Enhancing TikTok Content Success Prediction through Multimodal Fusion

This research introduces a Multimodal Ensemble Architecture for predicting TikTok content success by combining visual and audio features. The study utilizes a diverse dataset of TikTok videos, with detailed information about its size and characteristics. Visual embeddings are extracted through an unsupervised ConvLSTM Autoencoder, capturing spatial and temporal features. Simultaneously, audio embeddings are obtained using the Whisper ASR model, which transcribes spoken content. These embeddings are integrated into a dual Transformer-based regression model for comprehensive multimodal analysis.

The dataset, comprising a substantial number of TikTok videos, is processed and analyzed using PyTorch, Torch-vision, and Scikit-learn. Hardware resources include NVIDIA GPUs and CPUs, with considerations for VRAM limitations during the training phase. The challenges related to tensor sizes, given the variability in video lengths, are adeptly addressed through techniques such as fixed-size padding, average pooling, and sequential handling strategies.

In addition to the model architecture, this research introduces the standard normalization of target values to enhance model generalization. The evaluation metrics encompass both Mean Squared Error (MSE) and Mean Absolute Error (MAE), providing a comprehensive assessment of model performance. Furthermore, the research delves into the impact of outliers on MSE and MAE, highlighting the importance of robust loss functions in handling extreme values in the dataset.

The results showcase the effectiveness of the ConvLSTM Autoencoder in learning meaningful visual representations, with decreasing loss over epochs. The multimodal ensemble regression model outperforms baseline models, including a 3D convolution model and Swin Transformer, in terms of MSE and MAE. A classifier variation, transforming the regression problem into quartile-based classification, provides additional insights into the complexity of predicting video success.

This research not only contributes to the advancement of multimodal deep regression but also underscores the significance of handling real-world challenges posed by heterogeneous datasets. The detailed exploration of the dataset, along with model architecture and evaluation metrics, lays a robust foundation for future work in the realm of TikTok content success prediction and beyond. The implications extend to applications in audio-visual recognition and multi-source data integration.