Video Interpolation and Extrapolation using a Transformer with Relative Positional Embedding & Relative Global Attention

Autumn Beyer Mitchell Johnstone Sam Keyser Ryan Kruk
Tyler Schreiber Tillie Pasternak Michael Conner

Department of Electrical Engineering & Computer Science

Milwaukee School of Engineering

{beyera, johnstonem, keysers, krukr, schreibert, pasternakt, connerm}@msoe.edu

Abstract

Video frame interpolation is a popular technique used to increase the frame rate of a video sequence, resulting in smoother and more fluid playback. This process involves generating intermediate frames between existing ones, which fill in the gaps and produce a more natural and visually appealing video. The goal of this technique is to estimate motion information between frames and synthesize new pixels to fill in the gaps. This is typically achieved using convolutional neural networks (CNNs) that have been trained on large datasets of videos. In recent years, transformers have been used for video interpolation tasks, showing significant progress in this field. However, there is still limited knowledge on the use of relative positional embeddings to help capture more complex relationships between different sections of frames based on position. To address this gap, our research investigates the use of relative positional embeddings in video frame interpolation and extrapolation. Our goal is to capture complex spatial relationships between frames in a video sequence that can improve the accuracy and quality of interpolated frames. In addition to exploring the use of relative positional embeddings in video frame interpolation, we also investigate their effectiveness in video extrapolation. Video extrapolation involves generating new frames beyond the end of a given video sequence, which is a challenging task due to the lack of visual information available. By using relative positional embeddings, we aim to capture the spatial relationships between frames in both the forward and backward directions, which can lead to more accurate and realistic extrapolation results. To conduct our experiments, we use the Vimeo90K dataset, which is widely used in the field and allows for easy comparison with other models. Our research contributes to the growing body of knowledge on the use of transformers in video processing and

provides new insights into the potential benefits of relative positional embeddings. Video frame interpolation and extrapolation are essential techniques used in various applications, such as video compression, slow-motion effects, and video enhancement. Our research aims to improve the accuracy and quality of these techniques. In conclusion, we believe that our research will provide new insights into the use of relative positional embeddings in video processing and contribute to the development of more effective and accurate video frame interpolation and extrapolation methods.

1 Introduction

Video interpolation is the process of generating a frame in between two consecutive frames. This process increases the frame rate of a video, allowing users to create slow motion videos in the native frame rate or create a smoother video with more frames per second. Beyond the obvious applications to media, video interpolation also holds promise for better data compression. If high quality video interpolation can be achieved, videos can be stored as a set of several key frames, with the intermediate frames being recoverable via interpolation.

Video extrapolation, on the other hand, aims to generate a new frame after being given a sequence of some number of frames. The traditional application of video extrapolation is for video prediction, which is predicting the next image given a sequence of preceding images. Extrapolation can also be used for novel view synthesis (NVS), which is the general task of trying to generate a view of a scene given some number of complementary views [3]. An example might be trying to generate the side profile of a car, given a view of the car at 45 degrees.

Recently transformers have been applied to the problem of video interpolation. Zhihao et al. recently published a video interpolation model, the Video Frame Interpolation Transformer, built on top of the transformer architecture which achieved state of the art results. In this paper we will evaluate how their base model behaves for the task of video extrapolation, as well as adding relative global self-attention and relative positional encoding.

While video interpolation and extrapolation can produce impressive results, the quality of the output frames can be further improved with the use of advanced techniques such as relative global attention and relative positional embedding. Relative positional embedding helps the model establish an improved spatial relationship between relative location of pixels within a frame. By having this established relationship, the model's output will generate a more accurate results when frames contain multiple dynamic objects.

In Summary, the integration of relative global attention and relative positional embedding can improve the quality of output frames in video interpolation and extrapolation by capturing long-range dependencies and establishing improved spatial relationships between pixels within a frame, leading to more accurate predictions and smoother, more natural-looking video sequence.

2 Background

2.1 Transformer Architecture

The transformer is a deep neural network architecture originally introduced for sequence transduction in the 2017 paper "Attention is all you Need" by Vaswani et al [4]. The transformer uses the same encoder-decoder scheme used by previous sequence transducers

but introduces the concept of “self-attention”. Self-attention allows the model to weigh individual parts of the input when generating its prediction. This mechanism is expanded to multi-head attention, which computes the attention several times in parallel. This allows the network to “attend” to, or pay attention to, several parts of the sequence at once.

Position embeddings are a way of incorporating positional information into the input representation, by adding a fixed-length vector to each token's embedding that encodes its position within the sequence. These embeddings are typically learned during training and are often represented as sinusoidal functions of different frequencies and offsets.

Beyond basic self-attention, there are several variations on attention which have been introduced. We describe a few which we used in our own architecture.

Global self-attention is a variant of self-attention that computes attention weights between all pairs of positions in the input sequence, instead of just between adjacent positions. This enables the model to capture long-range dependencies between different parts of the sequence, which can be particularly important in tasks like language translation or summarization.

In contrast, cross-attention computes attention weights between positions in different sequences, allowing the model to attend to different parts of the key sequence based on the information in the query sequence. This mechanism is useful for tasks like machine translation or multimodal learning, where the model needs to attend to multiple sequences or modalities.

Causal attention is a variant of self-attention that only allows the model to attend to positions that come before the current position in the input sequence. This is important in tasks like language modeling or text generation, where the model must generate output one step at a time based on previous output and prevent the model from "cheating" by attending to positions that come after the current position.

2.2 Patches

Transformers are trained on a sequence of frames to create the next element. This makes sense in terms of image extrapolation, as we would be taking two images and predict the next.

However, evaluating the images without modification has some issues. Mostly, it’s that a computer has difficulty in maintaining context. For example, if the input images were evaluated wholly, it would be difficult to capture regional movements within an image. To improve this process, 16x16 pixel patches are extracted from the input images and fed to the transformer as the input sequence.

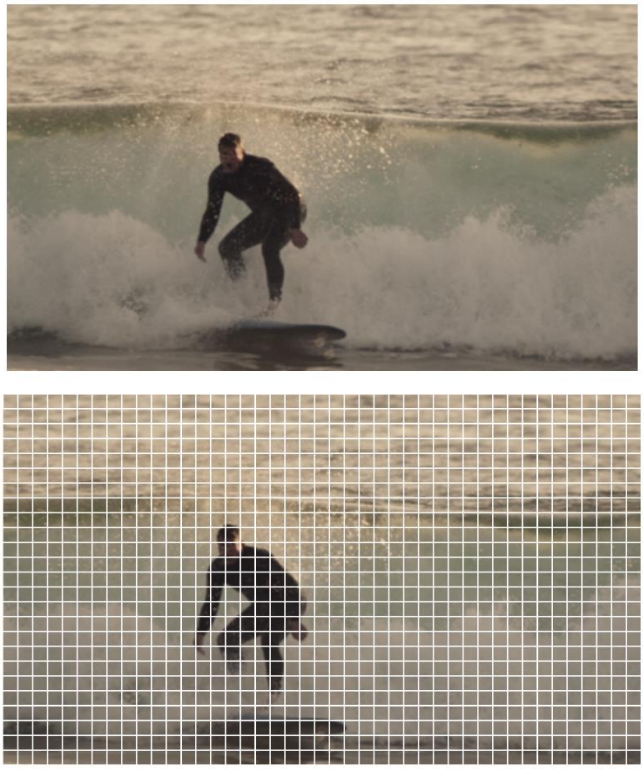

Figure 2: Sample input image before patch encoding, then after patch encoding

This allows the transformer model to capture regional movements more finely, as it can focus on individual components within the images.

2.3 Relative Positioning

Transformers by default use absolute positional embedding, which just encodes the absolute position of a token within the overall sequence. However, this approach loses the information encoded by the positions of each token relative to each other, as well as enforcing a maximum sequence length. Shaw et al. introduced a method for using the pairwise distances to create positional encodings in the paper "Self-Attention with Relative Position Representations" [2]. The pairwise distances get added to the keys during the computation of attention,

$$e_{ij} = \frac{x_i W^Q \left(x_j W^K + a_{ij}^K\right)^T}{\sqrt{d_z}} \quad (1)$$

and then again as a subcomponent of the values.

$$z_i = \sum_{j=1}^{n} \alpha_{ij} \left(x_j W^V + a_{ij}^V\right) \quad (2)$$

We don't apply the second formula in our implementation as it was found to not have a significant effect on our results.

Computing the relative positional embeddings requires O(L^2D) memory where L is the length of the sequence and D is the hidden state size.

Huang et al. use a trick called "skewing" to efficiently compute the relative positional embedding with ever expanding a [1]. We will not replicate their full method in this section, but they were able to lower the memory footprint to O(LD), which makes it feasible to train with a much longer sequence length. We use the skewing method in our implementation.

3 Our Architectures

The first model explored was a baseline transformer architecture. Custom implementations were created for the cross-attention, global-attention, and causal self-attention. From there, the rest of the transformer architecture was written to take the image sequence as patches. This gave a baseline model to base our results on.

To try to modify the models from prior implementations of image extrapolation and interpolation using a transformer model, relative attention was implemented to attempt to see if the relative positioning of the patches in the input images were relevant to the output images. The logic behind this decision was that the positioning of patches within an image would be important to their context, so incorporating the position of the patch in the attention equation should capture correlating movement relations.

Our model was inspired by the Relative Global Attention discussed in the paper "Music Transformer" [1]. While the context is different, as that particular paper focused on music note sequences and our model used image patch sequences, the application is relevant as both music notes and image patches use neighboring components to provide context.

4 Results

After comparing the two models, our current evaluation has yielded inconclusive results with regards to the performance of the two models under consideration. Specifically, we have not been able to identify any significant differences between their results. The model incorporates relative global attention, as opposed to the base attention. This does not appear to exhibit any substantial divergence in terms of its efficacy or accuracy when compared to its counterpart.

Moreover, our experiments have led us to believe that both models generate similar results when presented with the same set of data. While further tests may be necessary to validate these initial findings, our current analysis indicates that there is no marked discrepancy in the performance or accuracy once relative global attention is used.

5 Further Work

This paper explored some basic interpolation and extrapolation applications using basic and relative encoding attention. One extension that could be done using extrapolation is to prolong the input sequence by using the output image as the next term in the sequence. In doing so, one could propagate the extrapolation to generate entire videos from a starting sequence of images, continually updating the sequence in a recurrence style.

Similar to the extrapolation to generate a new image, repeated interpolation could provide higher quality slow-motion videos. These may not match up to the real motions of the object due to some sampling issues, but for some motions it may be able to properly describe the real function of the real actions.

The relative positioning incorporated in the relative global attention is useful for sequences of data, such as strings of words. One issue in the current setup is that the patches generated from the image are strung out, meaning that there is no vertical association between the patches. A potential exploration is modifying the relative positioning to incorporate the vertical position as well, potentially using a metric such as the Manhattan Distance between patch locations.

Otherwise, giving improper images may create interesting results. It would be interesting to see the different effects of extrapolation and interpolation on the same set of images. For example, if the input images were of a house and a rainbow, does the extrapolation make something completely different? Does interpolation create a morphing house and rainbow? More work will have to be done to see the effects of these processes.

References

[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, Douglas Eck: “Music Transformer”, 2018; arXiv:1809.04281.

[2] Peter Shaw, Jakob Uszkoreit, Ashish Vaswani: “Self-Attention with Relative Position Representations”, 2018; arXiv:1803.02155.

[3] Yunzhi Zhang, Jiajun Wu: “Video Extrapolation in Space and Time”, 2022; arXiv:2205.02084.

[4] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, Ming-Hsuan Yang: “Video Frame Interpolation Transformer”, 2021; arXiv:2111.13817.