MOSO: Decomposing MOtion, Scene and Object for Video Prediction

1. **とりあえず一言**

   VQVAE + Transformerの動画予測モデル。動画をモーション、シーン、物体に分けて処理した

2. **どんなもの？**

   - VQVAE
     - 入力された動画をモーション、シーン、物体に分ける
       - motionは現在のフレームと過去のフレームとの差
       - motionと閾値を設けて、よく動いているピクセルを物体とする
       - それ以外のピクセルをシーンとする

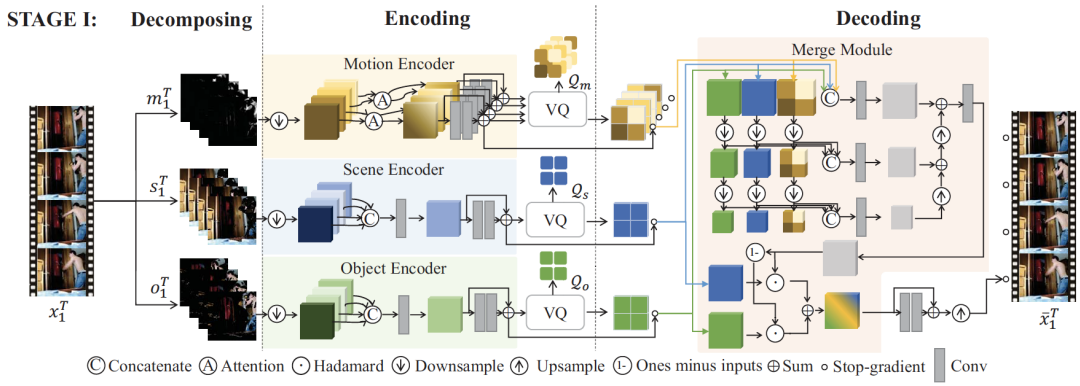
Algorithm 1 Preprocessing algorithm.

Input: Video frames x_1^T
Parameter: c_{lb}, c_{ub} and channel dimension d_c
Output: Motion, scene and object videos

Let $t = 1$, $x_s = x_1$.
while t $\leq$ T **do**
  $x_{nxt} = x_T$ if $t == T$ else x_{t+1}
  $m_t = 2x_t - x_s - x_{nxt}$
  $d_{pixel} = max(abs(m_t), dim = d_c)$
  $mask = (d_{pixel} \geq c_{lb}) \odot (d_{pixel} \leq c_{ub})$
  $o_t = mask \odot x_t$
  $s_t = (1 - mask) \odot x_t$
end while
return m_1^T, s_1^T and o_1^T

     - 分解されたモーション、シーン、物体は、それぞれ別々のEncoderを通して潜在変数となる。次にCodebookを通して離散化されTokenに変換す

る。最後にDecoderでこれらを融合して再構成する

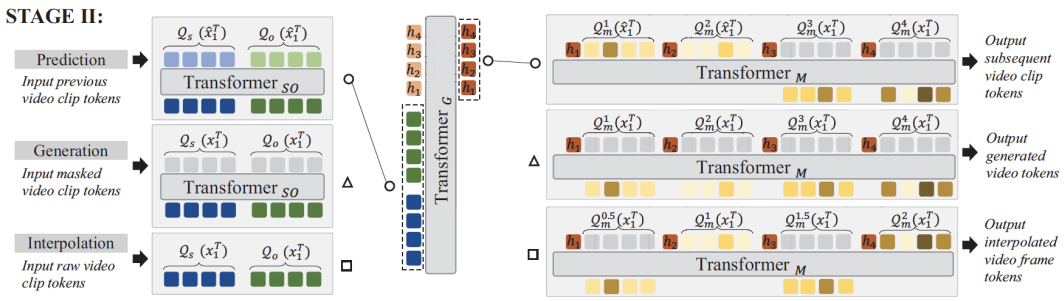


- 損失関数
  - recon loss + commitment loss

- Transformer
  - VQVAEの入力のシークエンス長は決まっているので、入力のシークエンス長を変更したい場合は、シークエンス長分最後のフレームをコピーする。

$$\hat{x}_1^T = \{x_1, ..., x_{K-1}, x_K, \underbrace{x_K, ..., x_K}_{T-K}\}, \quad K \leq T$$

3つのTransformerを用いて未来のmotion, scene, objectのTokenを予測する。



  - Transformer_SO
    - motion encoder, scene encoderの出力のTokenから未来のmotion, sceneのTokenを予測する

        - Transformer_G
            - Transformer_Mの予測で使用するhを出力する
        - Transformer_M
            - hとmotionのTokenから未来のmotionのTokenを出力する
    - 損失関数
        - Cross Entropy (実装を見るとマスクしているTokenをTransformerにかけて、マスクしていないTokenとのcross entropyを取っているが、詳細については不明)

3. **先行研究と比べてどこがすごい？**

   content, motionからscene, object, motionに拡張

4. **技術や手法のキモはどこ？**

   分割, transformerでrecon lossを取っていない

5. **どうやって有効だと検証した？**

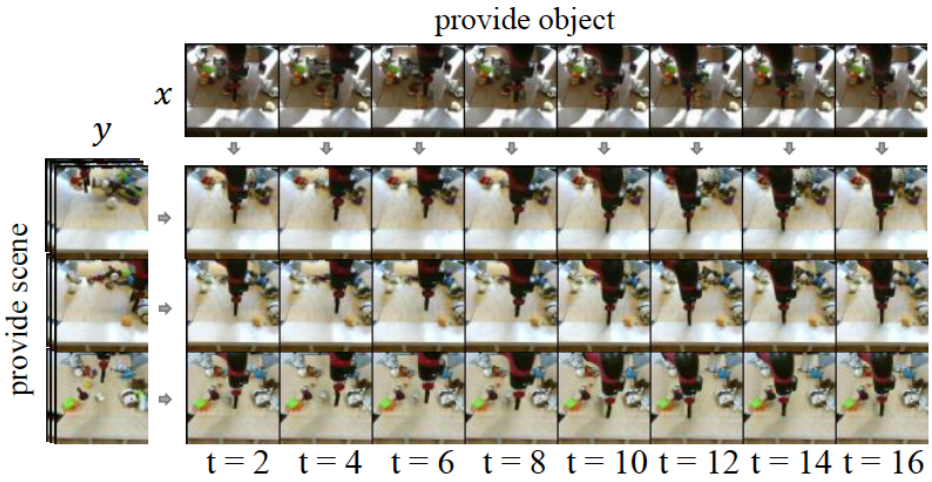
   BAIR, RoboNet, KTH, KITTI, UCF101で比較

   評価指標はFVD, FID, PSNR, SSIM, LPIPS

6. **議論はある？**

   unconditional video generation, video interpolation, video manipulationができる

   - video manipulation
       - 二つのシークエンスのmotion, scene, objectのtokenを組み合わせることで、別の動画を作成できる



7. **次に読むべき論文は？**

8. **参考文献**

9. **メモ**

Abstract :

Abstract :

> Motion, scene and object are three primary visual components of a video. In particular, objects represent the foreground, scenes represent the background, and motion traces their dynamics. Based on this insight, we propose a two-stage MOtion, Scene and Object decomposition framework (MOSO) for video prediction, consisting of MOSO-VQVAE and MOSO-Transformer. In the first stage, MOSO-VQVAE decomposes a previous video clip into the motion, scene and object components, and represents them as distinct groups of

> discrete tokens. Then, in the second stage, MOSO-Transformer predicts the object and scene tokens of the subsequent video clip based on the previous tokens and adds dynamic motion at the token level to the generated object and scene tokens. Our framework can be easily extended to unconditional video generation and video frame interpolation tasks. Experimental results demonstrate that our method achieves new state-of-the-art performance on five challenging benchmarks for video prediction and unconditional video generation: BAIR, RoboNet, KTH, KITTI and UCF101. In addition, MOSO can produce realistic videos by combining objects and scenes from different videos.