Edge-Aware Autoencoder Design for Real-Time Mixture-of-Experts Image Compression

Elvira Fleig, Jonas Geistert, Erik Bochinski, Rolf Jongebloed, and Thomas Sikora

Abstract—Steered-Mixtures-of-Experts (SMoE) models provide sparse, edge-aware representations, applicable to many use-cases in image processing. This includes denoising, super-resolution and compression of 2D- and higher dimensional pixel data. Recent works for image compression indicate that compression of images based on SMoE models can provide competitive performance to the state-of-the-art. Unfortunately, the iterative model-building process at the encoder comes with excessive computational demands.

In this paper we introduce a novel edge-aware Autoencoder (AE) strategy designed to avoid the time-consuming iterative optimization of SMoE models. This is done by directly mapping pixel blocks to model parameters for compression, in spirit similar to recent works on “unfolding” of algorithms, while maintaining full compatibility to the established SMoE framework. With our plug-in AE encoder, we achieve a quantum-leap in performance with encoder run-time savings by a factor of 500 to 1000 with even improved image reconstruction quality. For image compression the plug-in AE encoder has real-time properties and improves RD-performance compared to our previous works.

Index Terms—Image Compression, Image Processing, Autoencoder, Sparse Representation, Steered Mixture-of-Experts

I. INTRODUCTION

The Steered Mixture-of-Experts (SMoE) approach has been first presented as a promising regression framework for coding images [1] [2] [3] [4] [5] and has since expanded to other application fields such as video compression [6] and coding of higher dimensional data, like light field images and light field videos [7].

The intriguing property of the SMoE approach is its capability to model in-stationarities in images (such as edges or smooth transitions) efficiently with a sparse edge-aware representation. Swarms of steered kernels are employed to model image pixel properties. Once the model is optimized, all kernels are aligned to yield the best possible correlation between pixels – and afterwards collaborate for reconstruction of the pixel amplitudes at the decoder. Compression of images is performed in the image-domain rather than in a transform domain by quantizing the SMoE kernel parameters.

Model parameters are optimized at the encoder network for each block to be coded, i.e. using Maximum-Expectation [1] or Gradient-Descent algorithms. For best model quality the Gradient-Descent optimization is preferred [4] [5] [8]. An advantage of this approach is, that cost functions such as mean-squared-error or SSIM with flexible side conditions can be used. For quantization of the model parameters, strategies like “fake quantization” can be readily employed within the optimization process to approach Rate-Distortion bounds. The reader is referred to the work by Jongebloed et al. [8] for competitive compression results of SMoE compared to JPEG2000. Unfortunately, to achieve reasonable model quality, many training iterations are required which leads to a high computational cost. This makes the SMoE coding approach unfeasible for real-time applications.

In this paper we present a novel approach for estimating SMoE parameters via the design of a novel Autoencoder network with embedded SMoE decoder capabilities. The strategy we are suggesting provides a break-through in run-time performance with no compromise to the quality of the reconstructed images.

To show the benefits of our strategy for compression we investigate our novel SMoE Autoencoder in the context of the SMoE image coding framework of Tok et al. in [5]. As we shall see, we are able to achieve real-time capabilities with encoder run-time savings by gains of 500-1000 compared to Tok et al. without sacrificing rate-distortion performance.

II. THE SPARSE, EDGE-AWARE SMoE IMAGE MODEL

The SMoE image model describes an edge-aware, parametric, continuous non-linear regression function. We use Fig. 1 to illustrate the kernel model concept and the capabilities of the model to reconstruct sharp and smooth transitions in images without common ringing artifacts seen in JPEG-like coders. The SMoE model is able to reconstruct the $32 \times 32$ pixel block with excellent edge quality. In contrast to JPEG, JPEG2000 and HEVC-Intra at the same bit rate, typical blocking and ringing artifacts are avoided completely. Both objective quality measures as well as subjective quality are greatly enhanced. The locations and steering properties of the gaussian kernels after optimization, depicted in Fig. 1f, explain the correlation between pixels. The resulting 2D-gating functions provide for the edge-awareness of the model, which are displayed in Fig. 1g. Sharp edges are modelled with minimal overlapping gating functions while smooth transitions by highly overlapping gates. The reader is referred to [2] [7] [8] for more insight into the subject. This edge-aware sparse representation of SMoE’s appears to be attractive for a number of reasons for compression and beyond:

- SMoE’s easily extend to N-dimensional signals, including light-fields and beyond [9].
Modeling is not restricted to regular N-dimensional pixel grids, but can be applied to irregularly sampled imagery including arbitrarily shaped segments and point clouds of any dimension.

The sparse (coded) representation allows to easily extract N-dimensional image features from the (decoded) model, such as edges, intensity flow, etc. [1].

Once built, the continuous SMoE model allows to resample to any resolution in time and space, which includes edge-aware super-resolution and motion-interpolation.

III. THEORETICAL BACKGROUND

Steered Mixtures-of-Experts regress the luminance value at each position \( \mathbf{x} \) of an image by a weighted sum of \( K \) experts:

\[
y_p(\mathbf{x}) = \sum_{i=1}^{K} m_i(\mathbf{x}) \cdot w_i(\mathbf{x})
\]

with \( m_i(\mathbf{x}) \) being the experts and \( w_i(\mathbf{x}) \) the gating functions. In this work, we focus on constant experts as of Eq. 2. However, other functions like hyperplanes [1] or polynomials can be used instead.

\[
m_i(\mathbf{x}) = m_i
\]

Respectively the gating functions \( w_i(\mathbf{x}) \) are defined as weighted soft-max functions

\[
w_i(\mathbf{x}) = \frac{\mathcal{K}(\mathbf{x}, \mu_i, \Sigma_i)}{\sum_{j=1}^{K} \mathcal{K}(\mathbf{x}, \mu_j, \Sigma_j)}
\]

described by kernel functions \( \mathcal{K}(\mathbf{x}, \mu_i, \Sigma_i) \), such as laplacians or as is the case in this work, gaussian kernels:

\[
\mathcal{K}(\mathbf{x}, \mu_i, \Sigma_i) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right)
\]

With \( \mu_i \) the center position of the kernels and \( \Sigma_i \) the steering parameters. This work uses a simplified definition by having radial kernels that do not steer in any direction and using kernels of the same bandwidth \( S \). So the regression function in Eq 1 simplifies to:

\[
y(\mathbf{x}) = \sum_{i=1}^{K} m_i \cdot \exp\left(-S||\mathbf{x} - \mu_i||^2\right) \sum_{j=1}^{K} \exp\left(-S||\mathbf{x} - \mu_j||^2\right)
\]

Even with these simplifications, a good representation of smooth and sharp transitions can be achieved by adjusting the center positions of the kernels. The possibility to model complex patterns increases with the number of used kernels. The work of [5] introduced the gradient descent optimization (SMoE-GD) to achieve better results in comparison to the EM Algorithm. The optimization objective is to directly minimize the reconstruction error instead of maximizing the joint likelihood of the location and pixel amplitude [1] which is less sensitive to subjective image quality. The function \( \mathcal{L} \) is defined as loss to perform the mean-squared-error optimization on all \( N \) pixels in a block:

\[
\mathcal{L} := \frac{1}{N} \sum_{n=1}^{N} \left(y_n(\mathbf{x}) - \sum_{i=1}^{K} m_i(\mathbf{x}) \cdot w_i(\mathbf{x})\right)^2
\]

This optimization was used in Fig. 1 to determine the SMoE model. Its quality is highly dependable on the initialization of the kernels and can quickly get stuck in local minima. In addition, many training iterations are necessary to achieve a good quality reconstruction, as will be discussed below. This makes the iterative approach unfeasible for real-time applications.

IV. EDGE-AWARE AUTOENCODER DESIGN WITH EMBEDDED SMoE DECODER

A. Autoencoder Design Principles

In order to circumvent the above drawbacks of the gradient descent optimization we propose to design a SMoE Autoencoder network (SMoE-AE) that generates SMoE parameters at the bottleneck layer ready for compression and coding. The basic concept of the SMoE-AE network with embedded SMoE decoder is depicted in Fig. 2. As such, we propose to design a deep encoder network in combination with a fixed, shallow SMoE decoder. Once the SMoE-AE network parameters are optimized in training based on ground-truth imagery the SMoE-AE encoder network is fixed - and can be used to predict SMoE parameters for pixels of individual input blocks. In this way the aforementioned block-wise gradient-descent SMoE model building
process develops into a “Learnt Compression” scenario. It is our hypothesis that SMoE-AE networks with suitable network sizes and restricted number of network parameters can produce SMoE parameters resulting in comparable quality - with a processing time significantly less than those produced with gradient descent optimization. The envisioned SMoE-AE network is novel in concept because the fixed SMoE model is an integral part of the SMoE-AE end-to-end training process. This ensures that the AE encoder network produces meaningful bottleneck SMoE parameters. It is readily feasible to design the encoders such that the SMoE parameters are ready for coding - discrete and quantized to approach RD-bounds. Any objective measure usually used in neural network learnt-compression approaches (such as MSE, SSIM and beyond) can be used for Rate-Distortion optimization of the SMoE-AE encoders.

In this paper we introduce such an Autoencoder with a bottleneck layer restricted to produce parameters for the radial SMoE decoder according to Eq. 5 with $K = 4$ radial kernels/block and bandwidth $S$ as hyper parameter. This makes it possible to compare our results directly with those of [5]. To comply with [5], the autoencoder is built to provide parameters for gray-level pixels in blocks of size $16 \times 16$. The SMoE-AE encoder is designed to follow the concept of conventional Autoencoders [10] with a $16 \times 16$ dimensional input and decreasing layer sizes towards the 12-dimensional latent space output. For the decoder part, the shallow SMoE decoder network is used. The architectural structure of the encoder is depicted in Tab. I and can be fully implemented in accordance with the given table. The last fully-connected layer presents the bottleneck of the autoencoder, and the values are then passed on to the SMoE decoder. The total number of parameters in the encoder network is 68,855,116. The innermost layer comprises the kernel center values $\mu$, and the expert values $m_i$, to perform the SMoE reconstruction based on Eq.5.

### B. Training the SMoE-AE network parameters

The mobile subset of the clic dataset [11] was used to train the network. The test images used for coding were not part of the training set. All training images were converted to greyscale, cropped to $1024 \times 1024$ pixels and $16 \times 16$ pixel blocks were then extracted out of these images. Pixel values were scaled to the domain $[0..1]$. The training set thus contained over four million individual blocks. For training we did not categorize blocks into textured or non-textured blocks as in [5]. The order of the blocks were shuffled during training to ensure a suitable generalization of the network. Bottleneck parameters were also restricted to be in the domain $[0..1]$. All training parameters are shown in Tab.

| Epochs | Quantity 6 | Quantity 5 | Output 512, 256, 128, 64, 12 |
|--------|------------|------------|-------------------------------|
| LR     | 0.00005    |            |                               |
| Optimizer | Adam      |            |                               |
| Activation | ReLU      |            |                               |
| Loss   | MSE        | Kernel Size 3x3 |                     |

| Sequence | SMoE-GD | SMoE-AE |
|----------|---------|---------|
| Baboon   | PSNR 21.34, 22.65, 22.91, 22.92 | 22.95 |
| Cameraman| SSIM 0.71, 0.82, 0.85, 0.85  | 0.85 |
| Lena     | PSNR 24.06, 27.82, 28.83, 28.91 | 29.05 |
| Peppers  | SSIM 0.62, 0.76, 0.79, 0.79, 0.79 | 0.82 |

| Encoding Time [s] | 0.25 | 30.54 | 296.29 | 1191.2 | 0.25 |

### Table I: Encoder architecture and training parameter

| General | Conv. Layers | Dense Layer |
|---------|--------------|-------------|
| Epochs  | 30           | 6           | 5           |
| LR      | 0.00005      | 16, 32, 64  | 512, 256, 128, 64, 12 |
| Optimizer | Adam      |              |             |
| Activation | ReLU      |              |             |
| Loss    | MSE         | Kernel Size 3x3 |       |

### Table II: PSNR in [dB] and SSIM reconstruction quality

I. The training took about 20 hours on a Nvidia GTX 1070 Ti. After training the encoder, the decoder part was decoupled and the encoder was used to predict the center positions $\hat{\mu}$ and expert values $m_i$ for each block.

### V. Experiments/Evaluation

The prime purpose was to understand trade-offs between image quality and run-time properties of the SMoE-AE approach compared to the complex SMoE-GD in [5]. All SMoE experiments used the fixed number of four kernels and a bandwidth $S = 0.035$.

With known input pixels $y_p$ and the estimated gating $w_i$ the experts $m_i$ of Eq. 1 can be optimally recalculated by means of Ordinary Least Squares (OLS). The estimation of these experts is still beneficial for the encoder training as it regularizes the network towards finding better center positions. Omitting the expert estimation and solely relying on the optimal experts obtained by above mentioned OLS would degrade the overall performance by 0.1 dB in PSNR.

Tab. II shows a comparison of the reconstruction quality of SMoE-GD to the SMoE-AE network for the unquantized model output. The results are based on $512 \times 512$ pixel gray-level images. The reconstruction quality of the SMoE-GD depends on the number of training iterations during encoding while the run-time of the SMoE-AE network does not change as the SMoE parameters are estimated in a single pass. Most noticeably: our SMoE-AE network consistently outperforms the SMoE-GD approach in [5] in terms of PSNR with comparable SSIM while reducing the run-time drastically, as seen in Tab. II.

On average the SMoE-GD approach converges to an acceptable quality between 2500 - 5000 iterations, consuming between 150-300 seconds for each $512 \times 512$ image for encoding. This is brought down to 0.25 seconds/image with the novel SMoE-AE encoder network making the SMoE approach with SMoE-AE parameter prediction feasible for practical implementations. This results in a speed up gain by a factor of 500 to 1000 which we believe is remarkable given the even increased quality. Even with 20,000 iterations the quality of SMoE-GD is still inferior to SMoE-AE in terms of PSNR and on par in SSIM. In the case of the Cameraman test image, the SMoE-GD optimization does not even seem to converge to a satisfying quality, which is due to initialization problems of the parameters. As outlined above, the SMoE-AE network does not need to initialize the SMoE model parameters and
seems to deal with such situations robustly. Tab. III shows a comparison of the compression results between JPEG, SMoE-GD and SMoE-AE at very low bit rates. To be comparable to JPEG the SMoE parameters are quantized and the division into "non-textured" and textured blocks is used identical to the work in [5]. For textured blocks, the SMoE center positions $\mu_i$ are quantized with 3-4 bits/component and the expert values $m_i$ with 4-5 bits as in [5]. The average luminance value of non-textured blocks is coded with 8 bits. To distinguish between the block types one bit is added. Under quantization, the improvements reported in Tab. II are preserved, as depicted in the Rate-Distortion curves in Fig. 4 - maintaining the reported tremendous run-time savings of SMoE-AE. The above mentioned recalculation of the experts via OLS becomes more important when adapting to center positions after quantization and provides gains of up to 1dB in PSNR. Additional tests to examine the performance of SMoE-AE in comparison to JPEG were made with the Kodak dataset, which is not depicted for SMoE-GD as it was not part of [5]. In Tab. III the mean value for PSNR and SSIM over all 25 gray-scale images is displayed, with the images resized and cut to 512\times 512, to be comparable to the other test images. Even when analyzing a broader variety of images, SMoE-AE still outperforms JPEG in very low bit rates with a gain of over 1.3 dB on average while operating on a lower bit rate, as JPEG was unable to reach these very low rate for many images. Visual results for test image "Peppers" in Fig. 5 and test image "kodim07" in Fig. 3 reveal that both SMoE-GD and SMoE-AE significantly outperform JPEG at very low bit rates, with excellent edge reconstruction properties. Also, SMoE-AE reconstruction is slightly better than quantized SMoE-GD, even though visually hardly noticeable.

In comparison with other autoencoder-based image compression methods [12] [13] [14], it is important to identify a few key differences. An advantage of SMoE-AE is the shallow SMoE decoder, which ensures a faster decoding time (0.02s) compared to the encoding time, while simultaneously maintaining its physical interpretation of the parameters regarding edges, etc. Additionally, the parameters are quantized after optimization, allowing a flexible bit rate with only one trained model. The significant result is, however, that the novel SMoE-AE network provides a quantum leap in run-time improvement for SMoE compression approaches.

### Table III: Comparison JPEG, SMoE-GD and SMoE-AE after quantization

| Sequence | Rate [bpp] | PSNR [dB] | SSIM | Rate [bpp] | PSNR [dB] | SSIM | Rate [bpp] | PSNR [dB] | SSIM |
|----------|------------|-----------|------|------------|-----------|------|------------|-----------|------|
| Baboon   | 0.14       | 21.31     | 0.45 | 0.14       | 22.55     | 0.49 | 0.14       | 22.45     | 0.47 |
| Cameraman| 0.16       | 21.76     | 0.5  | 0.16       | 22.64     | 0.50 | 0.16       | 22.66     | 0.51 |
| Lena     | 0.15       | 26.45     | 0.73 | 0.08       | 26.47     | 0.80 | 0.08       | 26.69     | 0.79 |
| Peppers  | 0.17       | 27.88     | 0.81 | 0.09       | 26.30     | 0.81 | 0.09       | 26.88     | 0.80 |

#### VI. SUMMARY/CONCLUSION

In this paper we proposed an Autoencoder approach for training and coding Steered-Mixture-of-Experts models with the goal of speeding up encoding times for real-time applications, including compression. We emphasize, that such strategy has similarities and is in the spirit of "unfolding" approaches currently being discussed with Deep Neural Networks [15] [16]. We conclude that the SMoE-AE "unfolding" approach drastically reduces encoder run-time compared to the fully optimized SMoE-GD models using gradient-descent algorithms. Very importantly, SMoE-AE coding allows the additional initially-mentioned functionalities at the decoder, such the super-resolution-ready reconstruction, by resampling Eq. 5 to any desired resolution. By calculating gradients from Eq. 5 (or analysing kernel positions in a block) edge and correlation information is easily accessible for further analysis/postprocessing at the decoder

With the contributions in this paper the SMoE compression approach introduced in [5] is ready for real-time applications. The coding structure in its present form, with fixed number of 4 non-steering kernels/block, is however not competitive with state-of-the-art JPEG2000 or HEVC-Intra coding at higher bit rates. Our goal is to make the approach block-adaptive by employing variable number of kernels/block and/or blocks of different sizes, including steering kernels. To this end we will investigate more complex SMoE-AE encoder and decoder networks for advanced kernel parameter prediction.
REFERENCES

[1] R. Verhack, T. Sikora, L. Lange, G. Van Wallendael, and P. Lambert, “A universal image coding approach using sparse steered mixture-of-experts regression,” in 2016 IEEE International Conference on Image Processing (ICIP), pp. 2142–2146, IEEE, 2016.

[2] R. Jongebloed, R. Verhack, L. Lange, and T. Sikora, “Hierarchical learning of sparse image representations using steered mixture-of-experts,” in 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6, IEEE, 2018.

[3] B. Liu, Y. Zhao, X. Jiang, and S. Wang, “An image coding approach based on mixture-of-experts regression using epanechnikov kernel,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1807–1811, IEEE, 2019.

[4] E. Bochinski, R. Jongebloed, M. Tok, and T. Sikora, “Regularized gradient descent training of steered mixture of experts for sparse image representation,” in 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3873–3877, IEEE, 2018.

[5] M. Tok, R. Jongebloed, L. Lange, E. Bochinski, and T. Sikora, “An mse approach for training and coding steered mixtures of experts,” in 2018 Picture Coding Symposium (PCS), pp. 273–277, IEEE, 2018.

[6] L. Lange, R. Verhack, and T. Sikora, “Video representation and coding using a sparse steered mixture-of-experts network,” in 2016 Picture Coding Symposium (PCS), pp. 1–5, IEEE, 2016.

[7] R. Verhack, T. Sikora, G. Van Wallendael, and P. Lambert, “Steered mixture-of-experts for light field images and video: Representation and coding,” IEEE Transactions on Multimedia, vol. 22, no. 3, pp. 579–593, 2019.

[8] R. Jongebloed, E. Bochinski, L. Lange, and T. Sikora, “Quantized and regularized optimization for coding images using steered mixtures-of-experts,” in 2019 Data Compression Conference (DCC), pp. 359–368, IEEE, 2019.

[9] R. Verhack, T. Sikora, L. Lange, R. Jongebloed, G. Van Wallendael, and P. Lambert, “Steered mixture-of-experts for light field coding, depth estimation, and processing,” in 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 1183–1188, IEEE, 2017.

[10] M. Tschannen, O. Bachem, and M. Lucic, “Recent advances in autoencoder-based representation learning,” arXiv preprint arXiv:1812.05069, 2018.

[11] G. Toderici, W. Shi, R. Timofte, L. Theis, J. Balle, E. Agustsson, N. Johnston, and F. Mentzer, “Workshop and challenge on learned image compression (clic2020),” 2020.

[12] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, “Deep convolutional autoencoder-based lossy image compression,” in 2018 Picture Coding Symposium (PCS), pp. 253–257, IEEE, 2018.

[13] L. Theis, W. Shi, A. Cunningham, and F. Huszár, “Lossy image compression with compressive autoencoders,” arXiv preprint arXiv:1703.00395, 2017.

[14] L. Zhou, C. Cai, Y. Gao, S. Su, and J. Wu, “Variational autoencoder for low bit-rate image compression,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2617–2620, 2018.

[15] J. R. Hershey, J. L. Roux, and F. Weninger, “Deep unfolding: Model-based inspiration of novel deep architectures,” arXiv preprint arXiv:1409.2574, 2014.

[16] K. Zhang, L. V. Gool, and R. Timofte, “Deep unfolding network for image super-resolution,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3217–3226, 2020.