Discussion of Results

In this section, we briefly discuss the performance on all the tasks that we use to evaluate the proposed method. We compare the performance on natural language processing in Figure 1, on computer vision in Figure 2, and on vision-language tasks in Figure 3.

Overall, the proposed method does not outperform the methods to which we compare. However, recall that our approach was designed to be a **proof-of-concept** for the feasibility of creating cheap and efficient vision-language models. Since we achieve reasonable performance across all tasks, outperforming well-known baselines on tasks like WNLI and COCO image retrieval, we consider our approach to be successful.

It is further not realistic to expect the proposed method to outperform the state-of-the-art methods, as they are larger in every aspect: Parameters, data, and compute. A good impression on where S-SMKE ranks among the vision-language models we repeatedly compare to can be obtained from Figure 4.

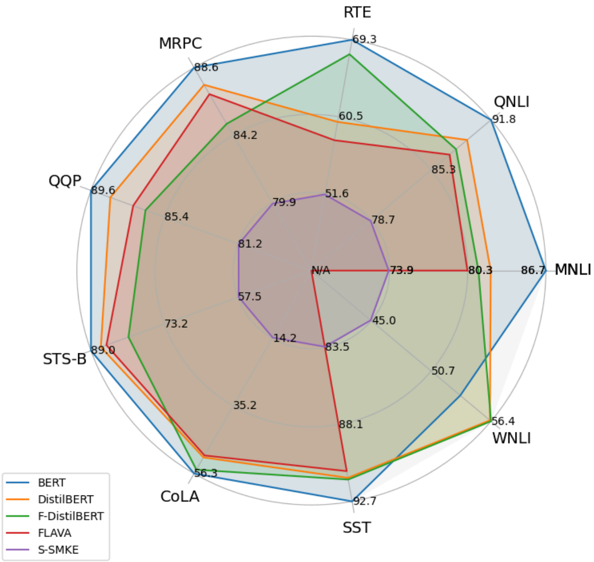

Figure 1:

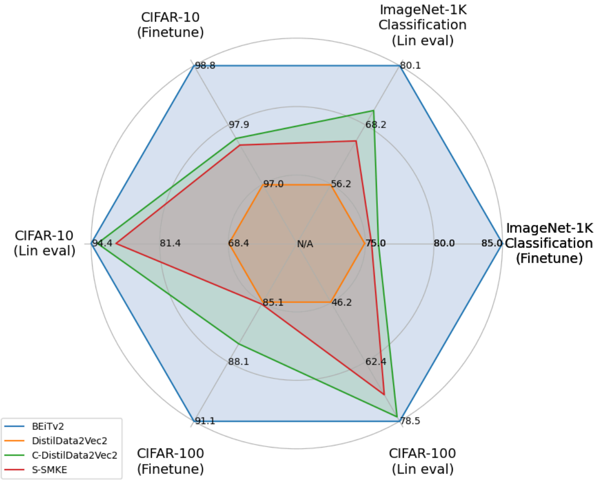

Figure 2:

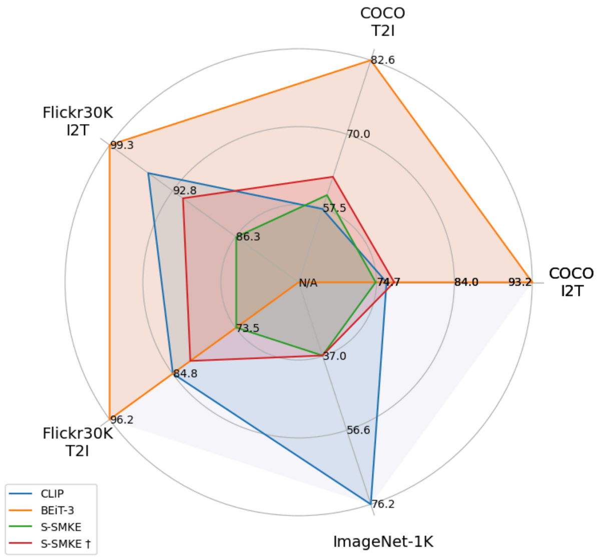

Figure 3:

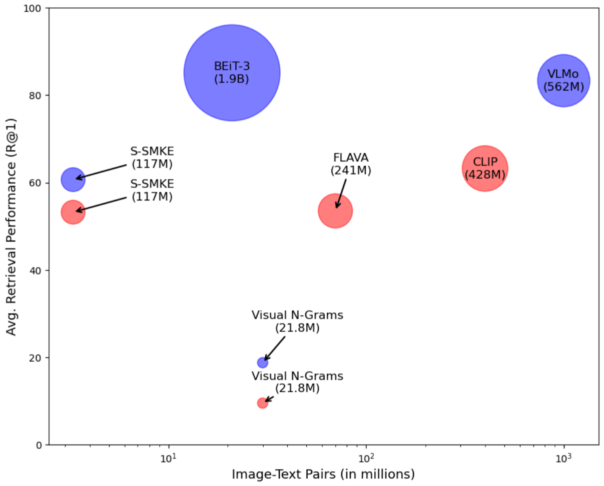

Figure 4: Overview of vision-language model landscape. Bubble sizes represent the number of parameters, also shown in parentheses next to the model name. The number of image-text pairs used is in log scale, and the R@1 retrieval performance is the average of the R@1 scores on COCO and Flickr30k.