2024/09/17 multimedia

- shindo

Entroformer: A Transformer-based Entropy Model for Learned Image Compression

(ICLR 2022)

Transformerを用いた最初の Learned Image Compression model.

Transformerを搭載する場所は hyper-prior path と context model.

Transformerを使用することのメリットは、画像のより広域の関連性の取得

新規性

3-1 Transformer architecture

一般的な、普通のattention module を搭載した、という内容

参考：attention is all you need (1706.03762 (arxiv.org))

3-2 Position Encoding

context の関連性の強さは、各位置を中心にダイヤモンド型に広がる

この特性を利用したcontext model の作成

attention weight を各コンテキスト間の距離を用いて作成することで達成

参考：relative position (1803.02155 (arxiv.org))

3-3 Top-k Scheme in self-attention

関連性のより高いtop-k attention weight のみを使用してattention算出

3-4 Parallel Bidirectional Context Model

checker board context model を参考に高速化を図る

- jin

ReNoise: Real Image Inversion Through Iterative Noising

(ECCV 2024)

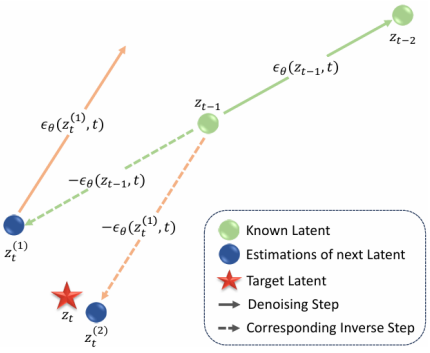
より良いadd noise過程を提案する.

$$z_t = \frac{z_{t-1} - \psi_t \epsilon_\theta(z_t, t, c) - \rho_t \epsilon_t}{\phi_t}, \qquad (2)$$

理想的なノイズ過程は、式(2)により得られる. しかし、この式は右辺においてもztを使用しており、近似を使用して解く必要がある. 近似式は以下により得られる.

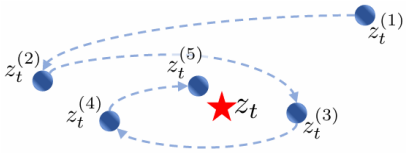
$$z_t^{(1)} = \frac{z_{t-1} - \psi_t \epsilon_\theta(z_{t-1}, t, c) - \rho_t \epsilon_t}{\phi_t}. \qquad (3)$$

右辺のztをzt-1に置き換えることで、計算を可能にする. 一方で、これはただの近似でしかなく、正確な値ではない. これを改善するための手法として本論文では、以下を提案している.



一度作成したzt(1)を用いて、zt(2)を作成、より本物のztに近くなるように計算を行う.

これは、以下の図のように、また、論文中の証明(4.Convergence Discussion)のように、収束することが示されている.



このステップをrenoising stepと呼び、反復しながら真のztを推測する.

Renoising step の導入の有無で、以下の様に生成結果に差が出る.

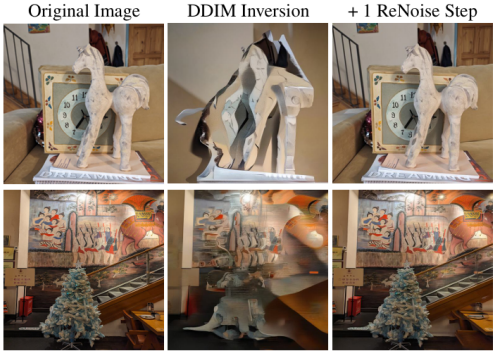


Figure 3. Comparing reconstruction results of plain DDIM inversion (middle column) on SDXL to DDIM inversion with one ReNoise iteration (rightmost column).

- yenan

Audio Match Cutting: Finding and Creating Matching Audio Transitions in Movies and Videos

(ICASSP 2024)

audio 情報を用いた映像接続に関する研究.

複数のvideoの中から、類似した音声を持つ映像２本を検索し、不自然無いように接続するタスク.

test 映像は公開：https://denfed.github.io/audiomatchcut/

実装方法についてはあまり詳しいことが書かれておらず.

音声特徴抽出器はCLAP, フリーズ状態で使用. 学習はprojection layer.

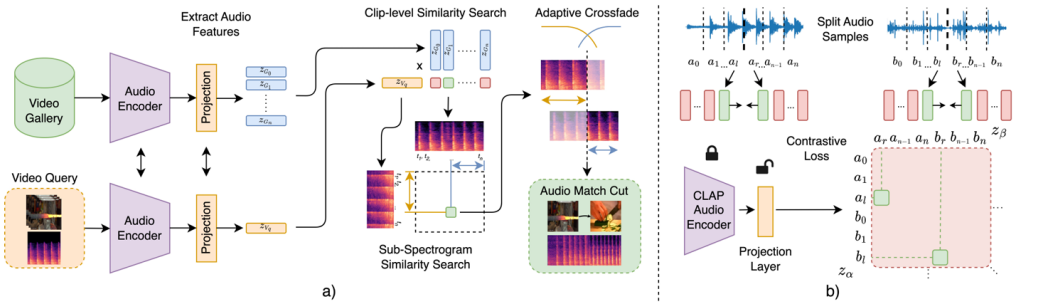


Fig. 2. a) Proposed Framework. Given a query video, we retrieve an audio match cut candidate from a video gallery and find the optimal transition point using a sub-spectrogram similarity search. Using the variance of the created similarity matrix, we adaptively select the crossfade length to blend both the query and match audio into a fluid audio match cut. b) Proposed "Split-and-Contrast" contrastive objective. Each audio sample is split at a randomly selected frame, then the adjacent frames of the split are contrasted towards each other.

結果は良いという主張だが、悪くはないけれど良いかどうかは怪しめ.