Within the realm of deep learning, I am especially curious and excited about foundational models for computer vision tasks. The foundational model term is often used to describe LLMs. The foundational model is created by learning via a gigantic corpus of data (the legality of using all of the data in the corpus being ignored for the moment). These foundational models learn the broad understanding of language structure. From there, these models are often fine tuned on a specific corpus (such as proprietary documents a company owns).

The concept of a foundational model can be applied to computer vision tasks as well. [The Segment Anything Model (SAM), developed by Meta AI](https://arxiv.org/pdf/2304.02643.pdf), is one such foundational model. In this case, it was created for image segmentation tasks. At its heart, SAM is an image encoder computes image embeddings while a prompt encoder embeds prompts. These two pipes of information are then combined via a mask decoder which generates the segmentation masks. Through this work, a vast number of diverse images can be segmented into component pieces with little effort. To put this in perspective, instead of having to iteratively select different radii for the Hough Transform, or carefully run through image preprocessing to ensure the Waterfall method found the correct segmentations, this foundational model could provide similar results with no additional forethought to the problem.

Additionally, SAM can then be fine tuned much like foundational models for LLMs are fine tuned. This allows for specific features which may not have been learned from the SAM training corpus, likely proprietary information or highly specific images like medical data, to be properly identified. Through their work, the Meta AI team found that SAM was able to perform well on the following zero-shot transfer experiments: 1) Single Point Valid Mask Evaluation, 2) Edge Detection, 3) Object Proposals, 4) Instance Segmentation, and 5) Text-to-Mask.

I believe that with the speed in which the machine learning community is able to create tools, such as SAM, and freely share them means that we will continue to see large leaps forward in the field. Foundational language models are trained separately and each has their own nuances as given by the language itself. This can make it difficult for one LLM to integrate others into it. I can image that with images, it may be easier to bundle many foundational models to create a ‘mega-model’. Such a system would be able to not only generalize, as in the case of SAM, but also make use of other tools (like [FocalClick, Chen et al., 2022](https://arxiv.org/pdf/2204.02574.pdf) or [YOLO, Redmon et al., 2015](https://arxiv.org/pdf/1506.02640.pdf)) which would allow for more precise analysis.

References:

Segment Anything: <https://arxiv.org/abs/2304.02643>

Focal Click: Towards Practical Interactive Image Segmentation: <https://arxiv.org/abs/2204.02574>

You Only Look Once (YOLO): Unified, Real-Time Object Detection: <https://arxiv.org/abs/1506.02640>