TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

“Human-in-the-loop Video Mining” - Public Version

Authors	Rozafë Llalloshi, Jaeyoung Cho
Mentors	M.Sc. Keesiu Wong (Design AI)
	M.Sc. Frederik Mattwich (Design AI)
Co-Mentor	M.Sc. Michael Rauchensteiner (Department of Mathematics)
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Jul 2021

Abstract

Videre AI is a product developed with the goal of speeding up and automating the video annotation process. The current product is powered by a Multi-modal transformer (MMT) model and targets the Dense Video Captioning Task, the task of identifying scenes on video and describing them using natural language sentences. MMT shows great performance on the ActivityNet dataset, surpassing state-of-the-art. However, sometimes extraction of higher-level information from video is needed ex. tagging a video with multiple labels on multiple semantic aspects such as objects seen on the video, human actions, events, concepts, etc. In our work, we use the powerful Multi-modal transformer encoder and propose an MMT tagging model that does video tagging by utilizing audio, video, speech, and optical flow modalities. We show that our model surpasses state-of-the-art on the Holistic Video Understanding (HVU) dataset. We research current research trends for the video tagging task and compare our MMT tagging model to current state-of-the-art models. Specifically, we train the TimeSformer model on the HVU dataset and compare it to our results.

Contents

1 Introduction

1.1 Motivation

Video content, be that for marketing, entertainment, education, etc. is massively growing. Companies of various industries rely on creating, accessing, and analyzing video content daily. Accessing and analyzing video-based content is not a trivial task, it requires videos to be manually annotated with information such as scenes, tags, captions, etc. Manual annotation, done by human annotators, is expensive and inefficient therefore it is often a productivity bottleneck for such companies. Videre AI is a product designed to increase video annotation efficiency, by providing automatic scene detection and captioning. Videre AI's intelligence comes from a multi-modal transformer (MMT) model trained for the Dense Video Captioning Task, the task of identifying scenes on video and describing them using natural language sentences. The MMT model reports state-of-the-art performance on the ActivityNet [13] dataset. For some companies, describing scenes on such a fine-grain level, with full sentences, is not needed, what they require to facilitate their daily operations are simple tags. Tagging a video based on its visual and audio information on different aspects such as objects it contains, surroundings, human actions seen on the video, etc. allows companies to get a higher-level understanding of the video content. Driven by the MMT dense video captioning model results we believe that the MMT model is very good at video understanding, therefore the goal of this project is to propose a novel MMT model for the video tagging task, the task of tagging a video with more than one label, which is known as *multi-label classification*.

1.2 Objectives

This project aims to adapt the current MMT model designed for dense video captioning for the multi-label classification task. To leverage this model we need a good dataset that supports Videre AI's product vision. Because the MMT model was designed based on research on the dense video captioning task, in our project we also want to research recent developments in video classification and compare our model to the current trends and state-of-the-art models. To achieve this we structured our work into the following phases:

- **Dataset research**: Research for multi-label video datasets that offer general labels which allow us to understand video on multiple different aspects and support the Videre AI product vision.
- **MMT tagging model**: Develop, train, and evaluate a tagging model which uses the current MMT dense video captioning model as a baseline
- **Research video classification trends**: Compare MMT model to current research trends in video classification

1.3 Our Approach

We identify the Holistic Video Understanding (HVU) [4] as a dataset that best matches our needs. HVU focuses on multi-task and multi-label video understanding by recognizing multiple semantic aspects of a video scene. These semantic aspects are defined on categories of scenes, objects, actions, events, attributes, and concepts that reflect the real-world scenarios. We create a multi-model transformer model for video tagging and show that our proposed model surpasses state-of-the-art results on the HVU dataset. We compare the current MMT model to Vision Transformers, a recent development in computer vision that targets image and video classification with transformer-based models.

2 Related Work

In this section, we will cover the research that our work is based on. This includes the MMT dense video captioning model that is the baseline for our MMT tagging model, research on Vision Transformers for video classification and HVU dataset used to train our tagging models.

2.1 Multi Modal Transformer

The Multi-Modal Transformer (MMT) is a Dense Video Captioning model which is the brain of the Videre AI product. The model segments a video into scenes and generates a natural language description for each proposed scene. MMT is based on the Bi-modal Transformer (BMT) [11] which generalizes the Transformer architecture for a bi-modal input, considering both video and audio modalities for the dense video captioning task. MMT takes this idea that considering other modalities contributes to the understanding of the video content a step further by extending the model to consider speech and optical flow alongside video and audio features.

2.2 Vision Transformers

We researched current trends in video classification and found that pure transformer based models for image and video classification have recently shown great success in computer vision. Previously, in computer vision, attention was applied in conjunction with convolutional network, Vision Transformer [5] (ViT) has shown that reliance on CNNs is not necessary and a pure transformer applied directly to image patches can perform very well on image classification and requires less computational power to train.

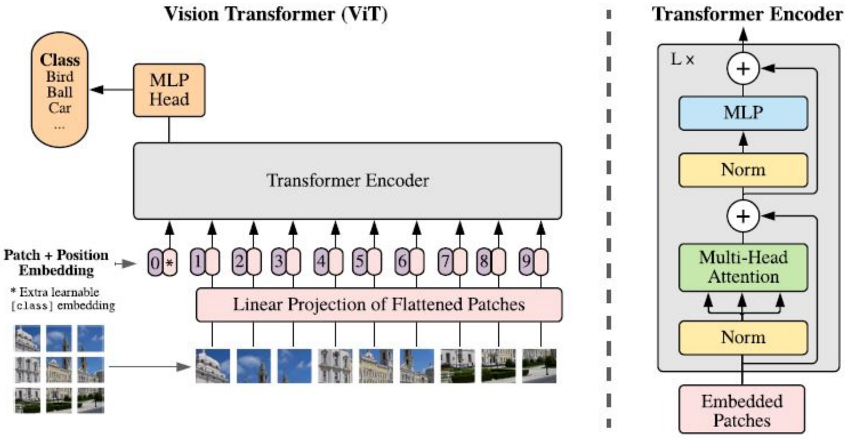


Figure 1: ViT Architecture and Transformer Encoder

The standard Transformer model receives a sequence of 1D token embeddings as input, as seen in Figure 1 ViT handles 2D image input $x \in \mathbb{R}^{H\times W\times C}$ by splitting it into fixed-size patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H,W) indicate the height and width of the original image, C is the number of channels, and (P,P) is the resolution of each image patch, so the resulting number of patches is $N = HW/P^2$. There patches are then mapped to D dimension patch embeddings with a trainable linear projection. To retail positional information, position embeddings are added to patch embeddings. The sequence of patch embeddings is fed into Transformer Encoder and the output for the extra learnable class embedding is used for classification. The Transformer encoder is composed of alternating layers of Multi-Head Attention and Multi Layer Perceptron(MLP) blocks. Norm layer is applied before every block and there are residual connections after every block.

Although 3D convolutional architectures [2, 6, 7] are still broadly used for video related task we are seeing a lot of research on full transformer based video classification fueled by ViT which we will discuss in the next session.

2.3 Dataset

To support Videre AI's vision the dataset should fulfill the following criteria:

- The covered topics should be general and not application-specific. Videre AI is a customer-centric product and our tagging model should not focus on a highly specific task but rather describe the overall content of a video.
- The dataset should contain multiple labels per video.
- To extract modality features for our MMT tagging model we should have access to the raw videos and it should be feasible in terms of computation power and time to extract the features for each modality.

While there are quite a few video datasets, there were surprisingly not many that satisfied our criteria. Almost all video dataset are targeting human action classification and/or provide only one label per video. Table 1 reviews a shortlist of datasets we considered.

Dataset	Topic	Multi-labels	Size	Public
Holistic Video Understanding	You8M+Kinetics	avg 6/video	580K	Yes
SOA	General	avg 6/video	562K	No
Youtube-8M	General	avg 3/video	8M	Yes
Kinetics-700	Human Actions	No	650K	Yes
ActivityNet Captions	Action	No	20K	Yes

Table 1: Overview of dataset shortlist

The MMT dense video captioning model was trained on the ActivityNet Caption dataset [13], it would be easy for us to continue working on the same dataset considering the modality features have already been extracted but ActivityNet only provides one label per video and the topic is human activity, therefore, it does not fulfill our criteria.

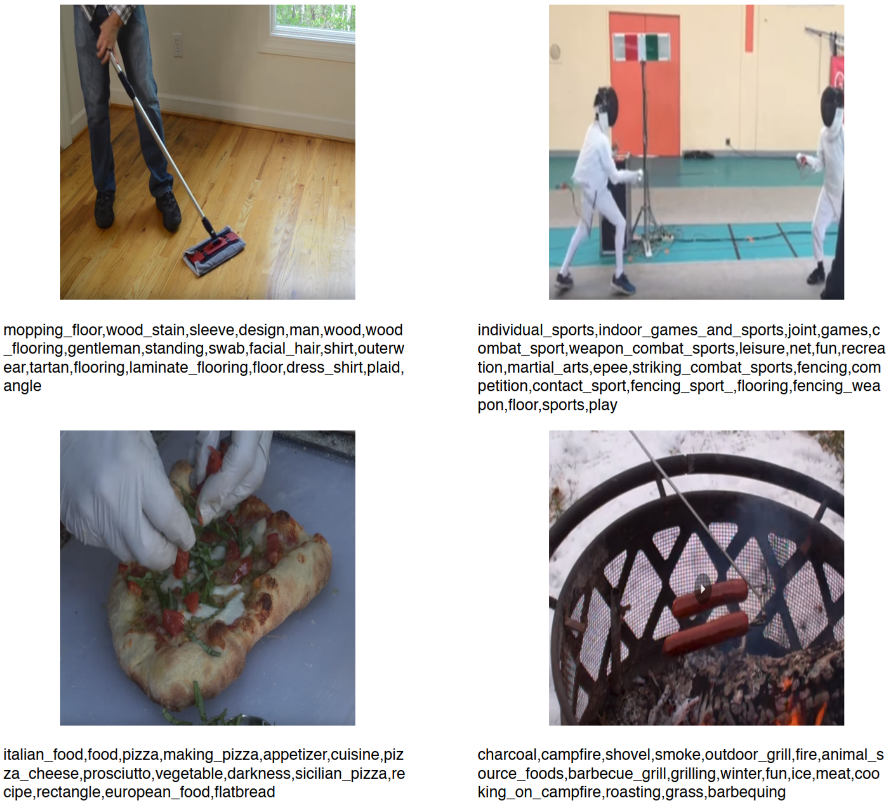
The Kinetics-700 dataset [3] is highly used in research and most of the models we want to compare our MMT tagging model to report their performance against it, but same as the ActivityNet dataset Kinetics also provides labels per video and the topic is limited to human actions.

The Youtube-8M [1] dataset is a large-scale video dataset that offers general labels and is suited for multi-label classification. Youtube-8M provides already extracted features for audio and RGB, as well as youtube ids that can be used for downloading the raw videos. The problem with the dataset is that it provides 8M videos and it would be infeasible in terms of time for us to download all videos and extract features for all four modalities.

Holistic Video Understanding (HVU) [4] dataset stood out and fulfilled our criteria. HVU focuses on multi-label and multi-task video understanding by organizing labels into 6 semantic aspects: scenes, objects, events, actions, attributes, and concepts allowing us to recognize multiple aspects in a dynamic scene. HVU was created by a mixture of videos from the Youtube-8M, Kinetics-600, and HACS [18] datasets, which were enriched with more annotations in a two-stage approach. First, they get rough annotations using Google Vision API [9] and Sensifai Video tagging API [16], around 30 labels per video, and in the second stage humans manually remove any incorrect labels and possibly add missing labels.
Table 2 shows dataset statistics for each of the HVU categories. There are a total of 3142 labels spread into 6 main categories. The category with the most labels is the object category, followed by the action category.
Figure 2 shows a few examples of video frame samples from the HVU dataset and the corresponding ground truth tags of different semantic categories. These examples show the diversity of the videos in the HVU dataset.

	Scene	Object	Action	Event	Attribute	Concept	Total
Labels	248	1678	739	69	117	291	3142
Annotations	672,622	3,418,198	1,473,216	245,868	581,449	1,108,552	7,499,905
Videos	251,794	471,068	479,568	164,924	316,040	410,711	481,417

Table 2: Statistics of HVU training set for different semantic categories



mopping_floor,wood_stain,sleeve,design,man,wood,wood_flooring,gentleman,standing,swab,facial_hair,shirt,outerwear,tartan,flooring,laminate_flooring,floor,dress_shirt,plaid,angle

individual_sports,indoor_games_and_sports,joint,games,combat_sport,weapon_combat_sports,leisure,net,fun,recreation,martial_arts,epee,striking_combat_sports,fencing,competition,contact_sport,fencing_sport_,flooring,fencing_weapon,floor,sports,play

italian_food,food,pizza,making_pizza,appetizer,cuisine,pizza_cheese,prosciutto,vegetable,darkness,sicilian_pizza,recipe,rectangle,european_food,flatbread

charcoal,campfire,shovel,smoke,outdoor_grill,fire,animal_source_foods,barbecue_grill,grilling,winter,fun,ice,meat,cooking_on_campfire,roasting,grass,barbequing

Figure 2: Examples of video frame samples from the HVU dataset with corresponding tags

Similar to HVU, the SOA [15] dataset also aims to recognize scenes, objects, and actions, but even though the paper was released in 2018, the dataset is not available for the public yet.

2.3.1 HVU state-of-the-art models

HVU is a fairly new dataset, being made public only about 15 months ago there are not many models that report their result for it. HATNet (Holistic Appearance and Temporal Network) [4] which was introduced along with the HVU dataset is the first model that reports its results for the HVU dataset. HATNet fuses 2D and 3D architectures extracting temporal and appearance information as two sources of information for video recognition. At the time of publishing, HATNet, pre-trained on HVU, improved state of the art performance on multiple datasets such as Kinetics-400, Kinetics-600 etc.

FrameExit [8] published on April 2021, sets a new state of the art for HVU. FrameExit aims at reducing computation costs by early-exiting, the main idea as that only a few frames are needed to tag "easy" videos and only a few "hard" videos require more temporal information. FrameExit extracts features using 2D image models pretrained on ImageNet as their backbone network. Temporal aggregation of features is done using average/max pooling.

In comparison to the MMT model, neither HATNet nor FrameExit are transformer models or learn from multiple modalities.

3 Downloading the HVU dataset

The HVU dataset provides CSV files for the training and validation set, these CSV files contain the youtube id for each video together with the corresponding ground truth tags and the start and end time of the labeled clip. Because HVU is an open challenge the test set is not publicly available, to report the results we use a subset of the validation set. We use the youtube ids to download the raw videos and clip them according to the start and end time. A downloader for the HVU dataset is already provided by HVU Research Group [10]. In this original downloader, the trimming process is done only after all videos have been downloaded. This is problematic because it requires a lot of storage and requires processing the whole dataset twice. As described in Figure 3 we modified the downloader to clip videos before saving them and to save the videos on AWS S3 object storage.

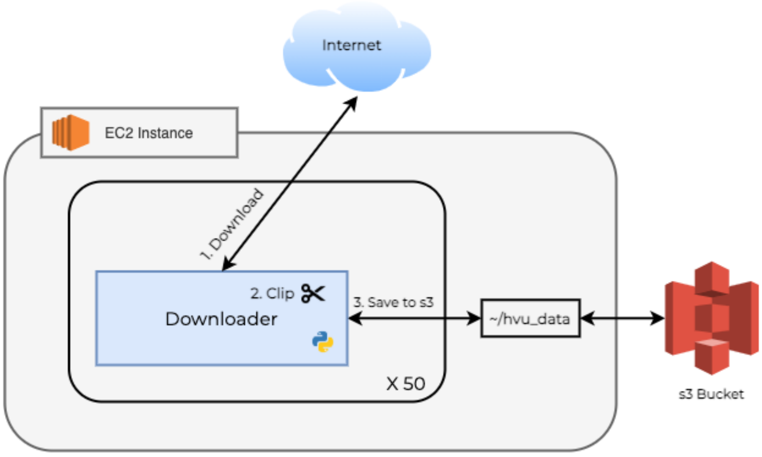


Figure 3: Proccess of downloading raw videos of the HVU dataset

In our HVU downloader, we use joblib [12] library for parallel computing, we launch 50 jobs, each job first downloads a single HVU video from the internet using youtube-dl [17], the video is then clipped according to the start and end time using moviepy [14] and finally, the clipped video is moved to AWS s3 object storage. With this change, our HVU Downloader is resilient from any unexpected interruptions, such as losing connection to the EC2 Instance, unhandled error from the code, etc.

Since the HVU dataset consists of public youtube videos, some of these videos might have been deleted by the owners or have been made private, therefore after downloading all the available videos we cleaned CSV files for train and validation set to remove unavailable videos. In the end, we here left with 432891 training samples and 27993 validation samples i.e we only possess approximately 90% of the training set.

4 Results and Discussion

4.1 Evaluation Metric

In multi-label classification, each sample can belong to more than one label, therefore we assume every label to be a Bernoulli random variable representing a different classification task. As such in the final layer, we use the sigmoid activation function instead of the softmax activation function used for multi-class classification. Therefore the output of our model is not a probability distribution over the possible labels, instead, each value in the output vector tells us the probability of the sample belonging to a particular label with possibly more than one probability value larger than 0.5. In multi-label classification, a prediction is no longer a hard wrong or right. The ground truth contains a list of labels for the sample and a prediction that predicts some of the ground truth labels should be considered better than one that does not predict any of them.

To quantify the quality of our model we use the mean Average Precision (mAP) measure. mAP is the mean of the Average Precision (AP) calculated for each label. Let n be the number of labels and AP_k the Average Precision for the k-th label, then:

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k$$

The Average Precision for a label summarizes the precision-recall curve as a weighted mean of precisions achieved at each threshold, where the weight is given by the increase in recall from the previous threshold.

4.2 Final Results

4.2.1 MMT tagging model

Our model is trained and tested on the HVU dataset. Since the HVU test set is not publicly available we report our results on a subset of the HVU validation set, a subset that is not seen by the model during training nor when using the validation set for hyperparameter tuning. We compare out MMT tagging model against other models that report their performance on the HVU dataset and show that even though we only posses 90% of the training dataset (explained in section 3) we significantly outperform the state-of-the-art. The results of the comparison are shown in table 3.

	Scene	Object	Action	Event	Attribute	Concept	Overall
HATNet (Multi-label)	55.8	34.2	51.8	38.5	33.6	26.1	40
HATNet (Multi-task)	57.2	35.1	53.5	39.8	34.9	27.3	41.3
FrameExit (Multi-label)	NK	NK	NK	NK	NK	NK	47.7
Ours (Multi-label)	54.7	**52**	**74.6**	**62.1**	**47**	**52.7**	**57**
Ours (Multi-task)	53.6	50	71.7	61.2	45.5	51.9	55.7

Table 3: State-of-the-art performance (mAP%) comparison on HVU.

Having a pre-trained MMT encoder for the dense video captioning task we also did a transfer learning experiment, initializing the encoder weights with the pre-trained encoder weights. Considering that the ActivityNet dataset is 25 times smaller than the HVU dataset we did not gain in performance.

Qualitative Analysis

In this section, we show a few prediction examples to get a better feel of what our model was able to learn. These samples are selected from the HVU test dataset and therefore never before seen by the model. For comparison, we show the ground truth tags alongside the predicted tags and use colors to indicate correctness. We use green if a tag is predicted correctly, red if a tag is an incorrect prediction, and yellow if the tag is on the ground truth but was missed by our prediction. These examples show the diversity of videos on the HVU dataset and how detailed tagging for multiple semantic aspects is. The example shown in Figure 4 goes to show that even labeling such a dataset in such detail is challenging, we can see that our model predicted the tag man, and even though we do see a few men on the video the tag is missing on the ground truth tags. In the example shown in Figure 5 we can witness the same, while we definitely see an arm in the video and our model predicts it, the tag is missing on the ground truth tags.
Most importantly these examples show how useful the predicted tags are and how our model is able to capture the meaning of the video in multiple semantic aspects.

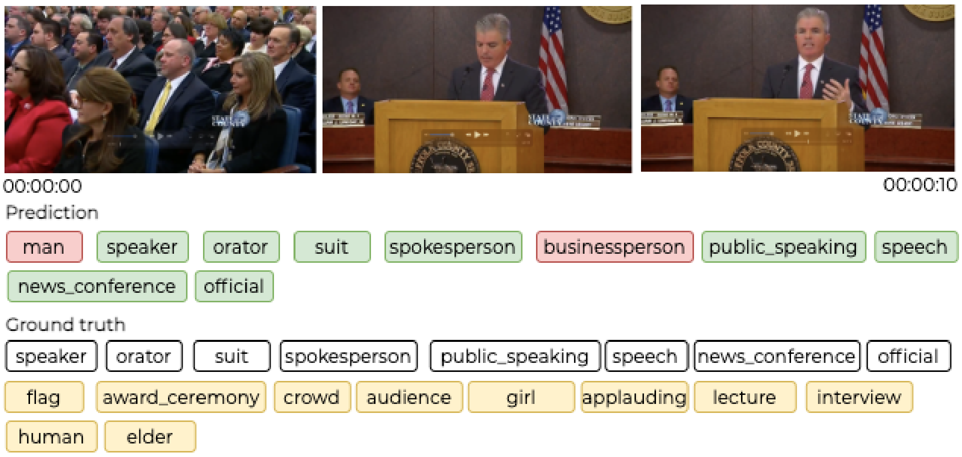


Figure 4: Prediction example 1. Video can be accessed here.

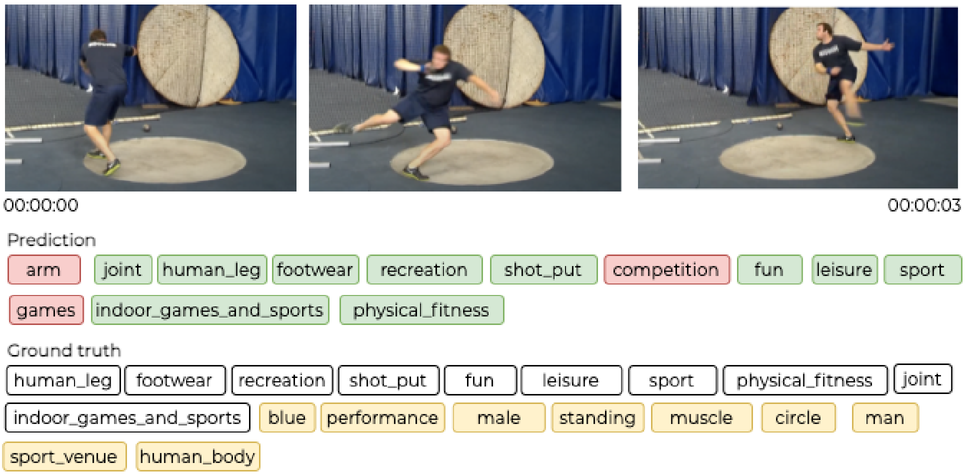


Figure 5: Prediction example 2. Video can be accessed here.

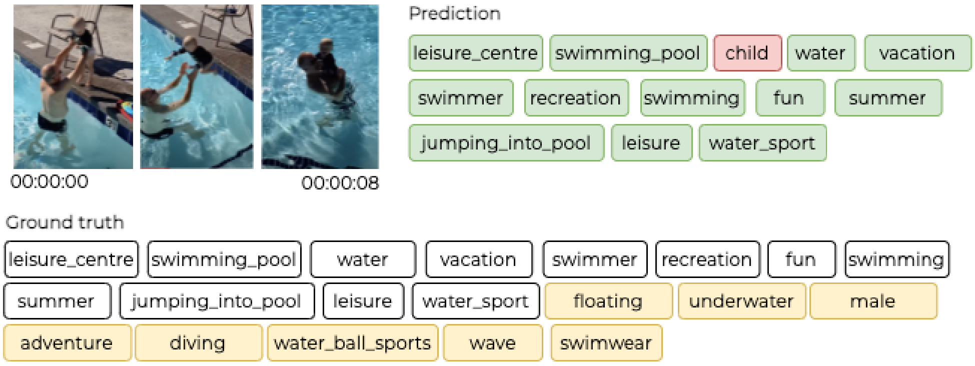


Figure 6: Prediction example 3. Video can be accessed here.

4.2.2 TimeSformer model

Training TimeSformer is quite challenging. Timesformer takes the raw videos as input and as such, it takes a lot of time and resources to train. Table 4 shows the reported results after training TimeSformer and compares them to our MMT tagging model for the same task, multi-task multi-label classification.

	Scene	Object	Action	Event	Attribute	Concept	Overall
Ours(Multi-task)	53.6	**50**	**71.7**	61.2	45.5	51.9	55.7
TimeSformer(Multi-task)	**58.6**	45.0	59.07	**65.2**	**51.8**	**56.3**	**56.0**

Table 4: TimeSformer vs MMT performance (mAP%) comparison for multi-task multi-label classification.

Although we did not manage to perform hyperparameter tuning as much as for our MMT tagging model, TimeSformer slightly surpasses our MMT tagging model on multi-task multi-label classification. We believe TimeSformer results can we be further improved by enabling data augmentation which we had disabled to get a fair comparison to the MMT tagging model which does not perform data augmentation.

4.2.3 Discussion

The MMT model takes as input modality features extracted with pre-trained models where one particular embedding for a specific modality represents a fragment of 2.56/0.96 seconds of the video/audio stream in the case of visual/audio features or a word in the transcript for speech features. To extract visual features 3D convolutions are used, which means the visual embeddings encode spatial and temporal information, as a result, we have a sequence of visual spatio-temporal embeddings each encoding 2.56 seconds of the full video. This sequence is fed to the transformer encoder, modeling pairwise interactions between the video fragment embeddings, computing attention temporally.

In contrast to this, the TimeSformer model acts directly on the raw video. Patching sampled 2D frames and linearly projecting them with a learnable projection. Patch embeddings are forwarded to the encoder with uses divided space-time attention where temporal attention and spatial attention are separately applied one after the other. In short, the MMT model computes attention temporally whereas in TimeSformer we have spatial plus temporal attention which we believe is the reason behind our results.

5 Conclusion

In this project, we introduced a novel video tagging model that tags a video for multiple semantic aspects. Our Multi-modal model makes use of audio, speech, visual and optical flow features which are fused together into a video representation vector which is then used for classification. We show that our model is able to understand video in multiple semantic aspects by surpassing state-of-the-art on the HVU dataset. Additionally, we research Vision Transformers, a new research area that explores transformer models for image and video understanding. To compare our MMT tagging model to Vision Transformers we train the TimeSfomer model on the HVU dataset and show that TimeSformer slightly surpasses the MMT model on the multi-task multi-label classification task.

Currently, our MMT tagging model does not use data augmentation and we experience overfitting after the 6-th training epoch. We strongly believe that using data augmentation techniques on the feature space would allow us to train longer and learn more and should be explored in the future.

Future work to improve results on the HVU dataset would be to apply the idea of using multiple modalities to learn video representations to a full transformer-based model. Or to further improve the Videre AI MMT dense video captioning model by using a full transformer based approach.