Robust Attentional Pooling via Feature Selection

Zhang, J.; Lee, T.-Y.; Feng, C.; Li, X.; Zhang, Z.

TR2018-124 August 25, 2018

Abstract

In this paper we propose a novel network module, namely Robust Attentional Pooling (RAP), that potentially can be applied in an arbitrary network for generating single vector representations for classification. By taking a feature matrix for each data sample as the input, our RAP learns data-dependent weights that are used to generate a vector through linear transformations of the feature matrix. We utilize feature selection to control the sparsity in weights for compressing the data matrices as well as enhancing the robustness of attentional pooling. As exemplary applications, we plug RAP into PointNet and ResNet for point cloud and image recognition, respectively. We demonstrate that our RAP significantly improves the recognition performance for both networks whenever sparsity is high. For instance, in extreme cases where only one feature per matrix is selected for recognition, RAP achieves more than 60% improvement over PointNet in terms of accuracy on the ModelNet40 dataset.

IEEE International Conference on Pattern Recognition (ICPR)

Robust Attentional Pooling via Feature Selection

Jian Zheng[†], Teng-Yok Lee[‡], Chen Feng[‡], Xiaohua Li[†], and Ziming Zhang[‡*]
[†]School of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902-6020
[‡]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139-1955
Email: {jzheng65, xli}@binghamton.edu, {tlee, cfeng, zzhang}@merl.com

Abstract—**In this paper we propose a novel network module, namely Robust Attentional Pooling (RAP), that potentially can be applied in an arbitrary network for generating single vector representations for classification. By taking a feature matrix for each data sample as the input, our RAP learns *data-dependent* weights that are used to generate a vector through linear transformations of the feature matrix. We utilize feature selection to control the sparsity in weights for compressing the data matrices as well as enhancing the robustness of attentional pooling. As exemplary applications, we plug RAP into PointNet and ResNet for point cloud and image recognition, respectively. We demonstrate that our RAP significantly improves the recognition performance for both networks whenever sparsity is high. For instance, in extreme cases where only one feature per matrix is selected for recognition, RAP achieves more than 60% improvement over PointNet in terms of accuracy on the ModelNet40 dataset.**

I. Introduction

In the past few years, we have witnessed an explosion of deep learning research [1], such as ResNet [2] for 2D image classification and PointNet [3] for 3D point cloud recognition that achieve the state-of-the-art performance in the fields, respectively. In these deep networks, the learned feature representations are aggregated into vectors before classification through conventional pooling, such as max or average pooling. Pooling plays an important role in deep neural networks as it can reduce data variance as well as computational complexity while extracting low-level features from the neighborhood. However, conventional pooling methods tend to learn a global feature vector without considering the relative importance of each individual feature.

In fact, many computer vision problems are spatially region related. Namely, some spatial locations may contain more information, while others include less or even irrelevant information. For example, to classify an object in an image, only the areas that are related to the object, both explicitly (*e.g.* object parts) and implicitly (*e.g.* indicative background), are crucial, while other areas can be considered as noisy background. Given an image, instead of looking at the entire image, humans can quickly understand the scene by attending to selective locations. Mimicking human visual attention, spatial visual attention has been extensively studied in recent years and deployed in various applications, such as image classification [4], image captioning [5][6][7], and visual question answering (VQA) [8][9][10].

While most attention models are developed for images and videos, fewer efforts have been made for tasks with 3D point clouds [11][12]. Point cloud is a collection of data points for object representation in the three-dimensional coordinate system. Modern technology advancement in 3D laser scanners empowers the proliferation of a large amount of point clouds data, rendering point clouds storage, transmission, and applications in practice. Among all the data points in a point cloud, a compact point descriptor is able to characterize the object shape and structure. For instance, a bench is depicted by a 3D point cloud with a large number (M) of 3D data points. Nevertheless, only a small set (K) of data points ($K << M$) are crucial to the object representation, such as those key points that capture the main characteristics of the bench (e.g., back, seat, and legs). Therefore, proper feature selection method is necessary to replace the original large point clouds with selected and simplified point clouds to facilitate the storage and transmission of 3D point clouds.

In this paper, we introduce Robust Attentional Pooling (RAP), which is simple and can be integrated with typical deep neural networks to conduct end-to-end training for classification. RAP consists of three major components, a feature adaptation module, an attention learning module, as well as a feature selection module. Specifically, the feature adaptation module is designed to transform the input data features to assist subsequent attention learning process. Then the attention learning module learns an attention vector which weights each data feature discriminatively. With the learned attention vector, feature selection is implemented to compress the data features attentionally by ranking and selecting attention weights and corresponding data features, generating a weighted feature vector for further classification.

To validate the effectiveness of RAP, extensive experiments have been conducted in both point cloud classification and image classification. The experiment results show that RAP is able to effectively select and aggregate input data features attentionally. The input data features can be compressed greatly with RAP by controlling attention weight sparsity, and the recognition performance with both point clouds and images can be significantly enhanced when the weight vector sparsity is high. Visualizations in both 3D and 2D domains are conducted, demonstrating that RAP is capable of compressing data features effectively by selecting key data features.

II. Related Work

In typical deep convolutional neural networks (ConvNets), pooling layers play an important role in nonlinear downsampling. Specifically, the spatial size of feature representation

[*] This work was done when the first author took internship at MERL.
[*] Corresponding author.

is reduced through pooling operations to further reduce the model and computational complexity. Overfitting problem can be alleviated through pooling. Max pooling [13][14][15], average pooling [2], and stochastic pooling [16] are widely used in various deep neural networks. In many deep networks, a global pooling operation is implemented over the learned feature representations in order to obtain a global feature map. For example, global average pooling proposed in [17] was applied in ResNet [2] prior to the final classifier. In 3D domain, a global max pooling layer was placed prior to the classifier in PointNet [3] for 3D point cloud classification. However, such global pooling operations fuse the learned feature representations without considering each location independently.

Incorporating visual attention into different computer vision tasks not only contributes more attention to selective spatial locations, but also facilitates the aggregation of data representations. While most of the research works are concerned with applying attention to 2D computer vision tasks with images [18][10][5][19] and videos [20][21], less attention mechanisms are developed for problems in 3D domain. Some methods have been proposed to deal with saliency detection of 3D surfaces [22][23][24]. However, it is nontrivial to extend those methods to unorganized 3D point clouds. A few works address saliency problem of point sets. For instance, [25] proposed a method to compute saliency maps by exploring local surface properties in different scales as well as depth information. A saliency detection technique in large point sets based on distinctness was proposed in [11]. Nevertheless, those methods are complex in terms of implementation and extra information are needed besides the original point clouds. Although a number of deep learning schemes have been developed on 3D point clouds [26][3][27], as far as we know, deep learning based 3D visual attention in point clouds has not been explored yet.

There are roughly three feature selection methods, i.e., filters [28][29], wrappers [30], and sparsity-inducing feature selection (SSFS) [31]. Filter methods select features as a pre-processing step by applying a statistical measure of each feature, without considering the interactions among features. Though filter methods are computationally fast, they may choose redundant features. Wrapper methods involve learning algorithms that optimize model performances, leading to better selection results than filters at the cost of high computational burden. SSFS methods take the structure of features into consideration while regularizing by the ℓ_1 norm for sparsity learning. However, SSFS methods suffer from the problem of high dimensionality. In contrast, we propose a deep learning based feature selection mechanism which takes visual attention into account.

III. Our Approach

A. Key Notations

We denote $\{(\mathbf{X}_n, y_n)\}_{n=1}^N$ as a set of training data where $\mathbf{X}_n \in \mathbb{R}^{d\times M_n}$ is the n-th feature matrix and $y_n \in \mathcal{Y}$ is its class label, $\mathbf{X}_{nm} \in \mathbb{R}^d, m \in \{1, \cdots, M_n\}$ as the m-th row in $\mathbf{X}_n$, $\phi : \mathbb{R}^d \to \mathbb{R}^D$ as a feature mapping function, $\mathbf{w}_n \in \mathbb{R}^{M_n}, \forall n$ as an *attentional weighting vector* corresponding to the n-th training sample that consists of scalars $\{w_{nm}\}_{m=1,\cdots,M_n}$, $f : \mathbb{R}^D \to \mathbb{R}$ as the classifier, $\Phi, \mathcal{W}, \mathcal{F}$ as the feasible solution spaces for the corresponding variables.

B. General Objective for Feature Selection

In this paper we are interested in optimizing the following nonconvex problem in general:

$$\min_{\phi\in\Phi,\{\mathbf{w}_n\in\mathcal{W}_n\},f\in\mathcal{F}} \Omega\Big(\phi, \{\mathbf{w}_n\}, f\Big) + C\sum_{n=1}^{N}\ell\Big(y_n, f\Big(\sum_{m=1}^{M_n} w_{nm}\phi\left(\mathbf{X}_{nm}\right)\Big)\Big), \quad (1)$$

where Ω denotes the regularizer over variables, ℓ denotes a loss function such as hinge loss, and $C \geq 0$ is a predefined constant. We call $f\Big(\sum_{m=1}^{M_n} w_{nm}\phi\left(\mathbf{X}_{nm}\right)\Big)$ *decision function.*

Note that $\Phi, \mathcal{W}, \mathcal{F}$ define the constraints on the variables. For instance, we can constrain each $\mathbf{w}_n$ to be non-negative, *i.e.* $w_{nm} \geq 0, \forall m$. The functionality of attentional vectors $\mathbf{w}$'s is to select informative features for recognition.

Discussion: From the perspective of problem definition, mathematically our objective in Eq. 1 can be considered as *generalization* of several classic machine learning problems such as multiple instance learning (MIL) [32], multiple kernel learning (MKL) [33], and boosting [34].

In addition, from the algorithmic perspective, Eq. 1 can be utilized as general objective in many algorithms as well by substituting the variables with different realizations, such as attentional pooling [35] and attentive pooling networks [36].

C. Robust Attentional Pooling (RAP)

In general Eq. 1 defines a nonconvex optimization problem. Traditionally this problem is solvable using alternating optimization, *i.e.* updating one variable while fixing the rest. Note that in order to compute the decision score, the variables in Eq. 1 have to be applied *sequentially*, *i.e.* $\phi \to \mathbf{w} \to f$, generally speaking without considering special substitution of variables.

This sequential behavior in computing the decision function in Eq. 1 inspires us to propose a new deep model based solver, as illustrated in Fig. 1. Specifically,

1) *Feature adaptation* module is for learning ϕ that transfers the original features into another space for generating discriminative weighted features;
2) *Attention learning* module is for learning $\mathbf{w}$ that weights informative features data-dependently. The weight vector $\mathbf{w}$ with the dimension of M is learned by passing the transformed features through a stack of two fully-connected layers, then ReLU is operated over $\mathbf{w}$;
3) *Feature selection* module is for selecting and aggregating the K most informative features. To be specific, we sort the attention vector $\mathbf{w}$ and select top K of them as well as the K corresponding features, where $0 < K \leqslant M$. Then we do a weighted sum with the selected weights and features, generating a single feature vector. The $\bigotimes$ in Fig. 1 denotes weighted sum;
4) *Classification* module is for learning f that makes decision based on the weighted feature vector from feature selection module.

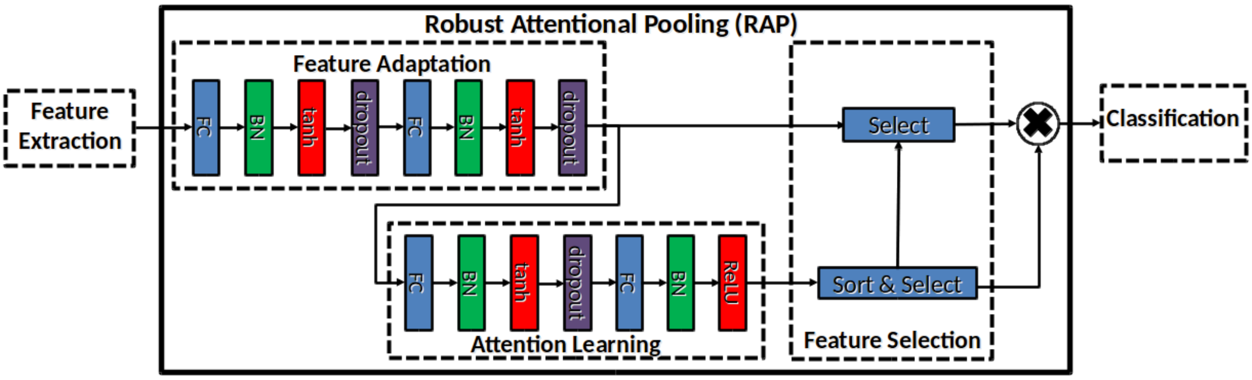


Fig. 1: RAP for visual recognition.

In the feature selection module, sorting can be written as a sequence of max functions, and thus it is differentiable in this sense. While selection serves as an operation to find indexes. Therefore, given the network realization of each module, the variables in Eq. 1 can be learned simultaneously using back-propagation.

Visual Object Recognition Networks with RAP: Many deep networks have been proposed for object recognition, such as ResNet [2] for 2D images and PointNet [3] for 3D point clouds. Such networks utilize the pooling layer to fuse the information from different spatial locations. Though pooling operates easily in terms of computation, it discards the difference in feature discriminality, leading to informative features faded away. In contrast, our RAP is developed intentionally to capture such difference for generating discriminative weighted features by training attention mechanism. The attentional weights can be utilized to select informative features for further usage such as data compression. Empirically we can introduce the attention mechanism into ResNet or PointNet by replacing their pooling layer with RAP for recognition.

Realization of RAP for Visual Object Recognition: Visual data has a remarkably large amount of redundancy for recognition. For instance, in 3D point clouds, corner and edge points will be more informative for recognition than the points on the surface. Here we are interested in the RAP that can locate such *key* points/locations in the visual data based on attention mechanism robustly without sacrificing recognition accuracy significantly, compared with using all the data information.

Therefore, we consider optimizing Eq. 1 under the constraints on $\mathbf{w}$'s so that

$$\mathcal{W}_n \overset{\text{def}}{=} \{\mathbf{w} \mid \mathbf{w} \geq \mathbf{0}, \|\mathbf{w}\|_0 \leq K\}, \tag{2}$$

where $\geq$ here is an element-wise operator, $\|\cdot\|_0$ denotes the ℓ_0-norm of a vector that counts the non-zeros in the vector, and $K \geq 0$ is a predefined constant.

Fig. 1 provides us a novel general framework to construct RAP. For instance, we can substitute each module with a network as a realization of RAP. In particular, we design a specific network architecture as shown in Fig. 1 as our RAP used in the experiments for visual recognition. We find that empirically this architecture has already achieved reasonably good performance and therefore we do not fine-tune the network architecture intensively.

IV. Experiment

To evaluate the effectiveness of proposed RAP, extensive experiments have been conducted for both point cloud classification and image classification. To verify the effectiveness of feature adaptation and attention learning modules, we propose a simplified RAP called AL, by removing feature adaptation module from RAP architecture as shown in Fig. 1.

A. RAP in point cloud classification

Datasets: RAP for point cloud classification is evaluated on three datasets, including 2D point clouds and 3D point clouds. Following [37], MNIST handwritten digits are converted to 2D point clouds by taking image coordinates of all non-zero pixels. Two 3D point clouds datasets are utilized for evaluation, including ModelNet40 [38] and ShapeNet [39]. ModelNet40 is a 3D CAD benchmark dataset, including 9843 training instances and 2468 testing instances from 40 object classes. ShapeNet benchmark contains 16881 instances of 16 categories of 3D shapes.

Configuration and training: In RAP based point cloud classification network, we apply feature extraction module to extract point features following [3]. Then the extracted point features are input to RAP for feature selection. We take PointNet for comparison, which selects K data points randomly from M data points ($M = 256$ for MNIST, $M = 1024$ for ModelNet40 and ShapeNet) for classification. While RAP and AL select K out of M data points attentionally. Specifically, the features of M data points are passed over to the feature adaptation module in RAP. Then, attention mechanism is implemented to learn attention vectors (each with M attention weights) based on transformed features. By ranking and selecting the top K attention weights, the features of K data points are selected and aggregated through weighted sum. The output of RAP and AL module is a global feature vector, which is fed to the classifier. Feature adaptation module and attention learning module are composed of a stack of fully-connected layers, as shown in Fig. 1. ReLU and tanh are used as activation functions, dropout and batch normalization are performed for regularizations. Training hyperparameters are set to be the same as that in PointNet [3], except that RAP and AL apply a different dropout value (0.5). Batch size is 32 and learning rate is 0.001. Each experiment is run for 250 epochs, taking 6 to 8 hours on a single NVIDIA GPU.

Experiment results: The experiment results of point cloud classification on MNIST, ShapeNet, and ModelNet40 are summarized in Table. I, Table. II and Table. III, respectively. It can be seen that when fewer data points are selected, RAP and AL outperform PointNet. Specifically, from Table. I and Table. II we can see that by selecting $K = 4$ data points, RAP and AL are capable of achieving comparable performance with PointNet when it utilizes all the M data points. Therefore, the experiment results indicate that RAP and AL can select point clouds effectively and robustly. RAP achieves better performance than AL when the weight sparsity increases. When the weight sparsity becomes increasingly high, AL tends to perform better than RAP in the end. RAP and AL are slightly inferior to PointNet when the weight sparsity is low (larger K).

K	PointNet	AL	RAP
256	**98.6 ± 0.0**	98.2 ± 0.0	98.2 ± 0.3
128	98.0 ± 0.0	98.3 ± 0.0	**98.3 ± 0.0**
64	96.3 ± 0.0	97.7 ± 0.0	**98.3 ± 0.0**
32	91.3 ± 0.1	98.1 ± 0.0	**98.4 ± 0.0**
16	80.3 ± 0.1	97.9 ± 0.0	**98.2 ± 0.0**
8	63.1 ± 0.2	97.8 ± 0.0	**98.3 ± 0.0**
4	44.8 ± 0.1	97.8 ± 0.0	**98.1 ± 0.0**
2	30.8 ± 0.1	**98.3 ± 0.0**	97.4 ± 0.1
1	21.4 ± 0.1	**98.2 ± 0.0**	95.7 ± 0.5

TABLE I: Point cloud classification results on MNIST

K	PointNet	AL	RAP
1024	**98.6 ± 0.0**	98.2 ± 0.0	98.2 ± 0.0
512	**98.5 ± 0.0**	98.3 ± 0.0	98.1 ± 0.1
256	**98.6 ± 0.0**	98.1 ± 0.0	98.0 ± 0.1
128	**98.4 ± 0.0**	97.5 ± 0.1	97.5 ± 0.0
64	**98.2 ± 0.1**	98.1 ± 0.0	98.0 ± 0.0
32	97.3 ± 0.2	97.8 ± 0.0	**97.9 ± 0.0**
16	95.4 ± 0.1	98.0 ± 0.1	**98.2 ± 0.1**
8	89.2 ± 0.2	**98.2 ± 0.0**	97.3 ± 0.1
4	74.0 ± 0.3	**98.1 ± 0.0**	97.9 ± 0.1
2	60.2 ± 0.4	**98.2 ± 0.0**	97.9 ± 0.2
1	48.6 ± 0.2	**97.9 ± 0.1**	94.1 ± 0.3

TABLE II: Point cloud classification results on ShapeNet

RAP visualizations on 3D point clouds: To further demonstrate that RAP has the capability of selecting key point features of point clouds, the selected data points with RAP on some examples from ModelNet40 are visualized in Fig. 2. Note that different colors in Fig. 2 imply different z coordinate values for each data point in the point cloud.

When fewer points are selected, for example, 256, 128, 64 and 32 points, the visualization results imply that RAP tends to select those crucial points that capture the key features of the objects' shapes. By taking the desk in Fig. 2 (second row) for example, RAP is able to select 128 out of 1024 data points which correspond to the desk's legs as the most informative parts.

K	PointNet	AL	RAP
1024	**88.9 ± 0.1**	88.4 ± 0.1	88.2 ± 0.0
512	**88.8 ± 0.1**	87.8 ± 0.1	88.1 ± 0.1
256	**88.6 ± 0.1**	87.6 ± 0.1	88.0 ± 0.1
128	87.3 ± 0.1	87.6 ± 0.1	**87.6 ± 0.1**
64	85.2 ± 0.1	85.6 ± 0.3	**87.0 ± 0.2**
32	82.2 ± 0.1	86.3 ± 0.2	**87.3 ± 0.1**
16	73.4 ± 0.1	85.4 ± 0.1	**86.3 ± 0.0**
8	54.6 ± 0.2	84.1 ± 0.1	**85.1 ± 0.2**
4	33.2 ± 0.2	**85.2 ± 0.2**	83.4 ± 0.4
2	19.6 ± 0.3	**86.4 ± 0.0**	78.4 ± 0.3
1	11.3 ± 0.1	**83.6 ± 0.2**	74.0 ± 1.8

TABLE III: Point cloud classification results on ModelNet40

B. RAP in image classification

Datasets: RAP based image classification is evaluated on CIFAR10 dataset [40], which consists of 50000 labeled training images and 10000 labeled test images from 10 classes. To evaluate the effectiveness of RAP on image classification, ResNet50 for image recognition [2] is deployed for image feature extraction. In our experiments, image features are extracted in two ways. The first set of image features are learned with ResNet50 which is pretrained on ImageNet [41]. In addition, we extract image features from finetuned ResNet50 on CIFAR10 with mean image subtracted.

Configuration and training: In RAP and AL based image classification network, the extracted image features with ResNet50 prior to global average pooling layer with the shape of $7 \times 7 \times 2048$ are reshaped to 49×2048 and input to the subsequent RAP and AL modules, here $M = 49$. We take ResNet50 as our baseline model, which selects the features of K ($0 < K \leqslant 49$) locations randomly and pools them with global average pooling. While RAP and AL select and aggregate image features data-dependently, generating a single feature vector for classification. Similarly, dropout and batch normalization are used for regularization, ReLU and tanh are applied for nonlinearity. Batch size is 250 and learning rate is 0.001. We run each experiment for 100 epochs, taking 4 to 6 hours on a single GPU.

Experiment results: Table. IV and Table. V present the experiment results of image classification. It can be seen from Table. IV that with finetuned ResNet50 for feature extraction, RAP achieves comparable results with ResNet50 and AL when all the image features are deployed for classification. When the features of K locations are selected, AL model achieves better performance than ResNet50 while RAP outperforms ResNet50 and AL model. As shown in Table. V, when with pretrained ResNet50 for feature extraction, AL model gains better performance than ResNet50 and RAP is superior to ResNet50 and AL model. The experiment results not only demonstrate that RAP and AL are effective and robust in attention learning and feature selection for image classifications, but also verify the effectiveness of feature

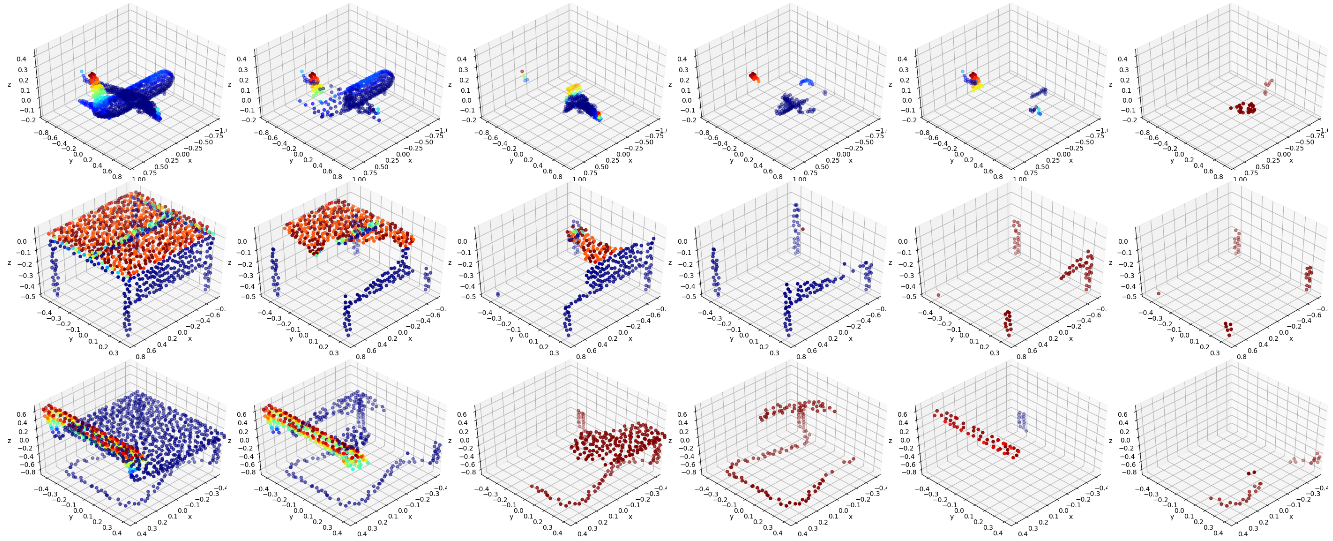


Fig. 2: RAP visualizations examples with point clouds on ModelNet40: left to right, $K = 1024, 512, 256, 128, 64, 32$.

adaptation module in RAP architecture for image classification.

K	PointNet	AL	RAP
49	**95.6 ± 0.0**	**95.6 ± 0.0**	**95.6 ± 0.0**
32	94.8 ± 0.0	95.2 ± 0.0	**95.5 ± 0.0**
16	86.2 ± 0.0	93.7 ± 0.1	**95.0 ± 0.0**
8	69.5 ± 0.0	88.8 ± 0.0	**94.7 ± 0.0**
4	63.0 ± 0.0	74.6 ± 1.9	**94.2 ± 0.0**
2	50.6 ± 0.0	60.4 ± 3.7	**74.5 ± 0.6**
1	42.8 ± 0.1	48.9 ± 0.0	**63.8 ± 0.3**

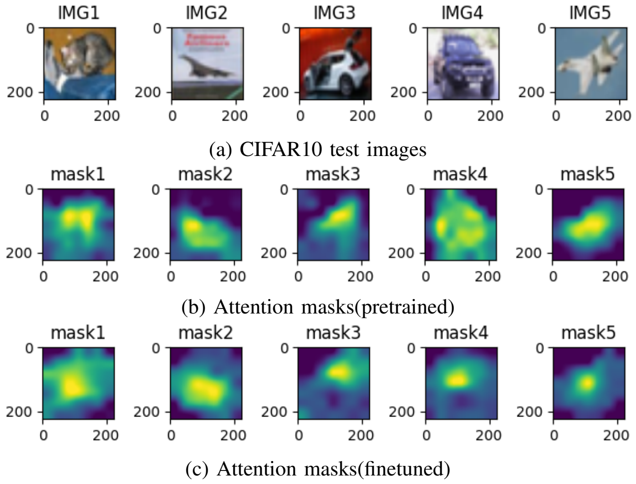
TABLE IV: Image classification results on Finetuned CIFAR10 features

K	PointNet	AL	RAP
49	90.5 ± 0.1	91.1 ± 0.1	**92.0 ± 0.0**
32	89.1 ± 0.0	90.7 ± 0.1	**92.1 ± 0.1**
16	74.2 ± 0.1	88.6 ± 0.1	**91.0 ± 0.1**
8	57.2 ± 0.0	80.8 ± 0.1	**89.7 ± 0.1**
4	50.3 ± 0.0	65.0 ± 0.1	**88.4 ± 0.1**
2	39.8 ± 0.1	52.8 ± 0.1	**85.7 ± 0.1**
1	34.6 ± 0.1	45.4 ± 0.1	**79.6 ± 0.1**

TABLE V: Image classification results on Pretrained CIFAR10 features

RAP visualizations on images: Visualizations of learned attention masks for CIFAR10 test images with RAP are presented in Fig. 3. Specifically, Fig. 3a shows five examples of CIFAR10 test images, while Fig. 3b and Fig. 3b show the learned attention masks for the 5 test images with pretrained and finetuned ResNet50, respectively. Attention heatmaps are used to present the attention masks, where the brighter the areas are, the more attention being paid to. From learned attention masks (both pretrained and finetuned cases) we can see that the important areas that describe the key objects in the images have received more attention, indicating that RAP can effectively learn good attention masks for 2D images.

V. Conclusions



(a) CIFAR10 test images

(b) Attention masks(pretrained)

(c) Attention masks(finetuned)

Fig. 3: RAP visualization examples with CIFAR10.

This paper presents Robust Attentional Pooling (RAP), a novel pluggable network module for feature selection in 3D and 2D visual recognition. RAP is characterized with a feature adaptation module, an attention learning module and a feature selection module. Attention learning module learns a data-dependent attention vector over transformed features output from feature adaptation module. Features are then selected in the feature selection module by sorting and selecting sparse attention weights and corresponding features, based on which a single weighted feature vector is generated for classification. Extensive experiments have been conducted for point cloud classification and image classification by plugging RAP to PointNet and ResNet50. Compared with conventional pooling functions in PointNet and ResNet, RAP is an attentional pooling module. Experiment results and visualizations validate the capability of RAP in effectively learning attention vectors and significantly compressing data features while maintaining good performance. When the sparsity of attention weights is high, RAP enhances the recognition performance in both tasks and outperforms PointNet and ResNet50, achieving much better classification performance.

References

[1] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *arXiv preprint arXiv:1612.00593*, 2016.

[4] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.

[5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[6] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," *arXiv preprint arXiv:1612.01887*, 2016.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[9] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[10] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 451–466.

[11] E. Shtrom, G. Leifman, and A. Tal, "Saliency detection in large point sets," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3591–3598.

[12] J. Leroy, N. Riche, M. Mancas, and B. Gosselin, "3d saliency based on supervoxels rarity in point clouds," *Hamburg, Germany*, 2015.

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

[14] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," *arXiv preprint arXiv:1301.3557*, 2013.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[18] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[19] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2956–2964.

[20] M. Zanfir, E. Marinoiu, and C. Sminchisescu, "Spatio-temporal attention models for grounded video captioning," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 104–119.

[21] C. Hori, T. Hori, T.-Y. Lee, K. Sumi, J. R. Hershey, and T. K. Marks, "Attention-based multimodal fusion for video description," *arXiv preprint arXiv:1701.03126*, 2017.

[22] X. Chen, A. Saparov, B. Pang, and T. Funkhouser, "Schelling points on 3d surface meshes," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, p. 29, 2012.

[23] R. Gal and D. Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 1, pp. 130–150, 2006.

[24] Y.-S. Liu, M. Liu, D. Kihara, and K. Ramani, "Salient critical points for meshes," in *Proceedings of the 2007 ACM symposium on Solid and physical modeling*. ACM, 2007, pp. 277–282.

[25] O. Akman and P. Jonker, "Computing saliency map from spatial information in point cloud data," in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2010, pp. 290–299.

[26] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 922–928.

[27] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[28] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC bioinformatics*, vol. 6, no. 1, p. 76, 2005.

[29] J. Zheng, W. Yang, and X. Li, "Training data reduction in deep neural networks with partial mutual information based feature selection and correlation matching based active learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2362–2366.

[30] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[31] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE transactions on neural networks and learning systems*, 2017.

[32] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2003, pp. 577–584.

[33] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *JMLR*, vol. 12, no. Jul, pp. 2211–2268, 2011.

[34] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[35] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *NIPS*, 2017, pp. 33–44.

[36] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[37] Y. Shen, C. Feng, Y. Yang, and D. Tian, "Neighbors do help: Deeply exploiting local structures of point clouds," *arXiv preprint arXiv:1712.06760*, 2017.

[38] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[39] L. Yi, V. G. Kim, D. Ceylan, I. Shen, M. Yan, H. Su, A. Lu, Q. Huang, A. Sheffer, L. Guibas *et al.*, "A scalable active framework for region annotation in 3d shape collections," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 210, 2016.

[40] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.