Multimodal Fine-tuning

Fine-tuning a model for your specific inputs of screen video and user's audio involves selecting a suitable multimodal model that can effectively handle both types of data.

The detailed guide on selecting and fine-tuning a model for this task:

Selecting a Multimodal Model

For the task of processing screen video and user's audio inputs, you would ideally choose a multimodal model that is capable of integrating both visual and audio information. Here are some models that can be considered:

1. **CLIP (Contrastive Language–Image Pre-training)**
    - **Description:** CLIP is trained to understand and integrate text and image inputs. While it primarily focuses on images, it can be adapted to handle video frames and audio embeddings.
    - **Usage:** Use CLIP with modifications to incorporate audio features alongside video frames.
2. **UNITER (UNiversal Image-TExt Representation)**
    - **Description:** UNITER is designed for integrating text and image data. It uses a transformer-based architecture that can potentially be adapted for video frames and audio.
    - **Usage:** Fine-tune UNITER to process both video frames and audio embeddings.
3. **ViT (Vision Transformer) + Wav2Vec 2.0**
    - **Description:** Combine ViT for visual processing with Wav2Vec 2.0 for speech-to-text conversion. This approach leverages separate models for each modality.
    - **Usage:** Integrate ViT and Wav2Vec 2.0 outputs at a higher level to fuse information from both modalities.

Fine-tuning Process

- **Initialize the Model:** Load the pre-trained weights of the selected model.
- **Define Optimizer and Loss Function:** Choose an optimizer (e.g., Adam) and a suitable loss function (e.g., cross-entropy for classification tasks).
- **Training Loop:** Iterate through batches of the annotated data. For each batch:
    - Forward pass: Feed the video frames and audio embeddings through the model.
    - Compute loss: Compare model predictions with ground truth labels.
    - Backward pass: Update model parameters to minimize the loss.
- **Validation:** Evaluate the model's performance on a validation set to monitor training progress and prevent overfitting.