AN INDEPENDENT COMPOSITIONAL SUBSPACE HYPOTHESIS FOR CLIP'S REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose the hypothesis that CLIP's representation disentangles compositional visual attributes into orthogonal, independent subspaces. We test the hypothesis on five core visual attributes: color, size, counting, camera view, and pattern. We empirically test in several experiments how these attributes are represented and used by CLIP and find that they are encoded in low-dimensional subspaces that are orthogonal to one another, as well as the subject of the image.

1 INTRODUCTION

CLIP-like foundation models are now among the best and most promising models for performing a range of vision tasks (Radford et al., 2021; Ilharco et al., 2021; Jia et al., 2021; Yu et al., 2022). In this work, we focus on understanding how CLIP has learned to organize and compose visual information from two billion image-caption pairs. Our hypothesis posits that, facilitated by natural language alignment, CLIP learns to decompose visual attributes of single-subject images in a way that appears natural humans. Specifically, we believe that in the latent space of CLIP's last layers (of both visual and text encoders), there exist independent subspaces which code for different single-subject compositional attribute types. The defining quality of a compositional attribute type subspace is based on the principle of mutual exclusion: if two compositional attributes can and do coexist, then they belong to different compositional attribute types and thus to different subspaces; if they cannot coexist, then they belong to the same compositional attribute type and thus to different subspaces.

To test our hypothesis, we focus on five compositional attributes types, namely counting (how many of the subject are present), color, size, camera view, and pattern (ex. plaid, zebra, leopard, polka dot). We choose to focus on these five because they align with fundamental compositional attributes for humans [1], though we believe there are many more to which our hypothesis could apply. Future work should explore potential other compositional attributes such as tactile texture, background information, and material, and technique (painting, digital rendering, sculpture, photo, etc.) and relational compositional attributes for multi-subject images. We provide empirical evidence which supports our hypothesis for the aforementioned five subspaces, as well as some interesting properties and consequences of the hypothesis, such as default subspace loading (see Section 4).

Previous work is mixed on whether CLIP's representation is compositional or not. Ma et al. (2022); Thrush et al. (2022) suggest that CLIP is not compositional in multi-object settings, while other work use complex methods to disentangle compositional attributes from subjects in CLIP (Nayak et al., 2022; Jiang et al., 2022; Xu et al., 2022). While most previous work analysed the behavior of CLIP on compositional combinations of attributes, we here perform a direct analysis of the subspaces and their properties of various visual attributes, without extra training. Most similar to our work are Lewis et al. (2022) and Zhou et al. (2022). Lewis et al. (2022) finds that CLIP is compositional in its simultaneous identification of compositional properties and attributes as well as subject matter in images. However, they do not analyse the representational subspaces or their independence. Zhou et al. (2022) also makes the independent subspace claim, and provides empirical evidence to support it, but their work is primarily focused on image manipulation on faces (and not on general understanding like ours), they do not rigorously test the independence of their subspaces, and they do not propose a general hypothesis for how CLIP organizes its latent space into fundamental compositional building blocks which mirrors human composition.

[1] https://www.getty.edu/education/for_teachers/building_lessons/elements.html

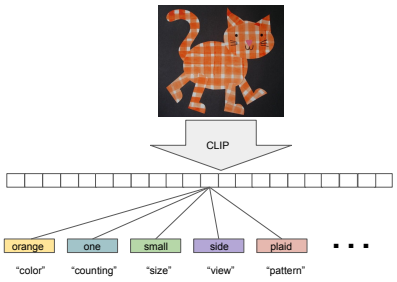


Figure 1: Overview of our hypothesis.

Attribute Type	Dimension
Counting	5
Color	4
Size	5
Camera View	5
Pattern	11

Figure 2: Attribute type subspace dimensions.

2 METHOD

In this paper, we use the following method to isolate attribute type subspaces which closely follows Zhou et al. (2022), though we note that other methods could exist as well.

First, we define a list of text prompts for an arbitrary subject which cover the range of the attribute type subspace we wish to find. For example, to define the size subspace, we use a list of text prompts such as [“a large subject.”, “a medium-sized subject.”, “a small subject.”]. Next, we generate text embeddings of each sentence using the text encoder. Because in our case we are only analyzing the modification of the subject, and not the subject itself, we subtract the embedding of the subject from the sentence, to obtain a representation for the modification of the subject. Then, we do PCA on the subject-subtracted embeddings to obtain an orthogonal basis for the attribute type subspace.

In all of our experiments, we use the OpenCLIP ViT-H-14 model trained on LAION-2B (Ilharco et al., 2021; Schuhmann et al., 2022).

3 EXPERIMENTS

Ideally, the subspaces we found using PCA should exhibit the following two properties: (1) they should be concentrated in a low-dimensional subspace, (2) should encode information only about its respective attribute and not the subject of the image, and (3) should be independent of all other subspaces in CLIP’s latent space. To test these properties, we perform three experiments, which we term the following: **subspace dimensionality** to test (1), **subspace generalization** to test (2), **attribute transfer** to test (2), and **inter-subspace independence** to test (3).

3.1 SUBSPACE DIMENSIONALITY

Following the methodology laid out in Section 2, we find that four to five dimensions are sufficient to describe more than 80% of the variance for Counting, Color, Size and Camera View, see Figure 2. For Pattern, a subspace with 11 dimensions is sufficient. All five visual attributes together thus fill roughly 2.93% of CLIP’s representation.

3.2 SUBSPACE GENERALIZATION

Design To test whether the subspaces we found using CLIP’s text encoder generalize to the image domain, we searched the web for images containing combinations of a 3-4 test attributes across 4 subjects, except for color. For color, we tested four basic shapes (plus, circle, square, and triangle) with four colors (blue, green, red, turquoise). For example, to test the pattern subspace, using an internet search we found 16 images across 4 patterns (zebra, leopard, polka dot, striped) and 4 subjects (dress, pants, shirt, socks) with each image combining a different pattern and subject. We encoded the images using the image encoder, projected the image embedding onto the respective attribute type subspace, and then computed pairwise similarity matrices between the 16 image embeddings before and after projection. If the subspace is indeed coding for the target attribute type, then the similarity matrices should display a diagonal block structure, with each block corresponding to one

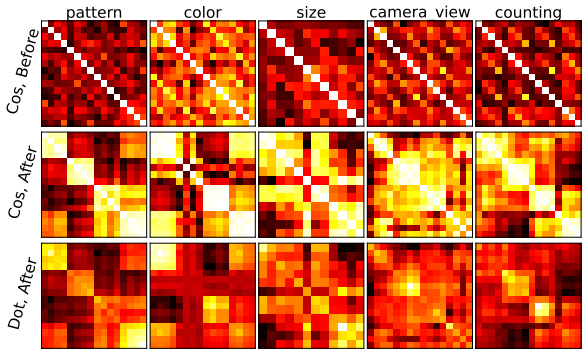


Figure 3: **Subspace Generalization**. The first row contains the pairwise cosine similarity matrices before projection onto the respective attribute type's subspace, the second row contains the pairwise cosine similarity matrices after projection, and the third row contains the pairwise dot product matrices after projection.

Attribute Type	Src Subject Match	Tg Subject Match	Src Attribute Match	Tg Attribute Match
Counting	0	0	-100.00	+89.58
Color	0	0	-91.67	+56.94
Size	-1.39	0	-61.11	+50.00
Camera View	0	0	-81.25	+81.25
Pattern	0	0	-100.00	+99.31

Table 1: **Attribute Transfer**. "Src Subject Match" refers to the change in percentage of times the source subject was predicted via vector arithmetic. "Tg Subject Match" refers to the change in percentage of times the target subject was predicted. "Src Attribute Match" refers to the change in percentage of times the source's attribute was predicted. "Tg Attribute Match" refers to the change in percentage of times the target's attribute was predicted.

of the 4 test attributes *only after projection*. We repeated this process for each of the attribute type subspaces. We only use test attributes that were present in the subspace basis creation process.

Results Results are displayed in Figure 3. In most cases, we find a fairly pronounced diagonal block structure, with each block corresponding to the same test attribute. Additionally, the similarity matrices also seem to show that hierarchical information is also encoded in the subspace bases. For instance, in the color subspace, the "turquoise" block is very similar to the "blue" and "green" blocks.

3.3 Attribute Transfer

Design In this experiment, for each attribute type subspace, we take the images found for the **subspace generalization** experiment, and encode them to get corresponding image embeddings. Then, we perform vector arithmetic to transfer the given test attribute of one image to another's. To this end, consider a projection $\mathbf{P}_a$ into the subspace of attribute a, a source embedding $\mathbf{z}_s$ and a target embedding $\mathbf{z}_t$. Then the transferred embedding is computed as

$$\mathbf{z}' \propto \mathbf{z} - \alpha \mathbf{P}_a^\top \mathbf{P}_a (\mathbf{z}_s - \mathbf{z}_t) \tag{1}$$

which we normalise to norm one. We then compute similarities between the transferred embeddings to text embeddings of captions combining every test attribute and subject before and after the attribute transformation. If the modifier subspaces are coding for their respective attribute types and act independently of content, then we should see that the predicted subject should *not* change before and after the vector arithmetic, but the predicted test attribute *should*. We throw out results for pairs containing the same attribute or content. In our experiments we found that $\alpha = 3$ worked well.

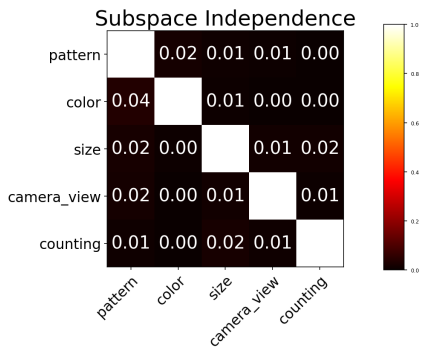


Figure 4: **Inter-Subspace Independence**. Columns explain variance for rows.

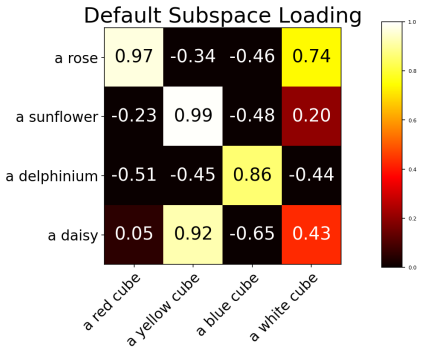


Figure 5: **Default Subspace Loading**. Attributeless flower text embeddings load a flower's default or most frequent colors into the color subspace.

Results Results are displayed in Table 1. In every case, the transformations successfully transfer an image embedding's attribute without significantly changing the subject.

3.4 Inter-Subspace Independence

Design To determine inter-subspace independence, we take ImageNet (Russakovsky et al., 2014) validation set embeddings, project them onto one attribute type's subspace, measure the variance, and then project the projections onto a second attribute type's subspace, measure the variance, and then take the ratio between the two variances. The intuition is that independence might be gauged by measuring how much of one subspace is explained by another.

Results Results for this experiment are displayed in Figure 4. All of the subspaces are highly independent by this metric.

4 Default Subspace Loading

In this section, we explore an interesting property of our hypothesis which we call default subspace loading. Default subspace loading occurs in the text encoder when a caption including a subject without specific attributes is embedded. In the absence of specific attributes, the text encoder loads default representations into the attribute subspaces which are determined during training. For example, when creating a text embedding of "a lion", CLIP will load a representation for yellow in the color subspace, because CLIP has learned to associate yellow with lions. To test this property, we create text embeddings of "a red cube", "a yellow cube", "a white cube", and "a blue cube", and then project these embeddings into the color subspace. Then we take text embeddings of "a rose" (red default), "a sunflower" (yellow default), "a daisy" (white default), and "a delphinium" (blue default), and project these embeddings onto the color subspace, and compute cosine similarities with each of the first set of projected colored cube embeddings. Results can be seen in Figure 5. This phenomenon explains why prompt engineering for zero-shot classification is so important: doing so would remove spurious correlations which exist because of the default associations and subspace loading.

5 Conclusion

In this paper, we posit that, facilitated by the natural language alignment objective, CLIP isolates visual information into orthogonal subspaces in a way similar to how humans compose images. According to the dimension of the subspaces we found, we explained about 2.93% of this CLIP model's final layer. Future work should try to increase this percentage by identifying more compositional attributes for single and multi-subject images.

REFERENCES

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Eric Wang. Comclip: Training-free compositional image and text matching. *ArXiv*, abs/2211.13854, 2022.

Martha Lewis, Qinan Yu, Jack Merullo, and Elizabeth-Jane Pavlick. Does clip bind concepts? probing compositionality in large image models. *ArXiv*, abs/2212.10537, 2022.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *ArXiv*, abs/2212.07796, 2022.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. *ArXiv*, abs/2204.03574, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2014.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

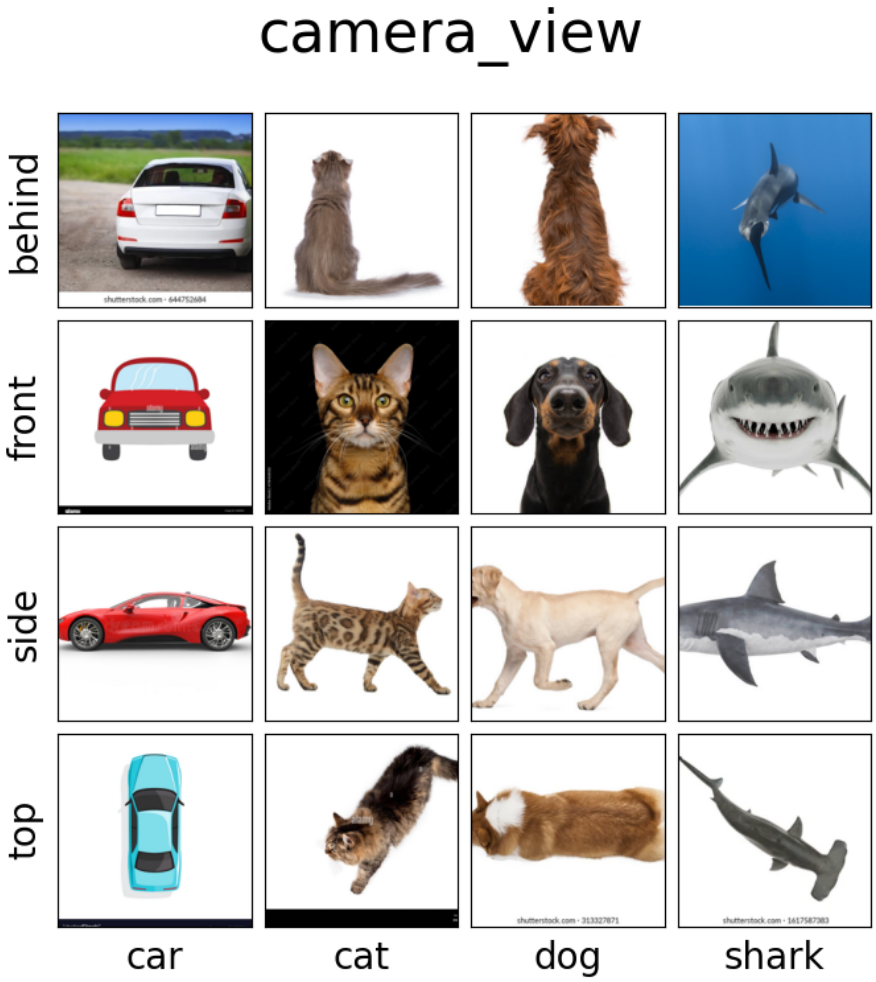
Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022.

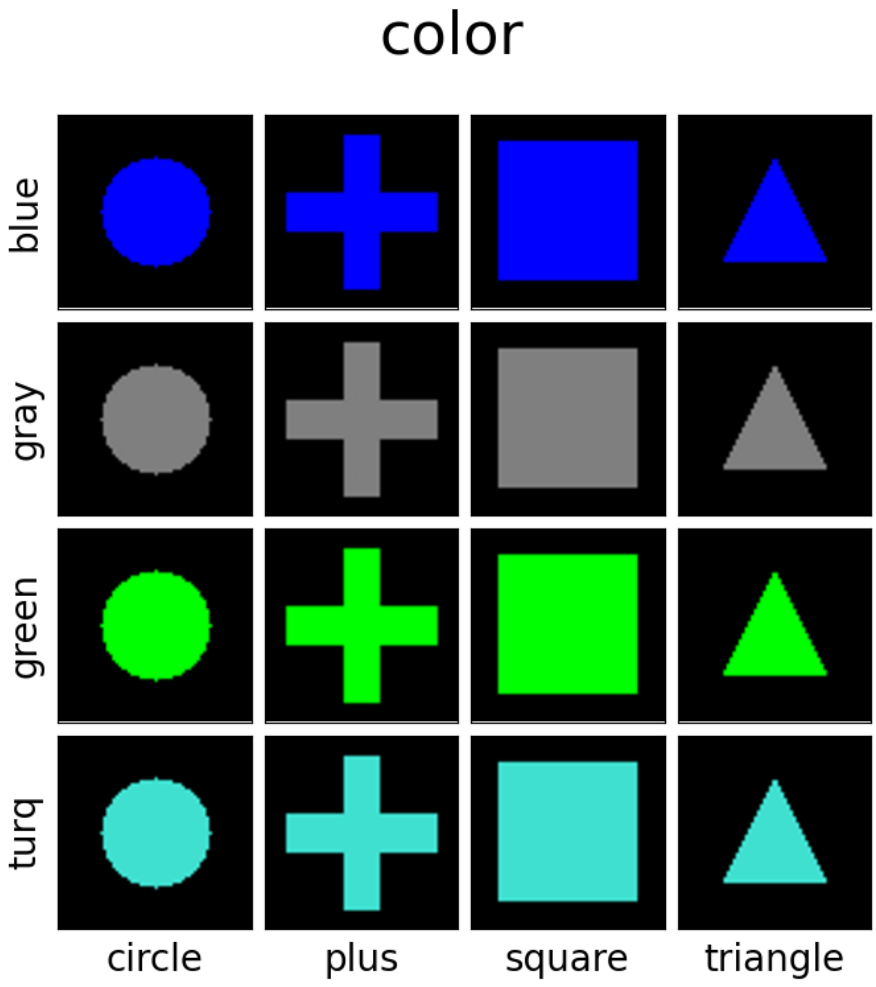
Guangyue Xu, Parisa Kordjamshidi, and Joyce Yue Chai. Prompting large pre-trained vision-language models for compositional concept learning. *ArXiv*, abs/2211.05077, 2022.

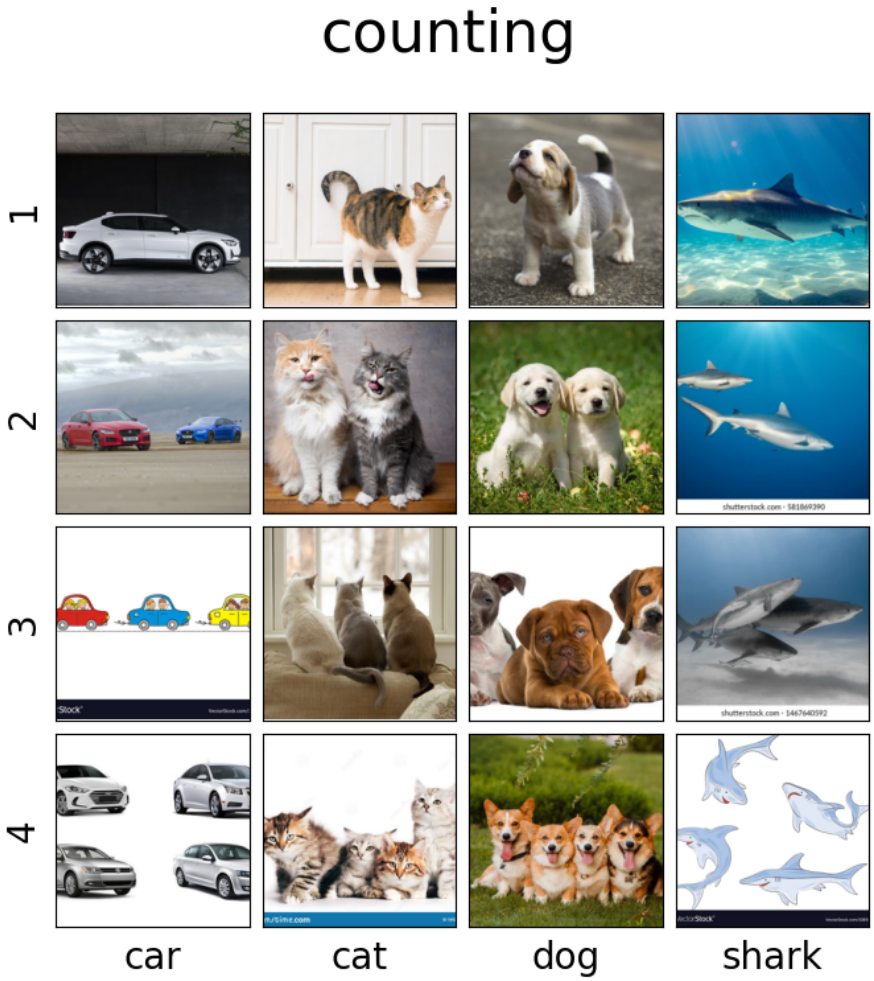
Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.

Chenliang Zhou, Fangcheng Zhong, and A. Cengiz Öztireli. Clip-pae: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable, and controllable text-guided image manipulation. *ArXiv*, abs/2210.03919, 2022.

A IMAGES USED FOR **SUBSPACE GENERALIZATION** AND **ATTRIBUTE TRANSFER**




color
blue
gray
green
turq
circle
plus
square
triangle


counting
1
2
3
4
car
cat
dog
shark


pattern
leopard
polka-dot
striped
zebra
dress
pants
shirt
socks


size
giant
medium
small
car
cat
dog
shark