2024/09/03 Multimedia Seminar Minutes

English follows Japanese

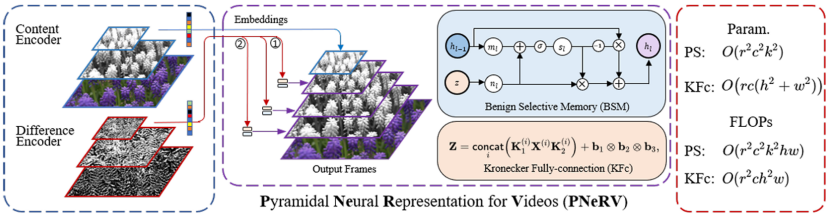
- Hayami

PNeRV: Enhancing Spatial Consistency via Pyramidal Neural Representation for Videos

(CVPR 2024)

NeRVの問題点⇒ 時間的な一貫性〇　空間的な一貫性✖

提案手法



1. ピラミッド型NeRV
2. Kronecker Fully-connected (KFc)レイヤー: アップサンプリングの工夫

   従来のNeRVはアップサンプリングにPixelShuffle(伸ばしたチャネル方向を縦横につぶす）を使用

   ⇒画素間の長距離関係のモデル化能力に欠ける

   KFcの計算式：

$$\mathbf{Z} = \underset{i}{\text{CONCAT}} \left(\mathbf{K}_1^{(i)} \mathbf{X}^{(i)} \mathbf{K}_2^{(i)} \right) + \mathbf{b}_c \otimes \mathbf{b}_h \otimes \mathbf{b}_w,$$

   Xは入力特徴量、Zは出力特徴量、$K_1^{(i)}$と$K_2^{(i)}$はchannel iの2つのカーネル、bc, h, wは3つのベクトルで、クロネッカー積⊗によってBIASを出力する
3. BSM：ピラミッド階層間の特徴量の融合方法

   入ってきた情報の重要度を学習で判断しながら、融合する

- Tatsumi

Learned Image Compression with Mixed Transformer-CNN Architectures

(CVPR2023)

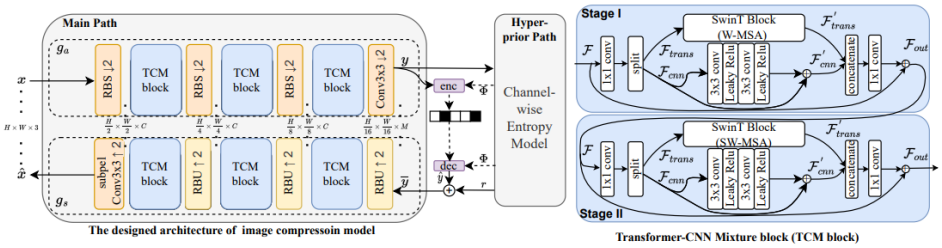
既存のLICモデルのほとんどはCNN-basedやtransformer-basedである。

CNNは局所的なモデリング、Transformerはより大まかなモデリングを得意とする。

提案手法

1. parallel Transformer-CNN Mixture (TCM) ブロックをもつLICフレームワーク

   residual networksとSwin-Transformer ブロックを並列に置くことで、複雑度を抑えつつ、CNNとTransformerの良いところを両立する。



The designed architecture of image compressoin model　　Transformer-CNN Mixture block (TCM block)

2. parameter efficient Swin-transformer-based attention (SWAtten)モジュールを持つchnnel-wise auto-regressive エントロピーモデル

- SWAttenモジュール自体が従来のattentionモジュールよりもパラメータが少ない
- attentionモジュールをmain pathではなくhyper-prior pathに置く

  ⇒モデルのcomplexityを削減

- Takabe

Multi-Scale 3D Gaussian Splatting for Anti-Aliased Rendering

(CVPR2024)

aliasingが原因で、低解像度画像での3D Gaussian Splattingのレンダリングはうまくいかない

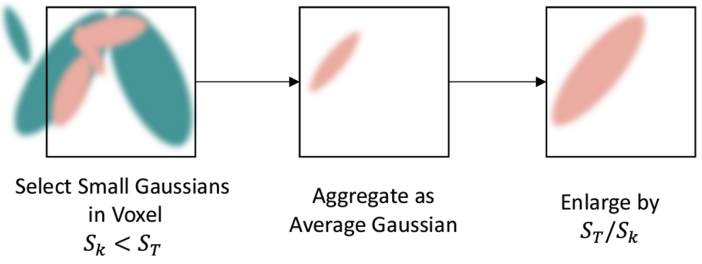
提案手法：pixel coverage

- 見たい解像度でのピクセルサイズと比較したときの、ガウシアンのサイズを反映する
  ⇒ つまり、見たい解像度に対して、あるピクセルにガウシアンがどれだけかかっているかを計算する

ここで、単純にナイキスト周波数でpixel coverageをフィルターすると、ただ小さい高周波のガウシアンを消すだけなので、画像内に欠けてしまう部分がでる。

そのため、まず解像度（ピクセルのサイズ）ごとに、小さすぎるガウシアンは平均をとってまとめて、pixel coverageを考慮して大きなガウシアンを作る。

（これにより、ガウシアンのうち5%未満のみが集約される）



新たに作った大きなガウシアンも含めて、学習を行う。

学習後、レンダリングのときには、レンダリングの解像度とガウシアンの大きさを考慮して、使うガウシアンを決める。

- 低解像度では高周波成分は使わず、新たに作った大きなガウシアンを使う。
- 高解像度では新たに作った大きなガウシアンは使わず、高周波成分をそのまま使う。

結果

解像度が低くなるにつれてクオリティと処理速度が3DGSよりもどんどん良くなる

しかし、線形的にはレンダリング速度は下がらない

（レンダリングの時に、ガウシアンが少しでもpixelにかかると計算に含めなければならないため）

- Tanaka

V3.0_ C3 High-performance and low-complexity neural compression from a single image or video

(CVPR2024)

C3 = COOL-CHICのversion 3

潜在変数をラウンディング関数で量子化すると、生じる量子化ノイズはステップ関数のような微分不可能な関数なので、これをそのまま学習に使用できない。そのため従来のCOOL-CHICでは、量子化ノイズの範囲を網羅するように一様ノイズを付加して、微分可能にしていた。

提案手法

学習の初期段階では、入力データのわずかな変動が大きな量子化誤差を引き起こして勾配が不安定になり、学習が困難になる可能性があるので、ノイズの範囲が大きい量子化を使用する。

その後段階的な量子化の近似を行う。

- 微分可能なsoft rounding 関数でノイズを滑らかに近似する
- soft rounding関数のTの値とkumaraswamy分布を調整して、kumaraswamyノイズの範囲を狭める⇒rounding関数に近似していく

単純な一様ノイズと比較するとより滑らかに量子化近似ができて、正確な復号が可能

English version

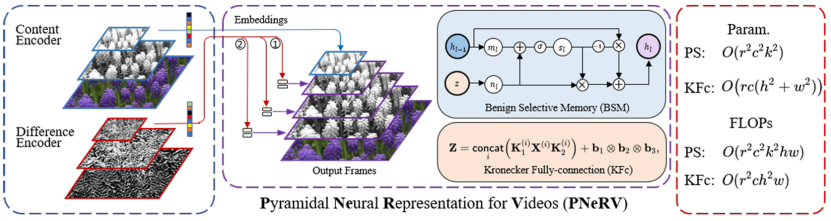
- Hayami

PNeRV: Enhancing Spatial Consistency via Pyramidal Neural Representation for Videos

(CVPR 2024)

Problems of NeRV ⇒ Good temporal consistency ; Bad spatial consistency

Proposed Methods



Pyramidal Neural Representation for Videos (PNeRV)

1. Pyramidal NeRV
2. Kronecker Fully-connected (KFc) Layer: A refinement in upsampling

   Traditional NeRV uses PixelShuffle (which redistributes expanded channel directions horizontally and vertically) for upsampling

   ⇒ Lacks the capability to model long-range relationships between pixels

   Formula for KFc:

$$\mathbf{Z} = \underset{i}{\text{CONCAT}} \left(\mathbf{K}_1^{(i)} \mathbf{X}^{(i)} \mathbf{K}_2^{(i)} \right) + \mathbf{b}_c \otimes \mathbf{b}_h \otimes \mathbf{b}_w,$$

   X is the input feature, Z is the output feature, $K_1^{(i)}$ and $K_2^{(i)}$ are two kernels for channel i, and b_c, b_h, b_w are three vectors that produce BIAS through the Kronecker product $\otimes$
3. BSM: Method for fusing features between pyramidal layers

   It fuses information while learning the importance of incoming information

---

- Tatsumi

Learned Image Compression with Mixed Transformer-CNN Architectures

(CVPR2023)

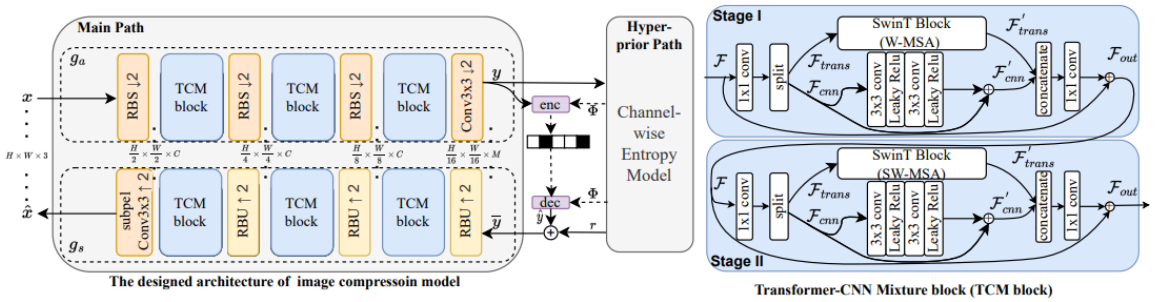
Most of the existing LIC models are CNN-based or Transformer-based.

CNN is good at local modeling while transformer is good at non-local modeling.

Proposed method:

1. LIC framework with parallel Transformer-CNN Mixture(TCM) blocks

   By placing residual networks and Swin-Transformer blocks in parallel, it achieves a balance between the strengths of CNNs and Transformers while maintaining low complexity.



The designed architecture of image compressoin model

Transformer-CNN Mixture block (TCM block)

2. Channel-wise auto-regressive entropy model with a parameter-efficient Swin-transformer-based attention (SWAtten) module

- The SWAtten module itself uses fewer parameters compared to traditional attention modules.
- The attention module is placed in the hyper-prior path rather than the main path.

  ⇒ Reduces model complexity.

- Takabe

Multi-Scale 3D Gaussian Splatting for Anti-Aliased Rendering

(CVPR2024)

Due to aliasing, rendering of 3D Gaussian Splatting (3DGS) in low-resolution images does not work well

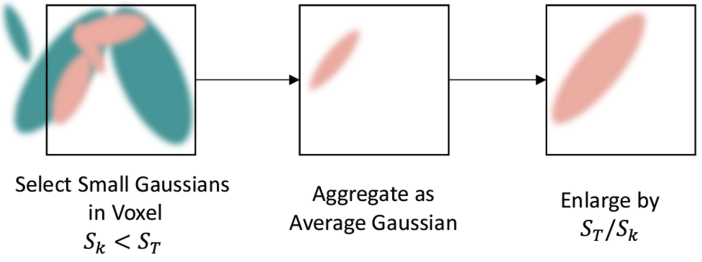
Proposed Method: Pixel coverage

Pixel coverage reflects the size of the Gaussian when compared to the pixel size at the desired resolution.
In other words, it calculates how much a Gaussian affects the given pixel at the desired resolution.

Simply filtering the pixel coverage at the Nyquist frequency would only eliminate small high-frequency Gaussians, resulting in missing parts in the image.

Therefore, first, for each resolution (pixel size), create large Gaussians by aggregating the small Gaussians in each voxel below the pixel coverage threshold, and then enlarge by the pixel coverage multiplier.

(This aggregates only less than 5% of the Gaussians.)



Include the newly created larger Gaussians in the training.

When rendering, the resolution and size of the Gaussians are considered to decide which Gaussians to use.

- Use newly created large Gaussians instead of high-frequency components at low resolutions.
- Use high-frequency components directly at high resolutions, not the newly created large Gaussians.

Results

As the resolution decreases, both quality and processing speed improve significantly compared to 3DGS.

However, the rendering speed does not decrease linearly.

(Because when rendering, any Gaussian that even slightly affects a pixel must be included in the calculations.)

- Tanaka

V3.0_ C3 High-performance and low-complexity neural compression from a single image or video

(CVPR 2024)

C3 = COOL-CHIC version 3

Quantizing the latent variables using a rounding function results in quantization noise, which is a non-differentiable function like a step function, and thus cannot be directly used for learning. Therefore, the traditional COOL-CHIC added uniform noise over the range of quantization noise to make it differentiable.

Proposed Method

In the early stages of learning, slight variations in input data can cause significant quantization errors, making gradients unstable and learning difficult, so a wide range of quantization noise is used.

Gradual approximation of quantization is then performed.

- Smoothly approximate noise using a differentiable soft rounding function.
- Adjust the T value of the soft rounding function and the kumaraswamy distribution to narrow the range of kumaraswamy noise, approximating it to the rounding function.

Compared to simple uniform noise, this allows for a smoother quantization approximation and accurate decoding.