

Supplementary Materials: Align2Concept: Language Guided Interpretable Image Recognition by Visual Prototype and Textual Concept Alignment

Anonymous Authors

1 INSTRUCTIONS IN PROMPT GENERATION

1.1 Instruction for Birds' Textual Concept

Here is the instruction for designing concept descriptions of bird texts, including two examples sampled from the pool of contexts.

Prompt for Constructing Birds' Textual Concepts.

[Task Name] : Generation of Bird Visual Prototype Concept Descriptions

[Role to Play] : You are an ornithology Expert. Given the [Bird Class], your task is to generate [Text description of Visual Prototype Concept] according to the Task Considerations and Standards provided below.

[Considerations for Task Completion] :

1. Accuracy: The descriptions must be based on ornithological knowledge to ensure they accurately reflect the characteristics of the specified bird species.

2. Focus on Visual Features: The generated phrases should emphasize describing the bird's visual features, such as color, size, shape, and feather patterns.

3. The phrases should represent the most representative prototype concepts of the category and be recognizable by visual prototype networks, not merely as textual concepts. This helps the visual prototype networks better understand the bird's features and attributes.

4. Scientificity: Use professional ornithological terminology to ensure the descriptions' professionalism and authority.

5. Avoid Redundancy: Avoid choosing repetitive or similar phrases to ensure each selected phrase contributes unique visual information about the bird species.

6. Integrate Expert Opinions: Consider bird knowledge experts' opinions when selecting phrases to ensure the scientific accuracy of the phrases.

[Task Standards]:

1. Clear Visualization Features: The generated 20 phrases should clearly describe the color, size, shape, and appearance of bird parts such as the bill, beak, belly, breast, crown, forehead, leg, eye, wings, tail, throat, etc. These features represented by the phrases should be recognizable by visual prototype networks.

2. Broad Coverage: The chosen phrases should cover as many different visual features of birds as possible, avoiding excessive focus on any single feature. There should be no repetition or extreme similarity among selected phrases, ensuring each provides unique information.

3. Semantic Clarity: Phrase descriptions must be semantically clear, avoiding vague or unclear descriptions.

4. Professional Accuracy: The selection of phrases should be endorsed by bird knowledge experts to ensure the descriptions' accuracy and scientific nature.

5. Phrase Conciseness: Each phrase should concisely and directly describe the external features of a prototype concept.

6. Descriptions should state the appearance directly without adding functional descriptions. Moreover, this only requires a description of their physical characteristics, without adding symbolic meanings and spiritual qualities.

[Examples] :

[Appropriate Phrases] :

1. "This bird species has white wings."

2. "The bird has a black body."

[Inappropriate Phrases] :

1. "This bird's song is particularly melodious."

2. "This bird's habitat is in the United States."

3. "Nest constructed from soil and vegetation."

4. "Longevity, with some individuals living over 50 years",

5. "Adapted to a diet of squid, fish, and crustaceans",

6. embodying the essence of pelagic seabirds."

[Output Format] :

Please complete the task according to these guidelines and list your answers with numbers.

[Note] :

Ensure the condensed features are as non-redundant as possible, and the improved condensed features cover as broad a range as possible while condensing. Ensure the condensed features must clearly describe the visual features of birds, making them easily recognizable by visual networks. You must Ensure that the phrases you generate are consistent with the language style of the examples provided. You must only generate answers directly, without adding any irrelevant responses.

[Bird Class] : common yellowthroat

[Text description of Visual Prototype Concept] :

1. "bright yellow throat makes it easy to identify",

2. "black mask through its eye",

3. "paler than the adults",

4. "small, warbler-like bird with a bright yellow throat and breast",

5. "small songbird with olive-green back and yellow breast",

6. "small songbird with a bright yellow throat",

7. "yellow bars on yellow wings",

8. "small songbird with a bright yellow breast and olive-green back",

9. "small songbird with a yellow body and black mask",

10. "bright yellow throat is distinctive",

11. "white throat and black mask",

12."small songbird with a bright yellow body and black throat",
13."small songbird with a yellow body and black markings on its head",
14."black mask that extends through its yellow eyes",
15."greenish-yellow body",
16."small songbird with a yellow body and black-and-white wings",
17."yellow mask around the eyes",
18."black mask covering its face",
19."striking bird with a bright yellow body and black facial markings",
20."distinctive yellow breast and black throat"
[Bird Class] : rhinoceros auklet
[Text description of Visual Prototype Concept] :
1."a black body and a white face",
2."black claws",
3."small seabird with a black body and a large white bill",
4."small seabird with a black body and white face",
5."plump seabird with a black body and white belly",
6."dark, sooty-black color",
7."black head with a white horn on its forehead",
8."small, chunky seabird with a large black bill",
9."plump little seabird with a black back and wings",
10."bill is black and stout",
11."single egg in a nest made of seaweed and other debris",
12."a large, orange bill",
13."orange with a black tip",
14."dark grey color",
15."black head with a white face",
16."white face with a black mask around the eyes",
17."orange feet with webbed toes",
18."black head",
19."brown with white spots",
20."dark, slate-gray color"
[Bird Class] : Black-footed Albatross category
[Text description of Visual Prototype Concept] : 1.

1.2 Instruction for Flowers' Textual Concept

Here is the instruction for designing concept descriptions of flower texts, including two examples sampled from the pool of contexts.

Prompt for Constructing Flowers' Textual Concepts.
[Task Name]: Generation of Visual Prototype Concept Descriptions for Flowers
[Role Play]: You are a botanist specializing in flowers. Given a [Flower Class], your task is to generate [Text description of Visual Prototype Concept] based on the task considerations and criteria provided below.
[Considerations for Task Completion]:
1. Accuracy: Descriptions must be based on botanical knowledge to ensure they accurately reflect the characteristics of the specified flower.
2. Focus on Visual Features: The generated phrases should emphasize the visual features of the flower, such as color,

size, shape, and pattern. Please do not include any floral scent descriptions.
3. Representative Phrasing: Phrases should represent the most iconic prototype concepts and be recognizable by a visual prototype network, not just as textual concepts. This helps the visual prototype network better understand the features and attributes of the flower.
4. Scientific Terminology: Use professional botanical terminology to ensure the descriptions are professional and authoritative.
5. Avoid Redundancy: Avoid choosing repetitive or similar phrases to ensure each selected phrase provides unique visual information about the flower.
6. Expert Opinion: Consider the opinions of botanical experts when selecting phrases to ensure the scientific accuracy of the phrases.
[Task Criteria]:
1. Clear Visual Features: The 20 generated phrases should clearly describe the flower's color, size, number, shape, appearance, and arrangement features in parts such as the stigma, petals, sepals, receptacle, and stem. These features represented by the phrases should be recognizable by a visual prototype network.
2. Broad Coverage: Selected phrases should cover a wide range of the flower's visual features as much as possible, avoiding excessive focus on any single feature. There should be no repetition or extreme similarity among selected phrases, ensuring each provides unique information.
3. Semantic Clarity: Phrase descriptions must be semantically clear, avoiding vague or unclear descriptions.
4. Professional Accuracy: Selected phrases should be endorsed by botanical experts to ensure the accuracy and scientific nature of the descriptions.
5. Conciseness of Phrases: Each phrase should concisely and directly describe the external features of the prototype concept.
6. Descriptions should directly state appearances without adding functional descriptions. Moreover, only physical features should be described, without adding symbolic meanings or spiritual qualities.
[Examples]:
[Suitable Phrases]:
1. "Purple-red color"
2. "Petals symmetrically and evenly distributed"
3. "Center is slightly pointy"
[Unsuitable Phrases]:
1. "Found in disturbed habitats."
2. "Its undeniable beauty is undeniable."
3. "Found in English gardens."
3. "Known as the common primrose, English primrose, or flower primrose," "Beloved by gardeners around the world."
4. "Pink color is associated with femininity, love, and romance."
5. "Member of the family Caryophyllaceae."
6. "Bloom continuously throughout the spring."
7. "National flower of Iceland"

8. "Symbol of good luck in Germany"
9. "Member of the borage family"
10. "Flower is the state flower of South Dakota"
11. "Also known as the American cone flower"
12. "Symbolizes happiness, good fortune, and delicate pleasure"
13. "Attracts bees, butterflies, and birds to the garden"
14. "Distinctive floral fragrance"
15. "No fragrance"

[Output Format]:

Please complete the task according to these guidelines and list your answers numerically.

[Note]: Ensure that the compressed features are as non-redundant as possible and that the improved compressed features cover a wide range while being concise. Ensure that the compressed features clearly describe the visual features of the flowers, making them easily recognizable by a visual network. You must not include any floral scent descriptions. You must ensure that the phrases you generate are consistent with the language style of the provided examples. You should only generate answers directly without adding any unrelated responses.

[Flower Class]: Canterbury bells

[Text description of Visual Prototype Concept]:

1. "long, slender neck that bends slightly at the top"
2. "large, green leaves that are shaped like a heart"
3. "delicate white trim"
4. "purple or blue color"
5. "range in colour from white to pink"
6. "beautiful, bell-shaped blossoms"
7. "narrower top"
8. "center of flower is filled with small, white seeds",
9. "wide, round bottom",
10. "grows in shades of blue and purple",
11. "deep red in color",
12. "gorgeous shade of blue",
13. "very deep and dark blue color",
14. "lighter blue hue on the inside of the petals",
15. "deep blue flower with a beautiful color",
16. "white, pink, or blue",
17. "beautiful and calming blue color",
18. "deep cup shape",
19. "very beautiful and soothing blue color",
20. "delicate blue color"

[Flower Class]: colt's foot

[Text description of Visual Prototype Concept]:

1. "dark green leaves that are covered in a white, fuzzy substance",
2. "still used today to treat respiratory conditions such as bronchitis",
3. "known as the cuckoo flower",
4. "so dark that it is almost black",
5. "intensely deep blue",
6. "small yellow flower",
7. "member of the borage family",

8. "five sepals that are fused together at the base",
9. "shaped like a small, yellow bell",
10. "related to the common daisy",
11. "calming light blue color",
12. "flower is also known as the lungwort",
13. "petals are edged in pink",
14. "member of the plant family asteraceae",
15. "yellow pollen",
16. "heart-shaped leaves",
17. "peaceful blue color",
18. "perfect for a spring or summertime wedding",
19. "small, yellow flowers",
29. "blue hue"

[Flower Class]: pink primrose

[Text description of Visual Prototype Concept]: 1.

2 EVALUATION METRIC OF INTERPRETABILITY

We evaluate our model by interpretability for fine-grained image recognition. It is worth mentioning that interpretability has not been quantified in the previous ProtoNet and the extension methods. Inspired by the previous part discovery for fine-grained recognition [1], we designed different quantitative metrics of interpretability schemes for datasets with different annotations. In the CUB-200-2011 dataset, there are 15 part landmarks for each image and we measure the object part localization error by comparing the response region of the learned semantic concept with the annotated part landmarks. The part localization error has been adopted by [1, 2]. For the datasets without part annotations such as Flower-102, we adopt the protocol of Pointing Game [4], which is a popular method to quantify interpretability in visualization methods [3, 4], and calculate the object localization error using the annotated segmentation. The detailed metrics are described in the supplement.

Specifically for part localization error in the ProtoPNet-based model, we first convert the heatmap of the semantic concepts of the ground-truth class to a set of landmark locations by learning a linear regression model from training set, which is similar strategy in [1, 2]. This linear regression model can establish the mapping from the 2D geometric centers of the concepts' heatmaps to the 2D object part landmarks in the image. The part localization errors are calculated by comparing the L2 distances between the predicted landmarks and the ground-truth part landmarks in the testing sets. The smaller the part localization errors, the more accurate the model discovers the part-level concepts. As for the pointing game in the Flower-102 dataset, we calculate the hit rate by counting the peak region of the concept's heatmaps inside the annotation segmentation. Since the semantic concepts in the ProtoPNet-based model are learned from both the foreground and background, the higher hit rate can only indicate that the model has learned less background concepts.

3 ADDITIONAL EXPERIMENTS

Implementation details of ProCoNet. First, we resize input images into 224×224 and adopt the offline data augmentation using random rotation, skew, shear, distortion, and left-right flip

ProtoPNet: Why is the bird classified as a Green Jay?

Evidence for this bird being a Green Jay:

Original image	Concept	Training image where concept comes from	Activation map	Similarity Score	Class connection	Contribution to logits
				2.496	$\times 0.676$	$= 1.687$
				2.245	$\times 0.558$	$= 1.252$
				2.226	$\times 0.846$	$= 1.883$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Total points to the Green Jay: 11.950

ProCoNet: Why is the bird classified as a Green Jay?

Evidence for this bird being a Green Jay:

Original image	Prototype	Training image where concept comes from	Activation map	Textual Concepts (display top 3)	Similarity Score	Class connection	Contribution to logits
				blue head black mask around the eyes blue feathers ...	0.2853 0.2769 0.2598	$\times 0.8587$ $\times 0.8280$ $\times 0.8387$...	$= 0.2450$ $= 0.2293$ $= 0.2179$...
				green wings with the primaries green and yellow feathers bright yellow underparts ...	0.2859 0.2777 0.2583	$\times 0.8612$ $\times 0.8357$ $\times 0.7921$...	$= 0.2462$ $= 0.2320$ $= 0.2045$...
				black markings on its face and throat black feet and legs black bill ...	0.2871 0.2790 0.2608	$\times 0.8268$ $\times 0.7479$ $\times 0.7770$...	$= 0.2373$ $= 0.2087$ $= 0.2026$...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Total points to the Green Jay: 15.317

Figure 1: The different reasoning process between ProtoPNet and ProCoNet.

Why is the bird classified as a Painted Bunting ?

Evidence for this bird being a Painted Bunting :

Original image	Prototype	Training image where concept comes from	Activation map	Textual Concepts (display top 3)	Similarity Score	Class connection	Contribution to logits
				bright blue head blue-gray head blue chin and throat ...	0.2837 0.2706 0.2574	$\times 0.7881$ $\times 0.8245$ $\times 0.8354$...	$= 0.2238$ $= 0.2231$ $= 0.2150$...
				green wings green feathers green-yellow nape ...	0.2856 0.2738 0.2574	$\times 0.8832$ $\times 0.7987$ $\times 0.7129$...	$= 0.2522$ $= 0.2186$ $= 0.1835$...
				vibrant red breast red rump a multicolored plumage of bright blue, green, and red ...	0.2897 0.2752 0.2695	$\times 0.8746$ $\times 0.7954$ $\times 0.7987$...	$= 0.2534$ $= 0.2188$ $= 0.2152$...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Total points to the Green Jay: 17.098

Figure 2: The interpretable reasoning process to identify the species of a bird.

Why is the flower classified as an English Marigold?

Evidence for this flower being an English Marigold:

Original image	Concept	Training image where concept comes from	Activation map	Textual Concepts (display top 3)	Similarity Score	Class connection	Contribution to logits
				bright yellow petals each petals has a slight curve vibrant golden color ...	0.2881 0.2773 0.2705	$\times 0.7918$ $\times 0.8326$ $\times 0.7793$...	$= 0.2281$ $= 0.2308$ $= 0.2108$...
				a darker orange center the center is raised round shape ...	0.2894 0.2796 0.2683	$\times 0.8559$ $\times 0.8803$ $\times 0.7692$...	$= 0.2477$ $= 0.2461$ $= 0.2063$...
				bright yellow petals vibrant golden color green stem with branches ...	0.2881 0.2705 0.2514	$\times 0.7918$ $\times 0.7793$ $\times 0.7062$...	$= 0.2281$ $= 0.2108$ $= 0.1775$...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Total points to the Pink Primrose: 14.139

Figure 3: The interpretable reasoning process to identify the species of a flower.

to enlarge the training set. We adopt the Adam optimizer with learning rate $3e-3$ for add on layers and basis concepts, and $1e-4$ for the vision and language encoder. The weight decay is $1e-3$. The coefficients of λ_1 , λ_2 are set to 0.8 and -0.08 respectively. For the hyper-parameter of manifold alignment, we set the scale of exponential moving average is 0.9 and the update frequency to be every 30 steps to update the orthogonal matrix according to the objective function of manifold alignment.

Training software and platform. We implemented our model using Pytorch and all experiments were run on 8 NVIDIA A4000 GPUs.

Comparison of the reasoning processes between ProCoNet and ProtoPNet. We provide a comparison of the different inference processes of ProtoPNet and ProCoNet, as shown in Figure 1. ProtoPNet infers solely based on the similarity between regions in the image and prototype images, while ProCoNet utilizes multi-modal information to infer based on the similarity between regions in the image and textual concepts.

ProCoNet's inference process. Figures 2 and 3 respectively illustrate the inference process of ProCoNet on other images of birds and flowers.

Our code is available in the supplementary.

REFERENCES

- [1] Zixuan Huang and Yin Li. 2020. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8662–8672.
- [2] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. 2019. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 869–878.
- [3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international*

conference on computer vision. 618–626.

- [4] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126, 10 (2018), 1084–1102.