https://doi.org/10.2352/EI.2022.34.14.COIMG-220

Recognition-Aware Learned Image Compression

Maxime Kawawa-Beaudan, Ryan Roggenkemper, Avideh Zakhor; Department of Electrical Engineering and Computer Science, University of California, Berkeley; Berkeley, CA, USA

Abstract

Learned image compression methods generally optimize a rate-distortion loss, trading off improvements in visual distortion for added bitrate. Increasingly, however, compressed imagery is used as an input to deep learning networks for various tasks such as classification, object detection, and super-resolution. We propose a recognition-aware learned compression method, which optimizes a rate-distortion loss alongside a task-specific loss, jointly learning compression and recognition networks. We augment a hierarchical autoencoder-based compression network with an EfficientNet recognition model and use two hyperparameters to trade off between distortion, bitrate, and recognition performance. We characterize the classification accuracy of our proposed method as a function of bitrate and find that for low bitrates our method achieves as much as 26% higher recognition accuracy at equivalent bitrates compared to traditional methods such as Better Portable Graphics (BPG).

Introduction

Image compression, the task of reducing the storage and transmission cost of images while preserving their quality, involves three steps: transformation, quantization, and bit allocation. Traditionally, each of these steps is manually engineered and inflexible, but in recent years, learned compression methods based on convolutional neural networks have proven their ability to outperform traditional codecs by optimizing rate-distortion losses [3, 4, 5, 6, 7]. These convolutional neural network based methods often leverage autoencoders, architectures which repeatedly downsample input images through convolution to yield low dimensional features called latents, which can be decoded to reconstruct the image [8, 9, 10].

Most deep learning methods seek optimal tradeoffs between compression efficiency and perceptual quality. As the intended consumer of the image is the human eye, compression research has focused on optimizing distortion metrics such as Peak Signal-to-Noise Ratio (PSNR) or Multiscale Structural Similarity (MS-SSIM). The bitrate, or the average number of bits required to encode a compressed image, is approximated using a model which learns to predict probability distributions over quantized latents. For a learned compression scheme, this bitrate can be approximated by the entropy of the distribution over the latents. Recent papers such as [11, 8, 12, 13] favor Gaussian Mixture Models (GMM) with learned means, variances, and mixing weights, to model the latent distributions. Quantizing the latents is a non-differentiable operation, which presents a challenge for deep learning based approaches, but widely adopted solutions to this problem include straight-through approximation, as in [14], and uniform noise approximation [5]. Hierarchical models, pioneered in [7], introduce a second level of compression, encoding the latents into hyper-latents which are transmitted as side information. Side information in learned compression schemes are additional bits used to improve the match between the estimated and real entropy of the latents. In GMM methods the hyperlatents are generally interpreted as the means, variances, and mixing weights for the constituent Gaussians. The bitrate of the hyper-latents must be accounted for in the loss and is usually estimated using a factorized entropy model, as introduced in [6].

The compression model used in our work incorporates all of these learned components: a factorized entropy model, a GMM, and a hierarchical structure. Our contribution is the addition of a task sensitivity. More and more, compressed images are consumed not by the human eye but by neural networks designed for tasks such as super-resolution or recognition. Such tasks may be sensitive to distortions not well represented by conventional distortion metrics such as PSNR, and as a result, task performance may suffer under compression by methods trained in a task-agnostic manner. Furthermore, compression methods trained using conventional metrics may be sub-optimal for a given task, allocating bits to features which, while salient for human perception, are irrelevant to task performance.

In this work we focus on the task of recognition. Some work relevant to recognition-aware image compression has been proposed, as in [15, 16]. These methods learn spatial quantization parameter maps for compressed images based on the response strengths of feature maps from recognition networks. [17, 18] present methods for image enhancement driven by classification. Images are pre-transformed by convolution layers which learn to enhance the aspects of the image conducive to recognition, before being passed to recognition models. While these methods induce no explicit compression, the end-to-end nature of the training schemes are similar in spirit to what we aim to implement. In [19], task-specific networks are optimized with augmented losses which penalize the entropy of learned features. This encourages models to learn compressible features which can then be encoded by existing compression methods. However, no tailored compression method is jointly learned with the task. No reconstructed image is generated: rather, the task output is immediately predicted from the features, doing away with the intermediate reconstructed image. The authors are thus able to do away with the distortion term in their loss.

Proposed Approach

In this paper we are interested in explicitly compressing an image and generating a reconstructed image which is passed to a recognition model. Learning the parameters of both models allows the networks to complement one another: The compression model is incentivized to allocate bits in a way which maximally preserves recognition accuracy. The recognition model is incentivized to fine tune its feature extraction layers to work efficiently with lower bitrate compressed images. As a result, we

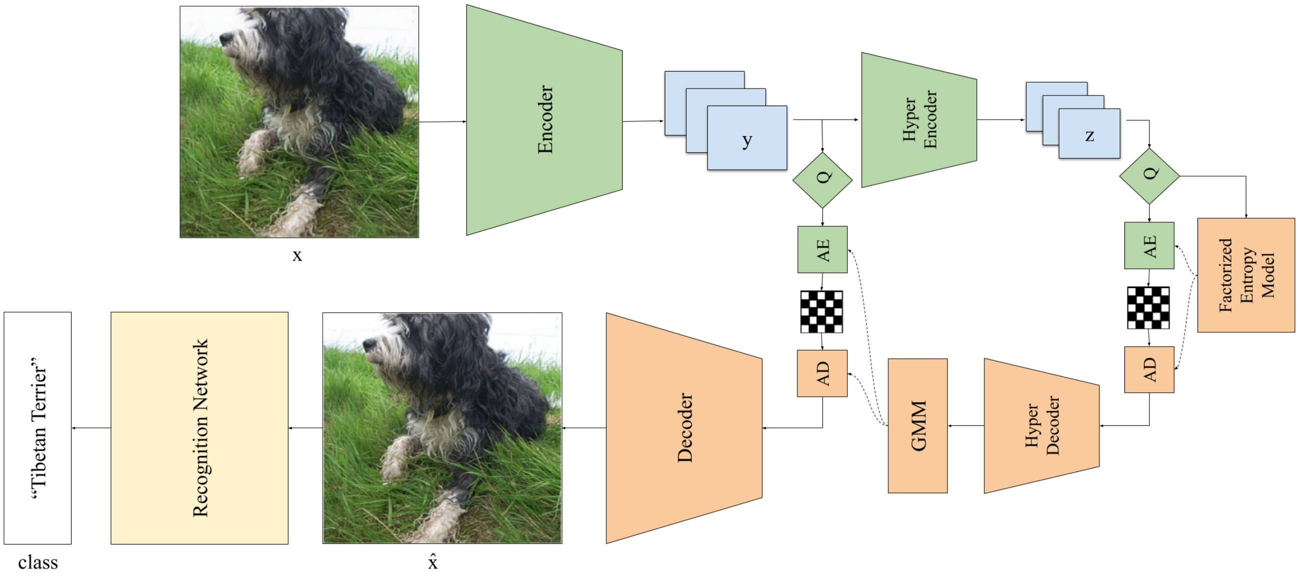

Figure 1. *The joint compression-recognition architecture, where the encoders, decoders, Gaussian Mixture Model (GMM), and factorized entropy model are as in [1]. The recognition network is an EfficientNet-B0 as in [2].* x *is an input image,* y *are the latents,* z *are the hyper-latents,* $\hat{x}$ *is the compressed image. AE and AD represent arithmetic encoding and decoding, respectively, and Q represents scalar rounding quantization. Dotted lines from component A to component B indicate that the outputs of A parameterize B.*

achieve higher recognition performance at lower bitrates compared to task-agnostic methods.

Most deep-learning methods optimize a problem of the form:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} R(\hat{x}) + \lambda D(x, \hat{x}) \quad (1)$$

over a set of neural network parameters θ, where x is the original image, $\hat{x}$ is the compressed image, $R(\cdot)$ is the bitrate of the compressed image, and $D(\cdot,\cdot)$ is some distortion metric, typically mean squared error (MSE) or MS-SSIM. λ is a Lagrange multiplier corresponding to the distortion term. We combine state-of-the-art compression and recognition models and train them jointly, learning the parameters of both models end-to-end. We optimize a three-part loss, balancing the traditional rate-distortion terms with a task-specific term added to induce a sensitivity to the recognition task. Our joint loss yields an optimization problem over the compression model's parameters θ and the recognition model's parameters ϕ of the form:

$$(\theta^*, \phi^*) = \underset{\theta,\phi}{\operatorname{argmin}} (1-\lambda)R(x) + \lambda D(x, \hat{x}) + \beta L_t(y, \hat{y}) \quad (2)$$

where y is the true task label, $\hat{y}$ is the model's predicted task label, and L_t is the task loss, in this case, cross entropy. The parameters λ and β allow us to control the emphasis placed on each of the constituent loss terms during training. By weighting the bitrate by $(1-\lambda)$ we couple the distortion and bitrate terms and bind λ to the range $[0,1]$. Note that any ratio of bitrate to distortion weighting achievable in the conventional loss with some setting λ_{CL} is achievable in our loss with the setting $\lambda = \lambda_{CL}/(1+\lambda_{CL})$. When λ is close to 1 the bitrate term is severely discounted and fidelity to the original image is prized. When λ is close to 0 distortion is ignored and the bitrate is optimized against accuracy.

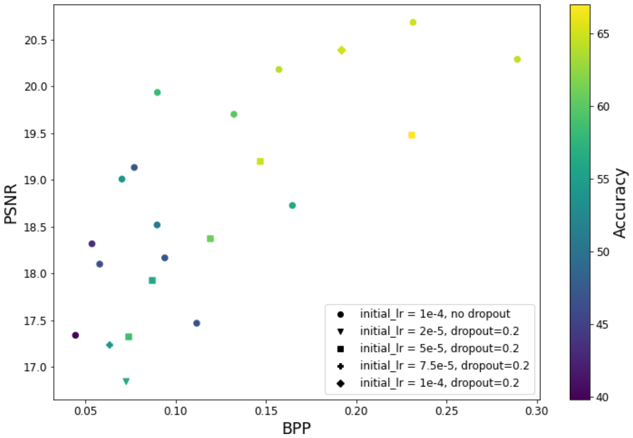

Figure 2. *Bitrate in bits per pixel (BPP), accuracy, and PSNR results for our joint model at various settings of* λ, β, *and training parameters. Dropout was not used during training unless specified. Markers indicate training scheme, as described in legend.*

Architecture Details

Our joint architecture is illustrated in Figure 1. The compression model is based largely on the architecture from [1], which achieves state of the art rate-distortion performance. We do away with the method's proposed decoder-side enhancement module, as it largely aims to improve perceived visual quality. For the sake of simplicity we also do away with the channel attention module in the encoder and hyperencoder. As in [1] we use a GMM with two Gaussians. We also adopt the uniform noise method of quantization, adding uniform noise to the latents during training to simulate the effects of rounding in a differentiable manner.

We add to this compression network an EfficientNet-B0

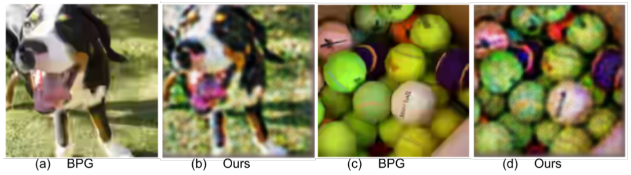
(a) BPG (b) Ours (c) BPG (d) Ours

Figure 3. *Sample images from BPG and our model, trained here with* $\lambda = 0.9, \beta = 1.0$, *initial learning rate of 5e-5, and dropout of 0.2. (a) has BPP=0.132, PSNR=21.55; (b) has BPP=0.119, PSNR=17.62; (c) has BPP=0.117, PSNR=26.35; (d) has BPP=0.119, PSNR=18.36.*

recognition model, as described in [2], chosen for its near state-of-the-art classification accuracy on ImageNet and low parameter count. The current state of the art on the ImageNet validation benchmark is a top-1 accuracy of 88.5%, achieved in [20] using a model with 480 million parameters. EfficientNet-B0 reaches a top-1 accuracy of 78.8% but comprises only 5.3 million parameters, making its outputs usable as a heuristic for recognition accuracy without slowing down training or inference unduly.

In the compression stage, input images are passed to an encoder, which uses downsampling convolutions and Generalized Divisive Normalization (GDN) [21] activation layers to yield latents – in our case, 192 feature layers of height and width 16. These latents are passed to a hyperencoder to repeat this process and yield hyperlatents. The latents and hyperlatents are quantized. At this stage in practice they would be encoded to a bitstream using arithmetic encoding. The quantized hyperlatents are passed to the factorized entropy model, which estimates their bitrate during training, before being decoded and sent to the GMM module, which uses them to generate the means, variances, and weights for the predicted probability distributions over latents. These predicted distributions are used to estimate the training bitrate of the latents, and in practice would be used for arithmetic encoding and decoding. The quantized latents are passed to the decoder to yield the reconstructed image $\hat{x}$, which is sent to the recognition network to yield a predicted class.

Experiments

We use Xavier initialization for the weights of our compression model, and initialize the EfficientNet with weights pretrained for ImageNet classification [22]. We train our model on a random subset of 500,000 of the 1.2 million images comprising the ImageNet dataset. For validation we use the full 50,000 image validation set from the Imagenet 2012 challenge, namely ILSVRC2012. We train for 9 epochs and use MSE as the distortion metric.

Figure 2 demonstrates our model's ability to reproduce the rate-distortion tradeoffs typical of compression methods. As the bitrate increases, PSNR increases and accuracy improves, a result which is indicated by the color gradient from blue to yellow. However, unlike in conventional rate-distortion curves with a one-to-one mapping between bitrates and PSNR values, our results illustrate the model's ability to trade off further between PSNR and accuracy. For a given bitrate it is possible to learn models with high PSNR and low accuracy or low PSNR and high accuracy, by altering β and training parameters such as dropout and learning rate. As in [2] we use dropout to combat overfitting in the recognition model, adopting the suggested value of 0.2. As seen in Figure 2, using dropout significantly improves bitrate and accuracy performance. In one experiment we train two models with identical learning rate and hyperparameter settings but use no dropout for one and dropout of 0.2 for the other. We find that adding dropout decreases the bitrate from 0.289 to 0.192 BPP and increases accuracy by 0.56%. Additionally, through most training we adopt the initial learning rate of 1e-4, as suggested by [1] and decrease the learning rate by half during the last epoch of training. We find, however, that in the high λ domain, e.g. $\lambda = 0.999$, stability during training becomes a challenge. Lowering the learning rate to 1e-5 in such cases improves model performance. In general, performance is highly sensitive to changes in initial learning rate. Learning rate experiments included in Figure 2, where the triangle, cross, and closest square marker represent models trained identically with the exception of learning rate, demonstrate this sensitivity.

Since our recognition model is initialized using weights pretrained on uncompressed ImageNet images, recognition performance is strongly correlated with low distortion. That is, the EfficientNet model does best when compressed input images are as close to the kinds of original, uncompressed images on which it was trained. If improvements in accuracy were due solely to lowered distortion, we would expect recognition accuracy to increase monotonically as PSNR improves. In this case, any non-joint method achieving higher PSNR at equivalent bitrates could be expected to achieve higher accuracy than our method at these points.

However, our model demonstrates the ability to produce images with low bitrate and low PSNR, yet competitive recognition accuracy. Sample output images from our model and BPG can be seen in Figure 3; while our model at this bitrate achieves an average PSNR of 18.37 compared to BPG's 25.22 on the ImageNet validation set, we achieve 16.28% greater accuracy. This result is repeated across bitrates, as illustrated in Figure 4, which compares our results to those of BPG, the state-of-the-art traditional or engineered codec. We attempt to match the bitrates produced by BPG using λ and β tuning, though this targeting is fairly imprecise. We observe higher recognition accuracy at roughly equivalent bitrates, with far lower PSNR. In the low bitrate domain in particular, our method vastly outperforms BPG, achieving 26.03% greater accuracy while producing images with PSNR lower by 6.47 on average. In this way our method makes more efficient use of allocated bits for the task at hand, optimizing for accuracy rather than visual distortion.

Our proposed system largely reduces to EDIC, the system in [1], when $\beta = 0$. That said, there are three differences between our system and that of EDIC: first, we use 192 channels in our convolutions rather than 320. Second we train on three times fewer images than [1]. Authors in [1] train their base model for 3,500,000 iterations with a batch size of 4, exposing the model to 14,000,000 images, while we train for 9 epochs on a dataset of 500,000 images, exposing our model to 4,500,000 images. The training dataset in [1] consists of 20,745 images from Flickr and their testing set is the Kodak PhotoCD dataset, while our training uses the aforementioned 500,000 images from ImageNet and our testing uses the full 50,000 image ImageNet 2012 validation dataset. Third, we have not implemented two blocks in [1], namely attention and decoder side enhancements, in our model. Replicating training in [1] in all other ways and running our sys-

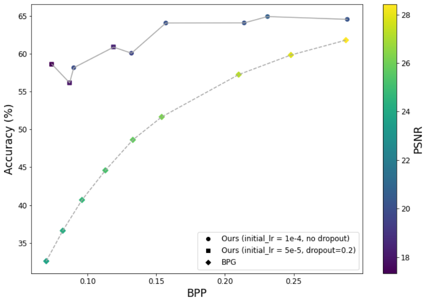

Figure 4. *Comparison between the state of the art traditional codec, BPG, and our joint model, in terms of bitrate, accuracy, and PSNR. By traversing the λ, β space we attempt to find equivalent or proximal bitrates to those achieved by BPG.*

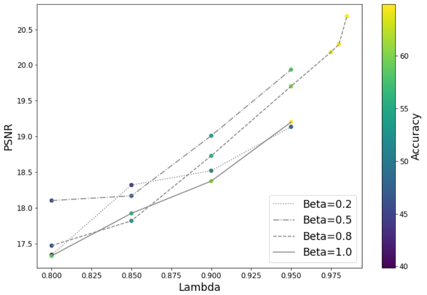

Figure 5. *A demonstration of distortion control using parameters λ, β. Models with $\beta = 1.0$ are trained with initial learning rate 5e-5 and dropout of 0.2; all others are trained using an initial learning rate of 1e-4 with no dropout.*

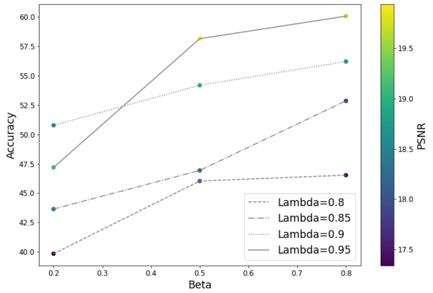

Figure 6. *A demonstration of accuracy control using parameters λ, β. Models are trained using an initial learning rate of 1e-4 with no dropout.*

tem at $\beta = 0$, i.e. with zero weight in the loss term for recognition accuracy, we achieve a bitrate of 0.35, PSNR of 25.57 and recognition accuracy of 42.85%. This PSNR is about 6.5dB less than the performance in [1] for similar bit rates. However, with non-zero weight for the recognition loss, e.g. $\beta = 0.2$, we achieve a higher recognition accuracy of 66.82%, at BPP of 0.43 and PSNR of 23.04. This demonstrates the trade off in our work between PSNR and recognition accuracy.

Our approach to bitrate and accuracy control using β and λ from our loss creates a two-dimensional hyperparameter search space. For a fixed β, increasing λ results in increased accuracy and lower distortion, and has an indeterminate effect on bitrate, as observed in Figure 5. As seen in Figure 6, we find that for a fixed λ, increasing β results in improved recognition accuracy at the cost of a higher bitrate, and has an indeterminate effect on distortion. Within each depicted group with shared λ, we see monotonically increasing accuracy among points with identical training schemes as β increases. We also find that changes in λ affect model performance more than changes in β. One explanation for this is that λ alters the model's emphasis on bitrate as well as distortion while β only indicates the emphasis on cross entropy.

Conclusion and Further Work

We present a joint approach to learned compression and recognition, training state-of-the-art models end-to-end to encourage the learning of complementary features. We demonstrate greater recognition accuracy results to those achieved by traditional methods like BPG, at equivalent bitrates. In future work we aim to extend our results to higher bitrates while remaining competitive with BPG in terms of accuracy.

References

[1] Jiaheng Liu, Guo Lu, Zhihao Hu, and Dong Xu, "A unified end-to-end framework for efficient deep image compression," *arXiv e-prints 2002.03370*, 2020.

[2] Mingxing Tan and Quoc V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019.

[3] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimized image compression," *arXiv e-prints 1611.01704*, 2016.

[4] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool, "Conditional probability models for deep image compression," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[5] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium (PCS)*, 2016, pp. 1–5.

[6] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 10771–10780.

[7] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[8] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy Image Compression with Compressive Autoencoders," *arXiv e-prints 1703.00395*, Mar. 2017.

[9] Lei Zhou, Chunlei Cai, Yue Gao, Sanbao Su, and Junmin Wu, "Variational autoencoder for low bit-rate image compression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[10] Oren Rippel and Lubomir Bourdev, "Real-time adaptive image compression," *arXiv e-prints 1705.05823*, 2017.

[11] Mohammad Akbari, Jie Liang, Jingning Han, and Chengjie Tu, "Generalized octave convolutions for learned multi-frequency image compression," *arXiv e-prints 2002.10032*, 2020.

[12] Jooyoung Lee, Seunghyun Cho, and Munchurl Kim, "An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization," *arXiv e-prints 1912.12817*, 2019.

[13] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020.

[14] Yoshua Bengio, Nicholas Léonard, and Aaron Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[15] Hyomin Choi and I. Bajic, "High efficiency compression for object detection," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1792–1796, 2018.

[16] L. Galteri, M. Bertini, L. Seidenari, and A. Del Bimbo, "Video com-

pression for object detection algorithms," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3007–3012.

[17] Vivek Sharma, Ali Diba, Davy Neven, Michael S. Brown, Luc Van Gool, and Rainer Stiefelhagen, "Classification-driven dynamic image enhancement," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[18] S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, "Image pre-transformation for recognition-aware image compression," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2686–2690.

[19] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, and G. Toderici, "End-to-end learning of compressible features," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3349–3353.

[20] Hugo Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy: Fixefficientnet," *ArXiv*, vol. abs/2003.08237, 2020.

[21] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "Density modeling of images using a generalized normalization transformation," *arXiv e-prints 1511.06281*, 2015.

[22] Ross Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2020.

Author Biography

Maxime Kawawa-Beaudan is a MS student in EECS at U.C. Berkeley advised by Professor Avideh Zakhor.

Avideh Zakhor is currently Qualcomm Chair and professor in EECS at U.C. Berkeley. Her areas of interest include theories and applications of signal, image and video processing and 3D computer vision.