HW 15

BLIP (Bootstrapped Language-Image Pre-training)[1,2] is a model that can process both images and text to solve tasks such as image captioning, visual question answering, image text matching, and zero-shot image classification. BLIP uses a vision transformer (ViT) to process and extract features from images., while also employing a transformer-based language model (similar to BERT or GPT) to process and understand textual input. It then aligns image and text features in a shared embedding space that allows a joint understanding of media from both modalities. In this homework, we will explore how BLIP can be used for Visual Question Answering.

Part A

- Download ten images of different objects. Three of the images need to contain cars.
- Import BLIP from the Transformers library.
- Ask BLIP a question about “What is contained in this image?” for one of the images (pick one randomly).

Part B

- Define the following query: “Is this an image of a car ?”
- Process the text query and images with the BLIP model.
- Compute Relevance Scores between the query and each image
- Sort images based on relevance scores
- Display the sorted list of images
- Open-ended question: In your comments, discuss how you would perform a performance evaluation of the BLIP model for the task described above.

References:

1. J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," arXiv preprint arXiv:2201.12086, 2022.
2. https://huggingface.co/docs/transformers/en/model_doc/blip