

MoNet: Moments Embedding Network

Mengran Gou¹ Fei Xiong² Octavia Camps¹ Mario Sznaiier¹

¹Electrical and Computer Engineering, Northeastern University, Boston, MA, US

²Information Sciences Institute, USC, CA, US

{mengran, camps, msznaiier}@coe.neu.edu

feixiong@ads.isi.edu

Abstract

Bilinear pooling has been recently proposed as a feature encoding layer, which can be used after the convolutional layers of a deep network, to improve performance in multiple vision tasks. Instead of conventional global average pooling or fully connected layer, bilinear pooling gathers 2nd order information in a translation invariant fashion. However, a serious drawback of this family of pooling layers is their dimensionality explosion. Approximate pooling methods with compact property have been explored towards resolving this weakness. Additionally, recent results have shown that significant performance gains can be achieved by using matrix normalization to regularize unstable higher order information. However, combining compact pooling with matrix normalization has not been explored until now.

In this paper, we unify the bilinear pooling layer and the global Gaussian embedding layer through the empirical moment matrix. In addition, with a proposed novel sub-matrix square-root layer, one can normalize the output of the convolution layer directly and mitigate the dimensionality problem with off-the-shelf compact pooling methods. Our experiments on three widely used fine-grained classification datasets illustrate that our proposed architecture MoNet can achieve similar or better performance than G²DeNet. When combined with compact pooling technique, it obtains comparable performance with the encoded feature of 96% less dimensions.

1. Introduction

Embedding local representations of an image to form a feature that is representative yet invariant to nuance noise is a key step in many computer vision tasks. Before the phenomenal success of deep convolutional neural networks (CNN) [18], researchers tackled this problem with hand-crafted consecutive independent steps. Remarkable works include HOG [6], SIFT [24], covariance descriptor [30], VLAD [14], Fisher vector [26] and bilinear pooling [3]. Al-

Table 1. Comparison of 2nd order statistic information based neural networks. Bilinear CNN (BCNN) only has 2nd order information and does not use matrix normalization. Both improved BCNN (iBCNN) and G²DeNet take advantage of matrix normalization but suffer from large dimensionality because they use the square-root of a large pooled matrix. Our proposed MoNet, with the help of a novel sub-matrix square-root layer, can normalize the local features directly and reduce the final representation dimension significantly by substituting the fully bilinear pooling with compact pooling.

	Ist order moment	Matrix normalization	Compact capacity
BCNN [22, 9]	✗	✗	✓
iBCNN [21]	✗	✓	✗
G ² DeNet [33]	✓	✓	✗
MoNet	✓	✓	✓

though CNNs are trained from end to end, they can be also viewed as two parts, where the convolutional layers are feature extraction steps and the later fully connected (FC) layers are an encoding step. Several works have been done to explore substituting the FC layers with conventional embedding methods in both two-stage fashion [4, 11] and end-to-end trainable way [22, 13].

Bilinear CNN was first proposed by Lin *et al.* [22] to pool the second order statistics information across the spatial locations. Bilinear pooling has been proven to be successful in many tasks, including fine-grained image classification [16, 9], large-scale image recognition [20], segmentation [13], visual question answering [8, 34], face recognition [28] and artistic style reconstruction [10]. Wang *et al.* [33] proposed to also include the 1st order information by using a Gaussian embedding. It has been shown that the normalization method is also critical to these CNNs performance. Two normalization methods have been proposed for the bilinear pooled matrix, $M = \frac{1}{n}X^T X$, where $X \in \mathbb{R}^{n \times C}$ represents the local features. On one hand,

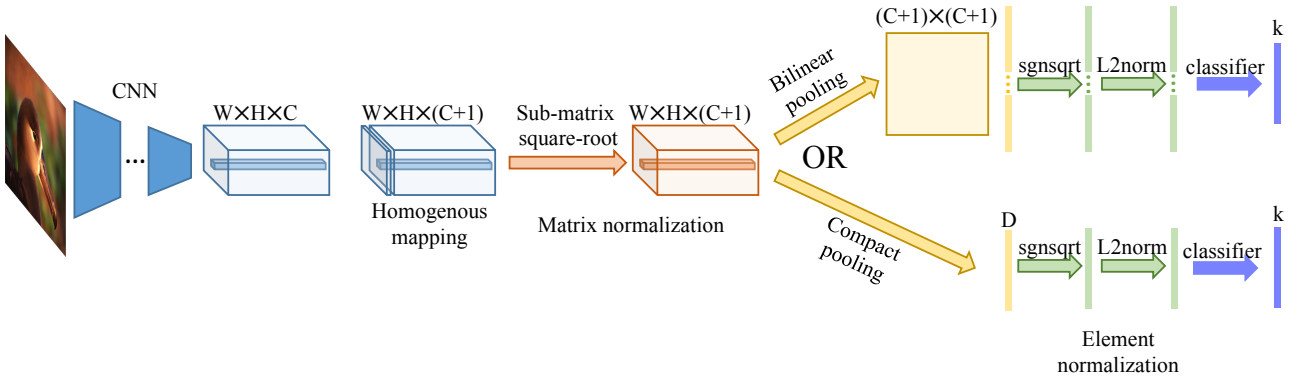


Figure 1. Architecture of the proposed moments-based network **MoNet**. With the proposed sub-matrix square-root layer, it is possible to perform matrix normalization before bilinear pooling or further apply compact pooling to reduce the dimensionality dramatically without undermining performance.

because \mathbf{M} is Symmetric Positive Definite (SPD), Ionescu *et al.* [13] proposed to apply matrix-logarithm to map the SPD matrices from the Riemannian manifold to an Euclidean space following $\log(\mathbf{M}) = \mathbf{U}_M \log(\mathbf{S}_M) \mathbf{U}_M^T$ with $\mathbf{M} = \mathbf{U}_M \mathbf{S}_M \mathbf{U}_M^T$. On the other hand, Wang *et al.* [33, 21] proposed matrix-power to scale \mathbf{M} non-linearly with $\mathbf{M}^p = \mathbf{U}_M \mathbf{S}_M^p \mathbf{U}_M^T$. In both works, matrix-power was shown to have better performance and numerical stability than the matrix-logarithm. In addition, Li *et al.* [20] provided theoretical support on the superior performance of matrix-power normalization in solving a general large-scale image recognition problem. Therefore, we propose to also integrate the matrix-power normalization into our MoNet architecture.

A critical weakness of the above feature encoding is the extremely high dimensionality of the encoded features. Due to the tensor product¹, the final feature dimension is C^2 where C is the number of feature channels of the last convolution layer. Even for relatively low $C = 512$ as in VGG-16 [29], the dimensionality of the final feature is already more than $260K$. This problem can be alleviated by using random projections [9], tensor sketching [9, 5], and the low rank property [16]. However, because the matrix-power normalization layer is applied on the pooled matrix \mathbf{M} , it is difficult to combine matrix normalization and compact pooling to achieve better performance and reduce the final feature dimensions at the same time.

In this paper, we re-write the formulation of G^2 DeNet using the tensor product of homogeneous padded local features to align it with the architecture of BCNN so that the Gaussian embedding operation and the bilinear pooling are decoupled. Instead of working on the bilinear pooled matrix \mathbf{M} , we derive the sub-matrix square-root layer to perform the matrix-power normalization directly on the (in-)homogeneous local features. With the help of this novel

layer, we can take advantage of compact pooling to approximate the tensor product but with much fewer dimensions.

The main contributions of this work are three-fold:

- We unify the G^2 DeNet and bilinear pooling CNN using the empirical moment matrix and decouple the Gaussian embedding from the bilinear pooling.
- We propose a new sub-matrix square-root layer to directly normalize the features before bilinear pooling layer, which makes it possible to reduce the dimensionality of the representation using compact pooling.
- We derive the gradient of the proposed layer using matrix back propagation so that the whole proposed *moments-based network* “**MoNet**” architecture can be optimized jointly.

2. Related work

Lasserre *et al.* [19] proposed to use the empirical moment matrix formed by explicit in-homogeneous polynomial kernel basis for outlier detection. In [12], the empirical moments matrix was applied as a feature embedding method for the person re-identification problem and it was shown that the Gaussian embedding [23] is a special case when the moment matrix order equals to 1. However, both of these works focus on a conventional pipeline and did not bring it to modern CNN architectures.

Ionescu *et al.* [13] introduced the theory and practice of matrix back-propagation for training CNNs, which enable structured matrix operations in deep neural networks training. Both [21] and [33] used it to derive the back-propagation of the matrix square-root and matrix logarithm for a symmetric matrix. Li *et al.* [20] applied a generalized p -th order matrix power normalization instead of the square-root. In our case, since we want to apply the matrix normalization directly on a non-square local feature ma-

¹We show that the Gaussian embedding can be written as a tensor product in sec. 3.2.1

trix, we cannot plug-in the equation directly from previous works.

Low dimension compact approximation of bilinear pooling has been explored recently. Gao *et al.* [9] bridged the bilinear pooling and a linear classifier with a second order polynomial kernel and adopted the off-the-shelf kernel approximation methods Random Maclaurin [15] and Tensor Sketch [27] to pool the local features in a compact way. Cui [5] generalized this approach to higher order polynomials with Tensor Sketch. By combining with bilinear SVM, Kong *et al.* [16] proposed to impose a low-rank constrain to reduce the number of parameters. However, none of these approaches can be easily integrated with matrix normalization because of the absence of a bilinear pooled matrix.

3. MoNet

The overview of the architecture of the proposed MoNet network is shown in Fig. 1. In this section, we will detail the design for each block.

For an input image \mathbf{I} , the output of the last convolution layer after the ReLU \mathbf{X} , consists of local features \mathbf{x}_i , across spatial locations $i = 1, 2, \dots, n$. Then, we map them to homogeneous coordinates by padding with an extra dimension with 1 and divide all elements by \sqrt{n} . After that, a proper sub-matrix square-root normalization is applied. Finally, a compact bilinear pooling layer pools all n features across all spatial locations, followed by element-wise square-root regularization and ℓ_2 normalization before the final fully-connected layer.

3.1. Homogeneous mapping layer

Assume $\mathbf{X} \in \mathbb{R}^{n \times C}$ corresponding to n features with dimension C and $n > C$. The homogeneous mapping of \mathbf{X} is nothing but padding an extra dimension 1. For the simplicity of the following layers, instead of applying the conventional bilinear pooling layer as in [22], we also divide the homogeneous feature by the square-root of the number of samples. Therefore, we have the forward equation of the homogeneous mapping layer as

$$\tilde{\mathbf{X}} = \frac{1}{\sqrt{n}}[\mathbf{1}|\mathbf{X}] \in \mathbb{R}^{n \times (C+1)} \quad (1)$$

The tensor product of $\tilde{\mathbf{X}}$ can be written as

$$\mathbf{M} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{1} & \\ \mu^T & \frac{\mu}{n} \mathbf{X}^T \mathbf{X} \end{bmatrix} \quad (2)$$

where $\mu = \frac{1}{n} \sum_1^n \mathbf{X}$. Because of $\frac{1}{n} \mathbf{X} \mathbf{X}^T = \Sigma + \mu \mu^T$, Eq. 2 is the Gaussian embedding method used in G²DeNet [33]. One can also show that the conventional bilinear pooling layer is equal to the tensor product of the in-homogeneous feature matrix.

3.2. Sub-matrix square-root layer

3.2.1 Forward propagation

Since both works [33, 21] showed with extensive experiments that the matrix square-root normalization is better than the matrix logarithm for performance and training stability, we also choose the matrix square-root normalization of \mathbf{M} :

$$\tilde{\mathbf{M}} = \mathbf{U}_M \mathbf{S}_M^{\frac{1}{2}} \mathbf{U}_M^T \quad (3)$$

where $\mathbf{M} = \mathbf{U}_M \mathbf{S}_M \mathbf{U}_M^T$. Since we have

$$\mathbf{M} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \mathbf{V} \mathbf{S}^T \mathbf{U}^T \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (4)$$

where $\tilde{\mathbf{X}} = \mathbf{U} \mathbf{S} \mathbf{V}^T$. Given $\mathbf{U}^T \mathbf{U} = \mathbf{I}$ and $\mathbf{S}^T \mathbf{S}$ is a square matrix, we can re-write Eq. 3 w.r.t. to $\tilde{\mathbf{X}}$ as

$$\tilde{\mathbf{M}} = \mathbf{V} (\mathbf{S}^T \mathbf{S})^{\frac{1}{2}} \mathbf{V}^T \quad (5)$$

Since $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times (C+1)}$, $n > C + 1$, we can decompose \mathbf{S} as

$$\mathbf{S} = \mathbf{A} \tilde{\mathbf{S}}, \mathbf{A} = [\mathbf{I}|\mathbf{0}]^T \quad (6)$$

where $\tilde{\mathbf{S}} \in \mathbb{R}^{(C+1) \times (C+1)}$ is a square diagonal matrix. Since we want the output of this layer \mathbf{Y} to satisfy $\tilde{\mathbf{M}} = \mathbf{Y}^T \mathbf{Y}$, we can set $\mathbf{Y} = \mathbf{A} \tilde{\mathbf{S}} \mathbf{V}^T$. Because for most modern CNNs, n cannot be much greater than C and the feature after ReLU tends to be sparse, $\tilde{\mathbf{X}}$ can easily go to rank deficient. Therefore, we only use the non-zero singular values and singular vectors. Then, the forward equation of the sub-matrix square-root layer can be written as

$$\mathbf{Y} = \mathbf{A}_{:,1:e} \tilde{\mathbf{S}}_{1:e}^{0.5} \mathbf{V}_{:,1:e}^T \quad (7)$$

where e is the index of the smallest singular value greater than ϵ^2 .

3.2.2 Backward propagation

We will follow the matrix back propagation techniques proposed by Ionescu *et al.* [13] to derive the equation of the back propagation path for the sub-matrix square-root layer.

Let $\tilde{\mathbf{X}} = \mathbf{U} \mathbf{S} \mathbf{V}^T$ and $\mathbf{U} \in \mathbb{R}^{n \times n}$. We can form \mathbf{U} using block decomposition as $\mathbf{U} = [\mathbf{U}_1|\mathbf{U}_2]$ with $\mathbf{U}_1 \in \mathbb{R}^{n \times (C+1)}$ and $\mathbf{U}_2 \in \mathbb{R}^{n \times (n-C-1)}$. The partial derivatives between a given scalar loss L and $\tilde{\mathbf{X}}$ are

$$\begin{aligned} \frac{\partial L}{\partial \tilde{\mathbf{X}}} &= \mathbf{D} \mathbf{V}^T + \mathbf{U} \left(\frac{\partial L}{\partial \mathbf{S}} - \mathbf{U}^T \mathbf{D} \right)_{diag} \mathbf{V}^T + \\ &2 \mathbf{U} \mathbf{S} (\mathbf{K}^T \circ \left(\mathbf{V}^T \left(\frac{\partial L}{\partial \mathbf{V}} - \mathbf{V} \mathbf{D}^T \mathbf{U} \mathbf{S} \right) \right)_{sym} \mathbf{V}^T \end{aligned} \quad (8)$$

²We will omit the subscript in the following for a concise notation

where \circ represents element-wise product, $(\mathbf{Q})_{sym} \doteq \frac{1}{2}(\mathbf{Q}^T + \mathbf{Q})$ and

$$\mathbf{D} = \left(\frac{\partial L}{\partial \mathbf{U}} \right)_1 \tilde{\mathbf{S}}^{-1} - \mathbf{U}_2 \left(\frac{\partial L}{\partial \mathbf{U}} \right)_2^T \mathbf{U}_1 \tilde{\mathbf{S}}^{-1} \quad (9)$$

$$\mathbf{K}_{ij} = \begin{cases} \frac{1}{s_i^2 - s_j^2} & i \neq j \\ 0 & i = j \end{cases} \quad (10)$$

From Eq. 7, we can compute the variation of \mathbf{Y} as

$$d\mathbf{Y} = \frac{1}{2} \mathbf{A} \tilde{\mathbf{S}}^{-\frac{1}{2}} d\tilde{\mathbf{S}} \mathbf{V}^T + \mathbf{A} \tilde{\mathbf{S}}^{\frac{1}{2}} d\mathbf{V}^T \quad (11)$$

Based on the chain rule, the total variation can be written as

$$\frac{\partial L}{\partial \mathbf{Y}} : d\mathbf{Y} = \frac{1}{2} \frac{\partial L}{\partial \mathbf{Y}} : \mathbf{A} \tilde{\mathbf{S}}^{-\frac{1}{2}} d\tilde{\mathbf{S}} \mathbf{V}^T + \frac{\partial L}{\partial \mathbf{Y}} : \mathbf{A} \tilde{\mathbf{S}}^{\frac{1}{2}} d\mathbf{V}^T \quad (12)$$

where $:$ denotes the inner-product. After re-arrangement with the rotation properties of inner-product, we re-write the above equation as

$$\frac{\partial L}{\partial \mathbf{Y}} : d\mathbf{Y} = \frac{1}{2} \tilde{\mathbf{S}}^{-\frac{1}{2}} \mathbf{A}^T \frac{\partial L}{\partial \mathbf{Y}} \mathbf{V} : d\tilde{\mathbf{S}} + \tilde{\mathbf{S}}^{\frac{1}{2}} \mathbf{A}^T \frac{\partial L}{\partial \mathbf{Y}} : d\mathbf{V}^T \quad (13)$$

Therefore, we have

$$\frac{\partial L}{\partial \tilde{\mathbf{S}}} = \mathbf{A} \frac{\partial L}{\partial \tilde{\mathbf{S}}} = \frac{1}{2} \mathbf{A} \tilde{\mathbf{S}}^{-\frac{1}{2}} \mathbf{A}^T \frac{\partial L}{\partial \mathbf{Y}} \mathbf{V} \quad (14)$$

$$\frac{\partial L}{\partial \mathbf{V}} = \left(\frac{\partial L}{\partial \mathbf{V}^T} \right)^T = \left(\frac{\partial L}{\partial \mathbf{Y}} \right)^T \mathbf{A} \tilde{\mathbf{S}}^{\frac{1}{2}} \quad (15)$$

Finally, substituting Eq. 14 and Eq. 15 into Eq. 8 and considering $\frac{\partial L}{\partial \mathbf{U}} = 0$, we have

$$\begin{aligned} \frac{\partial L}{\partial \tilde{\mathbf{X}}} = & \mathbf{U} \left(\frac{1}{2} \mathbf{A} \tilde{\mathbf{S}}^{-\frac{1}{2}} \mathbf{A}^T \frac{\partial L}{\partial \mathbf{Y}} \mathbf{V} + \right. \\ & \left. 2\mathbf{S} \left[\mathbf{K}^T \circ \left(\mathbf{V}^T \left(\frac{\partial L}{\partial \mathbf{Y}} \right)^T \mathbf{A} \tilde{\mathbf{S}}^{\frac{1}{2}} \right) \right]_{sym} \right) \mathbf{V}^T \end{aligned} \quad (16)$$

3.3. Compact pooling

Following the work in [9, 8], we adopt the Tensor Sketch (TS) method to approximate bilinear pooling, followed by a linear classifier because it has better performance and needs less computational time and space. Building up on count sketch and FFT, one can generate a tensor sketch function s.t. $\langle TS_1(x), TS_2(y) \rangle \approx \langle x, y \rangle^2$.

The back-propagation of a TS layer is given by in [9].

As shown in Tab. 2, with the techniques mentioned above, the proposed MoNet is capable to solve the problem with much less computation and memory complexity than the other BCNN based algorithms.

Algorithm 1 Tensor Sketch approximation pooling

Require: x , projected dimension D

- 1: Generate randomly picked but fixed two pair hash functions $h_t \in \mathbb{R}^D$ and $s_t \in \mathbb{R}^D$ where $t = 1, 2$ and $h_t(i), s_t(i)$ are uniformly drawn from $\{1, 2, \dots, D\}$ and $\{-1, +1\}$, respectively
 - 2: Define count sketch function $\Psi(x, h_t, s_t) = [\psi_1(x), \psi_2(x), \dots, \psi_D(x)]^T$ where $\psi_j(x) = \sum_{i: h_t(i)=j} s_t(i)x_i$
 - 3: Define $TS(x) = FFT^{-1}(FFT(\Psi(x, h_1, s_1) \circ (\Psi(x, h_2, s_2))))$ where \circ denotes element-wise multiply
-



Figure 2. Sample images from the fine-grained classification datasets. From left to right, each column corresponds to CUB, Aircraft and Cars, respectively.

4. Experiments

Aligned with other bilinear CNN based papers, we also evaluate the proposed MoNet with three widely used fine-grained classification datasets. The experimental setups and the algorithm implementation are described in detail in Sec. 4.1.2. Then, in Sec. 4.1.1, the experiment results on fine-grained classification are presented and analyzed.

4.1. Experimental setup

We evaluated MoNet on three widely used fine-grained classification datasets. Different from general objection recognition tasks, fine-grained classification usually tries to distinguish objects at the sub-category level, such as different makes of cars or different species of a bird. The main challenge of this task is the relatively large inter-class and relatively small intra-class variations.

In all experiments, the 13 convolutional layers of VGG-16 [29] are used as the local feature extractor, and their outputs are used as local appearance representations. These 13 convolution layers are trained with ImageNet [7] and fine tuned in our experiments with three fine-grained classifica-

Table 2. Dimension, computation and memory information for different network architectures we compared in this paper. H, W and C represent the height, width and number of feature channels for the output of the final convolution layer, respectively. k and D denote the number of classes and projected dimensions for Tensor Sketch, respectively. Numbers inside brackets indicate the typical value when the corresponding network was evaluated with VGG-16 model [29] on a classification task with 1,000 classes. In this case, $H = W = 13, C = 512, k = 1000, D = 10000$ and all data was stored with single precision.

	BCNN [22]	iBCNN [21]	iBCNN TS	G ² DeNet [33]	MoNet	MoNet TS
Dimension	C^2 [262K]	C^2 [262K]	D [10K]	$(C+1)^2$ [263K]	$(C+1)^2$ [263k]	D [10k]
Parameter Memory	0	0	$2C$	0	0	$2C$
Computation	$O(HWC^2)$	$O(HWC^2)$	$O(HW(C + D \log D))$	$O(HWC^2)$	$O(HWC^2)$	$O(HW(C + D \log D))$
Classifier Memory	kC^2 [1000MB]	kC^2 [1000MB]	kD [40MB]	$k(C + 1)^2$ [1004MB]	$k(C + 1)^2$ [1004MB]	kD [40MB]

Table 3. Basic statistics of the datasets used for evaluation

Datasets	# training	# testing	# classes
CUB [32]	5,994	5,794	200
Aircraft [25]	6,667	3,333	100
Cars [17]	8,144	8,041	196

tion datasets.

4.1.1 Datasets

Caltech-UCSD birds (CUB) [32] contains 200 species, mostly north-American, of birds. Being consistent with other works, we also use the 2011 extension with doubled number samples.

FGVC-Aircraft Benchmark (Aircraft) [25] is a benchmark fine-grained classification dataset with different aircrafts with various models and manufacturers.

Stanford cars (Cars) [17] contains images of different classes of cars at the level of make, model and year.

We use the provided train/test splits for all three datasets. Detailed information is given in Tab. 3 and Fig. 2 shows sample images.

4.1.2 Different pooling methods

Bilinear pooling (BCNN): The VGG-16 based BCNN [22] is utilized as the baseline pooling method, which applies the tensor product on the output of the conv_{5,3} layer with ReLU activation. The dimension of the final representation is $512 \times 512 \approx 262K$ and the number of the linear classifier parameters is $k \times 262K$, where k is the number of classes. To be fair, the latest results from the authors’ project page [1] are compared.

Improved bilinear pooling (iBCNN): In order to compare the iBCNN [21] with compact pooling technique, it is re-implemented with the sub-matrix square-root layer followed by the bilinear pooling layer. As explained in Sec.3.2.1 it is equivalent to the originally proposed bilinear pooling and matrix normalization layers [21]. Note that, the originally proposed iBCNN is without compact pooling and has the same dimensionality as BCNN. The classification performance from both the paper [21] and our re-implementation are reported.

Global Gaussian distribution embedding (G²DeNet):

Instead of fully bilinear pooling, G²DeNet pools the local features with a global Gaussian distribution embedding method, followed by a matrix square-root normalization. Since it includes the first order moment information, the dimension of the final feature is slightly greater than BCNN and iBCNN. The experiment results with “w/o BBox” configuration [33] are compared in this paper.

Proposed moment embedding net (MoNet): We implemented the proposed MoNet algorithm with structure as shown in Fig. 1 and fine-tuned the whole network in an end-to-end fashion. When using bilinear pooling, the feature dimensionality, computation and memory complexity are the same as G²DeNet.

Tensor Sketch compact pooling (TS): When building the network with compact pooling, the TS layer [9] is applied after the sub-matrix square-root layer. The projection dimension D is selected empirically for both iBCNN and MoNet.

4.1.3 Implementation details

Since the large enough number of samples is important to estimate stable and meaningful statistic moment information. The input images are resized to 448×448 in all the experiments, which produces a $28 \times 28 \times 512$ local feature matrix after conv_{5,3} for each image. Following common practice [33, 5], we first resize the image with a fixed aspect-ratio, such as the shorter edge reaches to 448 and then utilized a center crop to resize the image to 448×448 . During training, random horizontal flipping was applied as data augmentation. Different from [21] with VGG-M, no augmentation is applied during testing.

To avoid rank deficiency, the singular value threshold σ was set to 10^{-5} for both forward and backward propagation, which results in 10^{-10} for the singular value threshold of the tensor product matrix. The projected dimension in Tensor Sketch was fixed to 10^4 , which satisfies $C < D \ll C^2$.

As suggested by [21, 33], all pooling methods were followed by an element-wise sign kept square-root $\mathbf{y} = \text{sign}(\mathbf{y})\sqrt{\mathbf{y}}$ and ℓ_2 normalization $\mathbf{y} = \mathbf{y}/\|\mathbf{y}\|$. For the sake of a smooth training, the element-wise square-root [33] is also applied on local appearance features.

Table 4. Experiment results on fine-grained classification. Bilinear and TS represent fully bilinear pooling and Tensor Sketch compact pooling respectively. The best performance in each column is marked in red.

		CUB		Airplane		Car	
		Bilinear	TS	Bilinear	TS	Bilinear	TS
BCNN [22, 9]		84.0	84.0	86.9	87.2	90.6	90.2
iBCNN [21]		85.8	-	88.5	-	92.0	-
iBCNN re-implementation		86.0	85.7	86.7	86.7	90.5	90.3
G ² DeNet [33]		87.1	-	89.0	-	92.5	-
MoNet		86.4	85.7	89.3	88.1	91.8	90.8
Other higher order methods	KP [5]	-	86.2	-	86.9	-	92.6
	HOHC [2]		85.3		88.3		91.7
State-of-the-art	MA-CNN [35]		86.5		89.9		92.8

The weights of the VGG-16 convolutional layers are pretrained on ImageNet classification dataset. We first warm-started by fine-tuning the last linear classifier for 300 epochs. Then, we fine-tuned the whole network end-to-end with the learning rate as 0.001 and batch size as 16. The momentum was set to 0.9 and the weight decay was set to 0.0005. Most experiments converge to local optimum after 50 epochs.

The proposed MoNet is implemented with MatConvNet [31] and Matlab 2017a. Because of the numerical instability of SVDs, as suggested by Ionescu *et al.* [13], the sub-matrix square-root layer is implemented on CPU with double precision. The whole network is fine-tuned on a Ubuntu PC with 64GB RAM and Nvidia GTX 1080 Ti.

4.1.4 Experimental results

As shown in Tab. 4, the classification accuracy of each network are presented in a row. Bilinear and TS denote fully bilinear pooling and tensor sketch compact pooling respectively.

Comparison with different architectures: Based on [21], matrix normalization consistently improves the performance by 1-2% on all three datasets. Our re-implementation achieves slightly better classification accuracy (0.2%) on CUB dataset but performs worse on Airplane and Car datasets. We believe it is caused by the different approaches used to deal with rank deficiency. In our implementation, the singular value is hard thresholded as shown in Eq. 7, while iBCNN [21] deal with the rank deficiency by adding 1 to all the singular values, which is a relatively very small number comparing to the maximum singular value (10^6). By adding the 1st order moment information, G²DeNet outperforms iBCNN by around 1%, on all three datasets. By re-writing the Gaussian embedding with tensor product of homogeneous padded local features, our proposed MoNet can obtain similar or slightly better classification accuracy when comparing with G²DeNet. Specifically, the classification accuracy of MoNet is 0.3% higher on Airplane dataset, but 0.7% lower on both CUB and Car

datasets.

Comparison with fully bilinear pooling and compact pooling:

As shown in [9], compact pooling can achieve similar performance compared to BCNN, but with only 4% of the dimensionality. We also see a similar trend on re-implemented iBCNN network. The classification accuracy difference between the re-implemented iBCNN with its compact pooling version is less than 0.3% on all three datasets. However, the performance gaps are relatively greater when we compare the different pooling schemes on MoNet. Bilinear pooling improve the classification accuracy by 0.7%, 1.2% and 1% than compact pooling on CUB, Airplane and Car datasets, respectively. However, with compact pooling, the dimensionality of the final representation is 96% smaller. Although the final fully bilinear pooled representation dimensions of iBCNN and MoNet are roughly the same, MoNet utilizes more different order moments, which requires more count sketch projections to approximate it. Thus when fixing $D = 10,000$ for both iBCNN and MoNet, the performance of MoNet with compact pooling is degraded.

Comparison with other methods:

[5] and [2] are two other recent works that also take into account higher order statistic information. Cui *et al.* [5] applied Tensor Sketch repetitively to approximate up to 4th order explicit polynomial kernel space in a compact way. They obtained better results for CUB and Car datasets compared against other compact pooling results, but notably worse (1.2%) on the Airplane dataset. It may result from two factors. First, directly utilizing higher order moments without proper normalization leads to numerical instabilities. Second, approximating higher order moments with limited number of samples is essentially an ill-posed problem. Cai *et al.* [2] only utilize higher order self-product terms but not the interaction terms, which leads to worse performance in all three datasets. Finally, the state-of-the-art MA-CNN [35] achieves slightly better results on Airplane and Car datasets.

5. Conclusion

Bilinear pooling, as a recently proposed 2nd order moment pooling method, has been shown success in several vision tasks. G²DeNet extends it by adding 1st order moment information and matrix normalization. One key limitation of these approaches is the high dimension of the final representation. To resolve this, compact pooling methods have been proposed to approximate the bilinear pooling. However, the Gaussian embedding formation entangles the bilinear pooling and makes the compact pooling utilization non-trivial. In this paper, we reformulated the Gaussian embedding using the empirical moment matrix and decoupled the bilinear pooling step out. With the help of a novel sub-matrix square-root layer, our proposed network MoNet can take advantages among different order moments, matrix normalization as well as the compact pooling. Experiments on three widely used fine-grained classification datasets demonstrate that MoNet can achieve similar or better performance when comparing with G²DeNet and retain comparable results with only a 4% dimension number by compact pooling.

References

- [1] Bilinear CNNs project. <http://vis-www.cs.umass.edu/bcnn/>. Accessed: 2017-11-14. **5**
- [2] S. Cai, W. Zuo, and L. Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–520, 2017. **6**
- [3] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. *Computer Vision—ECCV 2012*, pages 430–443, 2012. **1**
- [4] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3828–3836, 2015. **1**
- [5] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie. Kernel pooling for convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. **2, 3, 5, 6**
- [6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. **1**
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. **4**
- [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. **1, 4**
- [9] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016. **1, 2, 3, 4, 5, 6**
- [10] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. **1**
- [11] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014. **1**
- [12] M. Gou, O. Camps, M. Szanier, et al. mom: Mean of moments feature for person re-identification. **2**
- [13] C. Ionescu, O. Vantzos, and C. Sminchisescu. Matrix back-propagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2965–2973, 2015. **1, 2, 3, 6**
- [14] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010. **1**
- [15] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *International conference on artificial intelligence and statistics*, pages 583–591, 2012. **3**
- [16] S. Kong and C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. **1, 2, 3**
- [17] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. **5**
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. **1**
- [19] J.-B. Lasserre and E. Pauwels. Sorting out typicality with the inverse moment matrix sos polynomial. In *Neural Information Processing Systems (NIPS 2016)*, 2016. **2**
- [20] P. Li, J. Xie, Q. Wang, and W. Zuo. Is second-order information helpful for large-scale visual recognition? In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. **1, 2**
- [21] T.-Y. Lin and S. Maji. Improved bilinear pooling with cnns. In *BMVC*, 2017. **1, 2, 3, 5, 6**
- [22] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. **1, 3, 5, 6**
- [23] M. Lovrić, M. Min-Oo, and E. A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000. **2**
- [24] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. **1**
- [25] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. **5**
- [26] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision—ECCV 2010*, pages 143–156, 2010. **1**

- [27] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013. 3
- [28] A. RoyChowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller. Face identification with bilinear cnns. *arXiv preprint arXiv: 1506.01342*, 2015. 1
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4, 5
- [30] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1713–1727, 2008. 1
- [31] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015. 6
- [32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 5
- [33] Q. Wang, P. Li, and L. Zhang. G2denet: Global gaussian distribution embedding network and its application to visual recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3, 5, 6
- [34] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *arXiv preprint arXiv:1708.01471*, 2017. 1
- [35] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, 2017. 6