Supplementary Material: Knowledge Distillation for GAN based Compression

Leonhard Helminger[1]
leonhard.helminger@inf.ethz.ch
Roberto Azevedo [2]
Abdelaziz Djelouah[2]
Markus Gross[1,2]
Christopher Schroers[2]

[1] Department of Computer Science
ETH Zurich, Switzerland
[2] DisneyResearch|Zurich

1 Knowledge Distillation for Video Compression with Latent Space Residuals

Let consider the sequence of frames $\mathbf{x}_0, \ldots, \mathbf{x}_{GOP}$. To compress the full sequence, the first frame, I-Frame or Keyframe, is compressed by an image compression codec to $\hat{\mathbf{x}}_0$. Given this compressed I-Frame, the codec computes the flow field $\mathbf{f} = (\mathbf{f}_y, \mathbf{f}_x)$ to the next frame. The decompressed flow field $\hat{\mathbf{f}}$ is then used to warp the previous frame $\hat{\mathbf{x}}_{t+1}^{warp} = \text{bilinear}\left(\hat{\mathbf{x}}_t, \hat{\mathbf{f}}\right)$. Similar as in [2] we use a *Motion Compensation* network to fix obvious errors of the warp, which eventually results in the final prediction $\mathbf{x}_{t+1}^{Pred}$. To compress the optical flow $\mathbf{f}$ we use the same auto encoder architecture as proposed in [1]. In our implementation we use spatial pyramid network [4] for flow field estimation.

To compute the latent residuals, we first encode both frames, $\mathbf{x}_{t+1}^{Pred}$ and $\mathbf{x}_{t+1}$ with the pre-trained HiFiC encoder and take the difference of the encodings $\mathbf{r}_{t+1} = \mathbf{y}_{t+1} - \mathbf{y}_{t+1}^{Pred}$. The probability distribution of the residuals is then learned with a Scale-Hyperprior [1]. By adding the decompressed residuals and the encodings of the prediction $\hat{\mathbf{y}}_{t+1} = \hat{\mathbf{r}}_{t+1} + \mathbf{y}_{t+1}^{Pred}$ we obtain the latents of the next frame $\hat{\mathbf{x}}_{t+1}$.

The training of the model is separated into four stages. In the first stage, we only train the motion vector compression network. The loss we optimize is defined as:

$$\mathcal{L}_{warp} = \lambda_w r\left(\hat{\mathbf{w}}_{t+1}\right) + MSE\left(\mathbf{x}_{t+1}, \hat{\mathbf{x}}_{t+1}^{warp}\right) \tag{1}$$

where $\mathbf{w}_{t+1}$ is the encoding of flow field $\mathbf{f}$ and λ_w is a hyperparameter controlling the trade-off between the distortion term and the rate term $r\left(\hat{\mathbf{w}}_{t+1}\right)$.

In the second phase, the *Motion Compensation* network is trained by optimizing the following loss:

$$\mathcal{L}_{mc} = \lambda_w r(\hat{\mathbf{w}}_{t+1}) + k_M MSE\left(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}^{Pred}\right) + k_M L_1\left(\mathbf{y}_{t+1}, \mathbf{y}_{t+1}^{Pred}\right). \tag{2}$$

Note, that the L_1 forces the model to learn predictions with latents close to latents of the ground truth image $\mathbf{x}_{t+1}$.

The loss of the third phase is defined as:

$$\mathcal{L}_{step} = \lambda_w \left(r(\hat{\mathbf{w}}_{t+1}) + r(\hat{\mathbf{r}}_{t+1})\right) + k_M MSE\left(\tilde{\mathbf{x}}_{t+1}, \hat{\mathbf{x}}_{t+1}\right) + k_p d_p\left(\mathbf{x}_{t+1}, \hat{\mathbf{x}}_{t+1}\right). \tag{3}$$

Important to note, in this phase we make use of the teacher-decoder. We minimize the loss between the compressed P-Frame $\hat{\mathbf{x}}_{t+1}$ and the *HiFiC* compressed frame $\tilde{\mathbf{x}}_{t+1}$.

In the fourth and final phase we optimize for $N = 3$ frames in one optimization step, to consider more temporal information and alleviate error accumulation [3]:

$$\mathcal{L}_{final} = \lambda \sum_{i=1}^{N} \left(r(\hat{\mathbf{w}}_i) + r(\hat{\mathbf{r}}_i)\right) + \sum_{i=1}^{N} k_M MSE\left(\tilde{\mathbf{x}}_i, \hat{\mathbf{x}}_i\right) + k_p d_p\left(\mathbf{x}_i, \hat{\mathbf{x}}_i\right), \tag{4}$$

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *ICLR*, 2018.

[2] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: an end-to-end deep video compression framework. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11006–11015. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01126. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Lu_DVC_An_End-To-End_Deep_Video_Compression_Framework_CVPR_2019_paper.html.

[3] Guo Lu, Chunlei Cai, Xiaoyun Zhang, Li Chen, Wanli Ouyang, Dong Xu, and Zhiyong Gao. Content adaptive and error propagation aware deep video compression. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 456–472. Springer, 2020. doi: 10.1007/978-3-030-58536-5_27. URL https://doi.org/10.1007/978-3-030-58536-5_27.

[4] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. *CoRR*, abs/1611.00850, 2016. URL http://arxiv.org/abs/1611.00850.