Report draft (Latex) view and edit through this link:

<https://www.overleaf.com/1731651235vcnthqmzvmqb>

# Human Action Recognition Based on Vision Transformer and L2 Regularization

# <https://dl.acm.org/doi/pdf/10.1145/3581807.3581840>

In this paper, based on attention mechanism of human action recognition method is studied, in order to improve the model accuracy and efficiency in VIT network structure as the framework of feature extraction, because video data includes characteristics of time and space, so choose the space and time attention mechanism instead of the traditional convolution network for feature extraction, In addition, L2 weight attenuation regularization is introduced in model training to prevent the model from overfitting the training data.

**DropDim: A Regularization Method for Transformer Networks**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9670702>

DropDim, a structured dropout method designed for regularizing the self-attention mechanism, which is a key component of the transformer. In contrast to the general dropout method, which randomly drops neurons, DropDim drops part of the embedding dimensions. In this way, the semantic information can be completely discarded. Thus, the excessive co-adapting between different embedding dimensions can be broken, and the self-attention is forced to encode meaningful features with a certain number of embedding dimensions erased.

**PolyViT: Co-training Vision Transformers on Images, Videos and Audio**

<https://arxiv.org/pdf/2111.12993>

Train a single transformer model capable of processing multiple modalities and datasets, whilst sharing almost all of its learnable parameters. By co-training different tasks on a single modality, this method is able to improve the accuracy of each individual task and achieve state-of-the-art results on 5 standard video- and audio-classification datasets. Co-training the model on multiple modalities and tasks leads to a model that is even more parameter-efficient, and learns representations that generalize across multiple domains.

**Unbox the Black-box: Predict and Interpret YouTube Viewership Using Deep Learning** <https://doi.org/10.1080/07421222.2023.2196780>

Existing interpretable predictive models face the challenges of imprecise interpretation and negligence of unstructured data. Following the design-science paradigm, we propose a novel Precise Wide-and-Deep Learning (PrecWD) to accurately predict viewership with unstructured video data and well-established features while precisely interpreting feature effects.

**Will You Dance To The Challenge? Predicting User Participation of TikTok Challenges**

<https://arxiv.org/pdf/2112.13384.pdf>

The uniqueness of the TikTok platform where both challenge content and user preferences are evolving requires the combination of challenge and user representation. This paper investigates social contagion of TikTok challenges through predicting a user’s participation. They propose a novel deep learning model to learn and combine latent user and challenge representations from past videos to perform this user-challenge prediction task.

**Multi-modal Representation Learning for Short Video Understanding and Recommendation**

<https://ieeexplore.ieee.org/document/8795067>

This work focuses on learning representations from different modalities to understand and recommend short videos effectively. The authors propose a multi-modal fusion framework that combines the features from each modality to capture the inherent relationships between them. They use deep neural networks for feature extraction and utilize a fusion strategy to combine these features for video understanding and recommendation tasks.

**Describing Videos using Multi-modal Fusion** ([https://dl.acm.org/doi/10.1145/2964284.2984065](https://ceur-ws.org/Vol-3102/paper2.pdf)

This work focuses on describing videos using multi-modal fusion techniques. The authors propose a deep neural network that takes both visual and textual information as input and fuses them at various stages in the network. The goal is to learn a joint representation of videos that effectively combines visual and textual information for video description tasks.

**Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach** <https://ceur-ws.org/Vol-3102/paper2.pdf>

The primary objective of this work is to predict the popularity of images and videos on Instagram. The authors propose a deep learning-based approach that leverages convolutional neural networks and long short-term memory networks to capture spatial and temporal information from images and videos, respectively. The extracted features from both modalities are then combined and fed into a regression model for popularity prediction.

**Instagram Popularity Prediction via Neural Networks and Regression Analysis**

[<https://cjqian.github.io/docs/instagram_paper.pdf>](https://cjqian.github.io/docs/instagram_paper.pdf)

In this work, the authors address the task of popularity prediction on Instagram using neural networks and regression analysis. To evaluate the predictive power of image composition on Instagram posts, they compare the popularity predictions of a neural network trained on aesthetic value to the predictions of regression models using social metadata.