The Multiple Input Segmenter (MIS)

Jichao Fang, Alphonse Mbinkar, Matthew Keisel

Background

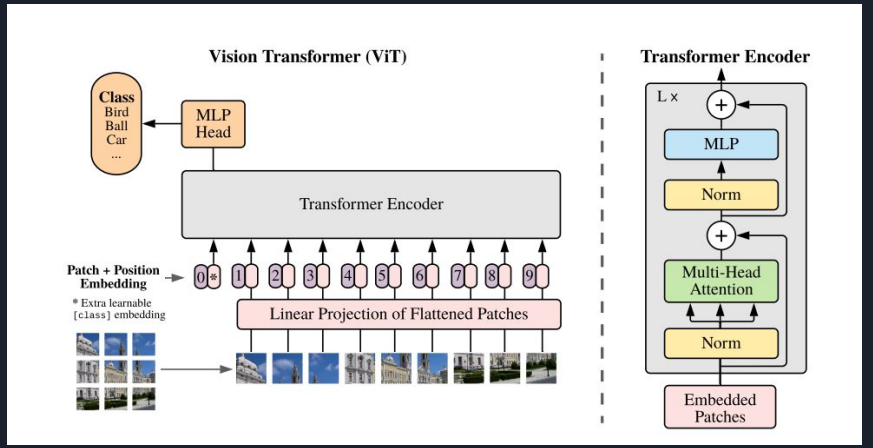
- An advancement in Artificial Intelligence and Natural Language Processors
- We have created an image segmenter that takes in 3 different inputs, creating a highly versatile and easy to use image segmenter for AI's such as chat gpt to use
- Text Input (optional)
- Audio input (optional)
- And image input (required)
- We will be examining the different models and methods we used to make this happen, along with the data we acquired

The purpose of our model

Our goal is to enhance SAM's functionalities by integrating text-based capabilities, enabling it to effectively read and comprehend paragraphs and academic papers. Additionally, we are planning to incorporate Google's advanced speech-to-text technology into SAM. By combining these two powerful features, SAM will become more versatile and user-friendly, catering to a wider range of user needs and preferences, whether they prefer typing or speaking. This upgrade will not only make SAM more efficient in handling various forms of input, but also enhance its potential as a tool for educational, professional, and personal use.

Vision Transformer, or ViT

- The process of formatting an image in a way that a computer is able to understand and comprehend.
- One of the main steps in image segmentation, image classification, object detection, and even image generation and enhancement
- Can also be used for facial recognition



Natural Language Processing (NLP)

- A field in artificial intelligence that revolves around translating the human language into something that a computer or an AI is able to understand.
- NLP and ViT often work together to combine textual and visual tasks, while the NLP processes the data, the ViT links this with the visual context allowing the AI to interpret the images and create a description based off the image.
- NLP is used in multiple different tasks such as text generation, text processing, reasoning, AI, and data extraction and mining.

Attention

- Attention is a filter used in encoding and transformers
- Its primary use is for feature enhancement
- It works by taking in a sentence or an image and weighing out the importance of the words or parts of the image
- Nouns are given a higher weight and relevance while filler words such as “and” or “or” are given a lesser weight.
- An attention filter is used to suppress noise and irrelevant information while enhancing or highlighting more relevant and important features.

Speech to text conversion

- Speech to text conversion is the process of taking an audio recording of someone speaking and transforming it into a readable text such as a string
- Once the audio is recorded, it is then converted and transcribed with the Google speech to text api which is used to extract the features of the text followed by writing and converting it into a string to be processed by the model

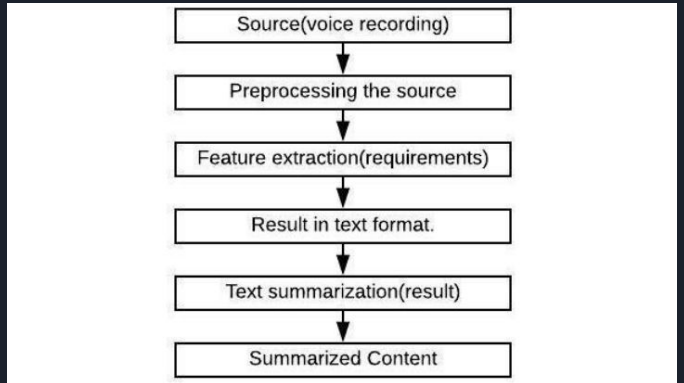


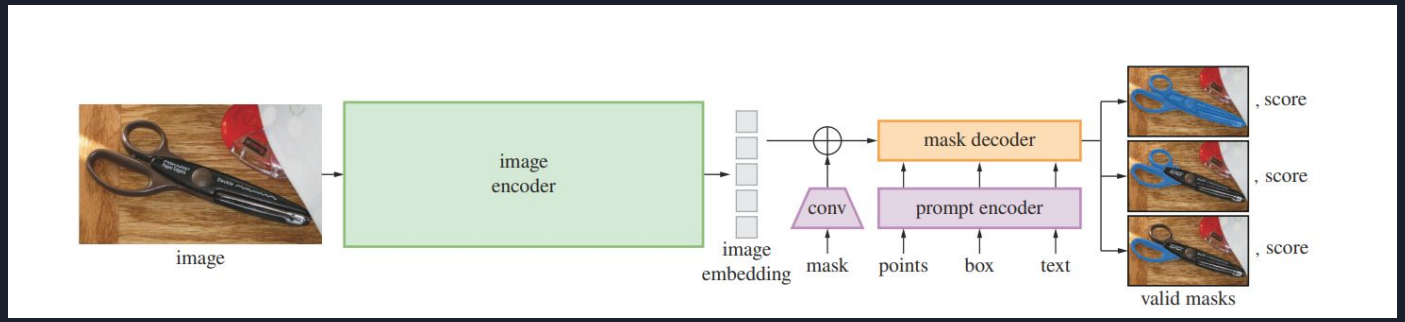
Figure 1. Speech recognition and text summarisation process flow

Segment Anything Model, or SAM

- The Image segmentation model, also the primary model we are trying to extend and improve upon
- This is the model we used to segment the images.
- Initially it worked only off of points and boxes as inputs
- SAM examines both the segmented data, the environment that is in, and the context clues of the environment regarding the segmented data. Segmented data can mean words in a paragraph, a video, audio, or time-series data. However, for now, SAM is only able to scan images.

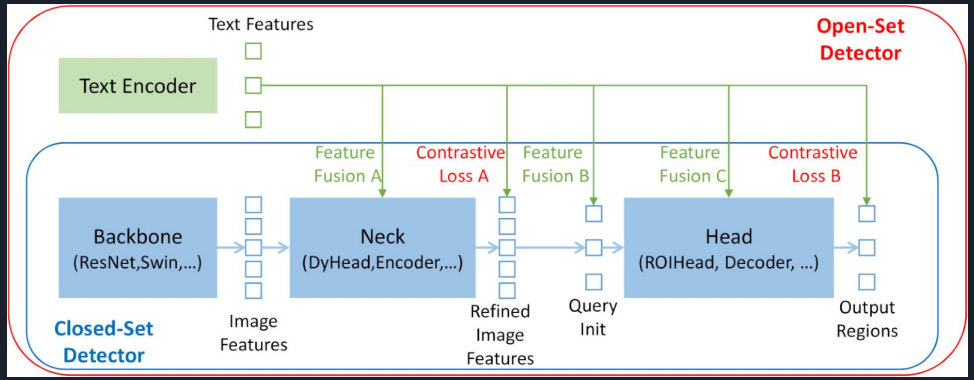
Sam continued

- SAM uses zero-shot prediction to find the borders and outline unknown objects in images that were not used previously as training data.
- SAM can also be used to segment specific pieces of information based off the inputs given.
- SAM uses an image and prompt encoder to take and convert human input into a format that is understood by the model, it then embeds the image, masks it, and uses a mask decoder to detect certain aspects of an image.



Grounding DINO

- Grounding DINO is described by its creators as an open-set object detector and is used to detect objects in an image based on text inputs
- It is a dual encoder single decoder with takes in an image and text pair to create output boxes around specified objects in an image.
- This is the model we used to allow SAM to take in text inputs



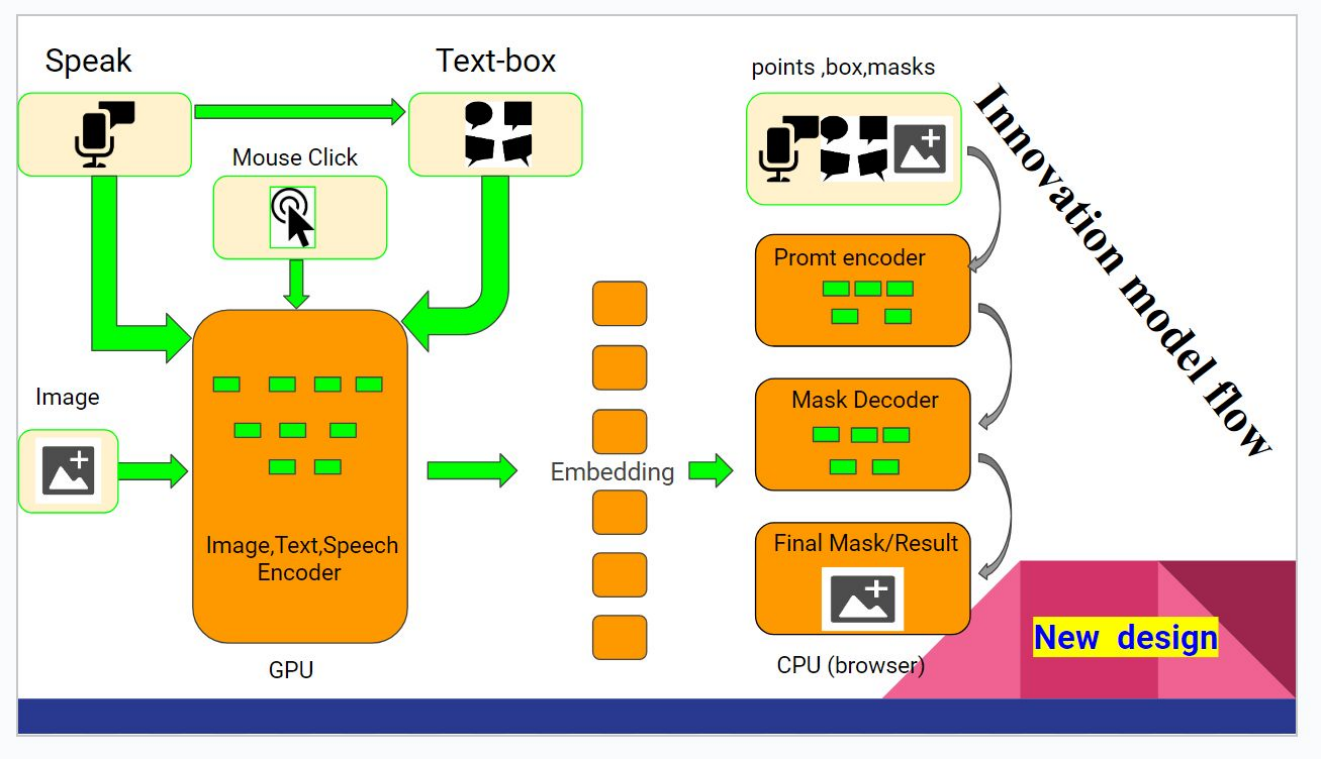
PartImageNet

- PartImageNet is a dataset that consists of a large quantity of high resolution images
- We used this dataset when testing our model to determine how well it performs object recognition and image segmentation
- It comes with around 24000 annotated images to be tested, we tested 1988 images

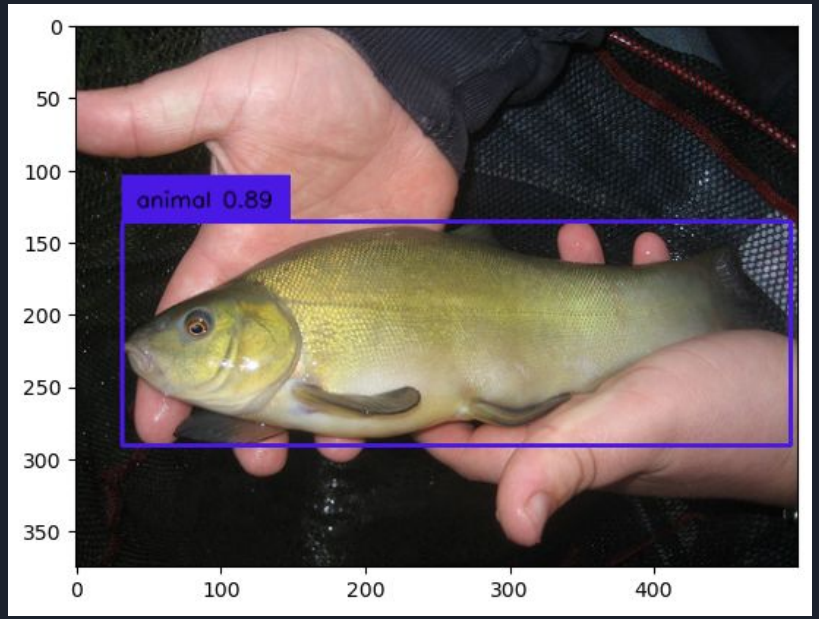
Google Speech to Text API

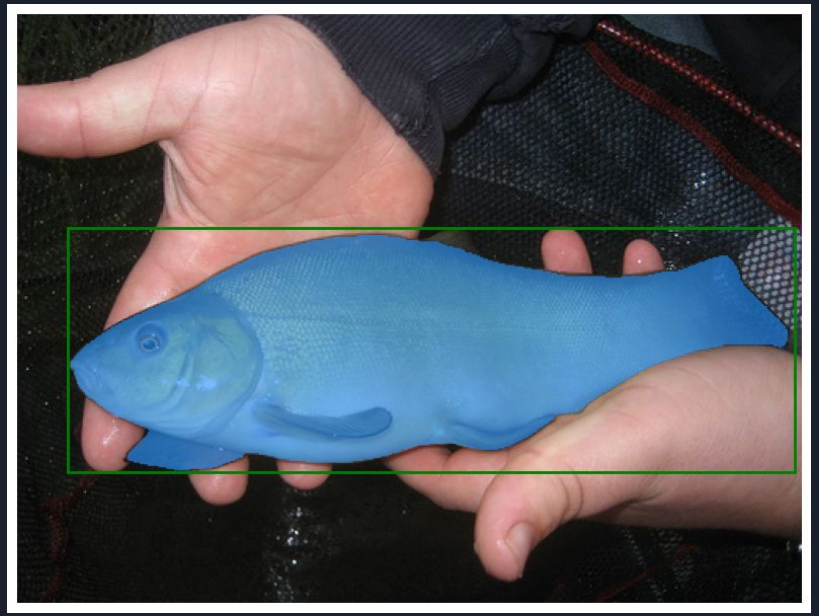
- Google's speech to text is a widely used audio transcriber which is what we used to convert an audio file to a string which can then be encoded and interpreted by our model
- This is what we used in our model to translate speech to text by first recording an audio clip of us speaking, followed by using Google’s API to convert it to a string that can be sent through a text encoder.

Our model layout



Just grounding Dino vs grounding dino + sam

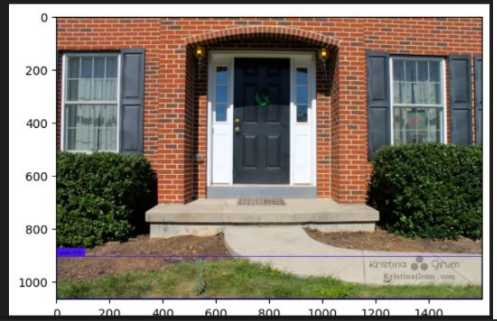


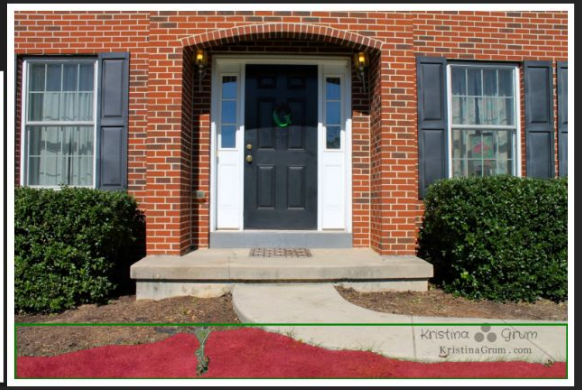


Multiple instance segmentation

Examples

```
Output for keyword 'grass':
tensor([[0.5001, 0.9246, 0.9970, 0.1480]])
tensor([0.6950])
['grass']
```

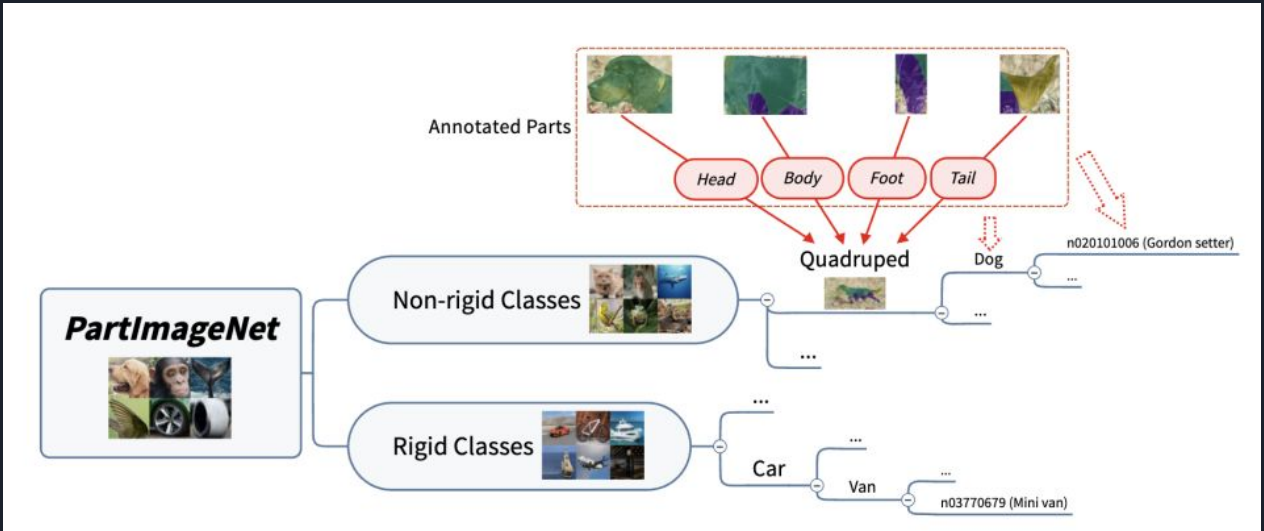




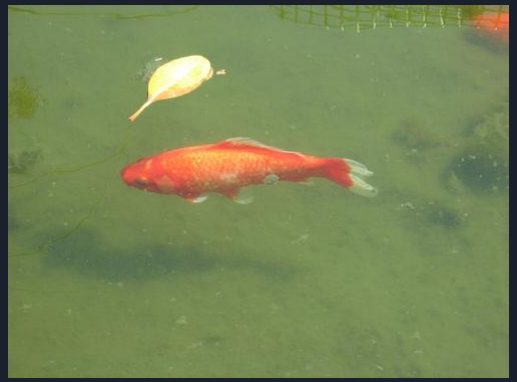

```
Output for keyword 'door':
tensor([[0.4816, 0.3708, 0.1362, 0.4500]])
tensor([0.6642])
['door']
```

Evaluation

- Dataset - PartImageNet, contains both rigid and non rigid object classes, with each object being annotated



Evaluation



Evaluation

- Evaluation metric - Intersection of Union (IoU)
- We tested on 1988 images from PartImageNet test data
- We prompt the model simply with “animal” to see how it segments the image
- We calculated IoU of each image to see the performance
- Out of 1988 images, 36 have IoU less than 0.2, which we consider as “missed” segmentation
- Over 1952 successfully segmented images, mIoU reached 0.89, a reasonably high score for instance segmentation

Resources

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv. https://arxiv.org/abs/2303.05499

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. arXiv. https://arxiv.org/pdf/2304.02643.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. Retrieved from https://arxiv.org/pdf/1706.03762.pdf

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Retrieved from https://arxiv.org/abs/2010.11929

Schopf, T., Arabi, K., & Matthes, F. (2023). Exploring the Landscape of Natural Language Processing Research. arXiv. https://arxiv.org/abs/2307.10652

A, Vinnarasu & Jose, Deepa. (2019). Speech to text conversion and summarization for effective understanding and documentation. International Journal of Electrical and Computer Engineering (IJECE). 9. 3642. 10.11591/ijece.v9i5.pp3642-3648.

He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.-N., Liu, S., Yang, C., Yu, Q., & Yuille, A. (2022). PartImageNet: A Large, High-Quality Dataset of Parts. arXiv. https://arxiv.org/abs/2112.00933