**AML Project Proposal**

**Team members:**

Andrew Fang (fafaaf61@vt.edu)

Kanad Naleshwarkar ([kanadn@vt.edu](mailto:kanadn@vt.edu)) (coordinator)

Sanket Bhujbal ([sanketb@vt.edu](mailto:sanketb@vt.edu))

**Project Topic:**

Zero-shot learning is an emerging area of research in machine learning that aims to enable machines to learn from new concepts and tasks without the need for explicit training data. One of the most challenging and promising applications of zero-shot learning is cross-modal retrieval, such as text-image and image-text. In this project, we plan to evaluate the performance of a newly proposed model called Distilled and Hard-negative Training (DiHT) on Google’s Conceptual Captions dataset and hope to discover any possible limitations as well as show future developments that can be made to the model for more accurate results.

Traditional machine learning models require large amounts of labeled data to learn, which can be a bottleneck in many real-world applications where data is scarce or expensive to obtain. Zero-shot learning aims to overcome this limitation by leveraging prior knowledge about the relationships between concepts and tasks and using it to generalize to new tasks and concepts. Thus, we wanted to explore recent advances in zero-shot learning that are outperforming benchmarks like CLIP.

The Distilled and Hard-negative Training (DiHT) approach has been recently proposed by Radenovic and others at Meta AI that puts forth three novel methods for improving contrastive pre-training: i) Complexity, Action, and Text-spotting (CAT) that is a filtering strategy to select only informative text-image pairs from noisy datasets ii) Concept Distillation that leverages pre-trained vision models and iii) Multimodal alignment with hard negatives. This approach is based on the existing pre-training paradigm for multimodal models, modifying the pre-training process by adding a filter to reduce noise, improving the “warm-start” of large-scale pre-training by training a linear classifier over an image encoder to predict distilled concepts, and changing from InfoNCS loss to model-based importance sampling techniques. This model has been evaluated on datasets like ImageNet1K, COCO and Flickr for T2I and I2T retrieval tasks and found to be outperforming benchmark models like CLIP and Open-CLIP.

Hence, we wanted to perform an evaluation on yet another dataset, Google’s Conceptual Captions, which provides ~3.3M images annotated with captions and represents a wider variety of styles. We are planning to divide this project into two phases, the first goal will be to apply the DiHT model to the GCC dataset and evaluate the performance and the second, we will explore possible approaches that can increase the performance of this implementation.

We are hoping to explore and understand the GCC dataset and complete the first phase by March end. This will give us enough time to perform a brief exploratory study to search for approaches that can help us improve the performance of our implementation. We are aiming to complete this model improvement phase by mid-April so that we have some time in hand as a buffer and also for documenting all of the things that we learned along the way which we believe will be very useful.