Supplementary: Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

1. Attention weights visualized

We visualize the attention weights learned by the CLIP-A-self for datasets Stanford Cars, UCF101, FGVC Aircraft, Oxford Flowers and CUB in Table 1. We notice that the self-attention mechanism in CLIP-A-self assigns more weight to visually descriptive sentences that are most relevant for discriminating between the classes of the dataset under consideration. For instance, we see that for discriminating images of birds species (CUB dataset) and flower species (Oxford Flowers) sentences describing the color of the head and wings of birds and petals of the flowers are important but for identifying different car or aircraft models sentences describing the color or livery is one of the least important. We also see that if the information being described by the VDT sentence is not clearly visible in the image, the attention weight assigned to it by CLIP-A-self is low. For instance, in CUB dataset, the the undersides of birds or the sepals in Oxford Flowers dataset are often not visible in the images, hence the VDT sentence correponding to this is is in the bottom 3 attributes picked by the learnt attention weights. It's also worth noting that, some of the VDT sentences do not have much variation between different classes and hence are not useful in dsicrimination between the classes of the dataset. For instance, in Oxford-flowers, the color of the leaves, the color of the stem are often green for most flowers in the dataset, which maybe why low attention score was learnt for this attribute.

2. Prompts for GPT-4

Throughout our experiments, we use a two-step prompting strategy in which we first ask the LLM to generate a list of attributes which will aid in visually distinguishing between the different classes in a particular dataset. The second prompt asks the LLM to create a description using the attributes provided by the first prompt and specifies the expected output format. We request a python dictionary as output with a list of sentences, each corresponding to one attribute. The output structure is simple to use (downstream), preserves attribute-level detail and encourages attribute richness.

Example of first prompt for the FGVC Aircraft dataset:

> *I am creating class attributes for a zero-shot image recognition algorithm to classify different images of a diverse set of aircrafts. The attributes are part of side information about the classes. List 20 attributes that can form part of a description of the class that will aid in distinguishing between the following list of classes visually: [list of classes in dataset e.g. 'Boeing 707-320', 'Boeing 727-200',..., 'Yakovlev Yak-42']*

GPT-4 response:

> *Manufacturer: The company that produces the aircraft (e.g., Boeing, Airbus). Model Number: Differentiates different aircraft within the same manufacturer. Aircraft Function: The primary role or use of the aircraft, such as commercial (passenger), cargo (freight transport), or military (fighter, bomber, etc). Engine Count: The number of engines on the aircraft. Engine Position: The location of the engines, either under the wings or at the rear of the fuselage. Wing Position: The position of the wings: high-wing, mid-wing, or low-wing design. Tail Type: The shape and type of the tail, such as traditional, T-tail, V-tail, etc. Body Length: The length of the aircraft's fuselage. Body Shape: The overall shape of the fuselage, like narrow, wide, double-decker, etc. Wing Shape: The shape of the wings, either swept or straight. Nose Shape: The shape of the nose or front of the aircraft. Landing Gear Type: The type of landing gear, like tricycle or taildragger. Cockpit Window Shape: The shape and number of cockpit windows. Color Scheme: Typical color schemes or liveries associated with the aircraft type. Number of Decks: Either single-deck or*

Dataset	Top 3 attributes selected	Bottom 3 attributes selected
FGVC	Unique visual identifier, presence of canards, tail type.	Color scheme, model number, commercial or cargo.
Cars	Body shape, fender description, spoiler description,	Interior description, brand logo description, color scheme
UCF-101	Equipment used, Posture of person, Interaction info.	Body muscles used, force involved, speed of motion
Oxford-Flowers	Shape of the flower, Color, shape and number of petals, Texture and description of veins in leaves	Stem color, Color of leaves, Description of sepals
CUB	Wings color and shape, Head color and shape, Beak color and shape	Color and description of legs, Underparts color, Tail shape and color

Table 1. The top 3 and bottom 3 attributes selected by the attention mechanism in GPT-A-self for 3 different datasets. For UCF101, We see that attention learns to pick visually descriptive sentences like posture and description of objects over temporal information like speed of motion and force applied.

> *double-deck (like the Airbus A380). Existence of Winglets: Whether or not the aircraft has winglets at the end of the wings. Existence of Canards: Whether or not the aircraft has canards, small wing-like surfaces located at the front. Propeller or Jet: Indicates if the aircraft is propeller-driven or jet-powered. Number of Seats: An estimate of the seating capacity of the aircraft. Range: The maximum distance the aircraft can travel without refueling. Unique Visual Identifier: Specific structures or features that are unique to a model or series of aircraft. Aircraft Type: Differentiates between commercial, cargo, or military aircraft.*

Example of the second prompt for the FGVC Aircraft dataset which uses GPT-4's response for the first prompt as part of the prompt:

> *Describe the following aircraft by adding one sentence about each attribute for the following aircraft: Airbus A340-200. Return the answer as a python dictionary with the aircraft variant as the key (i.e. remove the manufacturer) and the value is a list of sentences. Rewrite the attribute as a full sentence. Do not include the attributes as keys. Attributes: Manufacturer: The company that produces the aircraft (e.g., Boeing, Airbus). Model Number: Differentiates different aircraft within the same manufacturer. Aircraft Function: The primary role or use of the aircraft, such as commercial (passenger), cargo (freight transport), or military (fighter, bomber, etc). Engine Count: The number of engines on the aircraft. Engine Position: The location of the engines, either under the wings or at the rear of the fuselage. Wing Position: The position of the wings: high-wing, mid-wing, or low-wing design. Tail Type: The shape and type of the tail, such as traditional, T-tail, V-tail, etc. Body Length: The length of the aircraft's fuselage. Body Shape: The overall shape of the fuselage, like narrow, wide, double-decker, etc. Wing Shape: The shape of the wings, either swept or straight. Nose Shape: The shape of the nose or front of the aircraft. Landing Gear Type: The type of landing gear, like tricycle or taildragger. Cockpit Window Shape: The shape and number of cockpit windows. Color Scheme: Typical color schemes or liveries associated with the aircraft type. Number of Decks:*

> *Either single-deck or double-deck (like the Airbus A380). Existence of Winglets: Whether or not the aircraft has winglets at the end of the wings. Existence of Canards: Whether or not the aircraft has canards, small wing-like surfaces located at the front. Propeller or Jet: Indicates if the aircraft is propeller-driven or jet-powered. Number of Seats: An estimate of the seating capacity of the aircraft. Range: The maximum distance the aircraft can travel without refueling. Unique Visual Identifier: Specific structures or features that are unique to a model or series of aircraft. Aircraft Type: Differentiates between commercial, cargo, or military aircraft.*

The response of the second prompt constitutes the VDT information we utilise as side-information for Airbus A340-200 as an example:

> *"A340-200": ["The Airbus A340-200 is produced by Airbus, a renowned aircraft manufacturer.", "It differentiates itself from other aircraft within the Airbus family through its unique model number: A340-200.", "This aircraft primarily serves a commercial role, typically used for passenger transport.", "The Airbus A340-200 is equipped with four engines.", "These engines are situated under the aircraft's wings.", "The aircraft features a low-wing design, with wings positioned at the bottom of the fuselage.", "It has a traditional tail configuration, common to many large commercial aircraft.", "The A340-200 has a lengthy fuselage, extending to about 59.4 meters.", "The body of the Airbus A340-200 is wide-bodied, facilitating a larger passenger capacity.", "Its wings are swept back, a design aspect that improves fuel efficiency and performance at high speeds.", "The aircraft features a rounded nose shape, contributing to its aerodynamic design.", "The Airbus A340-200 uses a tricycle type landing gear, supporting stability during takeoffs and landings.", "Its cockpit windows are angular and include six panels, giving pilots a comprehensive view of their surroundings.", "Color schemes vary by airline, but the Airbus corporate livery features a predominantly white body with blue and teal accents.", "This model is a single-deck aircraft, focusing on width rather than height for passenger capacity.", "The A340-200 does not have winglets, differing from some newer Airbus models.", "There are no canards present on the Airbus A340-200, instead, it employs a more traditional aircraft design.", "As a jet-powered aircraft, the A340-200 uses high-speed jet engines for propulsion.", "The A340-200 typically accommodates around 260 passengers, though the exact number can vary with the configuration.", "With a range of approximately 7,800 nautical miles, the Airbus A340-200 can cover considerable distances without refueling.", "The aircraft's four-engine configuration and lengthy, wide-bodied design are unique visual identifiers of the A340-200 model.", "Classified as a commercial aircraft, the Airbus A340-200 is primarily used for passenger transportation."]*

GPT-4 generally adheres to the python dictionary output requirement in the *User* prompt, but tends to return additional explanations, motivations or clarifications. To encourage the LLM to only return a Python dictionary as requested, we add the following *System* prompt:

> *You are ChatGPT, a large language model trained by OpenAI. Return only the python dictionary, with no explanation.*

Conversely, OpenAssistant's [1] output requires manual cleaning and reformatting to get into Python dictionary format. GPT-3.5 performed slightly worse than GPT-4 in terms of adherence to the prompt, as it did not consistently return only a dictionary. In such cases, we simply called the API again. After repeated incorrect format responses, we manually cleaned those cases.

We primarily utilized GPT-4 via the ChatGPT Plus subscription plan at a cost of $20 since the GPT-4 API was not generally available during most of our experimentation phase. The GPT-4 API cost to create the VDT information for the SUN397 dataset was $14.90, as opposed to $1.94 using the GPT-3.5 API.

3. Comparing our VDT with GPT3

In Table 2, we compare the VDT generated by GPT-4 using our prompting technique with that of [2] who used GPT-3 to obtain visual descriptors for different classes of the dataset. Here we notice that, including a prompt step asking the GPT-4 for visual attributes necessary for classifying between images of the classes result in a fixed number of sentences per class, a fixed order guaranteeing that every class is accompanied by as much visual information as possible. By using GPT-4 we also get much richer and more accurate visual descriptions. For example, for the class industrial, our descriptions provide inforrmation about density of buildings, shadow in the image, road accessibility and layout while the description used by [2] is only 'evidence of human activity'. A similar phenomenon can be observed for DTD dataset. This explains the jump in performance for specialized datasets like DTD and Eurosat over DCLIP.

4. Generalizability at lower shots

In Figure 1, we compare the harmonic mean of Base and New accuracies of CLIP-A-self with that of CLIP-A over number of shots = 1, 5, 10, 16. Our CLIP-A-self demonstrates performance improvements at lower shots, outperforming CLIP-A on average by over $1.5\%/$ for the 1-shot case and over $2.5\%/$ for the 5-shot case. Our adapter shows higher improvements over CLIP-A in the higher shot scenario because of the number of parameters and the inherent difficulty in identifying the VDT sentences that are discriminative for the current classes in the low shot scenario. For instance, identifying the class from a single image is often difficult because of co-occuring objects, environment, background etc which can be resolved if we have more exmaple images from the same class. The largest improvements are for specialized and fine-grained datasets like StanfordCars, EuroSat Oxford Flowers, DTD and CUB. Oxford-pets and Food-101 results do not improve much because these datasets are relatively easy and already show good performance with default CLIP.

References

[1] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

[2] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *ICLR*, 2023.

Table 2. Comparing our VDT with that of descriptors from [2] for 2 random classes of datasets DTD and Eurosat

	Ours	**DCLIP[2]**
Stratified (DTD)	’The surface feels moderately smooth, with slight roughness due to the layered structure.’ ’There is no distinct pattern, but the layers create a natural, linear visual effect.’ ’The structure is characterized by multiple layers stacked upon each other.’ ’The texture has a two-dimensional feel, with the layers adding a sense of depth.’ ’The density varies, with some layers appearing closely packed while others are more sparse.’ ’The regularity of the texture is defined by the consistent layering.’ ’The texture is opaque, with no transparency between the layers.’ ’There are no significant surface defects, but minor irregularities may occur between layers.’	’a series of layers’ ’each layer is of a different material’ ’the layers are parallel to each other’ ’the layers may be of different thicknesses’ ’the layers may be of different colors’ ’the layers may have different textures’
Lined (DTD)	’The texture feels moderately smooth to the touch, not too rough nor too sleek.’ ’It exhibits a lined pattern, reminiscent of ruled notebook paper.’ ’The structure of the texture is stratified, with lines arranged one after the other.’ ’The texture has a two-dimensional quality, with no noticeable depth or relief.’ ’The lines are densely packed, leaving little space between them.’ ’The texture displays a high degree of regularity, with the lines evenly spaced and parallel.’ ’The texture is opaque, with no transparency or translucency.’ ’There are no noticeable surface defects, the lines are clean and uninterrupted.’	’a series of parallel lines’ ’can be straight or curved’ ’may be of different colors’ ’may be of different widths’ ’may be of different thicknesses’
Industrial (Eurosat)	’Industrial buildings have texture that is smooth, regular.’ ’Industrial buildings have shape that is rectangular, irregular.’ ’Industrial buildings have size (relative) that is large.’ ’Industrial buildings have pattern that is regular, dense.’ ’Industrial buildings have spectral reflectance that is high in visible spectrum.’ ’Industrial buildings have a shadow that is present (due to high-rise buildings).’ ’Industrial buildings have adjacent land features that is commercial, residential, roads.’ ’Industrial buildings have change over time that is stable.’ ’Industrial buildings have density that is high.’ ’Industrial buildings have proximity to water bodies that is variable.’ ’Industrial buildings have road accessibility that is high.’	’evidence of human activity’
Forest (Eurosat)	’Forest has texture that is rough.’ ’Forest has shape that is irregular.’ ’Forest has size (relative) that is large.’ ’Forest has pattern that is no pattern.’ ’Forest has spectral reflectance that is high in near-infrared.’ ’Forest has shadow that is present (due to trees).’ ’Forest has adjacent land features that is land, mountains, rivers.’ ’Forest has change over time that is mostly stable.’ ’Forest has density that is high.’ ’Forest has proximity to water bodies that is variable.’ ’Forest has road accessibility that is low.’	’a large area of trees’ ’green leaves’

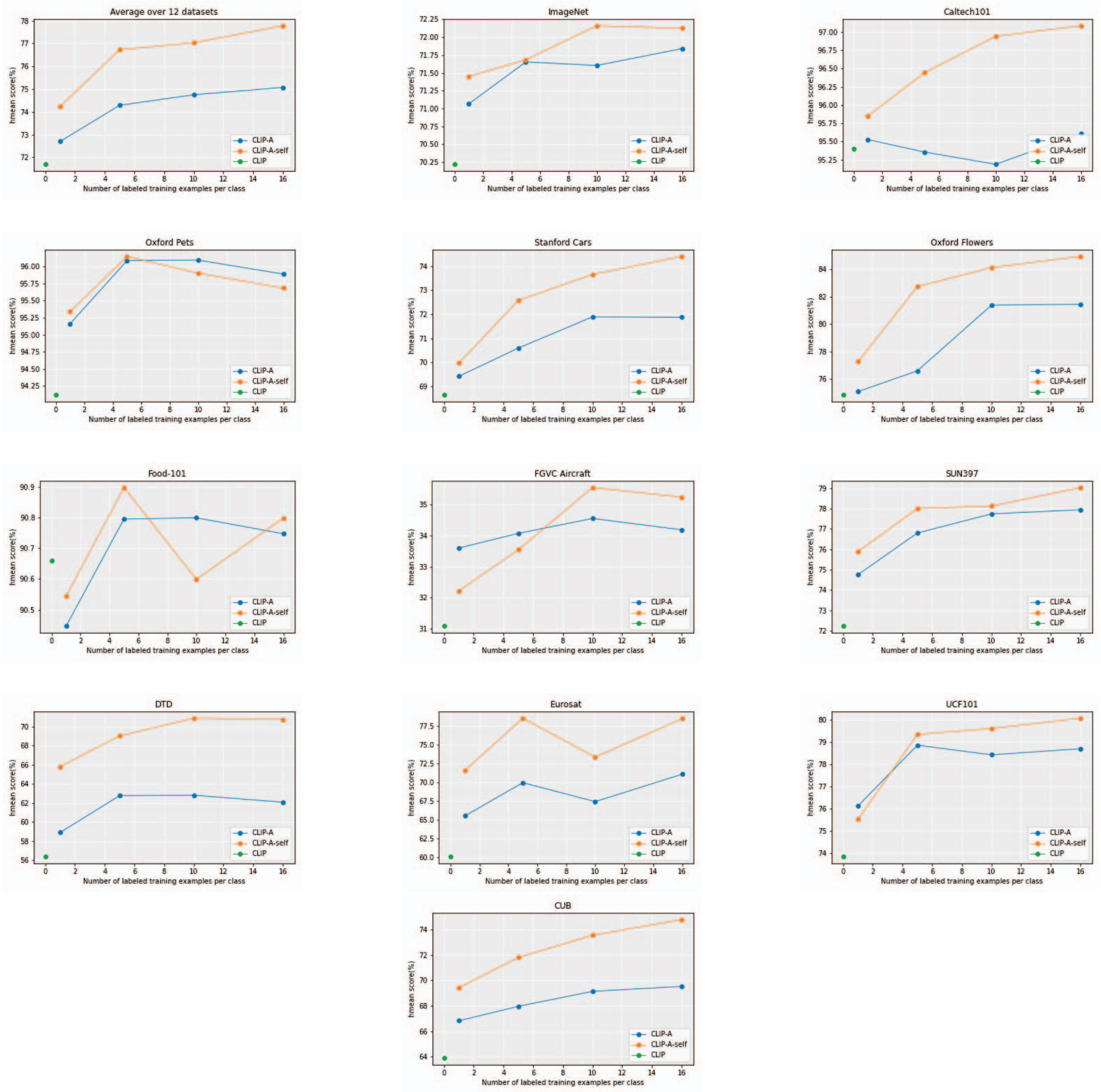

Figure 1. Main results of Base-to-New few shot learning on 12 datasets. CLIP-A-self consistently shows better performance over CLIP-A over different training shots, demonstrating the importance of Visually descriptive text in improving the generalizability of few-shot classifiers for CLIP.