Unsupervised Video Transformers

Dev Singh

CS Seminar: Machine Learning

May 1, 2024

The Problem and Overview

- Current residual networks are not ideal for video data with long temporal dependencies.
- Transformer networks have shown great promise in video classification and understanding tasks by reducing the dependency on recurrent networks, and instead using self-attention techniques.
- By using self-attention, a neural network can learn long-term dependencies with lower computational requirements and higher accuracy.

Basics of Artificial Neural Networks

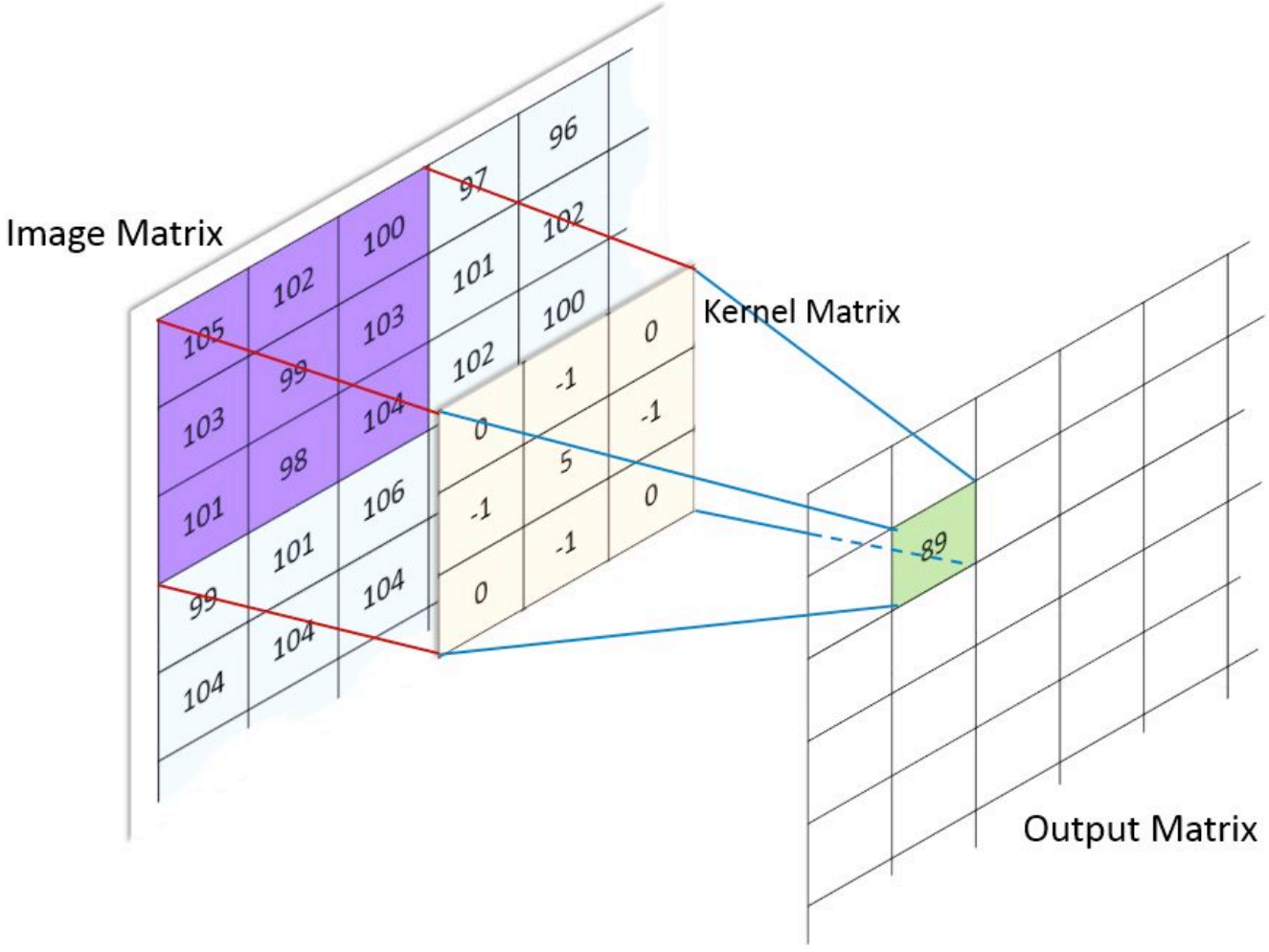
Supervised approach

- General goal: optimize the parameters of a function $f : \mathbb{R}^d \mapsto \mathbb{R}^n$ such that for some inputs $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$ (e.g., features where $x_i \in \mathbb{R}^d$) and their associated ground truth (e.g., a label for each input) $\mathbf{Y} = \mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}$ are close by some loss function $L(\mathbf{X}, \mathbf{Y})$
- Feed-forward neural networks consist of "layers" of neurons that take a linear combination of previous inputs ($l(\mathbf{x}) = \mathbf{wx} + \mathbf{b}$) and the output of a non-linear "activation function" designed to allow the network to model non-linear data.
- Network is trained by back-propagating the error ∇L.

Convolutional Neural Network (CNN)

- Overarching goal: to extract the most important spatial features from image or image-like data, by processing through a network of convolutional filters.
- Introduced for image classification by LeCun et al. (1989) and provided state-of-the-art performance in image recognition and object detection tasks.
- Filter w is convolved with the image X with chunks x, i.e., "slide the filter over chunks of the image, computing the dot products".
- Used for "feature extraction" to extract important traits of the provided image.

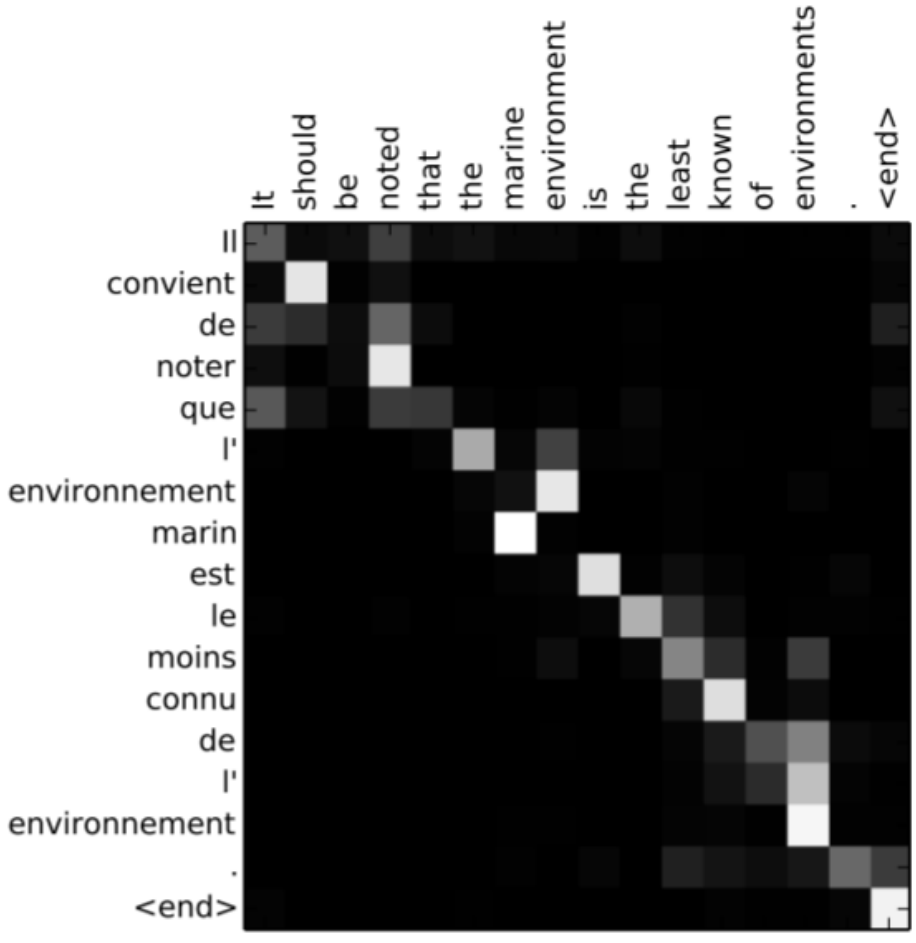
CNN Visualization



Attention Is All You Need

- Concept of "Attention" introduced in Vaswani et al. (2017).
- Solves recurrent architecture bottlenecks and allows the model to focus on the relevant parts of the input sequence as needed.
- Originally applied to Natural Langauge Processing for translation transformers.
- Attention(Q, K, V) = softmax($\frac{QK^T}{\sqrt{d_k}}V$), where d_k is the dimensions of the keys.
- There are various enhancements to basic attention, including Multi-Head Attention.

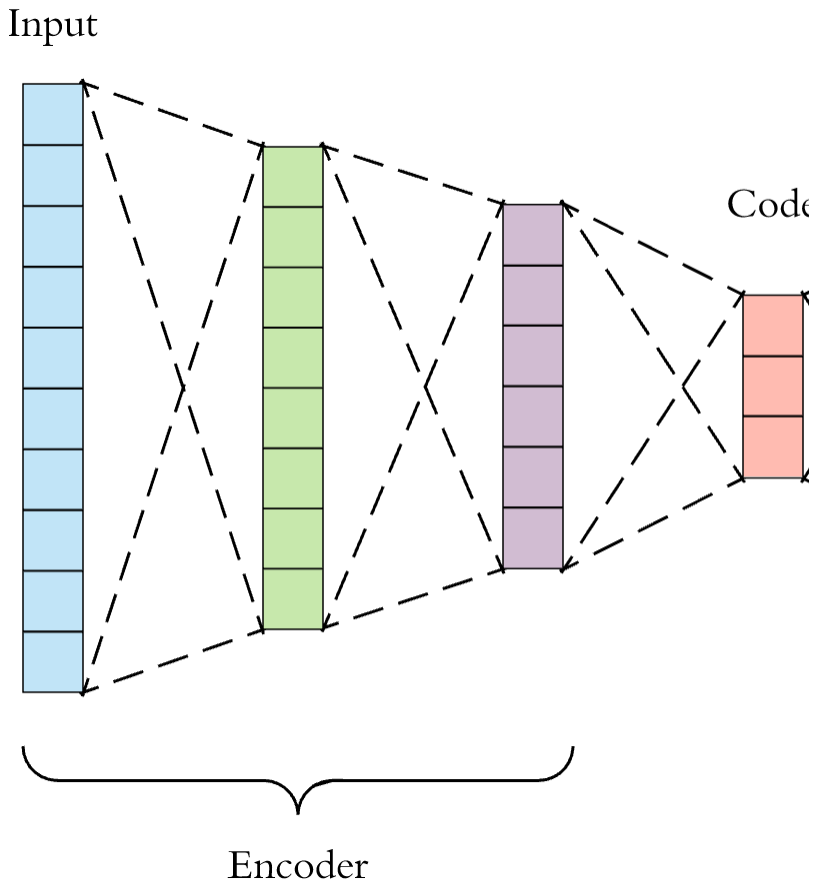
Visual Representation of Attention



Encoder Block

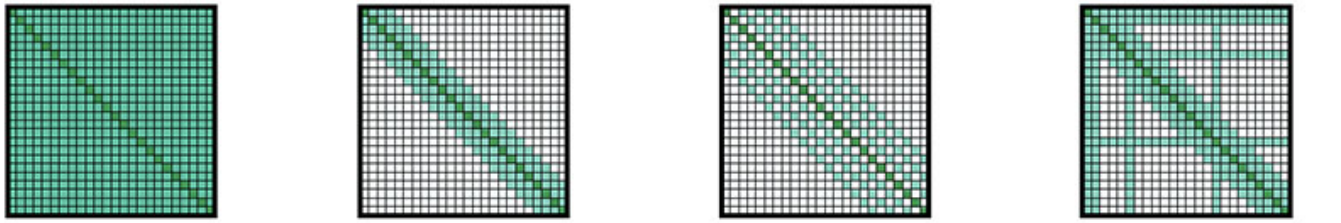
Unsupervised architecture

- Goal: $f : \mathbb{R}^a \mapsto \mathbb{R}^b, b << a$ (reduce dimensionality of data while retaining features).



Transformer Architecture

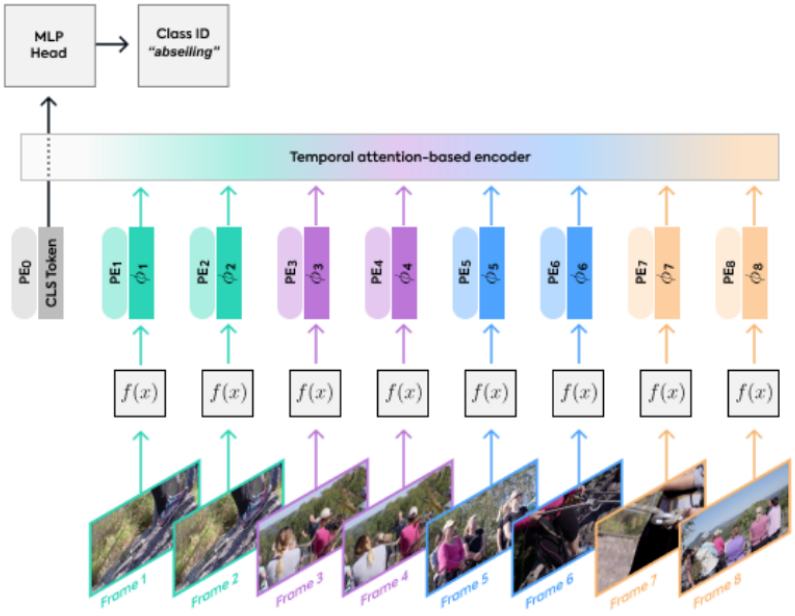
- Based on an attention encoder-decoder architecture.
- Longformer uses temporal encoder and a sliding-chunks attention window technique with a runtime and memory complexity of $O(n)$, in contrast to traditional full-attention encoders that have a runtime and memory complexity of $O(n^2)$.



(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

Transformers for Video Classification

- Basis work is the Video Transformer Network (VTN) as proposed by Neimark et al. (2021).
- Feature Extraction, temporal long-document transformer (Longformer) with encoder block, MLP classification head.



Contrastive Learning

- Subset of self-supervised learning.
- Learn the general features of the data without labels by teaching the model which data points are similar or different.
- Constrastive Loss: $L(i,j) = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1\ [k\neq i]}^{2N} exp(sim(z_i, z_j)/\tau)}$

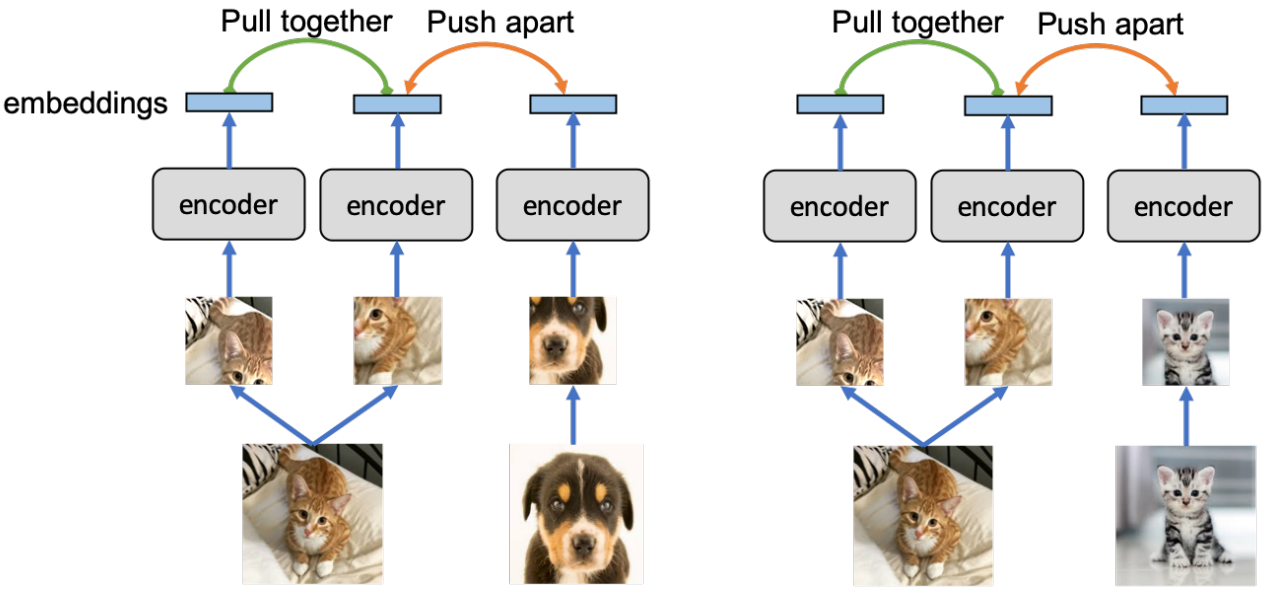
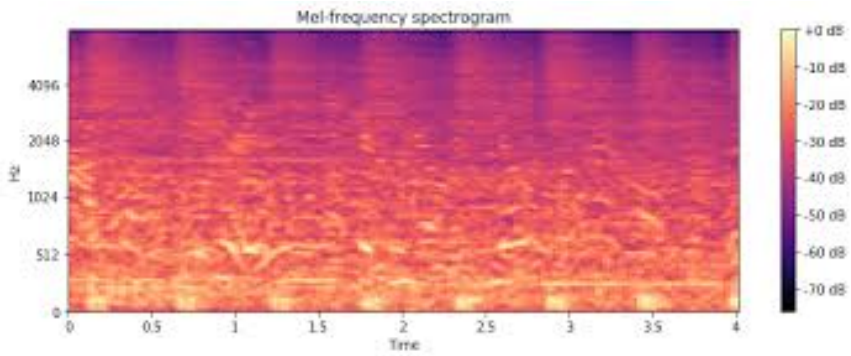


Figure 1 (a) (b)

Multi-Modal Learning

- Many video classification techniques do not incorporate audio information.
- Yet, audio is important for understanding video content - see human behavior.
- Spectrograms can be treated as "image-like" representations of audio with a given window size, and we can use CNNs to learn their features.



Contrastive Multi-Modal Video Transformers

- Use the information from one modality (video) as a supervisory signal for the other modality (audio), and vice-versa.
- Push together similar data and pull apart different modalities/data to train the feature extraction model in an unsupervised manner.
- Generate representations of data between frames of videos to extract structural relations in the data.