ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՄԱԹԵՄԱՏԻԿԱՅԻ ԵՎ ՄԵԽԱՆԻԿԱՅԻ ՖԱԿՈՒԼՏԵՏ

ԿԻՐԱՌԱԿԱՆ ՎԻՃԱԿԱԳՐՈՒԹՅՈՒՆ ԵՎ ՏՎՅԱԼՆԵՐԻ

ԳԻՏՈՒԹՅՈՒՆ ՄԱԳԻՍՏՐՈՍԱԿԱՆ ԾՐԱԳԻՐ

ԲԼԲՈՒԼՅԱՆ ԳՈՒՐԳԵՆ ՎԱՐԴԱՆԻ ՄԱԳԻՍՏՐՈՍԱԿԱՆ ԹԵԶ

ՏԵՔՍՏԻ ՍՏԱՑՈՒՄԸ ՎԻԴԵՈՅԻ ՀԻՄԱՆ ՎՐԱ

«Կիրառական վիճակագրություն և տվյալների գիտություն»

մասնագիտությամբ

վիճակագրության մագիստրոսի որակավորման աստիճանի համար

Երևան 2022

Ուսանող՝ _____

Բլբուլյան Գուրգեն

Գիտական ղեկավար՝ _____

ֆ.մ.գ.թ., Ամիրխանյան Գագիկ

«Թույլատրել պաշտպանության»

Մագիստրոսական ծրագրի ղեկավար՝ _____

ֆ.մ.գ.դ., ասիստենտ Քեռյան Կարեն

«_____»_____20__թ.

Տեքստի ստացումը վիդեոյի հիման վրա

Video-based text generation

Генерация текста из видео

ABSTRACT

In this work, we propose an end-to-end approach for text generation from a video by using a vanilla transformer architecture. We connect a pretrained vision encoder and language decoder into one transformer and finetune it. Using ViT as a decoder and GPT2 as an encoder gives high-quality text generation with deep understandings of video context.

Codes are available at this address

https://gitlab.com/blbulyangurgen/ysu-asds-thesis

Here is a deployed app in huggingface

https://huggingface.co/spaces/gurgenblbulyan/video-based-text-generation

Model is available at this address

https://huggingface.co/armgabrielyan/video-summarization

Table of Contents

INTRODUCTION

With the recent advances in self-supervised learning, pre-training techniques play a vital role in learning visual and language representations. With the development of transformers and the increase in computational power, scientists started creating more and more large models. Natural Language Processing (NLP) transformers are trained using huge amounts of data and some of them are trained on every text on the internet and understand natural languages very well. Transformers are not only used in NLP but also in Computer Vision (CV) for learning visual representations. Using multimodal transformers will help solve many problems, such as image captioning, video-text retrieval, visual question answering, Optical Character Recognition (OCR), video to text generation, and searching in video.

In this work, we try to solve the video-based text generation problem. The main goal of this problem is to generate a good quality textual representation of a video by understanding its semantic context. This can be useful for, for example, guiding drivers, sport commentary, searching in videos, healthcare.

RELATED PROBLEMS

VIDEO-TEXT RETRIEVAL AND TEXT-VIDEO RETRIEVAL

The objective of video retrieval is to select the video which corresponds to a given text query from a pool of candidate videos or vice versa. This plays a crucial role in multi-modal video-and-language understanding. Much of the recent work on video-text retrieval has applied transformer-based solutions on both video and text encoder ends, especially the family of CLIP [1] methods, such as CLIP4Clip [2], CLIP2Video [3] and CLIP2TV [4]. In these model architectures, CLIP is the backbone, serving as encoders on both sides, which are used to process both video and text data. More concretely, ViT is applied to encode raw visual information and a Transformer like BERT is used to encode raw text information.

CLIP is an object identification model that is jointly trained on a variety of (image, text) tuples of training examples. It consists of an image encoder and a text encoder that are trained on 400 million images and corresponding text data on the web to predict the correct mapping between image and text. It possesses zero-shot capabilities and can perform object identification without re-training.

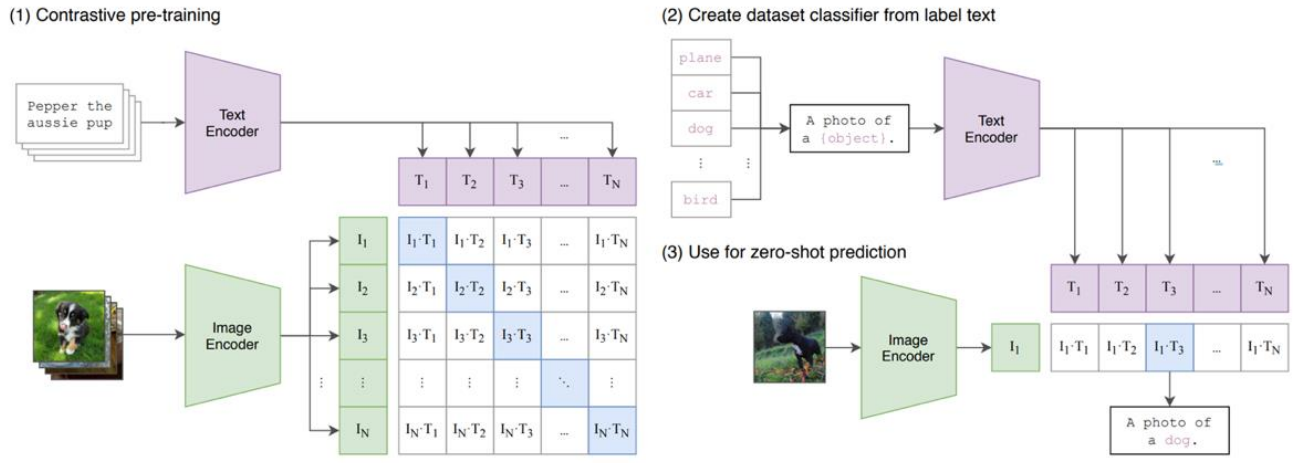


Figure 1: CLIP

CLIP2TV is a model that calculates similarities between corresponding videos and texts, basing its foundation on the framework of CLIP4Clip. CLIP2TV consists of text and video encoders. It tries to match video and caption embeddings with contrastive loss and cosine similarity because video frame features and caption features are projected into multi-modal space. As a result of these advancements, CLIP2TV achieves a state-of-the-art result of 52.9@R1 on the MSR-VTT dataset on the text-to-video retrieval task.

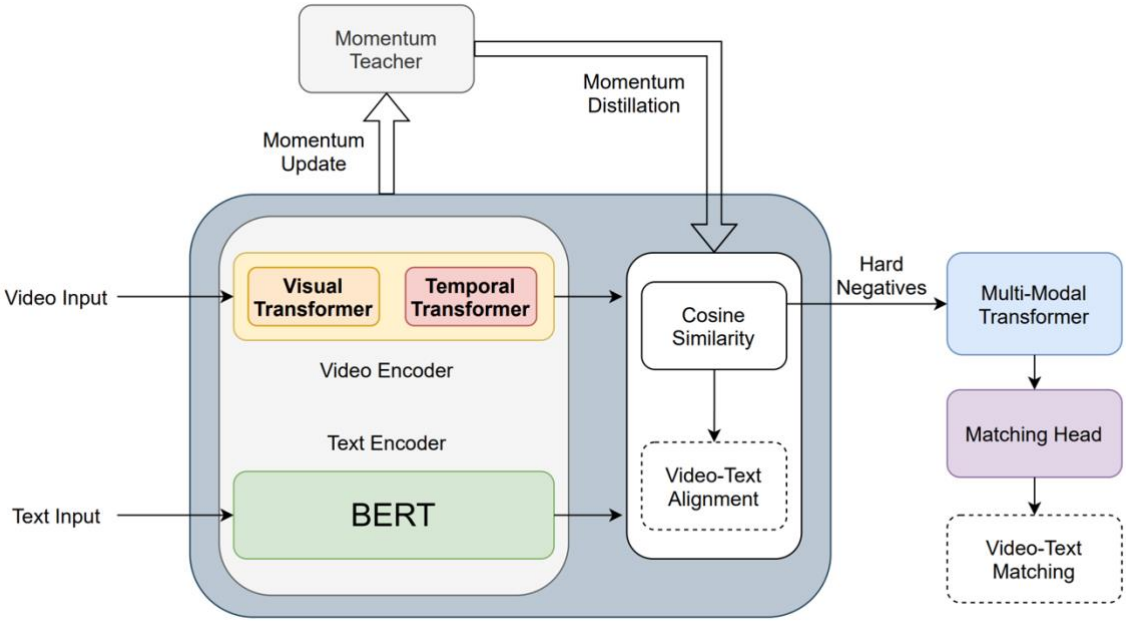


Figure 2: CLIP2TV

IMAGE CAPTIONING AND VISUAL QUESTION ANSWERING

Image captioning is the task of translating an image into a natural language description. It is usually a combination of computer vision and natural language processing methods. Image captioning models usually follow an encoder-decoder architecture wherein information about image features is encoded into a space that can be decoded and an image caption in the form of a text sequence can be generated.

The Simple Visual Language Model (SimVLM) [5] is a model that approaches the tasks of vision and language pre-training from a different perspective. It reduces training complexity by leveraging weak supervision as well as a prefix language modeling objective, which allows the model to be trained end-to-end. Its architecture is based on the fundamentals of the Transformer model and generally employs an encoder-decoder framework. The vision modality of SimVLM is significantly inspired by ViT. It maps its input raw image into a sequence of image patches, which becomes the input for the transformer block. To consider contextual information about image patches, a convolution block is applied instead of a trainable linear projection in ViT. The textual modality follows the de-facto standard of tokenizing input text sequences and learning textual embeddings. Also, positional embeddings are attached to both image patches and text sequences. SimVLM is trained on sufficiently large-scale text and image raw data crawled from the web. As a result, SimVLM achieves high results on a variety of vision and language tasks, e.g., image captioning and visual question answering.

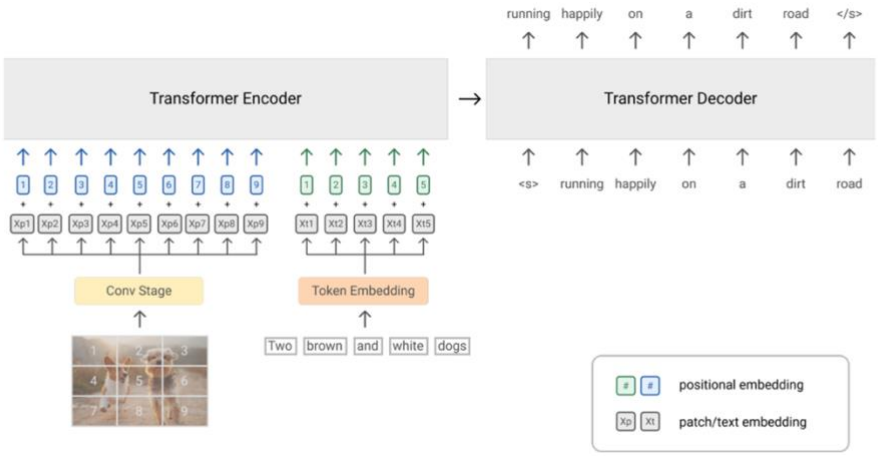


Figure 3: SimVLM

Transformer-based Optical Character Recognition (TrOCR) [6] is a model that tries to solve the problem of text recognition for textual document digitization, particularly converting handwritten or printed text into digital machine-encoded format. It is a Transformer-based encoder-decoder model that leverages pre-trained computer vision Transformer as an encoder to handle visual representations from images and pre-trained natural language processing Transformer as a decoder to generate corresponding textual information. TrOCR uses ViT as its visual encoder while using the original Transformer decoder for processing textual information. TrOCR achieves state-of-the-art results on both handwritten text recognition and printed tasks.

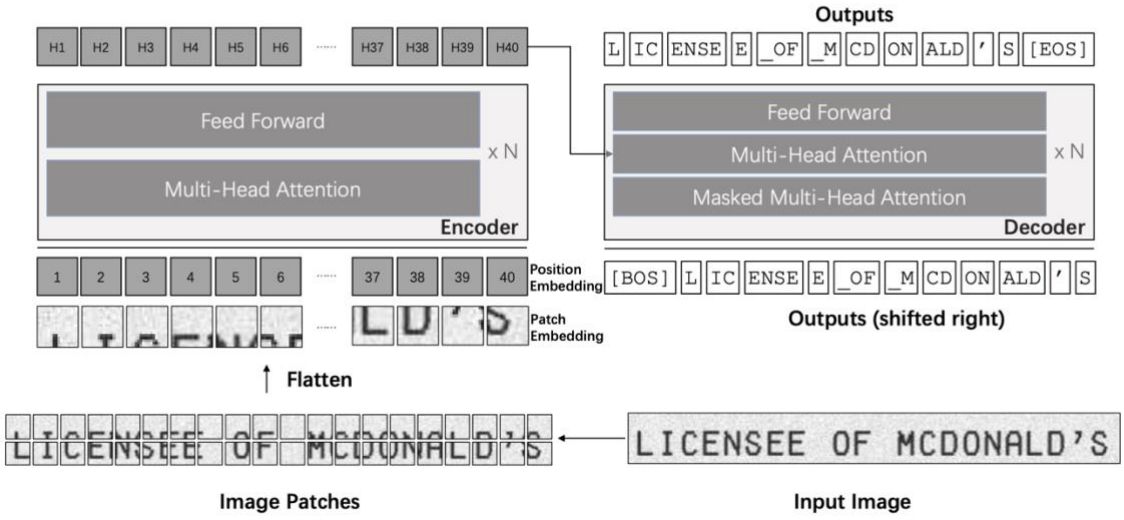


Figure 4: TrOCR

RELATED WORKS

VIDEO CAPTIONING

Video captioning aims to generate natural language descriptions automatically from the visual information in given videos. One of the influential models in video captioning is ORG-TRL [7]. It consists of 3 parts. The first part is an object relational graph (ORG) based encoder, which captures more detailed interaction features to enrich visual representation. The second part is the teacher-recommended learning (TRL) method to make full use of the successful external language model (ELM) to integrate the abundant linguistic knowledge into the caption model. The third part is teacher-enforced learning (TEL) which forces the caption model to learn the ground-truth word at each training step.

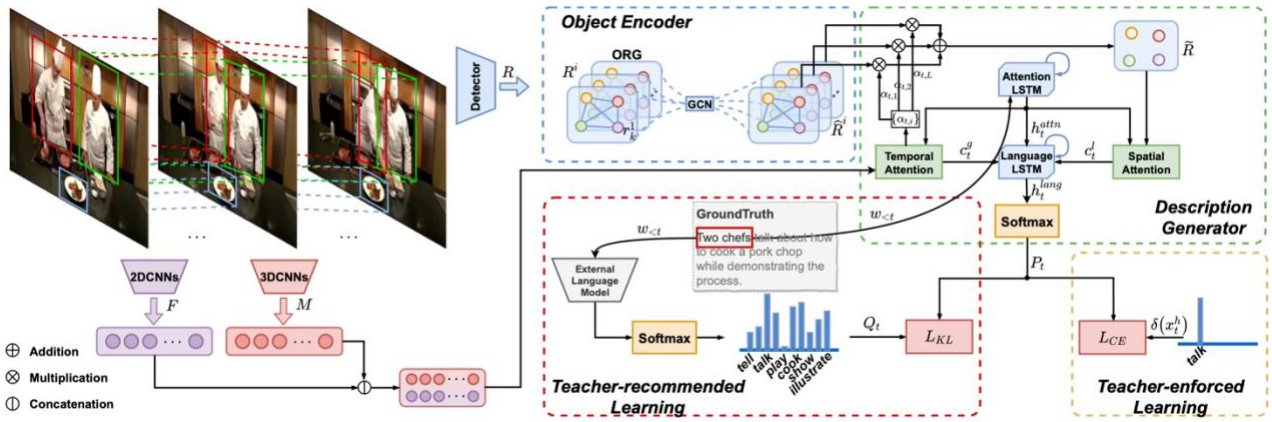


Figure 5: ORG-TRL

TEXT GENERATION FROM MULTIMODAL INPUTS

Here, text generation is done not only using video but also using speech, text, and audio. One of the influential models in this problem is VX2TEXT [8]. At first, each modality is converted into a set of language embeddings, then they are concatenated and fed into an encoder in an encoder-decoder architecture.

MODEL ARCHITECTURE

Because of our limited resources, we tried to solve the problem with pretrained models by keeping the Transformers encoder-decoder architecture. We take a transformer for video understanding and put it into the encoder part, and for the decoder part we put a language model transformer. In the decoder, after every attention layer, we put a newly initialized cross attention layer that takes KEYS and VALUES from the encoder and QUERY from the decoder. This is done to bring information from the encoder to the decoder. By training the encoder, we force the encoder to respond to the decoder's "textual" QUERY by transforming its "visual" KEYS and VALUES to "textual" KEYS and VALUES.

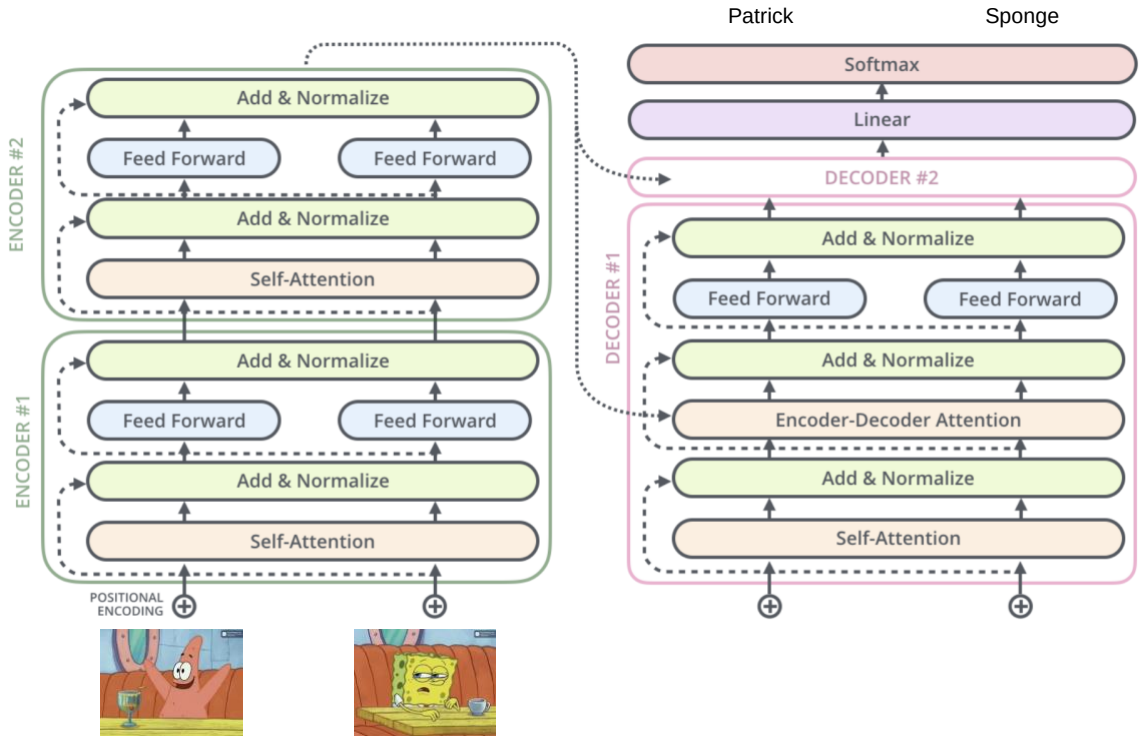


Figure 6: Vision-Language model

There are not many choices for the encoder part that were trained for video, so we decided to take a transformer that was trained for images and use the patch-embedding method. For that reason, ViT seems like a good choice here. We use video frames as patches and feed them into ViT. For the decoder part, we need a transformer that understands language well and can generate text well. There are many choices here like BERT, RoBERTa, and GPT.

Vision Transformer (ViT)

In general, convolutional neural network architectures are dominant in computer vision models, but Vision Transformer (ViT) [9] has demonstrated that the Transformer architecture can be comparable to state-of-the-art convolutional networks on computer vision tasks. Meanwhile, ViT can require less computational power for training. ViT model design closely adheres to the original Transformer architecture. It is designed to process a 2D image by dividing the image into a grid of square patches. Specifically, it reshapes an input image $x \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D image patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Here (H, W) represents the width and height dimensions of the input image, C is the number of image channels, (P, P) is the resolution of each image patch while $N = \frac{HW}{P^2}$ is the resulting number of patches. In the next step, each square patch is flattened by stacking its image channels. The output is mapped to the desired input dimension by introducing a linear projection with trainable units. The result of this trainable linear projection is called patch embeddings. Inspired by BERT's design, a learnable embedding is prepended to the patch embeddings. During both pre-training and fine-tuning, a classification head is attached to the state of this embedding, outputted by the Transformer encoder. In nature Transformer models are agnostic to the spatial and structural information of its inputs, thus it does not understand much about the relative location of square patches in the original image. However, it is important that the Vision Transformer model learns such relevant information about the input images. That is why learnable position embeddings are attached to each image patch, enabling the model to learn about the spatial and structural elements of input images and encoding them in the position embeddings.

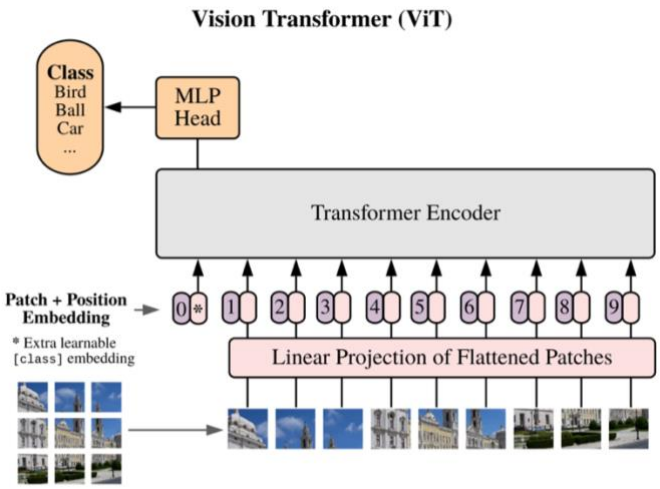


Figure 7: ViT

DECODER

BERT

BERT [10] is probably one of the most exciting developments in NLP in recent years. BERT stands for Bidirectional Encoder Representation of Transformer. It's the encoder part of the encoder-decoder transformer model, and it's also bidirectional in nature, which means that for any input it's able to learn dependencies from both left and right of any word. As a language model BERT can get away with the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective. The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. In addition to the masked language model, BERT also uses a "next sentence prediction" task that jointly pretrains text-pair representations.

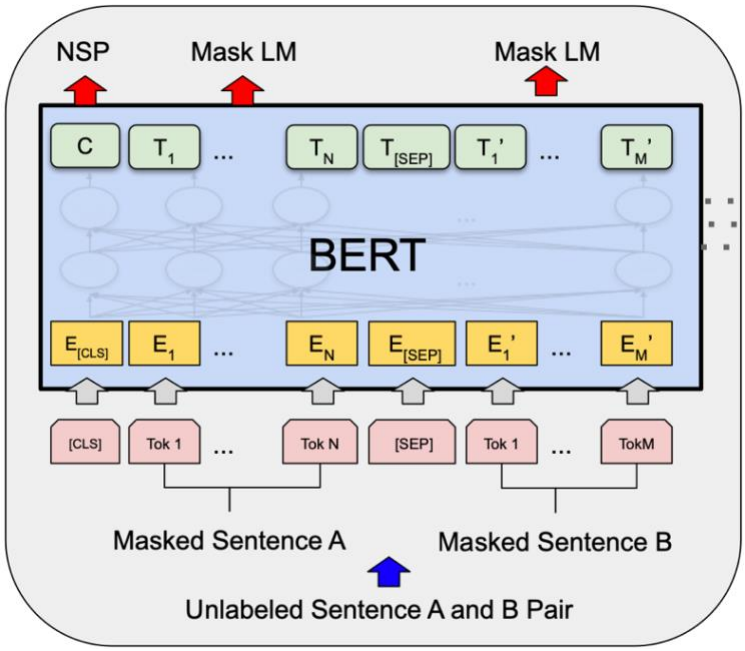


Figure 8: BERT

BERT can also be used for generating Natural Language but not very well. We can do it in an autoregressive manner, where the model always predicts the [MASK] that is at extreme right based on just left filled up context words.

RoBERTa

In RoBERTa [11], researchers found that BERT was significantly undertrained and proposed an improved recipe for training BERT models that can match or exceed the performance of all of the post-BERT methods. Here are their modifications: (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. RoBERTa produced state-of-the-art results on the widely used NLP benchmark, General Language Understanding Evaluation (GLUE).

GPT-2

GPT-2 [12] is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages(40GB). GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. Instead of BERT, The GPT-2 is built using transformer decoder blocks only. GPT-2 doesn't have a bidirectional nature. GPT-2 is better suited for text generation tasks because it is trained for that purpose instead of BERT like models that uses masked language model training objectives.

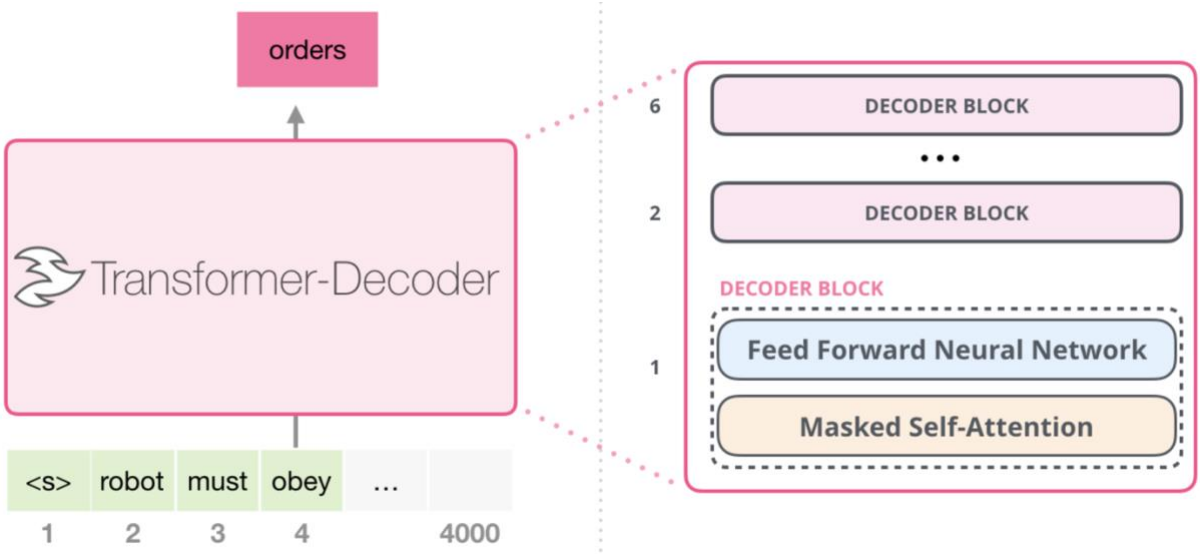


Figure 9: GPT

DATA PREPROCESSING

For feeding ViT, we uniformly sampled 49 frames from every video. Then we normalized every frame and resized them to 224x224. After that, we put together video frames side by side by making a square image and we resized the contracted image to 224x224. In the final input, each frame has a 32X32 size. That is the reason why 32X32 resolution patches are used in our ViT encoder to handle video data so that a sequence of video frames can be encoded as a sequence of image patches, empowering the model to learn video embeddings. For texts, we used the GPT2 tokenizer to tokenize them.

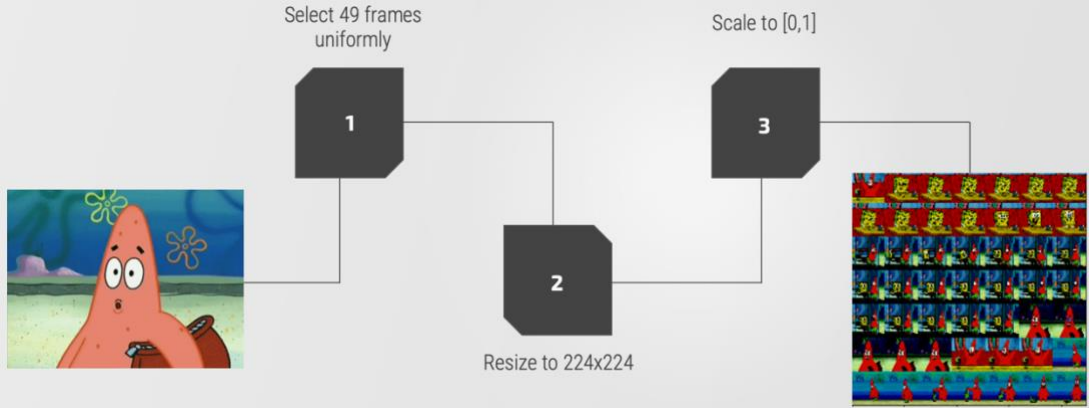


Figure 10: Data preprocessing

EXPERIMENTS

We have used ViT as an encoder for extracting visual embeddings from videos represented as an image consisting of video frames. We have used the base version of ViT with (224,224) resolution and have divided the image into a grid of (32,32) image patches.

For textual modality and generating text sequences we have leveraged several Transformer-based models, namely BERT, RoBERTa and GPT-2, using their base versions by default. As an initial experiment, we have used several fine-tuning variants and strategies and trained the model with each setup for 20 epochs.

We have used the pretrained versions of ViT and each transformer model and tried to fine-tune some of their layers. Specifically, we have fine-tuned either the last 10-20 layers of ViT or tried to train the whole model as an image encoder. On the other hand, we have experimented on the decoder model by first freezing the whole decoder model, fine-tuning 10-20 layers of either downstream, upstream layers or both ends as well as training the complete transformer model. We have demonstrated experimentally that RoBERTa performs better than BERT. As BERT-like models aren't created for text generation and they are only encoder transformers, we decided to use GPT-2 as a decoder and it works better than RoBERTa.

As we insert a cross attention layer in the decoder with random weights, we can't use GPT-2 to its full potential. We noticed this by experiment as well, when we were finetuning the decoder. After some epochs it started to predict only some words probably by overfitting our small dataset. That's why we decided to try another finetuning technique, by finetuning only the cross-attention layers and language model head in the decoder and the last 20 layers in the encoder. And as a result, it started to generate high quality meaningful texts.

After this we noticed that our model didn't catch context very well. This is because ViT was trained for images, and it understands patches as some part of images, not as a full image. From here we decided to finetune the whole encoder part. We also took ViT and GPT-2 with larger versions to catch more context and generate high quality texts.

RESULTS

As a result of experiments, the best model was the model with ViT-large (24 layers and 307M parameters) in the encoder and GPT-2-large (36 layers and 774M parameters) in the decoder. The number of our model parameters is about 1.3B. In this model, the encoder is fully finetuned and in the decoder, only the cross-attention layers and language model head are finetuned. The best model was chosen with early stopping by looking at the loss. The best model was gained at the 12075th iteration using a batch size of 16.

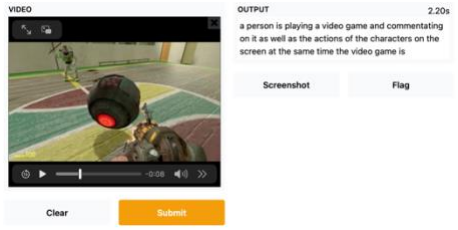
Our model has these results on MSR-VTT dataset compared with SOTA in video captioning.

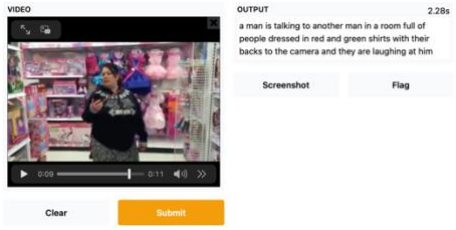
Model	*BLEU*	*METEOR*	*ROUGE-L*
ORG-TRL	43.6	28.8	62.1
OUR	5.6	4.8	12.5

ORG-TRL used huge and different datasets, FASTER-RCNN to detect objects, and calculated some features between them using a graph-based encoder. They also keep a big language teacher model, and 2 other pretrained big vision models to extract two types of features from video. Their model architecture is very complicated and the model is huge. Our model architecture is just a vanilla transformer architecture, not so many resources required, and can be usable. Here are some examples for different categories. On the left side are the best and on the right side are the worst examples according to METEOR.

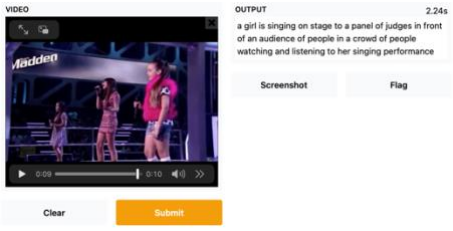
Best Worst

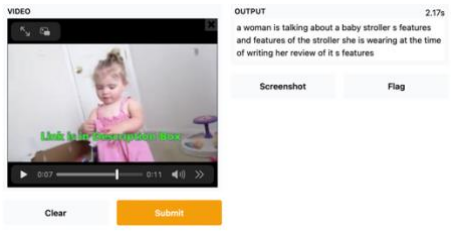
GAMING



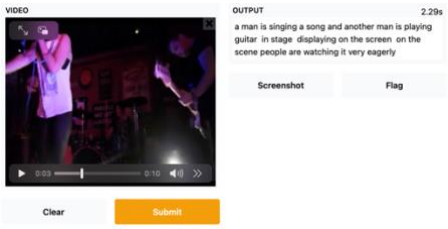


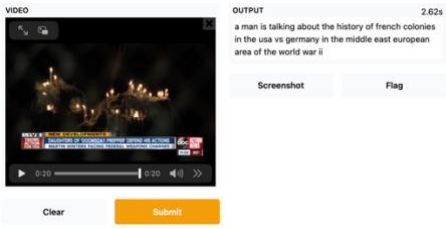
KIDS/FAMILY



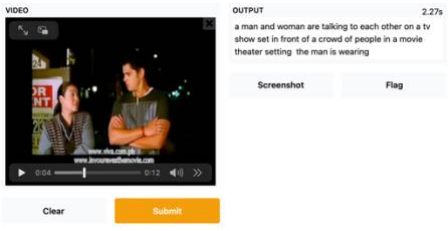


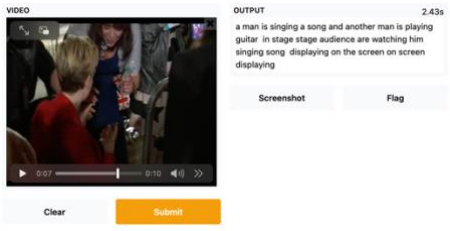
SPORT/ACTIONS





TV SHOWS





CONCLUSION AND SUGGESTIONS

In this paper, we have proposed a transformer model by connecting a vision encoder and a language model decoder. We have used a specific method of finetuning here by finetuning the whole encoder and cross-attention layers in the decoder to keep the advantages of the GPT transformer. As a result, we can generate high quality textual representation of video, which can be used in many applications, such as searching in video, helping blind people, in driver assistant and many other applications.

The results are good and practically usable, but many improvements can be applied. Because in ViT every frame appears as a 32x32 image, it can be small size to capture whole context of video. Other ViT models can be used, that have bigger input dimensions, and also before ViT we can use some pretrained convolutional layers to squeeze dimensions instead of resizing. We can also take ViT that trained with clip methodology, this is because its image representations will be close to textual representations and convergence might be more fast.

In the decoder part, the T5 model decoder can be taken, because T5 is a encoder-decoder transformer and it has cross attention layers in the decoder, and the new cross attention weights won't be initialized randomly at the beginning.

For future reference, audio can also be inputted to the model to get more information from the video.

BIBLIOGRAPHY

[1] Radford A. et al. Learning transferable visual models from natural language supervision //International Conference on Machine Learning. – PMLR, 2021. – С. 8748-8763.

[2] Luo H. et al. Clip4clip: An empirical study of clip for end to end video clip retrieval //arXiv preprint arXiv:2104.08860. – 2021.

[3] Fang H. et al. Clip2video: Mastering video-text retrieval via image clip //arXiv preprint arXiv:2106.11097. – 2021.

[4] Gao Z. et al. CLIP2TV: An Empirical Study on Transformer-based Methods for Video-Text Retrieval //arXiv preprint arXiv:2111.05610. – 2021.

[5] Wang Z. et al. Simvlm: Simple visual language model pretraining with weak supervision //arXiv preprint arXiv:2108.10904. – 2021.

[6] Li M. et al. Trocr: Transformer-based optical character recognition with pre-trained models //arXiv preprint arXiv:2109.10282. – 2021.

[7] Zhang Z. et al. Object relational graph with teacher-recommended learning for video captioning //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – С. 13278-13288.

[8] Lin X. et al. Vx2text: End-to-end learning of video-based text generation from multimodal inputs //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2021. – С. 7005-7015.

[9] Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale //arXiv preprint arXiv:2010.11929. – 2020.

[10] Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

[11] Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.

[12] Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – Т. 1. – №. 8. – С. 9.