Disentangling Embedding Vectors for Controllable Facial Video Generation

Anonymous Author(s)

ABSTRACT

The task of editing video aims to control the content whilst generating realistic and coherent videos. The embedding vectors of an encoding-decoding architecture can be manipulated to create novel videos with certain characteristics, but they are typically entangled, making editing difficult and generalization weak. In this paper, we propose a novel vision transformer architecture and contrastive training regime for facial video generation and editing. Our model is able to disentangle embedding vectors, which yields embeddings with semantic interpretations. This allows for manipulation of videos in a direct and intuitive manner. We show that our model is effective for facial video editing. This has many potential applications in the animation, gaming, and video editing industries.

KEYWORDS

Disentangled vectors, video editing, video generation

ACM Reference Format:

Anonymous Author(s). 2023. Disentangling Embedding Vectors for Controllable Facial Video Generation. In *Proceedings of The 15th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Video content editing is a challenging task, with many applications in the animation and gaming industries. One approach is to encode each frame, manipulate the encoding and then decode the frame to produce novel videos. However, the embedding vectors (or latent codes) are typically entangled, meaning that the axes of the vectors do not have clear semantic interpretations. This entanglement makes editing difficult and generalization to new scenarios weaker.

This paper proposes a novel contrastive training approach to disentangle embedding vectors, and demonstrates the effectiveness of this approach when applied to facial video editing.

2 METHODOLOGY

2.1 Model Architecture

In recent years, vision transformers (ViTs), [Dosovitskiy et al. 2020] have achieved state-of-the-art results in many image processing and vision tasks, such as image classification, object detection, and image segmentation. ViTs are particularly well-suited for video generation, as they can learn long-range dependencies and spatial-temporal contexts in videos. In this paper, we use a vision transformer architecture for video generation. Our architecture is based on the following key components:

- A spatial transformer encoder to learn spatial features from each frame in the video.
- A spatial transformer decoder to generate the output video frame by frame.
- A temporal transformer to learn temporal dynamics.
- A latent code editor to edit the embedding representation of the video before decoding.

Our spatial transformer encoder is a standard ViT encoder, which divides each frame into $32 \times 32 \times 3$ patches mapped onto 128-dimensional vectors and then uses self-attention to learn spatial features from these patches. Specifically, we have a 10-layer 8-headed transformer to perform encoding, and each of the patches uses additive position encoding. We adopt layer normalisation and use a 0.1 dropout during training to improve generalisation and help training.

Our decoder has the same ViT architecture as the spatial encoder with a 2-layer feed-forward network to cast the latent vector into image space. Note that as we want each dimension of our embedding vector to carry independent semantic information, we do not use dropout on the first layer of this decoder, as it encourages redundancy in the representation.

The temporal transformer works on the embedding vectors produced by the spatial encoder, generating vectors in the same space. As the present work focuses on the disentangling and is orthogonal to the temporal prediction, we will not detail the temporal transformer further.

The latent code editor simply maps each user action onto one specific dimension of the embedding vector before being decoded, and again is not the focus of this current work.

2.2 Training

The combined architecture is trained to minimize the reconstruction error between the generated video and the ground-truth video. Specifically, we use a SSIM similarity metric [Wang et al. 2004] to measure the reconstruction error, modified to emphasize the luma:

$$\mathrm{SSIM}(\mathrm{image}_1^{rgb}, \mathrm{image}_2^{rgb}) + \mathrm{SSIM}(\mathrm{image}_1^{gray}, \mathrm{image}_2^{gray})$$

A proprietary database of around 400 hours of video footage is used, with both natural conversational recordings as well as recordings of isolated facial movements. Heads are tracked and faces are extracted and normalized.

During training, contrastive learning [Chen et al. 2020] is applied. The recordings of isolated facial movements are used to disentangle the embeddings by constraining the model to associate only a single dimension of the embedding vector with each facial movement. Specifically, pairs of images from the same motion are presented

---

MIG '23, November 15—17, 2023, Rennes, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX