Fakultät IV NÜ Technische Universität Berlin

Erik Bochinski, Rolf Jongebloed, Michael Tok and Thomas Sikora

Regularized Gradient Descent Training of Steered Mixture of Experts for Sparse Image Representation

Prof. Dr.-Ing. Sikora

Introduction

The Steered Mixture-of-Experts (SMoE) framework targets a sparse space-continuous representation for images, videos, and light fields enabling processing tasks such as approximation, denoising, and coding. The underlying stochastic processes are represented by a Gaussian Mixture Model, traditionally trained by the Expectation-Maximization (EM) algorithm.

We instead propose to use the MSE of the regressed imagery for a Gradient Descent optimization as primary training objective. Further, we extend this approach with regularization terms to enforce desirable properties like the sparsity of the model or noise robustness of the training process. Experimental evaluations show that our approach outperforms the state-of-the-art consistently by 1.5 dB to 6.1 dB PSNR for image representation.

Original SMoE 34.63 dB JPEG 34.68 dB

Image quality comparison between regressed SMoE model trained by the proposed method and JPEG at the same PSNR level.

Steered Mixtures of Experts

Gaussian Mixture Models are used to define a multivariate joint density distribution for (spatial input) random vector $\boldsymbol{x}$ and (luminance output) random variable y as sum of K weighted Gaussian distributions (kernels)

$$p(\boldsymbol{x}, y) = \sum_{i=1}^{K} \pi_i \cdot \mathcal{N}(\boldsymbol{x}, y; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (1)$$

with mixing coefficients π_i, for which $\sum_{\forall i} \pi_i = 1$, covariance matrices and mean values (centers)

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{XX,i} & \boldsymbol{\Sigma}_{XY,i} \\ \boldsymbol{\Sigma}_{XY,i}^T & \sigma_{Y,i}^2 \end{bmatrix}, \ \boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{X,i} \\ \mu_{Y,i} \end{bmatrix}. \quad (2)$$

When the model parameters are trained (i.e. by EM) a 2D regression function $y_p(\boldsymbol{x})$ can be determined using the expected value of the conditional distribution of Y given $\boldsymbol{X}$

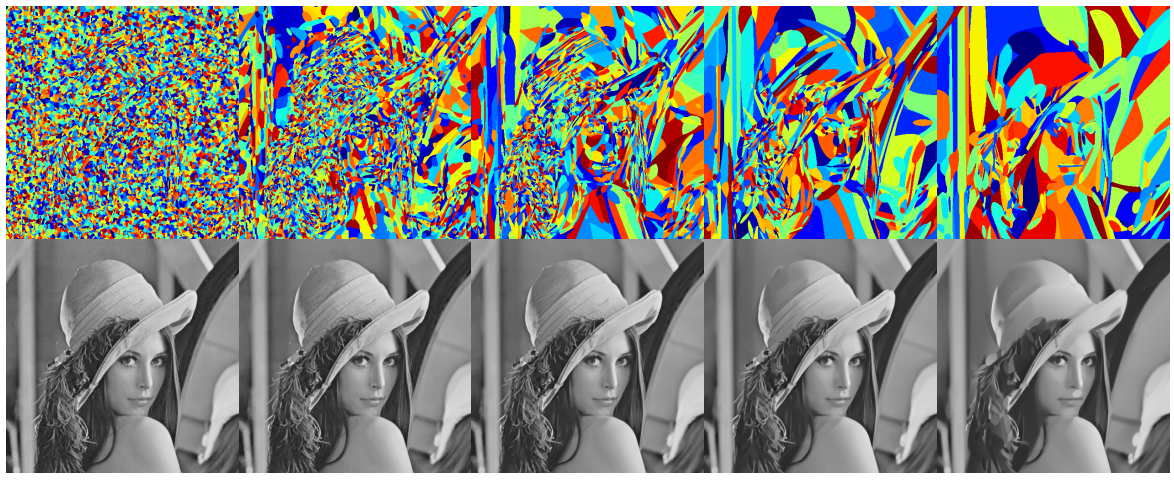
$$y_p(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X}] = \sum_{i=1}^{K} m_i(\boldsymbol{x}) \cdot w_i(\boldsymbol{x}) \quad (3)$$

with the so called (hyper-) plane components or simply experts

$$m_i(\boldsymbol{x}) = \mu_{Y,i} + \boldsymbol{\Sigma}_{XY,i}^T \boldsymbol{\Sigma}_{XX,i}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{X,i}) \quad (4)$$

and weighted soft max gating functions, called gates

$$w_i(\boldsymbol{x}) = \frac{\pi_i \cdot \mathcal{N}(\boldsymbol{x}, y; \boldsymbol{\Sigma}_{XX,i}, \boldsymbol{\mu}_{X,i})}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(\boldsymbol{x}, y; \boldsymbol{\Sigma}_{XX,j}, \boldsymbol{\mu}_{X,j})}. \quad (5)$$



$K = 13000$, 39.94 dB $K = 5000$, 37.29 dB $K = 2500$, 35.75 dB $K = 900$, 32.48 dB $K = 300$, 28.68 dB

Evolution of the regressed model trained using the proposed optimization method at different stages/values of the regularization parameter λ_S for sparsification.

GMM Adaptations for Gradient Descent Training

A closer look on equation (4) and (5) reveals the independence of experts and gates:

$$m_i(\boldsymbol{x}) = m_{0,i} + m_{1,i} \cdot x_{1,i} + m_{2,i} \cdot x_{2,i} \quad (6)$$

with $m_{0,i} = \mu_{Y,i} - \boldsymbol{\Sigma}_{XY,i}^T$, $[m_{1,i}\ m_{2,i}] = \boldsymbol{\Sigma}_{XY,i}^T \boldsymbol{\Sigma}_{XX,i}^{-1}$ and $[x_1\ x_2]^T = \boldsymbol{x}$. and gates can be trained independently if needed.

To enforce positive semidefiniteness of each covariance matrix $\boldsymbol{\Sigma}_{XX,i}$ and omit matrix inversion induced instabilities in the training process, we redefine each $\boldsymbol{\Sigma}_{XX,i}^{-1}$ by its Cholesky decomposition:

$$\boldsymbol{\Sigma}_{XX,i}^{-1} := \mathbf{A} \cdot \mathbf{A}^T \quad (7)$$

$$\mathbf{A} := \begin{pmatrix} a_{1,i} & 0 \\ a_{3,i} & a_{2,i} \end{pmatrix} \quad (8)$$

and optimize for $a_{1,i}$, $a_{2,i}$ and $a_{3,i}$ instead.

Multi-Task Optimization

Instead of maximizing the likelihood by the EM algorithm as quality metric, the MSE between image data y_n and parametric regression $y_p(\boldsymbol{x}_n)$ can form a more reasonable optimization criterion for gradient descent training:

$$\mathcal{L}^{\text{MSE}} := \frac{1}{N} \sum_{n=1}^{N} (y_n - y_p(\boldsymbol{x}_n))^2 \quad (9)$$

With an additional regularization loss:

$$\mathcal{L}^{\text{S}} := \lambda_S \cdot \sum_{i=1}^{K} \pi_i \quad (10)$$

promoting sparsity of the mixing coefficients π_i, gradually removing low contributing modes from the GMM.

Analogously, the bandwidth of each kernel can be maximized by:

$$\mathcal{L}^{\text{D}} := \lambda_D \cdot \sum_{i=1}^{K} \frac{1}{|\Sigma_{XX,i}|} = \lambda_D \sum_{i=1}^{K} (a_{1,i} \cdot a_{2,i})^2 \quad (11)$$

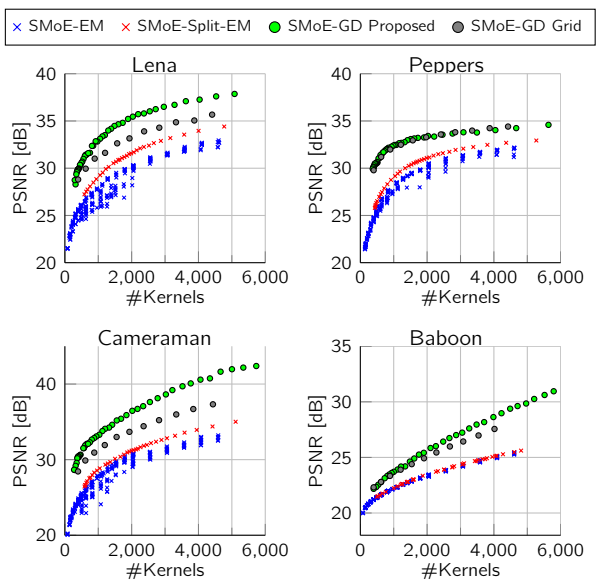
suppressing the modeling of too small details which are usually introduced by noise.

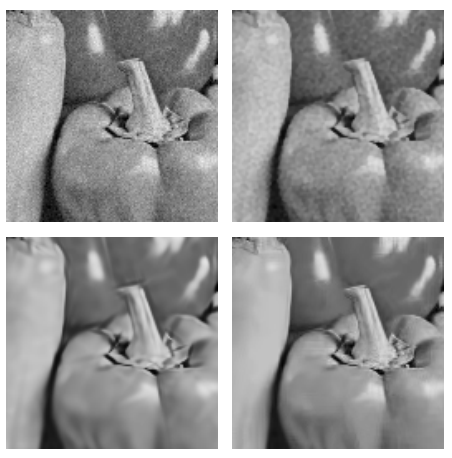
The final multi-task loss is composed as:

$$\mathcal{L} := \mathcal{L}^{\text{MSE}} + \mathcal{L}^{\text{S}} + \mathcal{L}^{\text{D}} \quad (12)$$

Experiments

Image	SMoE-GD Proposed Kernel	PSNR	SSIM	SMoE-GD Grid Kernel	PSNR	SSIM	GMM-Split-EM Kernel	PSNR	SSIM
Cameraman	-	-	-	13378	45.91	0.99	-	-	-
	3844	**40.07**	0.96	3845	36.69	0.96	3947	33.92	0.91
	1928	**36.02**	0.93	1933	33.67	0.93	1931	31.07	0.87
	880	**32.81**	0.88	880	30.93	0.88	878	28.29	0.83
Lena	-	-	-	13056	39.85	0.96	-	-	-
	3876	**37.17**	0.93	3876	35.04	0.92	4003	33.95	0.89
	1934	**35.27**	0.9	1934	33.13	0.9	1921	31.05	0.85
	893	**32.54**	0.87	893	31.02	0.86	854	28.48	0.79
Peppers	-	-	-	11899	36.36	0.91	-	-	-
	3868	33.96	0.84	3873	**34.24**	0.87	3805	32.47	0.81
	1971	**33.41**	0.83	1971	33.27	0.84	1988	31.08	0.79
	896	**31.94**	0.81	896	31.98	0.82	919	28.76	0.75
Baboon	-	-	-	11576	35.25	0.98	-	-	-
	3537	**27.96**	0.8	3539	26.97	0.84	3532	24.76	0.63
	1907	**25.46**	0.69	1907	24.92	0.72	1869	23.35	0.52
	889	**23.46**	0.56	889	23.34	0.58	766	21.90	0.41





Visualization of the SMoE noise reduction capabilities. Top Left: Nosiy Input, Top Right: SMoE without noise reduction, Bottom Left: SMoE with noise reduction, Bottom Right: BM3D

Conclusion and Further Work

- Gradient Descent outperforms EM for SMoE optimization, regularization can be used to control the desired number of kernels in the model
- Further work: Extension to other types of media (e.g. video or light field)

Contact

Web: www.nue.tu-berlin.de
Email: {bochinski,sikora}@nue.tu-berlin.de

Code available at:
https://github.com/bochinski/tf-smoe