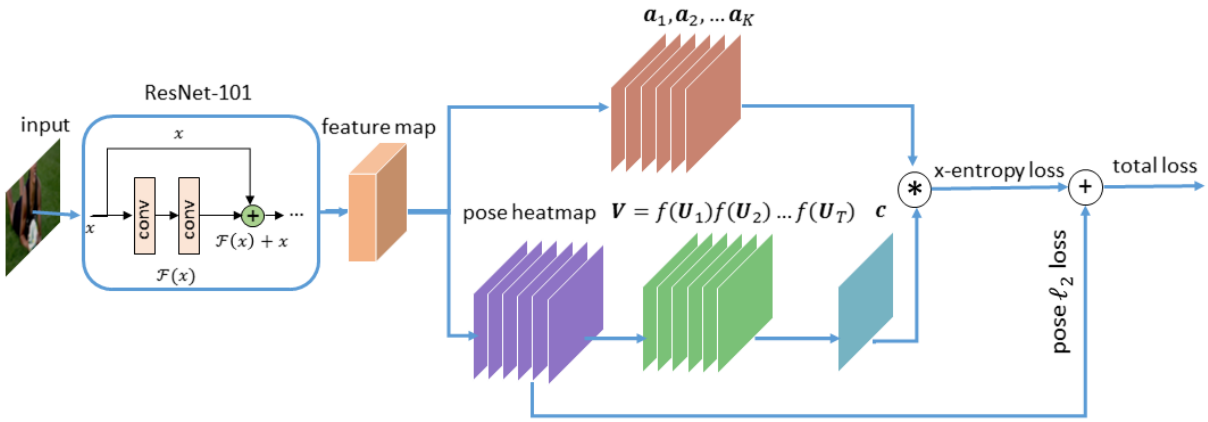

Fig. 1. Overview of the proposed GAP-CNN+Pose algorithm. Input images are fed into a ResNet-101 CNN (with the last pooling layer removed) to generate the feature map/matrix $\boldsymbol{X}$. Subsequently, two types of attention are imposed on the feature map, following [9]. The top branch denotes the top-down attention (*i.e.*, class-specific attention), which is constructed by multiplying the feature map/matrix $\boldsymbol{X}$ with a list of class-dependent vectors $\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_K$. On the bottom branch of the architecture, a series of T class-agnostic matrices $\boldsymbol{U}_1, \cdots \boldsymbol{U}_T$ are multiplied after nonlinear transformations $f(\cdot)$, *e.g.*, rectified linear unit (ReLU), followed by a class-agnostic vector $\mathbf{c}$ to represent the bottom-up attention, *i.e.*, the saliency-based attention. The additional human pose information is incorporated via the pose heatmaps and ℓ_2 regression.

[18]. For large scale still image action recognition datasets such as MPII [10] and HICO [19], the performance of popular baseline methods is unimpressive, *e.g.*, about 30% mAP on MPII dataset. Thanks to the extremely large number of classes (393 and 600 classes for MPII and HICO, respectively.) as well as high diversity[1], it is highly challenging to achieve high recognition accuracy on such datasets. On the contrary, popular video based action recognition datasets like UCF101 [7] and HMDB51 [8] are comparatively much smaller, with only 101 and 51 categories, respectively.

In this paper we focus on action recognition with still images. R*CNN [3] is a recent work in this field, in which R-CNN [9] is adapted to include one primary region and multiple proposal regions. The proposal region with the highest score is selected to cooperate with primary region to recognize the action in an image. Assisted with bounding boxes of the subject (*e.g.*, human), R*CNN achieves good result on the MPII dataset [10].

The most related work is [9], in which a rank-1 approximation of the weight matrix is proposed and attentional pooling is reformulated as low-rank second-order pooling. In [9], the attentional pooling is reformulated as a drop-in replacement for the popular mean pooling or max pooling near the end of CNNs. In contrast with [9], the proposed GAP extends the rank-1 approximation to a series of generalized non-linear factoring, and GAP can be incorporated after any layer in a CNN.

III. Formulation

The proposed GAP architecture is illustrated in Fig. 1. Let $\boldsymbol{X} \in \mathbb{R}^{n \times f}$ denote the reshaped output feature of a given layer, where n is the total number of spatial elements in the feature map, *i.e.*, the product of width and height of the feature map, and f is the number of channels. Conventional 1st order mean pooling and binary classification score computation can be formulated as

$$score^{bin}_{order1}(\boldsymbol{X}) = \frac{1}{n}\mathbf{1}^T\boldsymbol{X}\mathbf{w}, \tag{1}$$

with $\frac{1}{n}\boldsymbol{X}^T\mathbf{1}$ being the mean-pooled feature and $\mathbf{w}$ being a $f \times 1$ scoring weights.

Correspondingly, let matrix $\boldsymbol{W} \in \mathbb{R}^{f \times f}$ denote the scoring weight matrix after a second order pooling layer [11]. Following [9], the binary classification score is obtained by

$$score^{bin}_{order2}(\boldsymbol{X}) = Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}^T), \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times f}$ and $\boldsymbol{\Sigma} := \boldsymbol{X}^T\boldsymbol{X}$ is traditionally termed the sample covariance matrix[2] . Substitution of $\boldsymbol{\Sigma}$ into Eq. (2) yields

$$score^{bin}_{order2}(\boldsymbol{X}) = Tr(\boldsymbol{\Sigma}\boldsymbol{W}^T) = \sum_{i,j}\boldsymbol{\Sigma}_{i,j}\boldsymbol{W}_{i,j}, \tag{3}$$

where $\boldsymbol{\Sigma}, \boldsymbol{W} \in \mathbb{R}^{f \times f}$. From Eq. (3), matrix $\boldsymbol{W}$ can be interpreted as the element-wise weights of the sample covariance matrix $\boldsymbol{\Sigma}$.

Unlike the highly restrictive rank-1 approximation of $\boldsymbol{W}$ ($\boldsymbol{W} := \mathbf{a}\mathbf{b}^T$) in [9], we propose a gentler regularization by setting

$$\boldsymbol{W} := \mathbf{a}f(\boldsymbol{V}\mathbf{c})^T, \quad \mathbf{a} \in \mathbb{R}^{f \times 1}, \mathbf{c} \in \mathbb{R}^{r \times 1}, \boldsymbol{V} \in \mathbb{R}^{f \times r}, \tag{4}$$

where $f(\cdot)$ is an element-wise nonlinear transform function that keeps output dimensions as the input dimension, *e.g.*, rectified linear unit (ReLU). In addition, $\boldsymbol{V}$ can be further factorized into T matrices as

$$\boldsymbol{V} = \prod_{t=1}^{T} f(\boldsymbol{U}_t) = f(\boldsymbol{U}_1)f(\boldsymbol{U}_2)\ldots f(\boldsymbol{U}_T), \tag{5}$$

[1]In addition, it could be ambiguous to determine an action class based on a still image without temporal cues, *e.g.*, "sit down" versus "stand up".

[2]Sometimes sample mean values are subtracted before computing such sample covariance matrix. A constant factor of $1/(n-1)$ can also be included in the definition of Σ.

where $\boldsymbol{U}_1 \in \mathbb{R}^{f \times r_1}, \boldsymbol{U}_2 \in \mathbb{R}^{r_1 \times r_2}, \cdots, \boldsymbol{U}_T \in \mathbb{R}^{r_{(T-1)} \times r}$.

By the introduction of the matrix factorizations and nonlinear functions in Eq. (4)–(5), more structural information in the sample covariance matrix $\boldsymbol{\Sigma}$ could potentially be exploited. Practically, such factorization and nonlinearity are implemented as convolutional and ReLU layers, respectively. The optimal value of T is empirically decided to balance performance and model complexity[3]. Substitution of Eq. (4)–(5) into Eq. (2) yields a reformulation as the attentional score,

$$score_{att}^{bin}(\boldsymbol{X}) = Tr\left(\boldsymbol{X}^T \boldsymbol{X} f(\boldsymbol{V}\mathbf{c})\mathbf{a}^T\right) \quad (6)$$

$$= (\boldsymbol{X}\mathbf{a})^T \left(\boldsymbol{X} f(\boldsymbol{V}\mathbf{c})\right). \quad (7)$$

Eq. (7) indicates that the score can be seen as the inner product of two attentional heatmaps. Similarly, such derivations can be extended to K-class ($K \geq 3$) classifier. Let $\boldsymbol{W}_k$ be the class-specific weights for class k, $k = 1, \cdots, K$. Eq. (2) can be rewritten as,

$$score_{order2}^{Kclass}(\boldsymbol{X}, k) = Tr(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{W}_k^T), \quad (8)$$

with $\boldsymbol{W}_k \in \mathbb{R}^{f \times f}$. Parallel to Eq. (6)–(7), Let[4] $\boldsymbol{W}_k := \mathbf{a}_k(\boldsymbol{V}\mathbf{c})^T$, the class-specific attentional pooling and scoring is obtained as

$$score_{att}^{Kclass}(\boldsymbol{X}, k) = (\boldsymbol{X}\mathbf{a}_k)^T \boldsymbol{X}(f(\boldsymbol{V}\mathbf{c})). \quad (9)$$

In Eq. (9), the former terms $\boldsymbol{X}\mathbf{a}_k$ represent the class-specific top-down attentional feature maps; while the latter terms $\boldsymbol{X} f(\boldsymbol{V}\mathbf{c})$ denote the saliency-based, class-agnostic bottom-up attentional feature maps. As advocated in [20] and [9], the fusion of top-down and bottom-up attention maps is motivated biologically, and it is beneficial to modulate saliency maps with class-specific top-down information.

From [9], human pose regularization can contribute to the action recognition accuracy. Therefore, we incorporate human body keypoints heatmaps and use it as the regularization term for the cross-entropy loss in Fig. 1. Specifically, two additional convolutional layers are added after the last layer of the ResNet-101 CNN and a 16-channel regression layer to predict the pose keypoints. An l_2 loss is used to calculate the cost between the predicted heatmaps and the ground truth heatmaps.

The overall loss is calculated by weighted sum of this l_2 loss and a cross-entropy loss, making it possible to optimize the entire GAP-CNN+Pose network in an end-to-end manner.

IV. Experiments

Dataset. In this section, experiments are conducted on the challenging large-scale action recognition datasets, *i.e.*, the MPII still image dataset [10]. The MPII human pose dataset contains 15205 images in 393 action classes, grouped into a train split, a validation split and a test split, with 8218, 6987 and 5708 images, respectively. The dataset is also annotated with ground truth human body keypoints. We use the mean average precision (mAP) and classification accuracy as criteria to evaluate the performance of competing methods.

[3]More details are presented in Section IV.

[4]Note that $\boldsymbol{V}$ and $\mathbf{c}$ are bottom-up parameters, thus are class-agnostic.

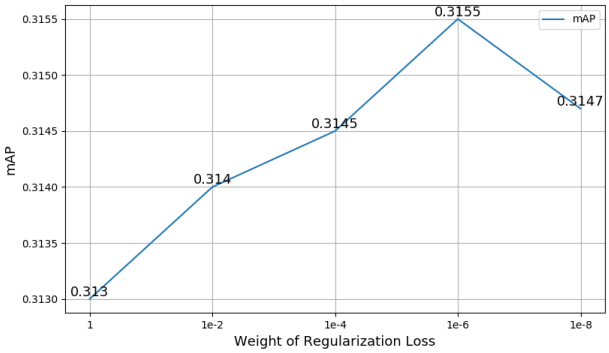

Fig. 2. Illustration of different mAP with respect to varying weights for the regularization pose ℓ_2 loss based on the validation split of the MPII dataset. X-axis is on inverted logarithmic scale while Y-axis is on linear scale.

Weight of Pose Regularization. Cross validation experiments are conducted to empirically determine the optimal weight of the regularization ℓ_2 loss from pose keypoints. Without loss of generality, the weight of the cross-entropy loss is fixed at constant value 1, and the weight of the pose regularization loss varies from 1 to 10^{-8}, as shown in Fig. 2. From Fig. 2, we observe that the mAP is insensitive to the choice of weight value for the pose regularization loss. The highest mAP is achieved with such weight at approximately 10^{-6}, thus 10^{-6} is chosen and fixed throughout the rest of the paper.

Number of Bottom-up Matrices. In this part we show the experiments designed to determine the optimal number of bottom-up matrices, *i.e.*, T in Eq. (5). Since convolution operations in CNNs are implemented by matrix multiplication, we take advantage of the existing convolution layers to implement matrix multiplication operations. We set $r_1 = 4096$ and determine the remaining values by induction as $r_{i+1} = r_i/2$, $i = 1, \cdots, T-2$. We use the convolutional layer $\boldsymbol{C}_i$ with the input r_{i-1} and output r_i to represent $\boldsymbol{U}_i$. ReLU layers are added between such convolution layers. Recognition accuracy and mAP are used as criteria in the choice of T based on the validation split of the MPII dataset, as shown in Fig. 3. We observe that both criteria reach plateau with T over 3. To keep the number of such convolutional layers as small as possible (for computational efficiency), T is fixed at 3 in the rest of this paper.

Attention Visualization. Figure 4 shows several typical examples of the GAP-CNN predicted attention heatmaps imposed on input images. We observe that the most informative parts of such input images are mostly highlighted in the corresponding heatmaps.

Comparison. Because the ground truth labels for the test split of the MPII dataset is not publicly available, the validation split is used for such evaluation. The comparison of the proposed GAP-CNN method with competing algorithms (without pose information) are summarized in the top half of Table I. Our proposed GAP-CNN method achieves both the highest mAP and the highest recognition accuracy. In addition,

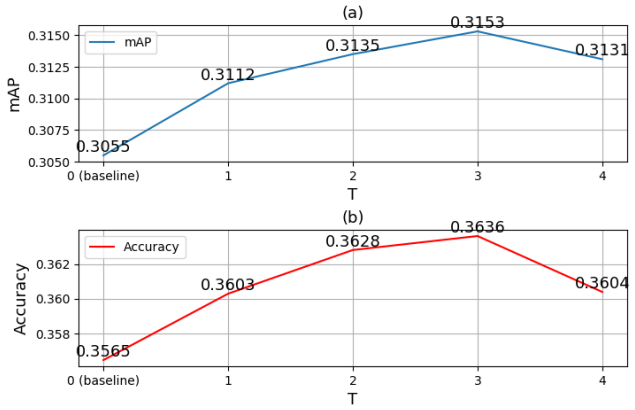

Fig. 3. Illustration of recognition accuracies and mAP with different T values based on the validation split of the MPII dataset. We observe that both mAP and accuracy reach plateau with T over 3. $T = 3$ is the choice to maximize mAP and accuracy.

Fig. 4. Examples of merged attentions on training images. All input images are color images with RGB values, which are only shown in grayscale in Fig. 4 to facilitate the visualization of heatmaps. We can find that our method can focus on the important parts in images.

TABLE I
PERFORMANCE COMPARISON ON THE VALIDATION SET OF MPII.

Method	mAP	Accuracy
VGG16, R-CNN [3]	16.5%	-
VGG16, R*CNN [3]	21.7%	-
ResNet-101 [9]	26.2%	-
Attn. Pool [9]	30.3%	35.3%
Proposed GAP-CNN	**30.6%**	**36.0%**
Attn. Pool.+Pose [9]	30.6%	35.7%
Proposed GAP-CNN+Pose	**31.6%**	**36.9%**

our proposed GAP-CNN+Pose algorithm also outperforms the pose-enhanced version of attentional pooling [9], supporting our speculation that the proposed GAP model could be complementary to hard attention.

V. CONCLUSION

In this paper, the Generalized Attentional Pooling based Convolutional Neural Network (GAP-CNN) algorithm is proposed for action recognition in still images. Empirical experiments are carried out to determine the practically optimal number of bottom-up pooling matrices. In addition, extra supervisions such as human pose keypoints are exploited. With the practically optimal number of bottom-up attentional pooling and a single top-down pooling, the proposed GAP-CNN algorithm outperforms 4 competing algorithms, including the original attentional pooling method [9]. Even after the incorporation of human pose keypoints information, the proposed GAP-CNN+Pose algorithm nevertheless achieves the state-of-the-art action recognition performance on the large-scale MPII still image dataset.

REFERENCES

[1] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.
[2] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*, 2010.
[3] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r*cnn," in *ICCV*, 2015.
[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.
[5] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.
[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *ECCV*, 2016.
[7] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV*, 2011.
[9] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *NIPS*, 2017.
[10] M. Andriluka, L. Pishchulin, P. Gehler, and S. Bernt, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.
[11] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *ECCV*, 2012.
[12] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.
[13] Q. Zhang and G. Hua, "Multi-view visual recognition of imperfect testing data," in *ACM MM*, 2015.
[14] Y. Wang, W. Zhou, Q. Zhang, X. Zhu, and H. Li, "Low-latency human action recognition with weighted multi-region convolutional neural network," *arXiv preprint arXiv:1805.02877*, 2018.
[15] X. Lv, L. Wang, Q. Zhang, Z. Niu, N. Zheng, and G. Hua, "Video object co-segmentation from noisy videos by a multi-level hypergraph model," in *ICIP*, 2018.
[16] J. Zang, L. Wang, Z. Liu, Q. Zhang, Z. Niu, G. Hua, and N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in *AIAI*, 2018.
[17] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.
[18] Q. Zhang, G. Hua, W. Liu, Z. Liu, and Z. Zhang, "Auxiliary training information assisted visual recognition," *IPSJ Trans. Comput. Vis. and Appl.*, vol. 7, pp. 138–150, 2015.
[19] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *ICCV*, 2015.
[20] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *CVPR*, 2006.