High-order Tensor Pooling with Attention for Action Recognition – Supplementary Material –

Lei Wang*,†,§ Ke Sun§,† Piotr Koniusz§,†
†Australian National University, §Data61/CSIRO

Below we provide the prerequisites on power normalization methods of high dimensional tensors.

A. Notations

$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 ... \times d_r}$ denotes r-order tensor. By default $r = 3$ meaning that $\boldsymbol{\mathcal{X}}$ is a third-order tensor with exceptions depending on the context. The i-th slice of this tensor is denoted as $\boldsymbol{\mathcal{X}}_{:,:,i}$, which is a $d_1 \times d_2$ matrix. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ and a vector $\mathbf{x} = (x_1, ..., x_{d_3}) \in \mathbb{R}^{d_3}$, $\boldsymbol{\mathcal{X}} = \mathbf{X} \uparrow \otimes \mathbf{x}$ gives a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, where the i-th slice of $\boldsymbol{\mathcal{X}}$ is given by $\mathbf{X} \cdot x_i$. A symmetric third-order tensor of rank one $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d \times d \times d}$ can be obtained from $\mathbf{x}$ as $\boldsymbol{\mathcal{X}} = \uparrow \otimes_3 \mathbf{x} \triangleq (\mathbf{x}\mathbf{x}^T) \uparrow \otimes \mathbf{x}$. $\boldsymbol{\mathcal{X}} = \uparrow \otimes_r \mathbf{x}$ means the r-order outer product of $\mathbf{x}$, where the $(i_1, ..., i_r)$-th coefficient of $\boldsymbol{\mathcal{X}}$ is given by $\mathcal{X}_{i_1,...,i_r} = x_{i_1} \cdot x_{i_2} ... \cdot x_{i_r}$. The Frobenius norm of a tensor $\boldsymbol{\mathcal{X}}$ is given by $\|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\sum_{i,j,k} \mathcal{X}_{ijk}^2}$, where $\mathcal{X}_{ijk}$ represents the ijk-th element of $\boldsymbol{\mathcal{X}}$. Similarly, the inner-product between two tensors $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ is given by $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{ijk} \mathcal{X}_{ijk} \mathcal{Y}_{ijk}$. The tensor product $\times_j$ in mode j between $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 ... \times d_r}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d'_1 \times d'_2 ... \times d'_{r^*}}$, where $d_j = d'_j$, is denoted as $\boldsymbol{\mathcal{X}} \times_j \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{d_1 ... \times d_{j-1} \times d_{j+1} ... \times d_r \times d'_1 ... \times d'_{j-1} \times d'_{j+1} ... \times d'_{r^*}}$. The $(i_1, ..., i_{j-1}, i_{j+1}, ..., i_r, i'_1, ..., i'_{j-1}, i'_{j+1}, ..., i'_{r^*})$-th coefficient of $\boldsymbol{\mathcal{X}} \times_j \boldsymbol{\mathcal{Y}}$ is given by $\sum_{i_j} \mathcal{X}_{i_1,...,i_j,...,i_r} \cdot \mathcal{Y}_{i'_1,...,i_j,...,i'_{r^*}}$. We denote the spaces of $d \times d$ Symmetric Positive Semi-Definite (SPSD) and Symmetric Positive Definite (SPD) matrices as $\mathcal{S}_+^d$ and $\mathcal{S}_{++}^d$, $\mathcal{I}_r$ is an index sequence $1, 2, ..., r$, $\mathbb{I}$ is the identity matrix, $\dagger$ is a Moore–Penrose inverse, $\{\boldsymbol{e}_i : i \in \mathcal{I}_d\}$ are the spanning bases of $\mathbb{R}^d$. Bold lowercase/uppercase letters denote vectors/matrices, bold uppercase mathcal letters denote tensors, and regular letters denote scalars.

B. Eigenvalue Power Normalization

The following proposition formalizes the notion of higher-order descriptors.

*This paper has been accepted for IEEE ICASSP 2024.

Proposition 1 ([3]). *Let $\boldsymbol{\Phi} \equiv \{\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N \in \mathbb{R}^d\}$ and $\boldsymbol{\Phi}^* \equiv \{\boldsymbol{\phi}_1^*, ..., \boldsymbol{\phi}_M^* \in \mathbb{R}^d\}$ be feature vectors extracted from two instances to classify,* e.g.*, video sequences, images, text documents,* etc. *Let $\boldsymbol{w} \in \mathbb{R}_+^N$, $\boldsymbol{w}^* \in \mathbb{R}_+^M$ be some non-negative weights and $\boldsymbol{\mu}, \boldsymbol{\mu}^* \in \mathbb{R}^d$ be the mean vectors of $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^*$, respectively. A linearization of the sum of polynomial kernels of degree r*

$$\langle \boldsymbol{\mathcal{X}}(\boldsymbol{\Phi}; \boldsymbol{w}, \boldsymbol{\mu}), \boldsymbol{\mathcal{X}}(\boldsymbol{\Phi}^*; \boldsymbol{w}^*, \boldsymbol{\mu}^*) \rangle = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} w_n^r w_m^{*r} \langle \boldsymbol{\phi}_n - \boldsymbol{\mu}, \boldsymbol{\phi}_m^* - \boldsymbol{\mu}^* \rangle^r, \tag{1}$$

yields the tensor feature map

$$\boldsymbol{\mathcal{X}}(\boldsymbol{\Phi}; \boldsymbol{w}, \boldsymbol{\mu}) = \frac{1}{N} \sum_{n=1}^{N} w_n^r \uparrow \otimes_r (\boldsymbol{\phi}_n - \boldsymbol{\mu}) \in \mathbb{R}^{d \times d ... \times d}. \tag{2}$$

$\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^*$ do not have to be zero-mean centered ($\boldsymbol{\mu} = \boldsymbol{\mu}^* = 0$) if one uses an auto-correlation matrix/tensor instead of covariance. The weights $\boldsymbol{w}$ and $\boldsymbol{w}^*$ can differ for each feature vector, *e.g.*, they may be the same within each group of feature vectors (image patch, video subsequence) but differ across different groups (patches of different sizes or subsequences of different lengths).

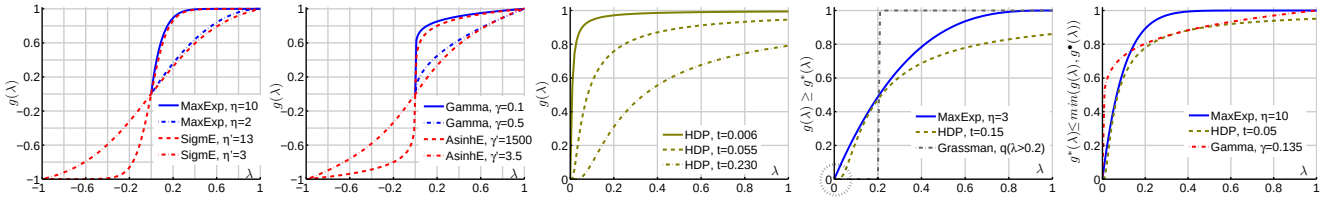
The EPN [4] performs a spectrum transformation on a given higher-order descriptor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times d_2 ... \times d_r}$, as detailed in the following steps

$$(\boldsymbol{\lambda}; \boldsymbol{U}_1, ..., \boldsymbol{U}_r) = \text{HOSVD}(\boldsymbol{\mathcal{X}}), \tag{3}$$

$$\hat{\boldsymbol{\lambda}} = g(\boldsymbol{\lambda}), \tag{4}$$

$$\boldsymbol{\mathcal{G}}(\boldsymbol{\mathcal{X}}) = ((\hat{\boldsymbol{\lambda}} \times_1 \boldsymbol{U}_1) ...) \times_r \boldsymbol{U}_r, \tag{5}$$

where HOSVD stands for Higher Order Singular Value Decomposition [7, 2], the small-case g acts on the so-called core tensor $\boldsymbol{\lambda} \in \mathbb{R}^{d'_1 \times d'_2 ... \times d'_r}$ in an element-wise manner, where $d'_i \leq d_i, \forall i$, and $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{d'_1 \times d'_2 ... \times d'_r}$ is the power-normalized counterpart of $\boldsymbol{\lambda}$. Moreover, $\{\boldsymbol{U}_i \in \mathbb{R}^{d_i \times d'_i}\}_{i \in \mathcal{I}_r}$ are r singular vector matrices. The uppercase mathcal notation indicates that $\boldsymbol{\mathcal{G}}$ is a spectrum-wise (*c.f.* element-wise) operator on $\boldsymbol{\mathcal{X}}$. As the input tensor $\boldsymbol{\mathcal{X}}$ is *super-symmetric* by Eq. (2), *i.e.*, we have $\boldsymbol{\mathcal{X}}_{i_1,i_2,...,i_r} =$



(a) MaxExp & SigmE (b) Gamma & AsinhE (c) HDP (d) Grassmann (e) Upper bounds

Figure 1: Different EPN functions and their profiles. Note the similarity of MaxExp ($\eta > 1$) from Fig. 1a, Gamma ($\gamma \in (0,1)$) from Fig. 1b and HDP ($t \in (0,1)$) from Fig. 1c to each other. Also, the circle in Fig. 1d highlights the region where the profile of HDP differs from MaxExp. Furthermore, note that the above EPN functions are soft approximations of the Grassmann map (the step function) in Fig. 1d. Finally, Fig. 1e shows that MaxExp and Gamma given by $g(\lambda)$ and $g^{\bullet}(\lambda)$ are upper bounds of HDP given by $g^{*}(\lambda)$ (see details in [5]).

$\boldsymbol{\mathcal{X}}_{\Pi(i_1,i_2,...,i_r)}$ for any indexes $(i_1,i_2,...,i_r)$ and any permutation Π, thus we have $\boldsymbol{U}_1 = \boldsymbol{U}_2 ... = \boldsymbol{U}_r$.

Popular variants of EPN pooling [6, 5] shown in Fig. 1, are given below, two for SPD/SPSD matrices and two for indefinite matrices (Krein spaces):

Gamma $g(\lambda;\gamma) = \lambda^{\gamma}$ (*e.g.*, $\lambda^{\frac{1}{2}}$), in matrix form, $\boldsymbol{\mathcal{G}}(\mathbf{X};\gamma) = \mathbf{X}^{\gamma}$ (*e.g.*, $\mathbf{X}^{\frac{1}{2}}$ *a.k.a.* matrix square root).

MaxExp $g(\lambda;\eta) = 1-(1-\lambda)^{\eta}$, in matrix form, $\boldsymbol{\mathcal{G}}(\mathbf{X};\eta) = \mathbb{I}-(\mathbb{I}-\mathbf{X})^{\eta}$.

AsinhE $g(\lambda;\gamma') = \mathrm{Arcsinh}(\gamma'\lambda)$ and $\boldsymbol{\mathcal{G}}(\mathbf{X};\gamma') = \mathrm{Log}\Big(\gamma'\mathbf{X} + (\mathbb{I}+\gamma'^2\mathbf{X}^2)^{\frac{1}{2}}\Big)$. AsinhE is the Arcus Hyperbolic Sine function.

SigmE $g(\lambda;\eta') = 2/(1+\mathrm{e}^{-\eta'\lambda}) - 1$ and $\boldsymbol{\mathcal{G}}(\mathbf{X};\eta') = 2\Big(\mathbb{I} + \mathrm{Exp}(-\eta'\mathbf{X})\Big)^{\dagger} - \mathbb{I}$. SigmE stands for a zero-centered Logistic *a.k.a.* Sigmoid function.

As shown in Figures 1a and 1b, SigmE/AsinhE extend Gamma/MaxExp to Krein spaces by reflecting function Gamma/MaxExp defined for non-negative eigenvalues λ by the vertical symmetry axis followed by the change of sign. Parameters $0 < \gamma \leq 1$, $\eta \geq 1$, $0 < \gamma' \leq 1$ and $\eta' \geq 1$ control effect/steepness of such non-linearities. Moreover, eigenvalues λ are typically normalized by $\sum_i |\lambda_i|$ and therefore $\forall i, -1 \leq \lambda_i \leq 1$. Figures 1c and 1d show the impact of PN by varying γ of Gamma. Figure 1d shows that Gamma performs whitening/evening out the spectrum of $\mathbf{X} \in \mathbb{R}^{d\times d}$.

The EPN induces a family of non-Euclidean distance $\|\boldsymbol{\mathcal{G}}(\boldsymbol{\mathcal{X}}) - \boldsymbol{\mathcal{G}}(\boldsymbol{\mathcal{Y}})\|_F$ in the SPSD/SPD cone. The Power-Euclidean (PowE) metric $\frac{1}{\gamma}||\mathbf{X}^{\gamma} - \mathbf{Y}^{\gamma}||_2$ is discussed by [1] who point out that as $\gamma \to 0$, the Power-Euclidean metric converges to the Log-Euclidean (LogE) metric $||\,\mathrm{Log}(\mathbf{X}) - \mathrm{Log}(\mathbf{Y})||_F$. Matrix Square Root (MSR) based distance is in fact close to the Cholesky-Euclidean (CholE) distance $||\,\mathrm{Chol}(\mathbf{X}) - \mathrm{Chol}(\mathbf{Y})||_F$ suggested by [1]. However, the best results for Power-Euclidean distance (whose underlying feature map is Gamma) are typically attained with $0 \ll \gamma \neq 0.5$ [8, 6], which means the above connections to the Log-Euclidean metric and 'robust covariance estimation' are somewhat loose. Gamma for element-wise matrix pooling (*c.f.* spectral/eigenvalue pooling) is connected [6] to an operator called MaxExp. Intuitively, MaxExp yields 'the probability of at least one co-occurrence event $(\phi_n \cap \phi'_n = 1)$ occurring in ϕ_n and ϕ'_n simultaneously, given $N \approx \eta$ Bernoulli trials and two event vectors $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \{0,1\}^N$. In fact, element-wise MaxExp/Gamma have similar profiles as Fig. 1 shows. As EPN whitens the eigenspectrum of signal, it also differs from the batch normalization of variance of gradient features.

References

[1] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009. 2

[2] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. 1

[3] P. Koniusz, Y. Tas, and F. Porikli. Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. In *CVPR*, pages 7139–7148, 2017. 1

[4] P. Koniusz, F. Yan, P. Gosselin, and K. Mikolajczyk. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *PAMI*, 2016. 1

[5] P. Koniusz and H. Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. In *TPAMI*. IEEE, 2020. 2

[6] P. Koniusz, H. Zhang, and F. Porikli. A deeper look at power normalizations. In *CVPR*, pages 5774–5783, 2018. 2

[7] L. D. Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Analysis and Applications*, 21:1253–1278, 2000. 1

[8] T.-Y. Lin and S. Maji. Improved Bilinear Pooling with CNNs. *BMVC*, 2017. 2