Video Classification

Final Presentation
- Team 3 -

Roadmap

./ Review - Initial approach

./ Dataset and Data preprocessing

./ Embedding analysis

./ Baseline models

./ Proposed method

./ Data augmentation and Ablation studies

./ Experimental results

./ Closing remarks

Brief Review of Our Initial Approach

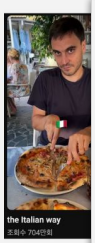
From ***niche classification*** to ***broad-topic modeling***

Task Definition

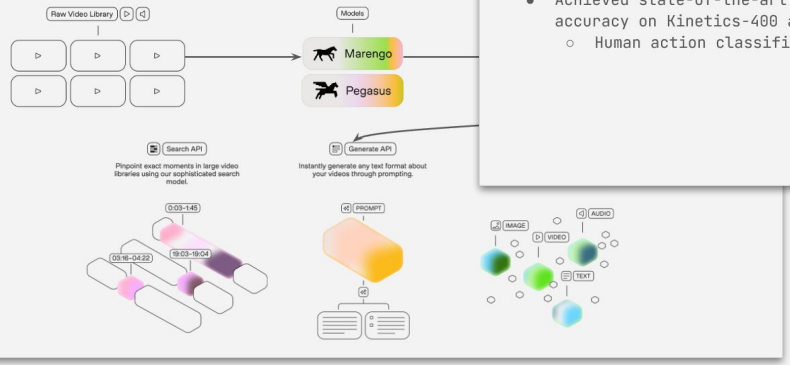
Cuisine-Type Classification for Short-form Video

Shorts

Korean?
Japanese?
Italian?
American?
...



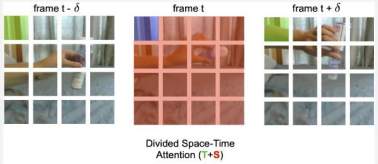
Data preprocessing > TwelveLabs



Related Works (Video classification)

TimeSformer

- A convolution-free video classification model based entirely on self-attention
- Adopted divided space-time attention
- Achieved state-of-the-art accuracy on Kinetics-400 and-600
  - Human action classification



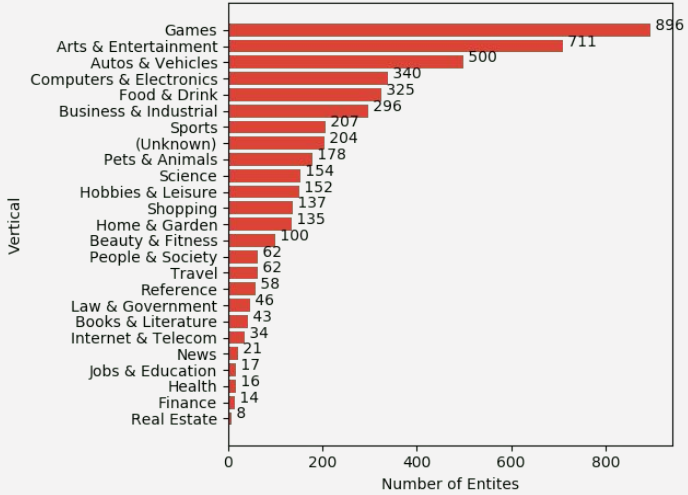
Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

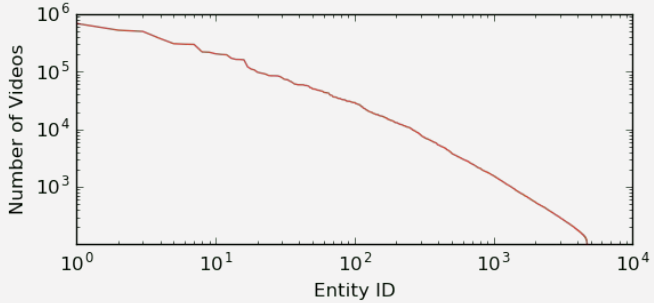
https://arxiv.org/abs/2102.05095

Dataset > YouTube-8M (ver. May 14th, 2018)

Technical Specs (full dataset)

- 6.1 M Videos
- 350 h Total video duration
- 2.6 B Audio/Video features
- 24 Categories
- 3862 Classes
- 3.0 Avg. classes per video





Dataset > YouTube-8M (ver. May 14th, 2018)

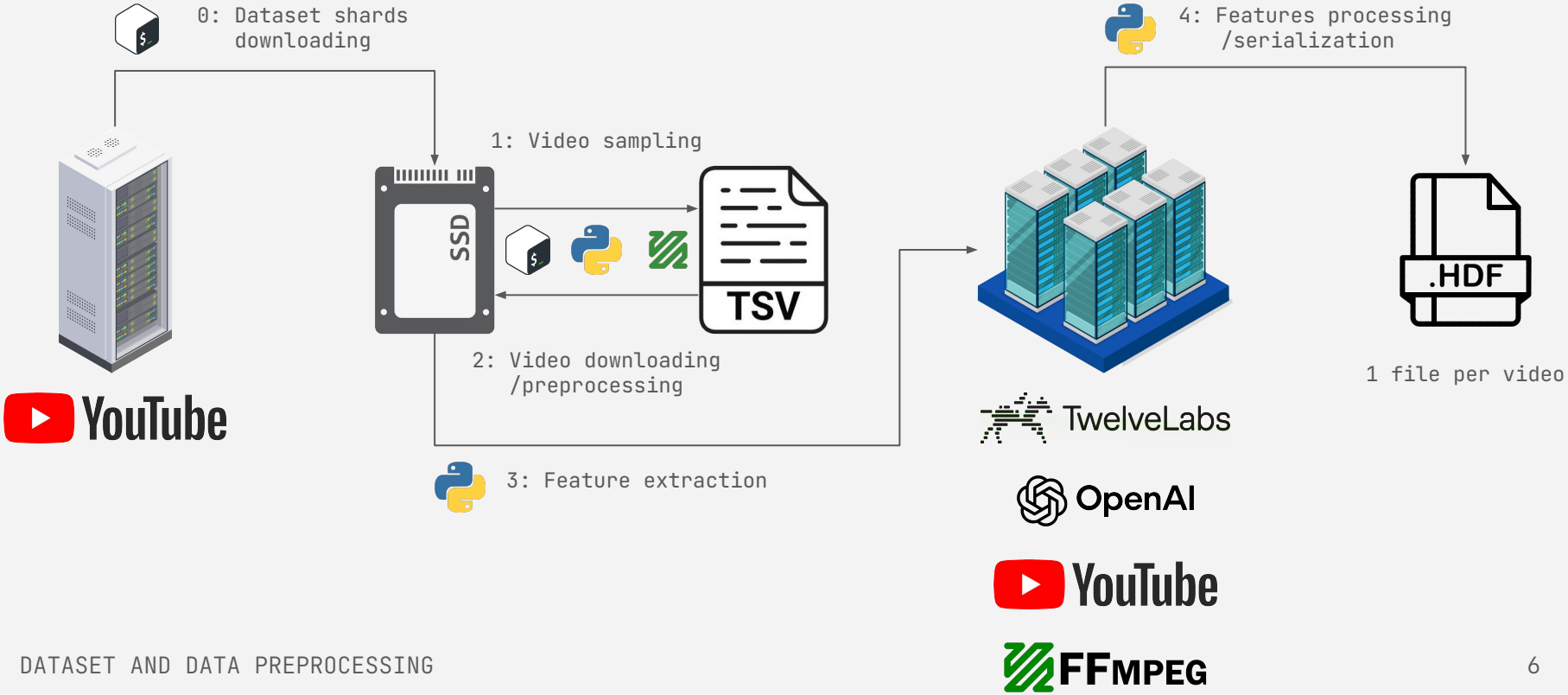
Technical Specs (working dataset)

- 1.8 K Videos
- 600 min Total video duration (20s/video)
- 10 Categories
- 2954 Classes
- 5.2 Avg. classes per video

Dataset splits

- 8:1:1 Split ratio
- 1440 Videos for training set
- 180 Videos for validation and test sets

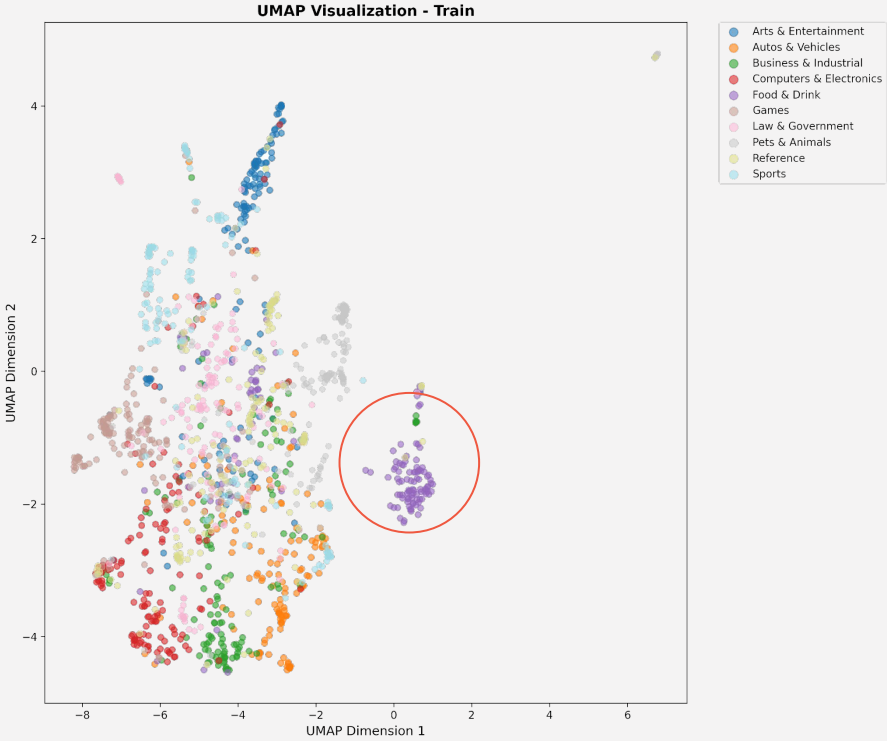
Dataset > YouTube-8M > Data Preprocessing

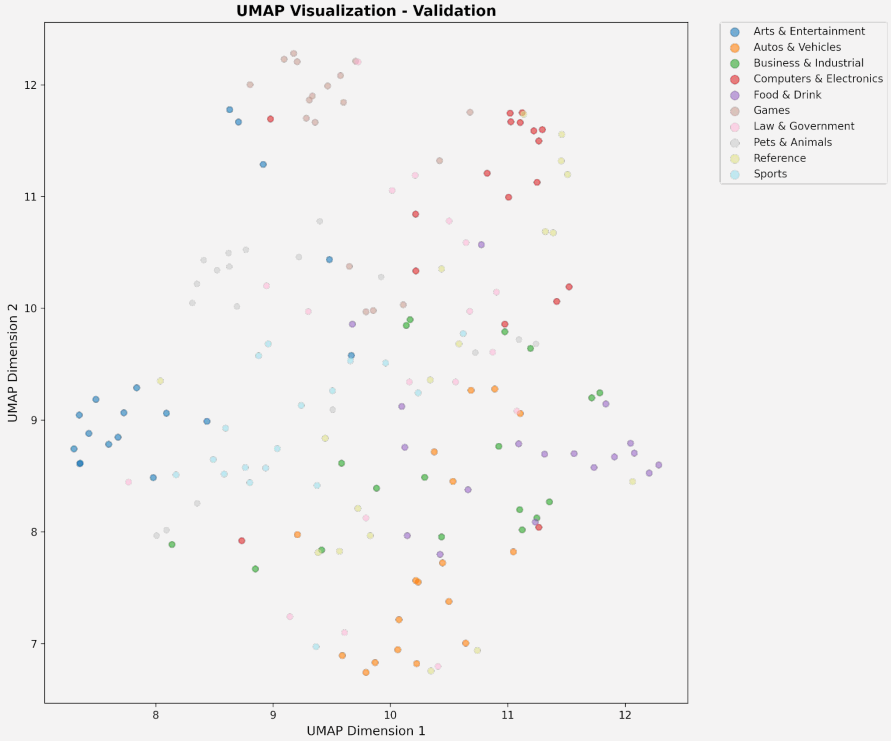


Dataset > YouTube-8M > Data Preprocessing > HDF5 file

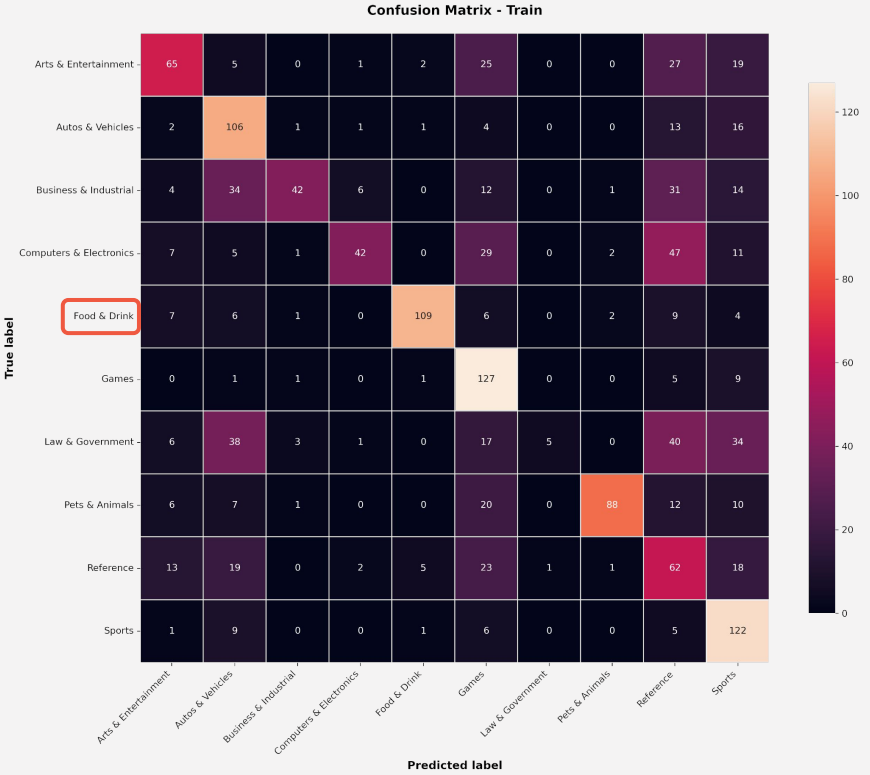
```
./embeddings/[train|validation|test]/<category_name>/
│
<video_id>.hdf5
│
├─ raw_features/                                  # original, "low-level" metadata
│   ├─ ffmpeg_numerical     float32[5]            # [width, height, fps, sample_rate, channels]
│   ├─ youtube_numerical    float32[5]            # [view_count, like_count, comment_count, favorite_count, duration_sec]
│   ├─ tags                 string[N_tags]        # variable-length list of tags
│   ├─ rating               string                # e.g. "ytAgeRestricted" or ""
│   ├─ dimension            string                # "2d" or "3d"
│   ├─ definition           string                # "hd" or "sd"
│   ├─ projection           string                # "rectangular" or "360"
│   ├─ video_codec          string                # e.g. "h264"
│   ├─ audio_codec          string                # e.g. "aac"
│   ├─ published_at         string                # ISO timestamp, e.g. "2025-05-27T03:14:15Z"
│   ├─ channel_title        string                # uploader's channel name
│   └─ default_language     string                # e.g. "en"
│
└─ embedded_features/                             # Third party embeddings and processed metadata
    ├─ audio_features                float32[1024]    # TwelveLabs audio embedding
    ├─ video_text_features           float32[1024]    # TwelveLabs visual-text embedding
    ├─ embedded_string_metadata      float32[3072]    # OpenAI embedding of string metadata
    └─ normalized_numerical_metadata float32[M_meta]  # L2-normalized numeric metadata vector
```

UMAP Visualization of Embeddings by Category





Baseline models > Zero-Shot Classification



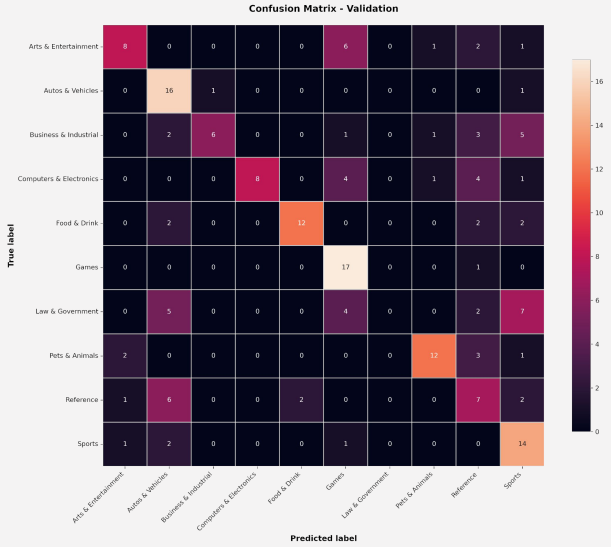


Table 1. Zero-Shot Classification

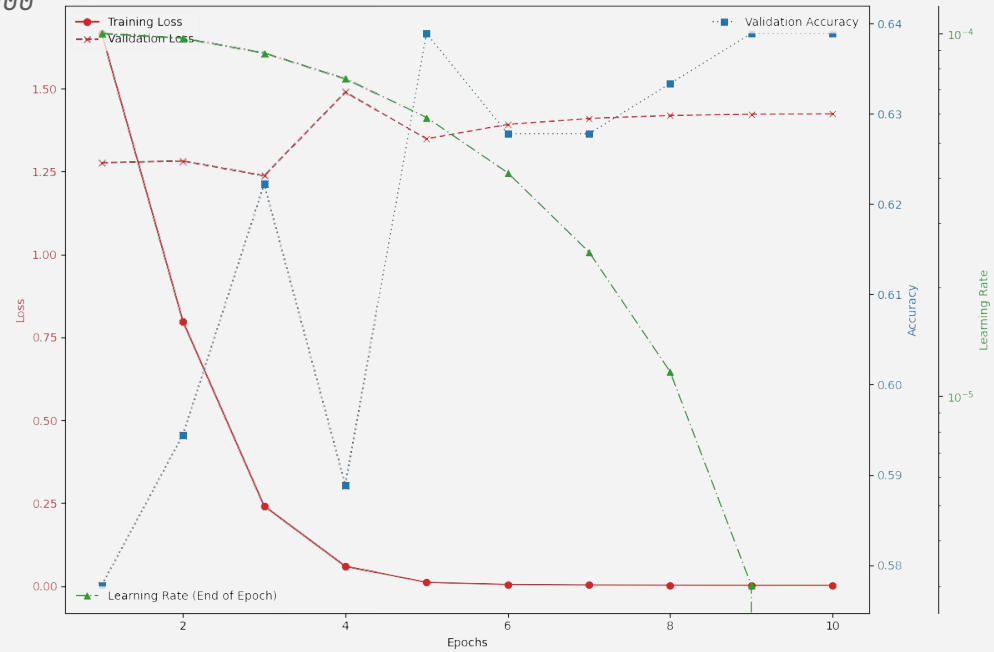
Metric	Score
Accuracy	55.6%
Precision	58.8%
Recall	55.6%
F1 Score	52.7%

Baseline models > TimeSformer-Based Classification

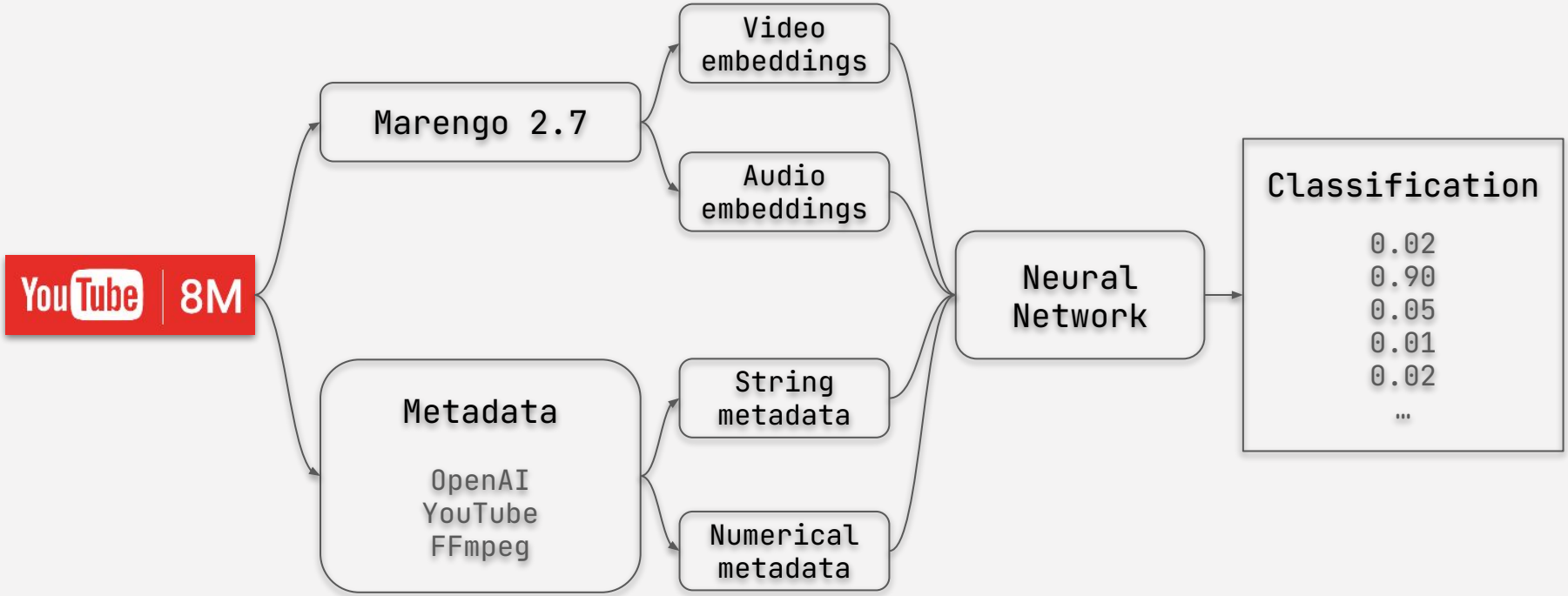
Training parameters

Fine-tuning TimeSformer pretrained for Kinetics-400

- Epochs 10
- Batch size 8
- Frame resized dim. 224x224
- Learning rate CosineAnnealingLR
  - Peak value 1e-4
  - Warm up length 1 epoch
- Total training time ~ 10h
- Accuracy 64.0%



Proposed Method > Overview

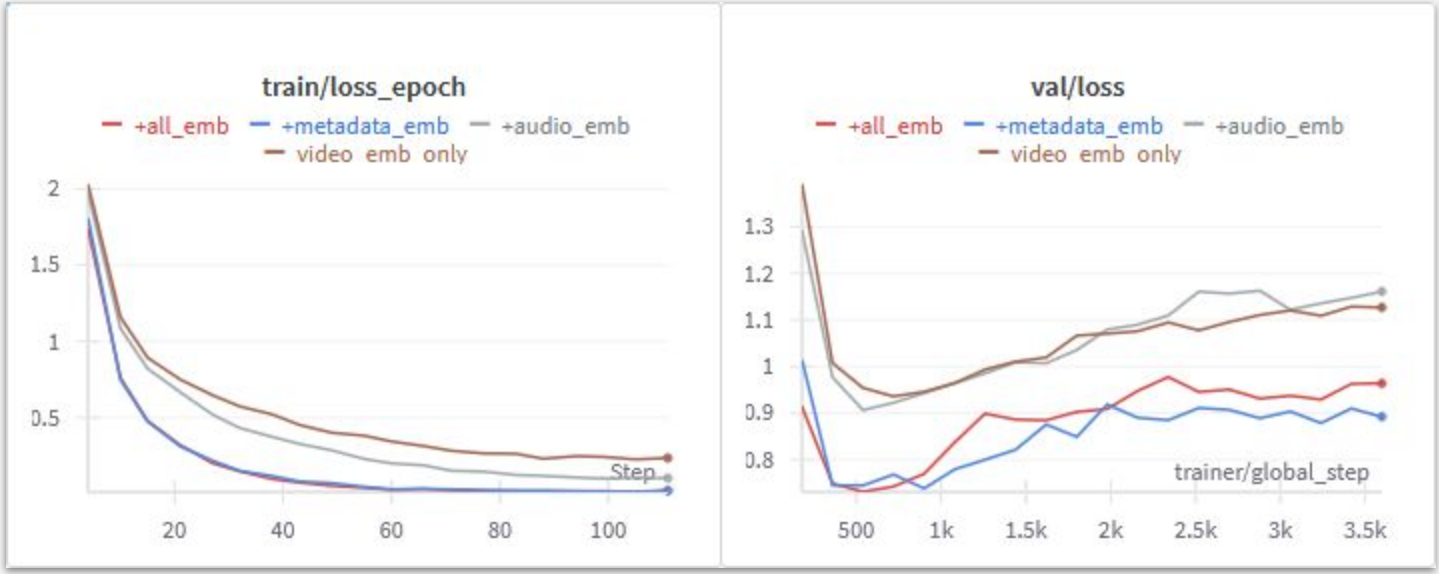


Proposed Method > Training

Training parameters

- Epochs 20
- Batch size 8
- Frame resized dim. 224x224
- Learning rate CosineAnnealingLR
  - Peak value 1e-4
  - Warm up length 1 epoch
- Total training time ~ 1m

Ablation Study > Input Features #1



Ablation Study > Input Features #2

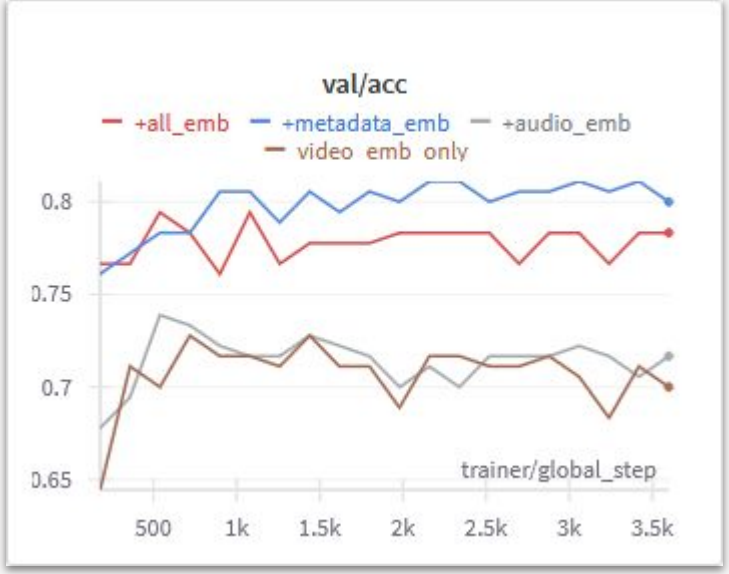


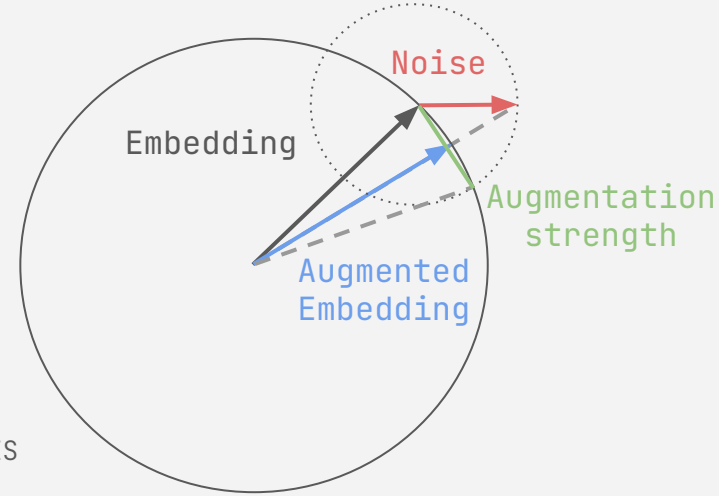
Table 1. Input ablation study

Mode	Validation accuracy
Video embedding	72.8%
+ Audio embedding	73.9%
+ Metadata embedding	**81.1%**
+ All embeddings	79.4%

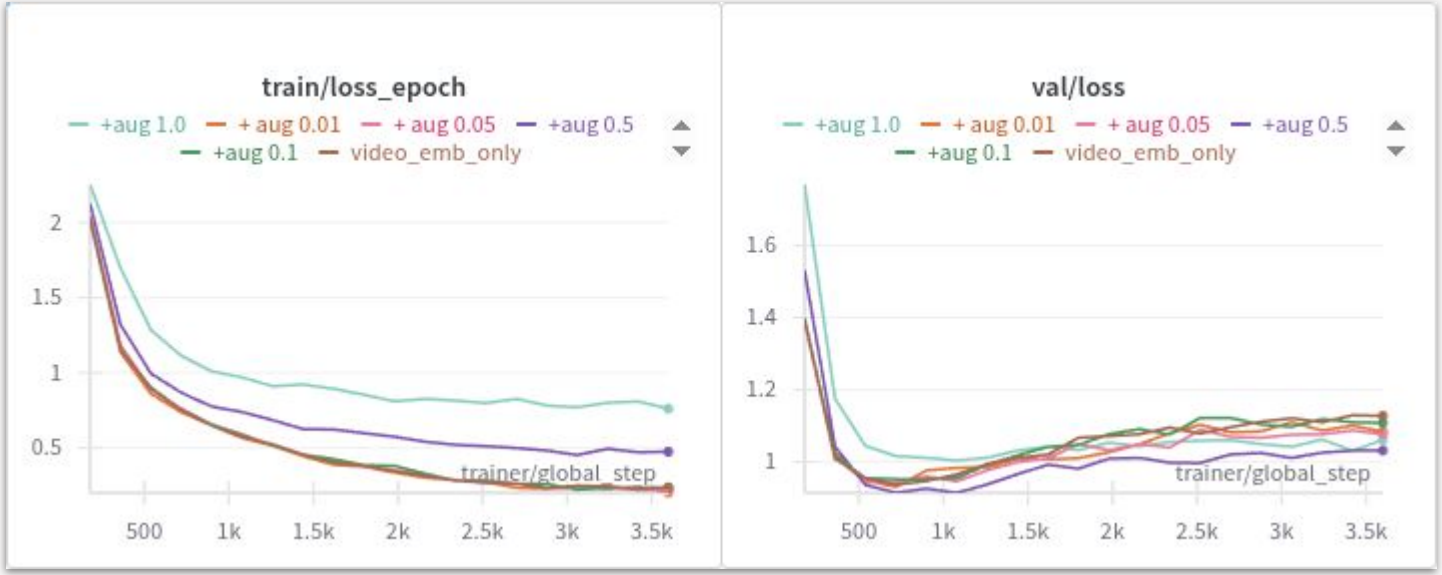
Data Augmentation Method

```python
import torch

noise = torch.randn_like(video_embedding)
noise = noise / (noise.norm(dim=1, keepdim=True) + 1e-8)
noise = noise * augmentation_strength

video_embedding = video_embedding + noise
video_embedding = video_embedding / (video_embedding.norm(dim=1, keepdim=True) + 1e-8)
```

https://github.com/usingcolor/20251R0136COSE47400/blob/main/model_training/model_pl.py#L32C25-L38C26



Ablation Study > Video Embeddings Augmentation #1



Ablation Study > Video Embeddings Augmentation #2

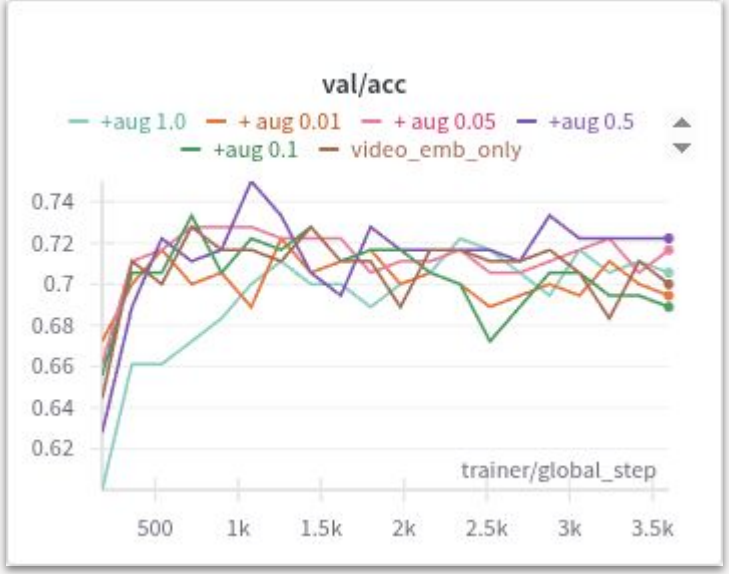
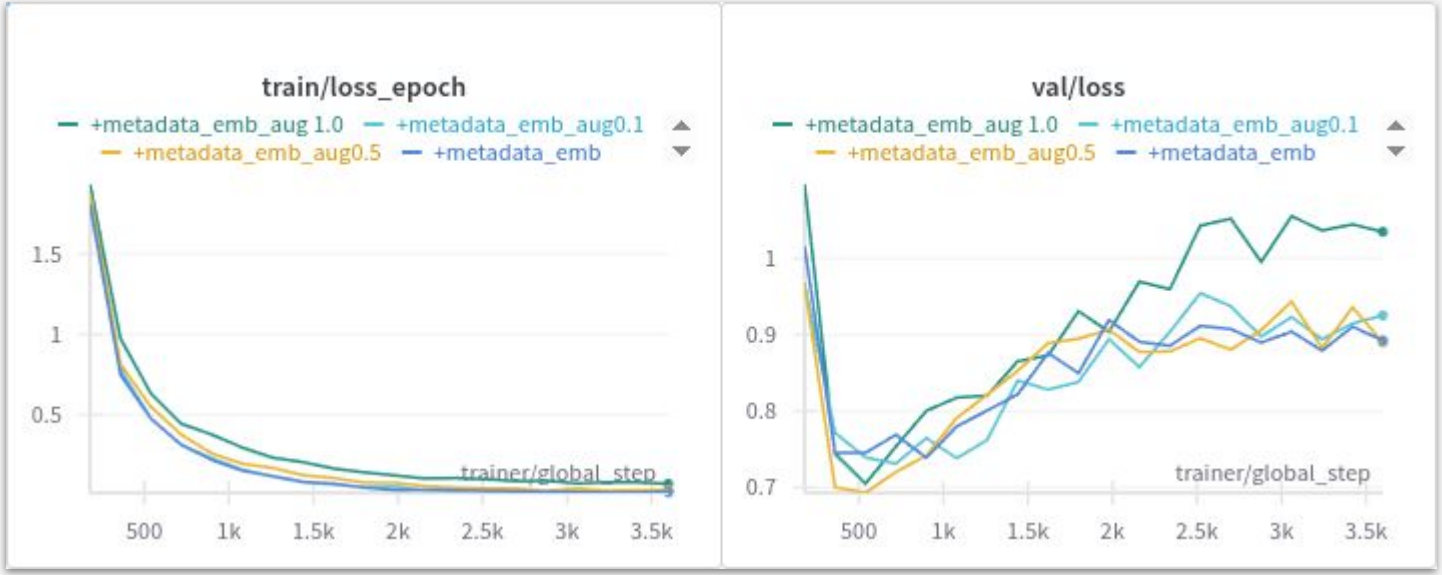


Table 1. Video embedding augmentation ablation study

Mode	Validation accuracy
Video embedding	72.8%
+ 0.01 aug. strength	72.2%
+ 0.05 aug. strength	72.8%
+ 0.1 aug. strength	73.3%
+ 0.5 aug. strength	**75.0%**
+ 1.0 aug. strength	70.0%

Ablation Study > Metadata Embeddings Augmentation #1



Ablation Study > Metadata Embeddings Augmentation #2

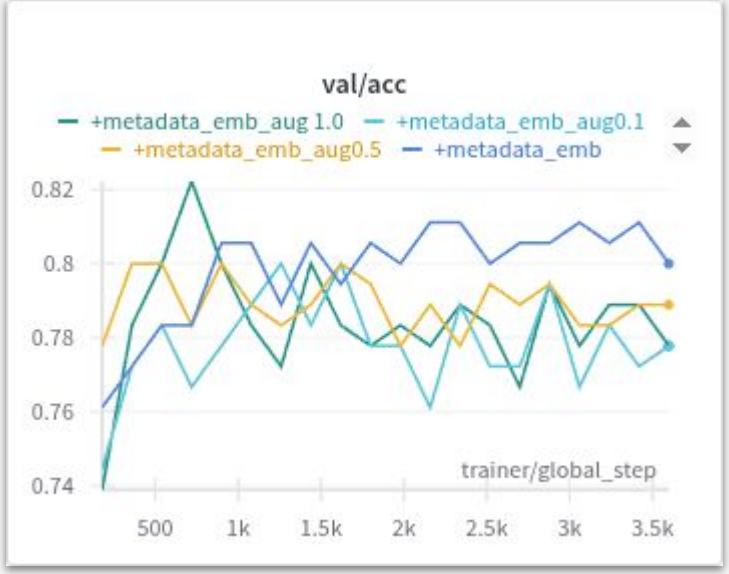
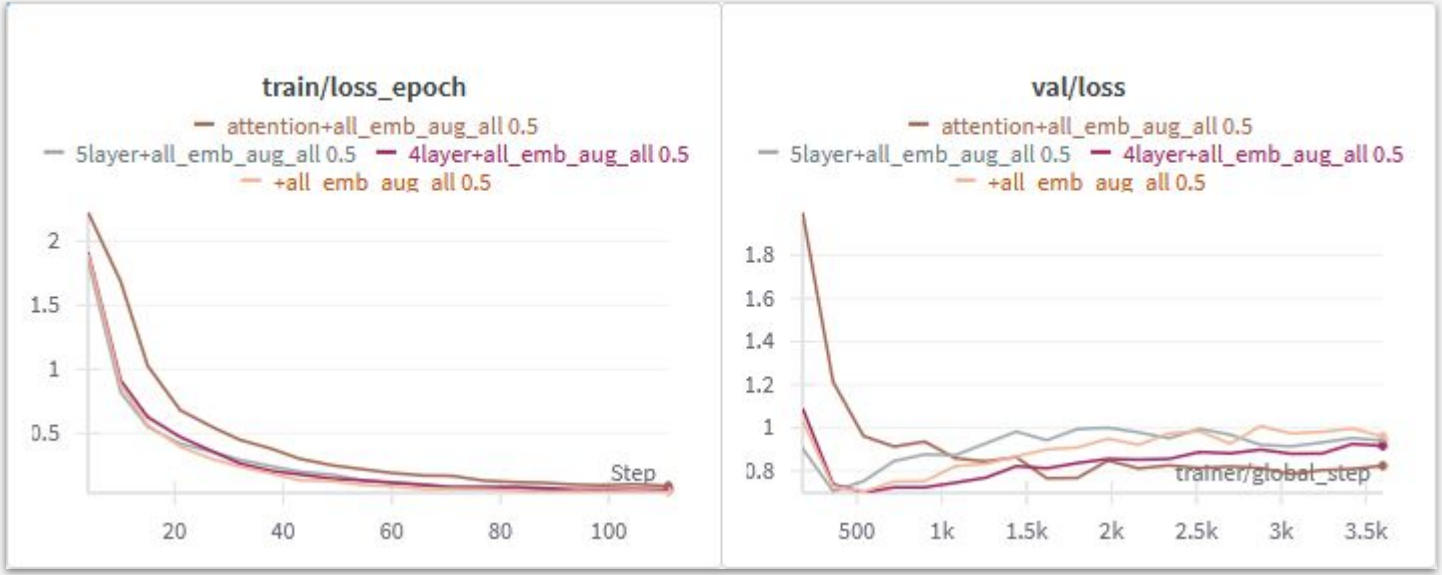


Table 1. Metadata embedding augmentation ablation study

Mode	Validation accuracy
Video emb. & Metadata emb.	81.1%
+ 0.1 aug. strength	80.0%
+ 0.5 aug. strength	80.0%
+ 1.0 aug. strength	**82.2%**

Ablation Study > Model Architecture #1



Ablation Study > Model Architecture #2

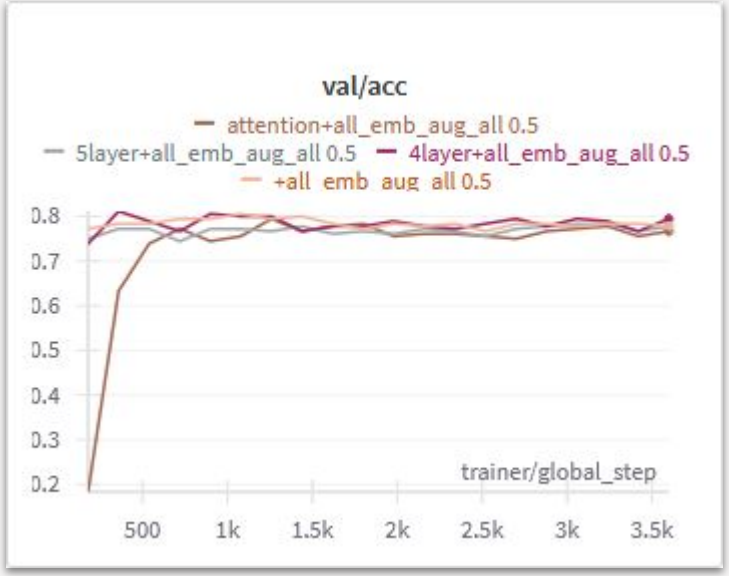
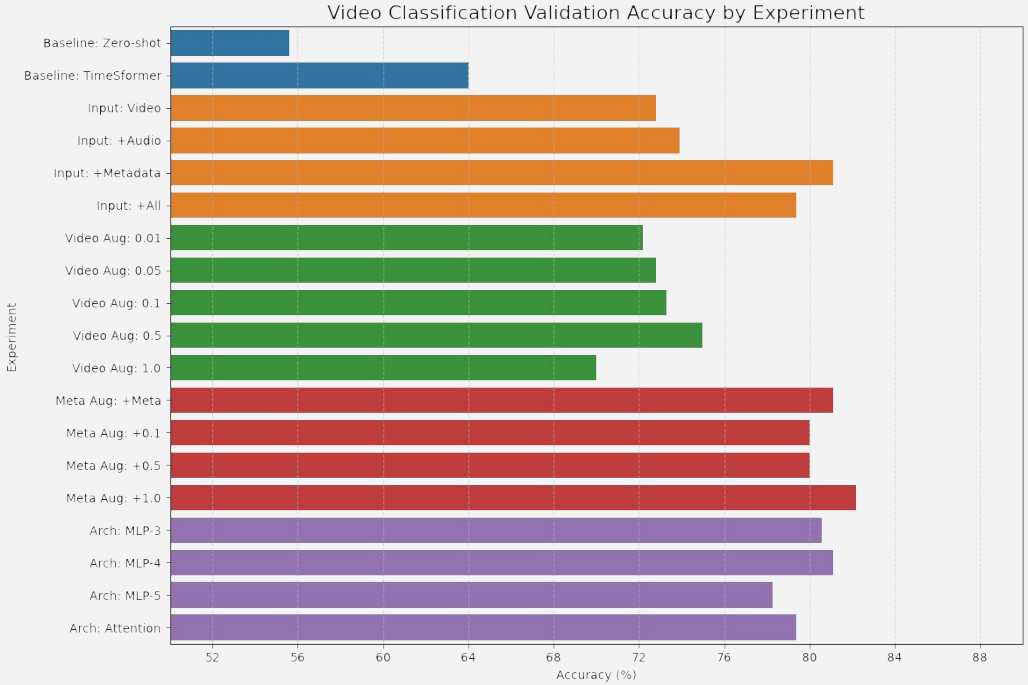
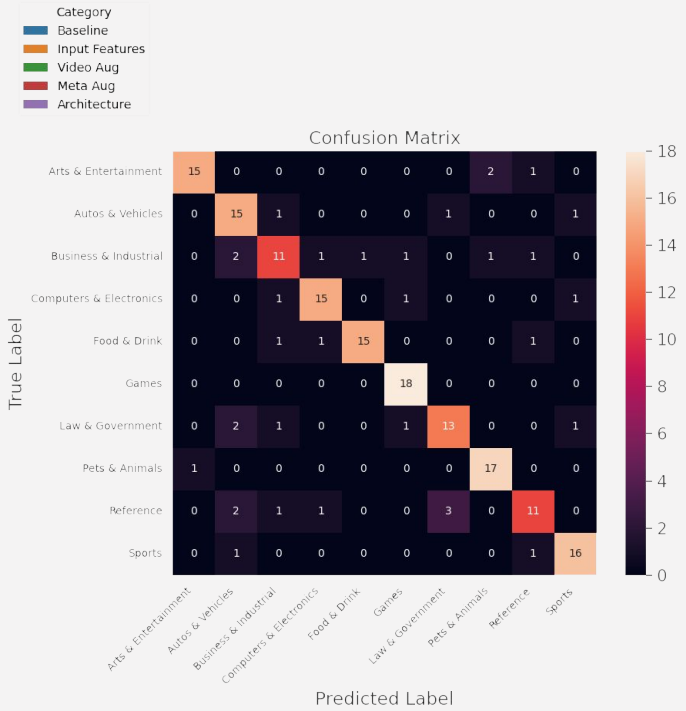


Table 1. Model architecture ablation study

Mode	Validation Accuracy
3 layers	80.6%
4 layers	**81.1%**
5 layers	78.3%
Attention	79.4%

Experimental Results





Closing Remarks

☞ **Our method** consistently **outperforms all baselines** across key metrics (e.g., accuracy), **with significantly reduced training time.**

☞ Our **novel augmentation strategy mitigates overfitting** in cosine distance space.

☞ Integrating additional **embeddings** (video, audio and metadata) **enhances model performance.**

☞ The **4-layer architecture** yielded the **best overall performance.**