\section\*{References}

\noindent[1] Jiaheng Xie, Yidong Chai, and Xiao Liu. "Unbox the Black-Box: Predict and Interpret YouTube Viewership Using Deep Learning." Journal of Management Information Systems, 2023, 541-579.\\

\noindent[2] Lynnette Hui Xian Ng, John Yeh Han Tan, Darryl Jing Heng Tan, Roy Ka-Wei Lee. Will you dance to the challenge? Predicting user participation of TikTok challenges. In Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM’21), The Hague, 2021, 356–360. New York: ACM.\\

\noindent[3] Qiliang Chen, Hasiqidalatu Tang, and Jiaxin Cai. "Human Action Recognition Based on Vision Transformer and L2 Regularization." In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition (ICCPR '22), 2022, 224-228.\\

\noindent[4] Daya Guo, Jiangshui Hong, Binli Luo, Qirui Yan, Zhangming Niu. "Multi-modal representation learning for short video understanding and recommendation." In 2019 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019, 687-690.\\

\noindent[5] Jin Q., Chen J., Chen S., Xiong Y., and Hauptmann A. "Describing videos using multi-modal fusion." In ACM MM, 2016.\\

\noindent[6] Massimiliano Viola, Luca Brunelli, and Gian Antonio Susto. "Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach." Università degli Studi di Padova, Padova, IT.\\

\noindent[7] Qian C., Tang J., Penza M., and Ferri C. "Instagram Popularity Prediction via Neural Networks and Regression Analysis," 2017, 2561-2570.\\

\noindent[8] Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. "Early versus late fusion in semantic video analysis." In Proceedings of the Annual ACM International Conference on Multimedia, Singapore, 2005, 399-402.\\

\noindent[9] Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020, 1–6.\\

\noindent[10] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows: https://arxiv.org/abs/2103.14030 \\ \hl{FIX FORMAT, ADD NAME, ETC.}

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

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

[11] Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting

<https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf>

Xingjian Shi Zhourong Chen Hao Wang Dit-Yan Yeung. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst. 28, 802–810 (2015)

This paper introduces a ConvLSTM model, which extends the fully connected LSTM with convolutional structures in input-to-state and state-to-state transitions.

[12] Robust Speech Recognition via Large-Scale Weak Supervision

<https://arxiv.org/pdf/2212.04356.pdf>

[Alec Radford](https://arxiv.org/search/eess?searchtype=author&query=Radford%2C+A), [Jong Wook Kim](https://arxiv.org/search/eess?searchtype=author&query=Kim%2C+J+W), [Tao Xu](https://arxiv.org/search/eess?searchtype=author&query=Xu%2C+T), [Greg Brockman](https://arxiv.org/search/eess?searchtype=author&query=Brockman%2C+G), [Christine McLeavey](https://arxiv.org/search/eess?searchtype=author&query=McLeavey%2C+C), [Ilya Sutskever](https://arxiv.org/search/eess?searchtype=author&query=Sutskever%2C+I)

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022).

The researchers have developed a new approach called "Whisper" for weakly supervised speech recognition.

[13] Attention Is All You Need

<https://arxiv.org/pdf/1706.03762.pdf>

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Introduces the Transformer, a model that relies exclusively on attention mechanisms, completely discarding the use of recurrence and convolutions.

[14] Self-attention generative adversarial networks

<https://arxiv.org/pdf/1805.08318.pdf>

Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena

Proceedings of the 36th International Conference on Machine Learning, PMLR 97:7354-7363, 2019.

The Self-Attention Generative Adversarial Network (SAGAN) is introduced, enabling image generation tasks with attention-driven, long-range dependency modeling.

[15] Video Prediction using ConvLSTM Autoencoder (PyTorch)

Andreas Holm Nielsen 2020

<https://holmdk.github.io/2020/04/02/video_prediction.html>

# Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition

# <https://link.springer.com/article/10.1007/s00138-021-01249-8>

# This paper tries to address the defects of each fusion category by first investigating more deeply the early-stage fusion that has been poorly explored in the literature. Second, intermediate fusion protocols operate on the feature map, irrespective of the particularity of human action, we propose a new scheme where we optimally combine modality-wise features. Third, as most of the late fusion solutions use handcrafted rules, prone to human bias, and far from real-world peculiarities, this paper adopt a neural learning strategy to extract significant features from data rather than assuming that artificial rules are correct. This paper validated our findings on two challenging datasets.

# Early versus late fusion in semantic video analysis

# <https://dl.acm.org/doi/10.1145/1101149.1101236>

# This paper compares early and late fusion techniques in the context of semantic video analysis. Based on an experiment on 184 hours of broadcast video using 20 semantic concepts, it concludes that late fusion methods are often superior for tasks that involve diverse and complex multimodal information.

# Early or Late Fusion Matters: Efficient RGB-D Fusion in Vision Transformers for 3D Object Recognition

# <https://arxiv.org/pdf/2210.00843.pdf>

# This paper explores which depth representation is better in terms of resulting accuracy and compare two methods for injecting RGB-D fusion within the ViT architecture (i.e., early vs. late fusion). Our results in the Washington RGB-D Objects dataset demonstrates that in such RGB → RGB-D scenarios, late fusion techniques work better than most popularly employed early fusion.

# Early vs Late Fusion in Multimodal Convolutional Neural Networks

# <https://ieeexplore.ieee.org/document/9190246>

# This paper determines whether the fusion of different modalities can provide an advantage as compared to uni-modal approaches, and whether a more complex early fusion strategy can outperform the simpler late-fusion strategy by making use of statistical correlations between the different modalities. Our results show a clear performance improvement by multi-modal fusion and a substantial advantage of an early fusion strategy.

# Multi-attention Recurrent Network for Human Communication Comprehension

# <https://arxiv.org/pdf/1802.00923>

# This paper compares early fusion and late fusion approaches for multimodal sentiment analysis. They propose a multi-attention recurrent network that can effectively integrate text and visual modalities. The results of their experiments suggest that the performance of early fusion and late fusion can vary depending on the task and dataset.

# Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

# <https://arxiv.org/pdf/2103.14030.pdf>

# This is the original Swin Transformer paper, which introduced the model architecture and provided details on how it was trained and optimized. This paper presents the vision Transformer, Swin Transformer, that capably serves as a general-purpose backbone for computer vision. Challenges in adapting Transformer from language to vision arise from differences between the two domains, such as large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text. To address these differences, this paper proposes a hierarchical Transformer whose representation is computed with \textbf{S}hifted \textbf{win}dows.

# Optimizing and Evaluating Swin Transformer for Aircraft Classification: Analysis and Generalizability of the MTARSI Dataset

# https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9996354

# This paper tests whether the performance of the Swin-Transformer on general-purpose image classification translates to domain-specific aircraft classification. It also investigates the effect of training procedure vs. model selection on the validation score.

# A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay

# <https://arxiv.org/pdf/1803.09820.pdf>

# Most trainings in deep learning are with suboptimal hyper-parameters, requiring unnecessarily long training times. Setting the hyper-parameters remains a black art that requires years of experience to acquire. This paper proposes several efficient ways to set the hyper-parameters that significantly reduce training time and improves performance. This paper shows how to examine the training validation/test loss function for subtle clues of underfitting and overfitting and suggests guidelines for moving toward the optimal balance point. It discusses how to increase/decrease the learning rate/momentum to speed up training. The experiments show that it is crucial to balance every manner of regularization for each dataset and architecture. Weight decay is used as a sample regularizer to show how its optimal value is tightly coupled with the learning rates and momentums.

# Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour

# <https://arxiv.org/pdf/1706.02677.pdf>

# Larger networks and larger datasets result in longer training times that impede research and development progress. Distributed synchronous SGD offers a potential solution to this problem by dividing SGD minibatches over a pool of parallel workers. This paper empirically shows that on the ImageNet dataset large minibatches cause optimization difficulties, but when these are addressed the trained networks exhibit good generalization.

# Revisiting Small Batch Training for Deep Neural Networks

# <https://arxiv.org/pdf/1804.07612.pdf>

# While the use of large mini-batches increases the available computational parallelism, small batch training has been shown to provide improved generalization performance and allows a significantly smaller memory footprint, which might also be exploited to improve machine throughput.The collected experimental results for the CIFAR-10, CIFAR-100 and ImageNet datasets show that increasing the mini-batch size progressively reduces the range of learning rates that provide stable convergence and acceptable test performance. On the other hand, small mini-batch sizes provide more up-to-date gradient calculations, which yields more stable and reliable training.

# 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.