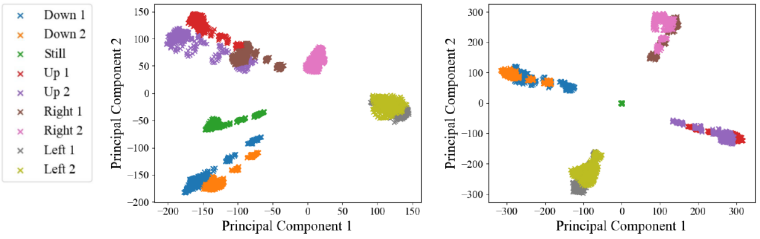
Multimediaゼミ2024年11月19日

- watanabe
  - アブスト、イントロ
    - PCAなどを通して動画の拡散モデルがモーションを認識しているか確認し、モーションを取り出す方法を提案した
    - コンテンツの情報を除去することが、Unetなどの動画の拡散モデルの特徴量におけるモーション情報を際立たせるのに有効であることを示した。これに主成分分析（PCA）を適用した結果、主成分とビデオのモーションとの間に強い相関が存在することを確認した。さらに調査を進めると、特徴の特定のチャネルがモーションの方向を決定する上で他のチャネルよりも重要な役割を果たしていることが明らかになった。
    - さらに、MOFTに基づいた新しいトレーニング不要のビデオモーション制御フレームワークを提案した。異なるアーキテクチャやチェックポイントごとに独立したトレーニングが必要な従来のトレーニングベースの方法とは異なり、訓練不要なので我々の方法はさまざまなアーキテクチャやチェックポイントに容易に適用可能。
  - MOtion FeaTures (MOFT)
    - 拡散モデル中の特徴量からモーション情報を抽出するのは他のセマンティックな情報（物体の形など）なども含まれているために難しい
      - このようなセマンティックな情報は全てのフレームに共通していると考えた。そこで、以下のように各フレームの特徴量からその平均を引いたものを求めた。ここにモーションの情報があるのではないかと考えた。

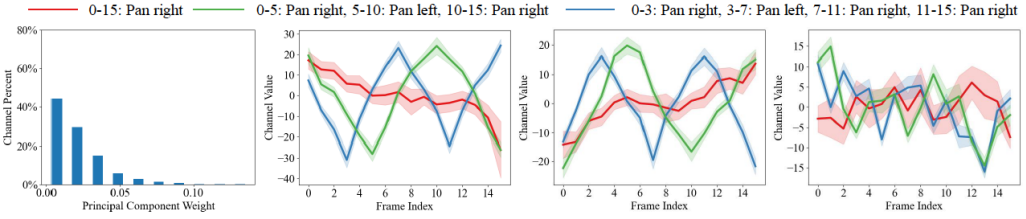
$$\mathcal{X}^{\text{norm}} = \mathcal{X} - \frac{1}{F}\sum_{i=1}^{F}\mathcal{X}_i,$$

      - これを検証するために、様々な動きがある動画を用意して、元のXとX_normでPCAを行い、動きとの相関性を見たところ、X_normがモーションをよく捉えていることがわかった



(a) PCA of vanilla feature (b) PCA of feature after correlation removal

- また、モーションの情報を持っているのは、一部のチャンネルではないかと考え、各チャンネルの主成分の構成度合いについても調査した。以下から分かるように一部のチャンネルが主成分を構成しており、これらがモーションに関わっていると判断した。（モーションチャンネル）それを検証するために、主成分の重みが大きいチャンネルと小さいチャンネルで比較した。全てのフレームで右方向や左方向に動いてる動画や右にいったり左にいったりしている動画で相関があるか検証し、モーションチャネルの方が相関があるとわかった。



(a) PC weight hist. on channels (b) Value of a motion channel (c) Value of a motion channel (d) Value of non-motion channel

- 上記を使用して、格差モデルの特徴量からモーションの情報を取り出すスキームを提案
  - チャンネルごとにはPCAでモーションチャンネルを抽出し、フレームごとには平均を引いて差分だけ残したもの

$$\mathcal{M} = (\mathcal{X}_{[j]} - \frac{1}{F}\sum_{i=1}^{F}\mathcal{X}_{i,[j]}), \quad j \in \mathcal{C},$$

- Fig4は、MOFTが一番よくて、取ってくる特徴量はUp Block-1が一番よく、いろんな動画の拡散モデルに適応可能なことを示している
- Motion Control
  - 推論時のデノイジングステップの際に以下の損失関数でノイズを最適化する

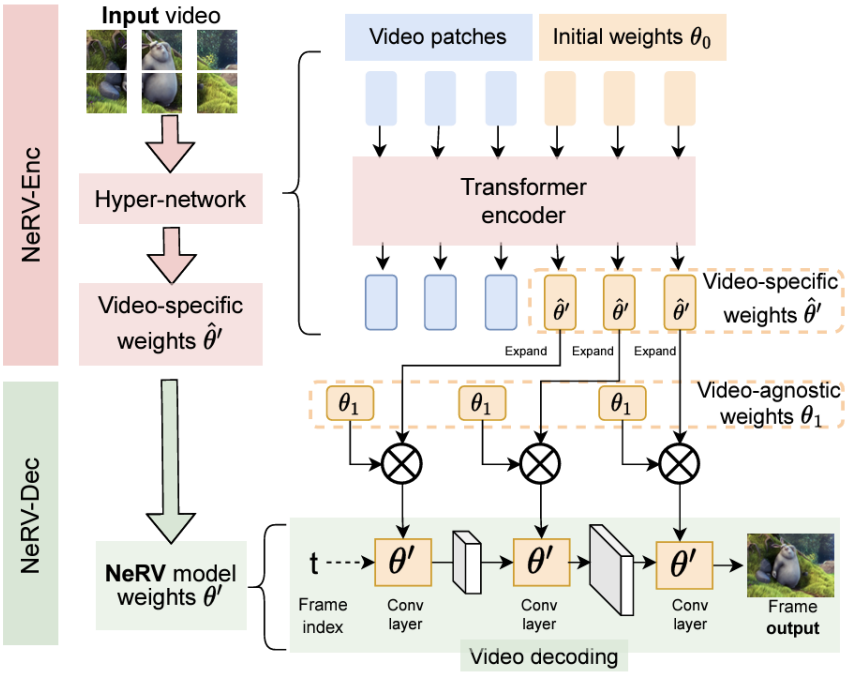
$$\mathcal{L}^c = \frac{1}{|\mathcal{R}|} \sum_{(i,j)\in\mathcal{R}} ||\mathcal{M}_{i,j} - \mathcal{M}^r_{i,j}||,$$

R - モーションをコントロールしたい領域：特徴量単位だとわからなくないか？全部？

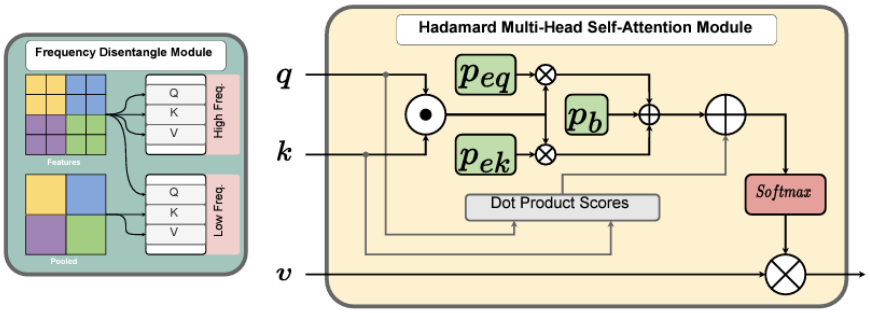
M^r - reference MOFT：参照動画のMOFTかチャンネルの値の最小値や最大値がPCAからわかるからその情報を使用したMOFT(?)

- Trajectoryを使用した方法よりも良い

- hayami
  - INRにおける動画表現手法はエンコード（学習）時間が長い
  - エンコードとデコードの速度改善に焦点を当てている
    - NeRVのEnc：Transformer-baseのHyper-network
    - NeRVのDec：並列デコーダ
  - INR
    - 座標ベース：ピクセル単位で学習・推論
    - フレームベース：フレーム単位で学習・推論
    - 動画のINRはフレームベースの方が品質・デコード速度の面で優れる
    - Hyper-Networks
      - Hyper-networkにより別のNeural Networkのモデルの重みを生成
  - 提案手法
    - Enc
      - transformer-baseのhyper-network
        - video-specific weightsを学習し、video-agnostic weightsと掛け合わせる
        - video-agnostic weightsは様々なデータで学習した重み
    - Dec
      - 掛け合わせた重みをconvolutionの重みとしてフレームインデックスからフレームを作成する
      - 複数NeRVを組み合わせることで並列化が可能



- tatsumi
  - 人が見る際には高周波数ではなく、低周波数を見ている
  - 特徴量を高周波と低周波に分ける
  - 提案手法
    - Transformerの中のMMAを二つに分ける
      - low frequencyとhigh frequencyに分けてconcatする
      - channel wise
        - yをチャネル方向に分けて順番に復号していくもの
        - low frequencyとhigh frequencyに分けることで、最初の方のスライスにlow frequencyが含まれ、最後の方のスライスにはhigh frequencyが含まれる。low frequencyを人が見ているので、そこに大きいビットを割り当てる



(c) Frequency disentangle module (left) / Hadamard augmented self-attention module (right).

- Transformerは画像中の離れた情報の関係性をとれる
  - local self-attentionがそれを妨げているので、以下のようなアダマール積の式を利用してアテンションスコアを計算した

$$v_i' = \sum_{j \in \Phi} \frac{e^{q_i^T k_j + (q_i \odot k_i) p_{ek}^o + p_{eq}^o (q_j \odot k_j) + p_b^o}}{\sum_{j \in \Phi} e^{q_i^T k_j + (q_i \odot k_i) p_{ek}^o + p_{eq}^o (q_j \odot k_j) + p_b^o}} v_j,$$