Knowledge Distillation for Learned Image Compression

Yunuo Chen[1*] Zezheng Lyu[2*] Bing He[1] Ning Cao[3] Gang Chen[3] Guo Lu[1 ✉] Wenjun Zhang[1]
[1] Shanghai Jiao Tong Unversity [2] Massachusetts Institute of Technology
[3] E-surfing Vision Technology Co., Ltd.

1. More Detailed Derivation

Theorem 1 (KL Divergence Comparison between models) *For the teacher model and the student model within the LIC framework described above, let M_1 denote the stage-wise method, M_2 denote the joint training method, J_i denote the absolute determinant value of the Jacobian Matrix of each block, $P_T(\boldsymbol{y}) = P(\boldsymbol{y}^T|\boldsymbol{\Theta}_a)$, and $P_S(\boldsymbol{y}) = P(\boldsymbol{y}^S|\boldsymbol{\Psi}_a)$. Assume the following:*

1. *Each block is differentiable and invertible.*
2. *Regardless of how the student model is trained, each J_i^{-1} follows an invariant distribution with fixed mean and variance, and is an unbiased estimator of the target teacher model's block.*
3. $\mathbb{E}_{M_2}[\prod_{i=1}^{3}(J_i^{-1})^2] \geq \prod_{i=1} \mathbb{E}_{M_1}(J_i^{-1})^2$.
4. $\mathbb{E}_{M_2}[\prod_{i=1}^{3} J_i^{-1}] \leq \prod_{i=1}^{3} \mathbb{E}_{M_1} J_i^{-1}$.

We can interpret Assumptions 3 and 4 as, in joint training, dependencies increase each block's co-movement in magnitudes but do not increase the absolute mean of each block's product.

Then, we state that:

$$D_{KL}(P_T(\boldsymbol{y})\|P_S(\boldsymbol{y}))_{M_1} < D_{KL}(P_T(\boldsymbol{y})\|P_S(\boldsymbol{y}))_{M_2}.$$

Proof First, we have:

$$J_T^{-1} := (J_1^T J_2^T J_3^T)^{-1}, \quad J_S^{-1} := (J_1^S J_2^S J_3^S)^{-1}.$$

Thus:

$$P_T(\boldsymbol{y}) = P(\boldsymbol{x}) \cdot J_T^{-1}, \quad P_S(\boldsymbol{y}) = P(\boldsymbol{x}) \cdot J_S^{-1}.$$

The KL divergence between the teacher and student latent distributions is defined as:

$$\begin{aligned} D_{\mathrm{KL}}(P_T\|P_S) &= \mathbb{E}_{P_T}\left[\log \frac{P_T(\boldsymbol{y})}{P_S(\boldsymbol{y})}\right] \\ &= \mathbb{E}_{P_T}\left[\log \frac{P(\boldsymbol{x}) \cdot J_T^{-1}}{P(\boldsymbol{x}) \cdot J_S^{-1}}\right] \\ &= \mathbb{E}_{P_T}\left[\log \frac{J_T^{-1}}{J_S^{-1}}\right] \\ &= -\mathbb{E}_{P_T}\left[\log J_S^{-1}\right] + \log J_T^{-1}, \end{aligned}$$

We perform a second-order Taylor expansion of $\log(J_S^{-1})$ around the deterministic Jacobian J_T^{-1}:

$$\log(J_S^{-1}) \approx \log(J_T^{-1}) + \frac{1}{J_T^{-1}}(J_S^{-1} - J_T^{-1}) - \frac{1}{2(J_T^{-1})^2}(J_S^{-1} - J_T^{-1})^2.$$

Taking expectations under P_T:

$$\mathbb{E}_{P_T}[\log(J_S^{-1})] \approx \log(J_T^{-1}) + \frac{1}{J_T^{-1}}\mathbb{E}_{P_T}[J_S^{-1} - J_T^{-1}] - \frac{1}{2(J_T^{-1})^2}\mathbb{E}_{P_T}[(J_S^{-1} - J_T^{-1})^2].$$

Since we assume that on average the student training is unbiased around the teacher distribution, we have:

$$\mathbb{E}_{P_T}[J_S^{-1} - J_T^{-1}] = 0.$$

Thus, the expectation simplifies clearly to:

$$\mathbb{E}_{P_T}[\log(J_S^{-1})] \approx \log(J_T^{-1}) - \frac{1}{2(J_T^{-1})^2}\mathrm{Var}(J_S^{-1}).$$

Substitute back into the original KL expression:

$$\begin{aligned} D_{\mathrm{KL}}(P_T\|P_S) &= -\mathbb{E}_{P_T}[\log(J_S^{-1})] + \log(J_T^{-1}) \\ &\approx -\left(\log(J_T^{-1}) - \frac{1}{2(J_T^{-1})^2}\mathrm{Var}(J_S^{-1})\right) + \log(J_T^{-1}). \end{aligned}$$

The $\log(J_T^{-1})$ terms cancel neatly, giving explicitly:

$$D_{\mathrm{KL}}(P_T\|P_S) \approx \frac{\mathrm{Var}(J_S^{-1})}{2(J_T^{-1})^2}.$$

Thus:

$$\begin{aligned} \frac{D_{\mathrm{KL}}(P_T\|P_S)_{M_2}}{D_{\mathrm{KL}}(P_T\|P_S)_{M_1}} &= \frac{\mathrm{Var}(J_S^{-1})_{M_2}}{\mathrm{Var}(J_S^{-1})_{M_1}} \\ &\geq 1 \quad \text{(Assumption 3 and 4)}, \end{aligned} \tag{1}$$

which completes the proof.

∗ Equal Contribution ✉ Corresponding Author

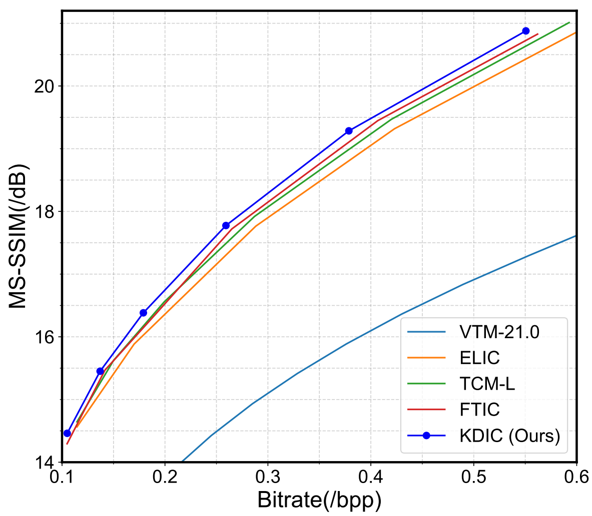

Figure 1. RD curves of MS-SSIM on CLIC dataset.

2. RD-Curve or MS-SSIM

As shown in Fig. 1, we provide more RD curves about MS-SSIM on the CLIC dataset. We compare our KDIC model with VTM-21.0 [1], FTIC [2] and TCM [3].

3. Settings of VTM-21.0

We utilize VTM-21.0 [1] and demonstrate sample bash commands for encoding and decoding a YUV format image with VTM-21.0.

```
VTM-21.0/bin/EncoderAppStatic -i tmp.yuv -c VTM-21.0/cfg/encoder_intra_vtm.cfg -q 61 -o /dev/null -b tmp.bin -wdt 768 -hgt 512 -fr 1 -f 1 --InputChromaFormat=444 --InputBitDepth=8 --ConformanceWindowMode=1

VTM-21.0/bin/DecoderAppStatic -b tmp.bin -o tmp.yuv -d 8
```

References

[1] A. Browne, Y. Ye, and S. Kim. Algorithm description for versatile video coding and test model 21 (vtm 21), document jvet-af2002. In Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 32nd Meeting, Hannover. 2

[2] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. In The Twelfth International Conference on Learning Representations, 2024. 2

[3] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14388–14397, 2023. 2