imsa.edu

Contrastive Multi-Modal Video Transformer

Dev Singh

Advisors: Chengxiang Zhai[1] Ismini Lourentzou[2]

[1]Donald Biggar Willett Professor in Engineering
Department of Computer Science
University of Illinois at Urbana-Champaign

[2]Assistant Professor of Computer Science
Virginia Tech

IMSAloquium 2021

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

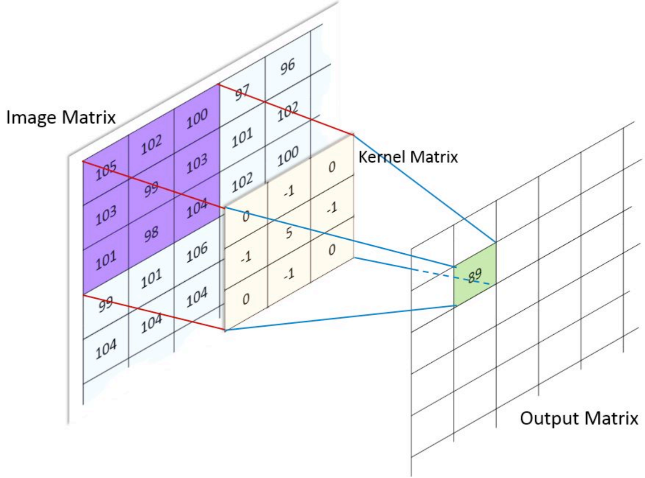
- General goal: optimize the parameters of a function $\mathrm{f} : \mathbb{R}^{\mathrm{d}} \mapsto \mathbb{R}^{\mathrm{n}}$ such that for some inputs $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$ (e.g., features where $\mathrm{x_i} \in \mathbb{R}^{\mathrm{d}}$) and their associated 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 $\mathrm{L}(\mathbf{X}, \mathbf{Y})$
- Feed-forward neural networks consist of "layers" of neurons that take a linear combination of previous inputs ($\mathrm{l}(\mathbf{x}) = \mathbf{wx} + \mathbf{b}$) 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 $\nabla \mathrm{L}$.

imsa.edu

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

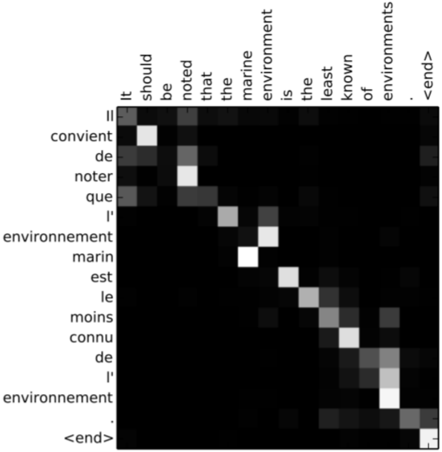


IMSA
imsa.edu

Attention Is All You Need

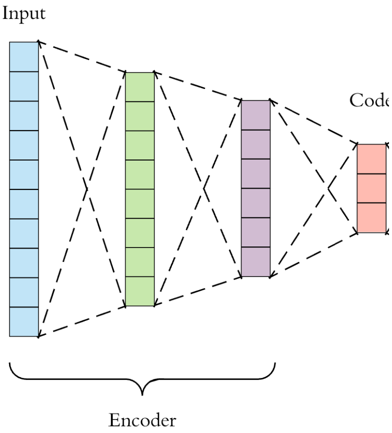
- Concept of "Attention" introduced in Vaswani et al. (2017).
- Solves recurrent architecture bottlenecks and allows the model to focus on the relevant parts of the input sequence as needed.
- Attention(Q, K, V) = softmax($\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\mathrm{V}$), where $\mathrm{d_k}$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

Visual Representation of Attention



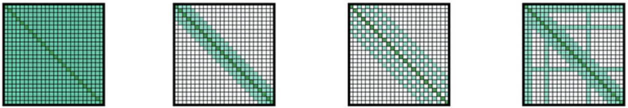
Encoder Block

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



Transformer Architecture

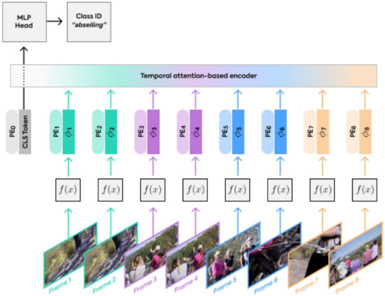
- Based on an attention encoder-decoder architecture.
- Our work based on Longformer encoder as proposed by Beltagy et al. (2020).
- Longformer uses temporal encoder and a sliding-chunks attention window technique with a runtime and memory complexity of $\mathrm{O}(\mathrm{n})$, in contrast to traditional full-attention encoders that have a runtime and memory complexity of $\mathrm{O}(\mathrm{n}^2)$.



(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.



IMSA
imsa.edu

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: $\mathrm{L(i,j)} = -\log \dfrac{\exp(\mathrm{sim}(\mathrm{z_i}, \mathrm{z_j})/\tau)}{\sum_{\mathrm{k=1}}^{\mathrm{2N}} {}_{[\mathrm{k \neq i}]} \exp(\mathrm{sim}(\mathrm{z_i}, \mathrm{z_j})/\tau)}$

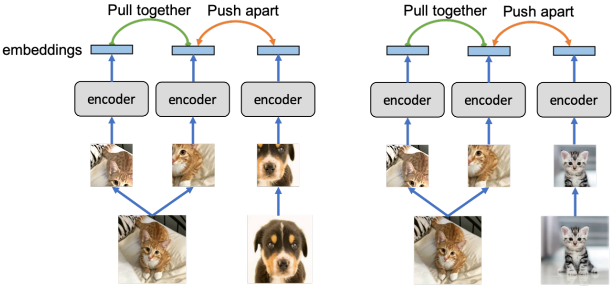
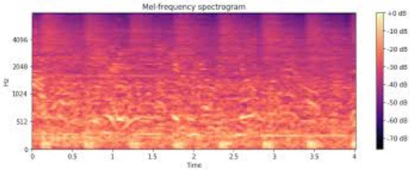


Figure 1 (a) (b)

IMSA
imsa.edu

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.
- We aim to make use of both signals; video and audio.



IMSA
imsa.edu

Contrastive Multi-Modal Video Transformer

Our Work

- Use the information from one modality (video) as a supervisory signal for the other modality (audio), and vice-versa as proposed in Alwassel et al. (2020).
- We cluster the 2D video and audio representations (using k-means clustering) and contrast the prototype representations.
- We are unaware of any applications of contrastive multi-modal learning with video transformer architectures for video classification tasks.

Upstream Tasks

- Kinetics-400 dataset - video classification on video clips of 400 human action classes.
  - Includes human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging.
- (Potentially) COVID video testimonial classification.

Next Steps

- Train models for proposed architecture
  - CNN feature extraction for audio.
  - VTN-like transformer for audio.
  - MLP classification head for video and audio.
- Ablation studies
  - Local vs. global attention.
  - Video vs. various video-audio techniques.