Towards Interpretable Visual Topics via Foundation Models

Shreya Havaldar*, Weiqiu You*, Lyle Ungar, Eric Wong

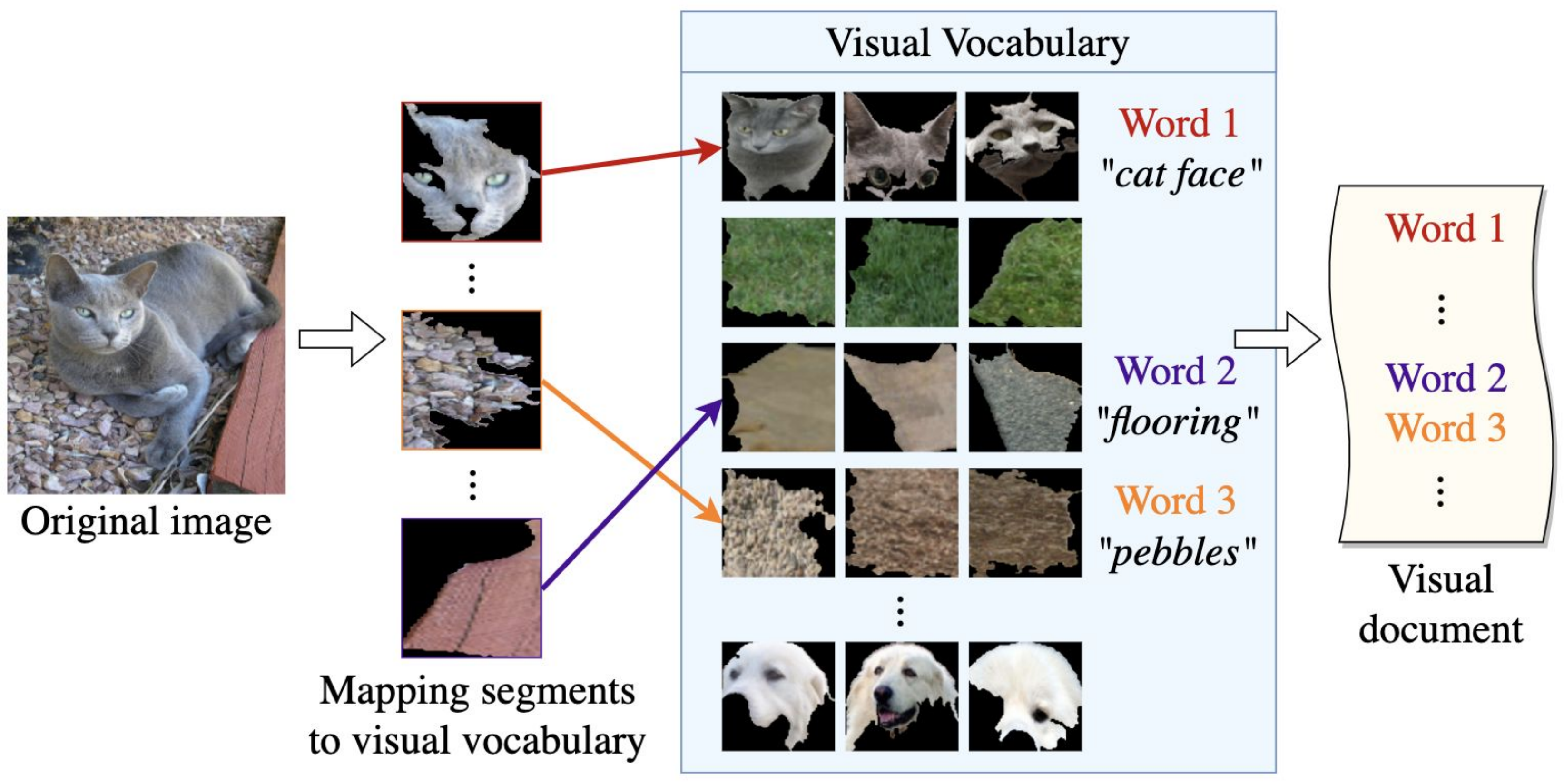
Text topics are useful. What about vision?

- Topic modeling is a technique to explain *relationships* in a language dataset
    - Example topics: *{sports, ball, field, jersey}, {government, war, Obama, state}, etc.*
- Written language is composed of letters → words → documents. But, no such structure exists in images!
    - We can create a visual vocabulary to mimic this structure.

We use foundation models to create high performing visual topics for image datasets.

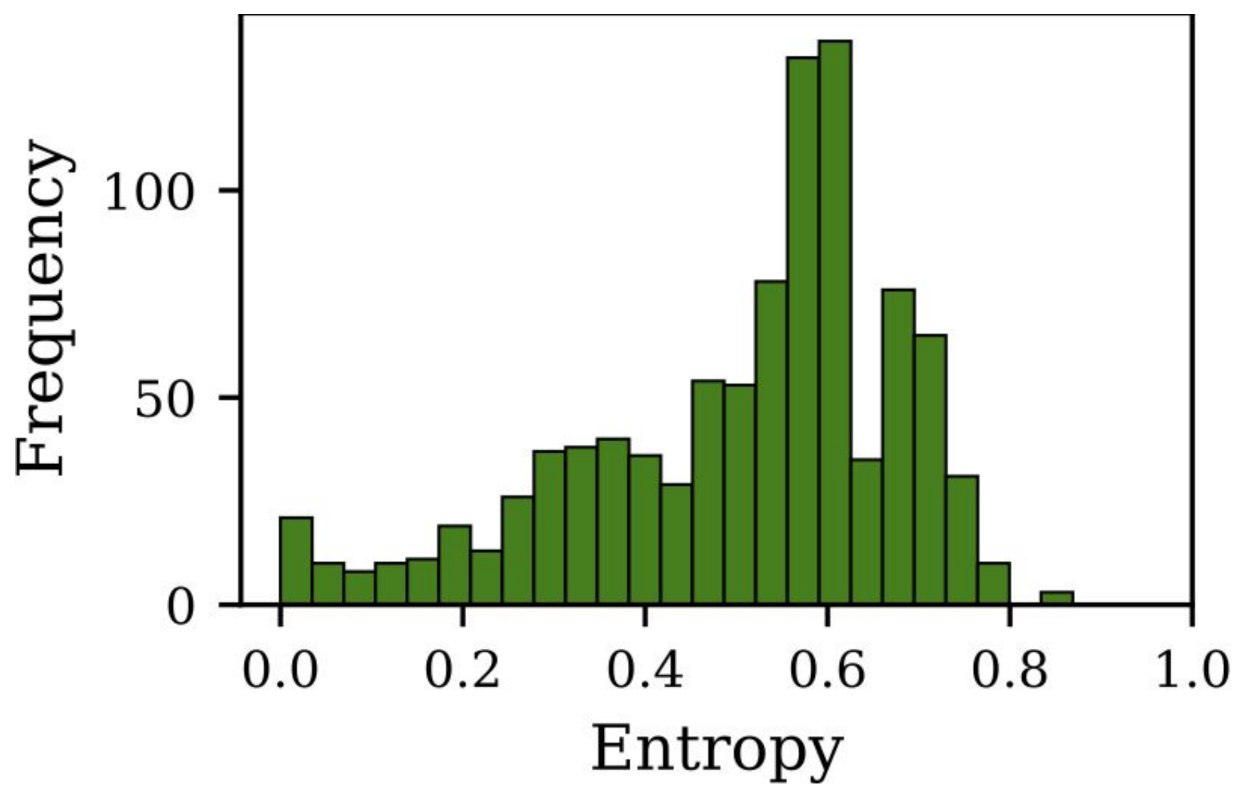
Visual topics from foundation models

1. Each image is broken into segments using a segmentation model.
2. We cluster segment embeddings from Vision Transformer to create a visual vocabulary. Each segment is then mapped to a "visual word" in this discrete visual vocabulary.
3. We can transform each image into a "visual document" containing its corresponding visual words, mapping an image dataset to a set of visual documents.



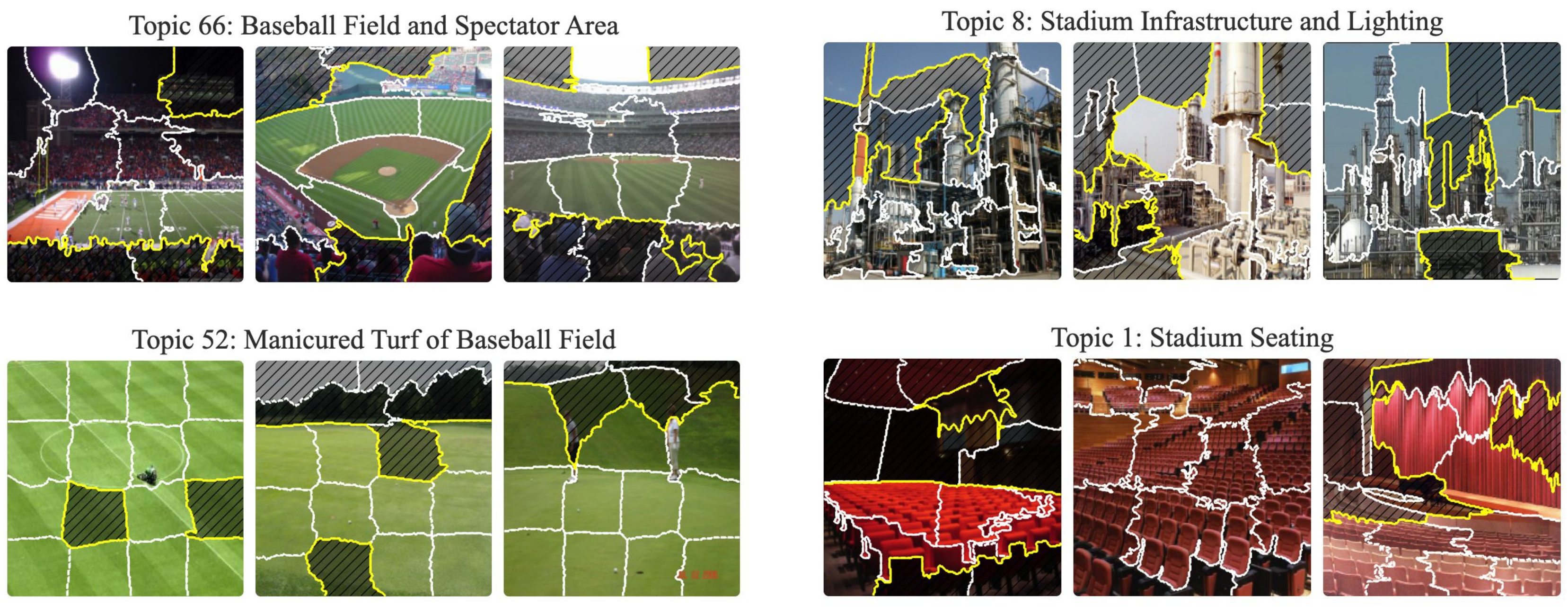
4. With the visual documents created above, we can directly apply any topic modeling algorithm an image dataset. (we use LDA in this work).

Topics are not the same as clusters! Clusters are visually similar and often from the same classes. Topics have high entropy across classes (right).



Topics can summarize visual classes

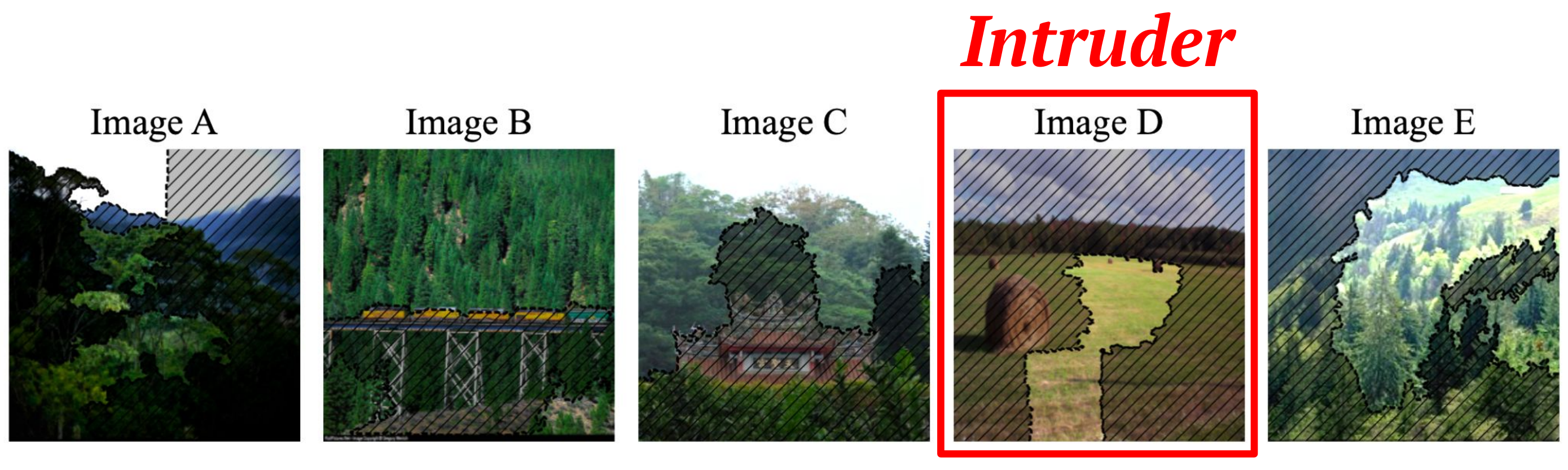
- Just like text topics, visual topics can summarize and explain image datasets.
- Summary of the baseball stadium class from the SUN 397 dataset. Instead of using 1,000 images to visualize a class, we can use **the top 4 topics it contains.**


Topic 66: Baseball Field and Spectator Area
Topic 8: Stadium Infrastructure and Lighting
Topic 52: Manicured Turf of Baseball Field
Topic 1: Stadium Seating

Areas not hatched out in the above images are the "words" in the topics. The topics are all different components of a baseball stadium: field, infrastructure, turf, seating, etc.

User study for interpretability

- We conduct the first visual intrusion detection study to assess the interpretability and utility of our topics.
- **Visual Word Intrusion Task:** Users are asked to identify which image contains highlighted visual words that do not belong in the topic.



12 users analyze 150 topics, and detect the correct intruder with 87.7% accuracy! *(baseline = 20%)* This is at par with intrusion detection scores for NLP topics.

Takeaways + future work

- We use SOTA foundation models to create more interpretable visual topics.
- We develop visual word intrusion detection as a way to measure interpretability of visual topics.
- High-quality visual topics can potentially explain image datasets in fields like medicine and the social sciences.

Paper