1. Baseline:
2. (ResNet) CNN + MLP = MSE Loss: 10
3. ResNet + RNN = MSE Loss: 5
4. I3D (Inflated 3D ConvNet) + LTSM = MSE Loss: 4
5. Our Solution:
6. MSE Loss = 0.01
7. Unsupervised Pretraining
9. ConvLSTMAutoencoder
10. \_\_\_Visual---> O-\ /-O
11. / O--\ /--O
12. / O---O---O
13. / O---O---O
14. / O---O---O
15. / O--/|\--O
16. / O-/ | \-O
17. / | Dual Transfer Learning
18. / Embedding Supervised Fine-tuning
19. Video< \
20. \ \-------> Transformer(LTSM/MLP) --> Regression <-- MSE\_Loss
21. \ /
22. \ /
23. \ Embedding
24. \ ^
25. \ |
26. \\_\_Audio----> DeepSpeech(Mozilla) --> Text
28. Prebuilt Pretraining

Baseline:

(ResNet) CNN + MLP = MSE Loss: 10:

A combination of a Convolutional Neural Network (CNN) based on the ResNet architecture and a Multi-Layer Perceptron (MLP). The model is trained using Mean Squared Error (MSE) loss, with a loss value of 10.

ResNet + RNN = MSE Loss: 5:

Another variant of the ResNet architecture combined with a Recurrent Neural Network (RNN). The model is trained with MSE loss, yielding a loss value of 5.

I3D (Inflated 3D ConvNet) + LSTM = MSE Loss: 4:

I3D (Inflated 3D ConvNet) is a 3D extension of the ConvNet, along with a Long Short-Term Memory (LSTM) network. The model is trained with MSE loss and achieves a loss value of 4.

Model:

1. The model begins by breaking down the input video into separate visual and audio tracks.

2. **Unsupervised Pretraining - Visual Track**: The visual track of the video is processed using a ConvLSTM Autoencoder. This architecture combines convolutional layers with LSTM (Long Short-Term Memory) layers to capture both spatial and temporal dependencies in the visual data. The autoencoder is trained in an unsupervised manner, meaning it learns to reconstruct the input data without any specific labels. The output of this process is an encoded representation or embedding of the visual context of the video.

**Unsupervised Pretraining - Audio Track**: The audio track of the video is processed using a pretrained model called DeepSpeech, developed by Mozilla or the open-source Whisper. The model converts the audio into a corresponding textual transcript, providing additional information about the content of the video. This step is prebuilt, meaning it uses a pre-existing model rather than training from scratch.

3. **Embedding Generation - Visual Embedding**: The encoded visual representation from the ConvLSTM Autoencoder is extracted as an embedding vector, capturing the visual context of the video.

**Embedding Generation - Audio Embedding**: The DeepSpeech/Whisper model generates textual output from the audio track, representing the spoken content. This output can be considered as an embedding of the audio information.

4. **Concatenation and Feature Fusion**: The visual and audio embeddings are concatenated or combined into a single feature representation and fed into a transformer-based regression model. This fusion of modalities allows the model to leverage both visual and audio information for subsequent analysis. The transformer consists of LSTM or MLP layers.:

5. **Transformer-Based Regression**: The concatenated feature representation is input into a transformer-based regression model. The transformer architecture, commonly used in natural language processing tasks, is adapted here to handle the combined visual and audio data. It learns the semantic and non-linear relationships between the input features and the video creator success metrics, such as video views. The regression model is trained to predict these metrics based on the fused feature representation.

6. **Loss Calculation**: The model is trained using a mean squared error (MSE) loss function, with the goal of minimizing the discrepancy between the predicted values and the actual values. The final MSE loss target is 0.01, indicating a high level of accuracy in predicting the video creator success metrics. The output of the regression model is compared to the ground truth video creator success metrics using the mean squared error (MSE) loss function. The aim is to minimize the difference between the predicted values and the actual values.

7. **Baseline and Evaluation**: To evaluate the effectiveness of the proposed model, a baseline model is established using a simpler architecture or approach. The baseline is compared with the main implementation in terms of the achieved MSE loss. A lower MSE loss value indicates better performance in predicting video creator success metrics.

Overall, this model leverages unsupervised pretraining, pretrained audio models, and a transformer-based regression model to predict video creator success metrics, with a focus on video views. The combination of visual and audio information allows for a more comprehensive understanding of the video content and improves the accuracy of the predictions. The model architecture includes elements of both unsupervised pretraining and supervised fine-tuning, combining the benefits of both approaches. The model also incorporates a process called dual transfer learning, where both the visual and audio embeddings contribute to the final prediction.

The video will first be broken down into visual and audio tracks, with our primary focus on the video visual.

The video visual will undergo an autoencoder process, utilizing a convolution-based network architecture, ConvLSTM Autoencoder, to unsupervised pre-training from scratch. This process will encode the context of the video into embedding vectors.

Subsequently for the audio, we will leverage a pretrained model, such as Mozilla's Deep Speech or the open-source Whisper, which will be used to create a transcript for the audio, supplementing the project.

Afterward, embeddings will be extracted from the visual branch and the audio, which will then be concatenated and input into a transformer-based regression model.

The aim is for this model to learn the semantic and non-linear relationships needed to predict video creator success metrics, such as video views. Finally, we will establish a baseline using a less complex model and compare it with our main implementation to evaluate its success.

diagram, key point connection to other research, summarize the math/LaTex.

Try to use a software, make sure things is editable, not flat image, in case architecture change later. Well, likely not, but maybe minor change.

(ResNet) CNN + MLP = MSE Loss: 10

1. Visual image – diagram
2. Small summary of each point

Explain

Compare with other study (google doc)

Latex math

HyperParameter Tuning

https://www.kaggle.com/discussions/getting-started/253300