Homework 4 Report: CV Winter 2025

Vinayak Agrawal (2022574)

April 20, 2025

1 Question 1: CLIP vs. CLIPS Comparative Analysis

This section details the comparison between the original CLIP model and the enhanced CLIPS framework based on their ability to score the similarity between an image (depicting a human and a dog) and various textual descriptions.

1.1 Installation and Loading (CLIP)

The baseline CLIP model was accessed using the Hugging Face `transformers` library. Dependencies were installed as required by the library.

1.2 Pre-trained Weights (CLIP)

The pre-trained weights for the `openai/clip-vit-base-patch32` model were loaded using the `CLIPModel.from_pretrained` method from the `transformers` library.

1.3 Similarity Scores (CLIP)

Using the loaded CLIP model and processor, similarity scores (logits) were computed between the sample image (human and dog) and 10 different textual descriptions. The results are presented in Table 1.

1.4 Installation and Loading (CLIPS)

Dependencies for CLIPS were installed, primarily `open_clip_torch`. The CLIPS model was loaded directly from the Hugging Face Hub using the identifier `hf-hub:UCSC-VLAA/ViT-L-14-CLIPS-Recap-DataComp-1B` leveraging the `open_clip` library's functionality.

1.5 Pre-trained Weights (CLIPS)

The pre-trained weights corresponding to the `ViT-L-14-CLIPS-Recap-DataComp-1B` model were automatically downloaded and loaded via the `open_clip.create_model_from_pretrained` function.

1.6 Similarity Scores (CLIPS)

Similarity scores (logits) were computed between the same sample image and the 10 textual descriptions using the loaded CLIPS model, processor, and tokenizer. The results are presented alongside the CLIP scores in Table 1.

Table 1: Comparison of CLIP and CLIPS Similarity Scores (Logits)

Description	Baseline CLIP	Enhanced CLIPS
A person and a dog sitting on the grass.	23.7514	5.0195
A human petting their golden retriever.	24.2050	8.3750
A man's best friend.	26.5426	23.2969
Outdoor scene with a canine and its owner.	24.2815	10.2266
A happy dog next to a person.	26.2364	16.2656
Two cats playing indoors.	24.3914	9.9062
A landscape photo of mountains.	14.0426	3.3008
A portrait of a dog.	22.4430	12.8047
Someone relaxing outside with their pet.	24.7869	8.8281
A close-up of a dog's face.	15.2554	8.8516

1.7 Comments on Results

The results presented in Table 1 highlight key differences between the baseline CLIP (`ViT-Base/32`) and the enhanced CLIPS (`ViT-Large/14`) models:

- **Score Magnitude:** The absolute logit scores differ significantly between the models (CLIP scores range roughly 14-27, while CLIPS scores range 3-24). This is expected due to differences in model architecture, training data (CLIPS uses synthetic captions), and the learned temperature scaling factor. Direct comparison of absolute values is not meaningful.

- **Relative Ranking:** Both models assign the highest scores to captions that are semantically relevant to a human-dog image, such as "A man's best friend." and "A happy dog next to a person.".

- **Distractor Rejection:** This is where CLIPS shows a marked improvement. Baseline CLIP assigns a surprisingly high score (24.39) to the clearly irrelevant caption "Two cats playing indoors.", indicating potential confusion or noise learned from web data. CLIPS correctly assigns this caption a much lower score (9.91). Similarly, CLIPS assigns a very low score (3.30) to "A landscape photo of mountains.", whereas baseline CLIP's score (14.04) is higher, although still low relative to relevant captions.

- **Model Capacity and Training:** The superior performance of CLIPS in distinguishing relevant from irrelevant captions can be attributed to both its larger ViT-L/14 backbone (providing greater representational capacity) and its training methodology involving synthetic, potentially cleaner captions, which likely reduces noise and improves fine-grained understanding compared to CLIP's reliance on raw web data.

In conclusion, while both models capture the core theme, the enhanced CLIPS model demonstrates significantly better zero-shot performance in terms of accurately assessing semantic relevance and rejecting clearly incorrect textual descriptions for the given image, likely due to its architecture and refined training strategy.

2 Question 2: Visual Question Answering (BLIP)

This section evaluates the BLIP model's capability for visual question answering using the same human and dog image from Question 1.

2.1 Installation and Loading

The BLIP model for VQA was accessed using the Hugging Face `transformers` library. The specific model used was `Salesforce/blip-vqa-base`. Dependencies were installed as required.

2.2 Question 1: Dog Location

Question: ”Where is the dog present in the image?”
Generated Answer: `in man ’ s arms`

2.3 Question 2: Man Location

Question: ”Where is the man present in the image?”
Generated Answer: `living room`

2.4 Comments on Output and Accuracy

The outputs generated by the BLIP VQA model provide insights into its understanding of the (implied) image content and spatial relationships:

- **Dog Location Answer (`in man ’ s arms`):** This answer suggests a close interaction between the human and the dog. Without seeing the specific image used, it’s plausible if the man is holding the dog. However, it lacks broader context (e.g., ”on the grass”, ”outdoors”). Its accuracy depends heavily on the visual input.

- **Man Location Answer (`living room`):** This answer seems less likely to be accurate for a typical image depicting a person and a dog outdoors (which many of the captions in Q1 implied). It suggests the model might be hallucinating a scene context (”living room”) that is not present or misinterpreting the background. This indicates a potential limitation in scene understanding or a bias towards indoor environments.

Overall, the BLIP VQA model provided answers that are grammatically coherent but potentially factually inaccurate or lacking specific detail depending on the actual image. The answer for the man’s location seems particularly suspect and might indicate a failure in grounding the question to the specific visual evidence.

3 Question 3: BLIP Captioning vs. CLIP/CLIPS Evaluation

This section explores using BLIP for image captioning and subsequently evaluating the generated captions using both baseline CLIP and enhanced CLIPS for semantic alignment with the source images.

3.1 Loading BLIP Captioning Model

The BLIP model pre-trained for image captioning (`Salesforce/blip-image-captioning-base`) was loaded using the Hugging Face `transformers` library.

3.2 Generated Captions

Captions were generated using the loaded BLIP model for 10 sample images. The generated captions are listed in Table 2.

3.3 CLIP Evaluation of Captions

The baseline CLIP model (`openai/clip-vit-base-patch32`) was used to compute the similarity score (logit) between each sample image and its corresponding BLIP-generated caption. These scores reflect how well the caption aligns with the image according to baseline CLIP. The scores are presented in Table 2.

3.4 CLIPS Evaluation of Captions

The enhanced CLIPS model (`UCSC-VLAA/ViT-L-14-CLIPS-Recap-DataComp-1B`) was also used to compute the similarity score (logit) between each image and its BLIP-generated caption. These scores provide an alternative measure of semantic alignment from a potentially more robust model. The scores are presented in Table 2.

Table 2: Evaluation of BLIP-Generated Captions using CLIP and CLIPS

Image Filename	Generated BLIP Caption	CLIP Score	CLIPS Score
ILSVRC2012_test_00000003.jpg	a small dog walking on a green carpet	31.5660	27.0156
ILSVRC2012_test_00000004.jpg	a small dog running across a green field	32.7026	27.7500
ILSVRC2012_test_00000018.jpg	a family sitting in a pool with a towel	31.3365	20.6562
ILSVRC2012_test_00000019.jpg	a small bird sitting on a plant	28.9383	24.0625
ILSVRC2012_test_00000022.jpg	a small dog standing on a stone ledge	31.0347	21.0938
ILSVRC2012_test_00000023.jpg	a man riding a bike down a wet street	30.8345	23.0312
ILSVRC2012_test_00000025.jpg	a brown butterfly sitting on a green plant	28.9163	24.0625
ILSVRC2012_test_00000026.jpg	a man in a suit and tie sitting on a couch	28.8953	17.7812
ILSVRC2012_test_00000030.jpg	a duck drinking water from a pond	30.5252	23.8906
ILSVRC2012_test_00000034.jpg	a coffee machine with two cups on it	27.9613	22.7500

Analysis Note: Observing Table 2, the BLIP captions appear generally relevant to the likely content of the images (based on filenames). Both CLIP and CLIPS assign relatively high similarity scores, suggesting they perceive reasonable alignment between the generated text and the (unseen) images. As in Q1, the absolute score scales differ, but the scores provide a quantitative measure of this alignment. For instance, the captions for the dog images (0003, 0004, 0022) receive high scores from both models. The lowest scores are associated with the coffee machine (0034) and the man in the suit (0026), perhaps indicating slightly less typical or more complex scenes for the models to align.

3.5 Metrics for Quantifying CLIP/BLIP Alignment

Several metrics can be used to quantify the alignment between image-text pairs, which is relevant when evaluating how well BLIP's generated captions align with an image according to CLIP/CLIPS, or how well CLIP/CLIPS scores align with other evaluations of BLIP's captions.

1. **Direct Similarity Score (Logits/Cosine Similarity):**

   - *What it is:* The raw output score from CLIP/CLIPS, representing scaled cosine similarity between image and text embeddings.
   - *Usefulness:* Directly measures alignment strength perceived by the specific CLIP/CLIPS model. Good for comparing different captions for the *same* image or comparing CLIP vs. CLIPS on the *same* image-caption pair. Provides a continuous alignment score.
   - *Example:* Using the scores in Table 2 to say that, according to CLIP, the caption for image ‘0004.jpg‘ is better aligned (32.70) than the caption for ‘0034.jpg‘ (27.96).

2. **Ranking Metrics (Recall@K, Mean Reciprocal Rank - MRR):**
   - *What they are:* Evaluate alignment in a retrieval context. Given an image and multiple captions (the BLIP caption + distractors), how highly is the BLIP caption ranked by CLIP/CLIPS score?
   - *Usefulness:* Assesses if the BLIP caption is discriminatively better than alternatives according to the similarity model. More robust to absolute score scale variations.
   - *Example:* For image ‘0003.jpg‘, calculate CLIP scores for ”a small dog walking on a green carpet” (BLIP’s output), ”a cat walking on grass”, and ”a dog swimming”. If the BLIP caption ranks first, it indicates good discriminative alignment. Calculate Recall@1 across many images.

3. **Correlation Coefficients (Spearman, Pearson):**
   - *What they are:* Measure statistical correlation between two sets of scores (e.g., CLIP scores vs. human ratings).
   - *Usefulness:* Assess if CLIP/CLIPS scores for BLIP captions align with human judgments of caption quality/relevance across a dataset. Can also compare if CLIP and CLIPS scores correlate highly with each other for the same set of BLIP captions.
   - *Example:* Collect human ratings (1-5) for the BLIP captions in Table 2. Calculate Spearman correlation between these ratings and the CLIP scores to see if CLIP’s assessment matches human perception for these specific generated captions.

4. **Text Generation Metrics (BLEU, ROUGE, METEOR, CIDEr - Contextual):**
   - *What they are:* Standard metrics comparing generated text (BLIP caption) to human-written reference captions.
   - *Usefulness (Indirect Alignment):* Primarily evaluate the linguistic quality of the BLIP caption itself. High scores on these metrics *combined* with high CLIP/CLIPS scores suggest strong overall performance (good visual grounding and linguistically sound). They don’t directly use the CLIP score but provide essential context.
   - *Example:* If reference captions exist for the sample images, calculate the CIDEr score for each BLIP caption. Compare images where BLIP achieves high CIDEr and high CLIP score versus cases where one is high and the other is low.

The choice of metric depends on the specific evaluation goal: direct alignment strength (Similarity Score), discriminative ability (Ranking), correlation with external judgments (Correlation), or linguistic quality (Text Generation Metrics). Often, a combination of metrics provides the most comprehensive evaluation.

4 Question 4: Referring Image Segmentation (CLIPSeg Adaptation)

This section addresses the task of Referring Image Segmentation (RIS). The original goal was to use the LAVT model ([?] - Note: Add actual citation if available). However, persistent dependency installation failures were encountered, specifically with older required versions of MMLab libraries (`mmcv-full==1.3.12`, `mmsegmentation==0.17.0`) in the available Kaggle environment. Multiple installation strategies, including using specific versions, latest versions via 'mim', and attempting builds, were unsuccessful.

Due to these technical challenges, the CLIPSeg model [?] (`CIDAS/clipseg-rd64-refined`), a readily available model capable of performing RIS, was used as a substitute to demonstrate the core concepts of the task.

4.1 Installation and Loading (CLIPSeg)

Dependencies for CLIPSeg, primarily the Hugging Face `transformers` library, were installed using `pip`. The CLIPSeg processor (`CLIPSegProcessor`) and model (`CLIPSegForImageSegmentation`) were loaded from the Hugging Face Hub using the identifier `CIDAS/clipseg-rd64-refined`. The model was successfully loaded onto the available CUDA device.

4.2 Segmentation with Provided References

The task requires segmenting an object based on a natural language description. For each sample image, the corresponding reference text (e.g., "The walking dog in the picture" for `...0003.jpg`) was provided as input to the loaded CLIPSeg model along with the image.

The CLIPSeg model outputs logits, which represent the model's confidence for each pixel belonging to the object described by the text prompt. These logits are typically visualized as a heatmap overlaid on the original image. The intensity of the heatmap indicates the predicted location of the referred object.

Example Results (Descriptive):

- For image `...0003.jpg` with the prompt "The walking dog in the picture", the CLIPSeg model produced a heatmap strongly highlighting the area occupied by the dog.
- For image `...0023.jpg` with the prompt "The guy in white shirt on the bicycle", the model generated distinct heatmaps, one localizing the person (likely focusing on the white shirt) and another localizing the bicycle when prompted separately (or a combined/stronger heatmap if prompted together, depending on implementation).
- For image `...0034.jpg` with the prompt "The white coffee cups on the coffee machine", the heatmap likely focused on the cups, potentially less strongly on the machine itself, depending on the model's focus.

(Note: Actual output images/plots were generated by the script but are described textually here.)

4.3 Feature Map Visualization ("Y1" Simulation)

The original task asked for plotting the "Y1 feature map" from the LAVT paper's Figure 2. As LAVT could not be run, and CLIPSeg has a different architecture, we adapted this requirement.

For CLIPSeg, the most relevant "feature map" representing the model's localization based on the text is the output logit map itself (before activation like sigmoid).

Therefore, for this part, the same output logits/heatmaps generated in Q4.2 were visualized again, potentially using a different colormap (e.g., 'inferno' or 'viridis') to emphasize the "feature map" aspect, representing the spatial distribution of the model's prediction scores for the given text prompt. These visualizations directly show where the model "sees" the object described by the text.

4.4 Failure Case Segmentation

To explore the model's limitations, segmentation was performed using deliberately incorrect or irrelevant text prompts for each image. Examples of failure prompts used include:

- For `...0003.jpg` (dog): "an airplane taking off", "a stack of pancakes"
- For `...0019.jpg` (bird): "a red fire truck", "a bowl of soup"
- For `...0025.jpg` (butterfly): "a desktop computer monitor", "a bridge over a river"

In these cases, the expected output from CLIPSeg would be either:

- A very diffuse, low-confidence heatmap across the image, indicating it couldn't confidently locate the described object.
- A heatmap incorrectly highlighting an unrelated object that might share some vague visual or semantic feature (less common with good models).
- A near-zero activation map.

Visualizing these failure cases helps understand the model's robustness and specificity.

4.5 Concluding Remarks on Q4

While the specific LAVT model could not be run due to environment incompatibilities with its older dependencies, the core task of Referring Image Segmentation was successfully demonstrated using the CLIPSeg model. CLIPSeg effectively localizes objects described by text prompts by generating spatial heatmaps. The analysis of its output for both correct references and designed failure cases provides insight into the capabilities and limitations of this approach to vision-language grounding for segmentation.