11-777 Report 1: Dataset Proposal and Analysis

Geerisha Jain* **Madhura Deshpande*** **Shaurya Singh*** **Shrey Madeyanda***
{geerishj, mvdeshpa, shauryas, smadeyan}@andrew.cmu.edu

1 Problem Definition and Dataset Choice

Compositionality, the understanding that "the meaning of the whole is a function of its parts" (Chen et al., 2020), is an important aspect of intelligence. In natural language, a sentence is made up of its words. For vision, we can consider an image, which is made up of parts like objects, their attributes, and their relationships (Hudson and Manning, 2019) (Janssen and Partee, 1997). For instance, compositionality allows people to differentiate between a photo of "a man in a yellow shirt facing a wall painted white" and "a man in a white shirt facing a wall painted yellow".

Today's vision-language models, pretrained on large-scale image-caption datasets, are being widely applied for tasks that benefit from compositional reasoning, including retrieval, text-to-image generation, and open-vocabulary classification (Ma et al., 2023).

Our choice of dataset, SUGARCREPE (Hsieh et al., 2023), provides a benchmark for the compositional capability of vision-language models. Other existing benchmarks exhibit inherent biases, which are most apparent when text-only models, with no access to the images, outperform vision language models on these benchmarks. Assessing other state-of-the-art models on SUGARCREPE revealed that their advancements were overestimated. Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) struggle with binding correct attributes to the correct objects, understanding relations between objects, generalizing systematically to unseen combinations of concepts and to larger and more complex sentences (Singh et al., 2023).

We aim to implement an approach that beats SUGARCREPE's compositionality benchmark, with a focus on attribute binding.

In the context of vision-language models and compositionality, attribute binding refers to the process of correctly associating descriptive attributes (such as colors, shapes, sizes, etc.) with their corresponding objects or entities depicted in an image. This task involves understanding the spatial relationships between objects and their attributes as described in textual descriptions. For example, in the sentence "a red apple on a table," attribute binding involves correctly associating the attribute "red" with the object "apple" and understanding the spatial relationship between the apple and the table.

Attribute binding is crucial for generating accurate and meaningful descriptions of visual scenes in natural language. It requires the model to not only recognize objects and their attributes within an image but also to understand how these attributes are related to the objects and their spatial configurations. Achieving effective attribute binding contributes to the overall compositional capability of vision-language models, allowing them to generate coherent and contextually relevant descriptions of complex visual scenes.

A concern with the SUGARCREPE dataset is that it is meant to be used as a test-only benchmark, it is not suitable to use for training a model. We will instead make use of the Flickr30k dataset for training our model, making use of its images and corresponding captions. Hence, our data analysis will be focused on images and captions from the Flickr30k dataset.

A drawback of the Flickr30k dataset for our purposes is that it would have only positive text for each of its images (the captions accurately describe the image with varying levels of detail), and to train a model with robust compositional capability,

*Everyone Contributed Equally – Alphabetical order

we will also require hard negative text, which will differ minimally from the positive text.

To generate these hard negative texts, we will consider the scene graph representations of the image and text contained in the Flickr30k dataset. We will utilize hard-negative mining techniques using graph transformations of the text scene graphs, that are seamlessly coupled with our text decomposition strategy, and applied over any text as outlined in (Singh et al., 2023).

The following is an example of a postive and hard-negative caption pair, with minimal difference:
Positive: A woman in a white dress staring at a grey wall.
Hard-Negative: A woman in a grey dress staring at a white wall.

1.1 What phenomena or task does this dataset help address?

By providing a benchmark with reduced dataset biases and a diverse set of fine-grained hard negatives, SUGARCREPE aims to facilitate a more faithful evaluation of vision-language models' compositionality.

This dataset would help us assess our models' ability to comprehend and describe visual scenes accurately by testing it's understanding of atomic concepts and relationships within the scenes. Specifically, we'll be focusing on attribute-binding.

1.2 What about this task is fundamentally multimodal?

This task will involve two modalities - text and images. We will examine the spatial relationships between objects and their attributes in the positive text, and attempt to align image representation with these relationships. The aligned representations can be used to generate a single representation using fusion, which can help us train a classifier to distinguish between a positive and hard-negative text (with minimal differences from the positive text) for any given image.

Our aim is to assess whether the relationships expressed between the attributes and the objects in the text modality actually reflects the attribute binding depicted in the image.

1.3 Hypothesis

We believe the following are places cross-modal information can be used or improved

1. Spatial relationships between attributes and objects in the text can be aligned with the corresponding image representation. We refer to the approach followed in (Rassin et al., 2024). Their focus is on text-conditioned image generation, where they aim to create a strong mapping between linguistic binding of entities and modifiers in the prompt and visual binding of the corresponding elements in the generated image. While this is a text-to-image retrieval task, our scope will be focused on compositionality benchmarks formulated as image-to-text retrieval task.

2. A text based cross-attention module can be used to encourage large overlaps between attributes and objects in the text. A representation learnt from this cross-attention module can then be used along with image representations in an image-text cross-attention module which would enable us to selectively attend to different parts of the image and the corresponding attribute-object bindings, facilitating the discovery of latent alignments and capturing fine-grained interplay between visual and textual information.

1.4 Expertise

We have the following expertise in the underlying modalities required by this task:

1. Geerisha Jain: Industry experience in NLP. Took 10-601 in Fall 2023. Taking 11-785, 11-711, 11-777 in Spring 2024.

2. Madhura Deshpande: Research experience in NLP and CV. Took 10-601 in Fall 2023. Taking 11-785, 11-711, 11-777 in Spring 2024.

3. Shaurya Singh: Research experience in NLP and CV. Industry experience in ML. Took 10-601 in Fall 2023. Taking 11-785, 11-711, 11-777 in Spring 2024.

4. Shrey Madeyanda: Worked in Data Science and Software Engineering roles in industry. Took 10-601 in Fall 2023. Taking 11-785, 11-711, 11-777 in Spring 2024.

2 Dataset Analysis

2.1 Dataset properties

1. Size (in GB) : 9
2. Number of Images : 31,000
3. Number of captions: 158,000
4. Number of bounding boxes (manually annotated): 276,000
5. Number of coreference chains: 244,000

2.2 Compute Requirements

1. Files (can fit in RAM?) : Yes, we expect the dataset files to fit within the RAM.
2. Models (can fit on GCP/AWS GPUs?): Yes, the models can fit on GCP/AWS GPUs.

2.3 Modality analysis

1. Total Number of Words - To gain insight into the dataset's scale and the verbal content's density, we count the total words across all captions on a sample of 40,460 captions.

   Value: 436,210

2. Total Number of Unique Words - To get an idea of the variety of words used across the captions, we find the total number of unique words across all captions on a sample of 40,460 captions.

   Value: 9243

3. Lexical diversity - The lexical diversity provides a snapshot of the richness of the language used in the dataset. We calculate this metric on a sample of 40,460 captions.

   Value: 0.0212

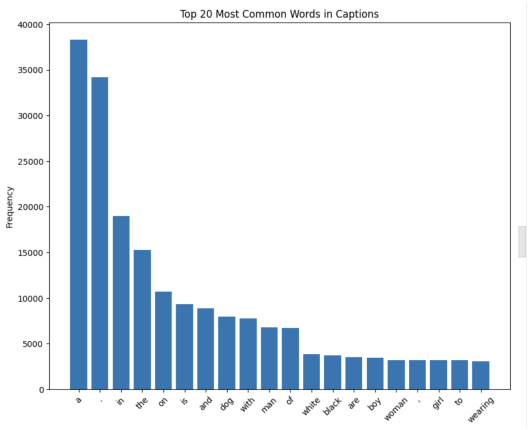
4. Average Number of Pairings Over All Captions - We evaluate the average number of word pairings across all captions, which assists in understanding the complexity and variability of language patterns used in the dataset.

   Value: 1.146158463385354

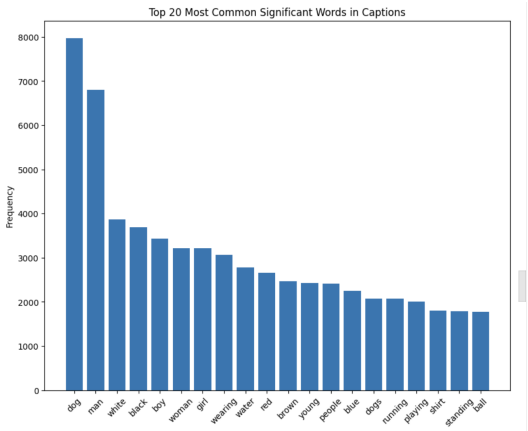
5. Average of the captions with the Highest Number of Pairings - For each of the 2000 images in the dataset, we identify the caption that exhibits the highest number of word pairings, providing insight into which captions are most linguistically rich or complex for each image. Such captions are the most relevant for getting representations to improve attribute binding.

   Value: 2.1025

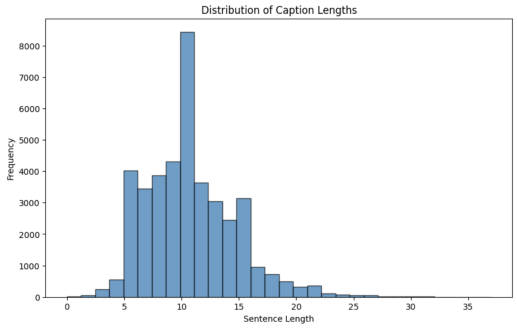
6. Plot the Most Common Words (with stopwords) - We provide a visualization of the frequency of all words, including common language stopwords, over a sample of 40,460 captions.



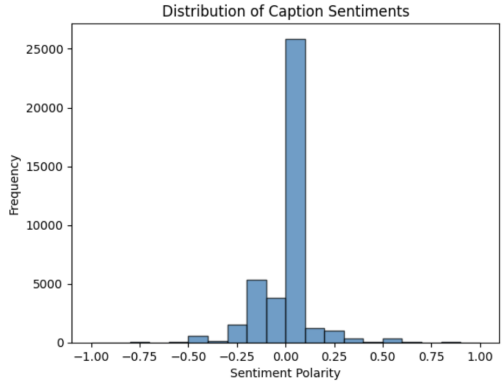
7. Top 20 Most Common Significant Words in Captions (without stopwords) - This visualization highlights the most frequent meaningful words in captions (after removing stopwords), over a sample of 40,460 captions, focusing on content-specific vocabulary.



8. Distribution of Caption Lengths - Our analysis on the lengths of captions, performed over a sample of 40,460 captions, helps to understand their variability and commonality, illustrating how concise or descriptive the captions tend to be.

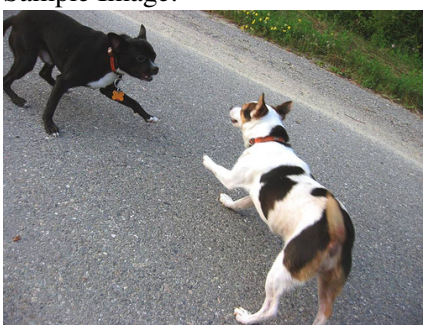


9. Caption Sentiments Polarity - We analyze the sentiments of a sample of 40,460 captions using polarity scores to assess the emotional tone conveyed by the text, ranging from positive to negative.



10. We showcase object recognition within a single sample image, highlighting the complexity and detail present in individual dataset images. Objects detected for the sample image include Greater Swiss Mountain dog, Walker hound, and toy terrier.

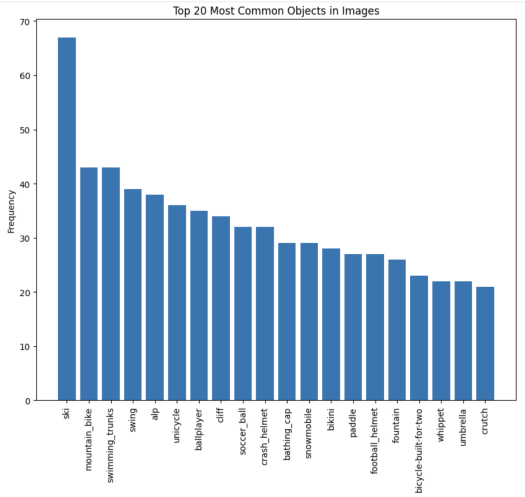
    Sample Image:



    Objects detected for the sample image:

    - Greater Swiss Mountain dog
    - Walker hound
    - toy terrier

11. Plot the Most Common Objects - Our plot identifies and visualizes the most frequently occurring objects, across a sample of 1000 images, providing insight into common themes or subjects within the visual data.



12. Average Number of Objects Detected Per Image - We calculate the average count of distinct objects recognized per image, across a sample of 1000 images, indicating the typical complexity and detail level in the dataset's images.

    Value: 3.0

2.4 Metrics used

We intend to focus on two aspects of compositionality - *systematicity* and *productivity*.

Systematicity measures how well a model is able to represent seen versus unseen atoms and their compositions (Ma et al., 2023). Put simply, it refers to the model's ability to understand and process combinations of elements in ways it has and hasn't explicitly been trained on.

Productivity studies how well a model can comprehend an unbounded set of increasingly complex expressions (Ma et al., 2023) i.e. more elements are added to the image the model is evaluating, defined by the number of atoms (objeccts/attributes) present in the text.

SUGARCREPE has been based on the image-text paris available in the COCO dataset (Lin et al., 2015), which consists of images in various contexts and of different layers of complexity, allowing us to test for systematicity and productivity. To test compositionality, we need these aspects reflected in the associated captions as well. SUGARCREPE achieves this by taking the (positive)

caption from COCO and using ChatGPT to generate hard-negatives while explicitly instructing it to avoid commonsense (logical) and fluency (grammatical) errors. The results are further filtered for false negatives with human validation (for example, replacing child with girl may still result in a correct caption). This is futher de-biased through adversarial refinement, to ensure that performance improvements on SUGARCREPE cannot be achieved by exploiting the identified nonsensical and non-fluent biases (Hsieh et al., 2023).

The benchmark considers three types of hard-negatives, for which performance will be measured (Hsieh et al., 2023):

1. REPLACE: Given a positive text describing a scene, a REPLACE hard negative is generated by replacing an atomic concept (object, attribute or relation) in the original text with a new concept that makes the text mismatch with the original scene.

2. SWAP: A SWAP hard negative is generated by swapping two atomic concepts (object or attribute) of the same category in the positive text, without introducing new concepts in the hard negatives.

3. ADD: Instead of replacing an atomic concept with a new one, an ADD hard negative is generated by adding a new atomic concept (object or attribute) to the positive text that makes it mismatch with the original scene.

Our focus will be on REPLACE, SWAP and ADD for *attributes*.

For measuring compositional capability, the positive captions will be paired with multiple hard negatives of each of the forms described above, and our model's performance will be based on its *accuracy score* across these forms i.e. the percentage of positive (correct) captions identified, when combined with multiple hard-negatives. A concern with using accuracy score as a metric is that it might not give an accurate estimate of the model's ability to handle novel capabilities and scenarios. It is also susceptible to biases in the dataset. SUGARCREPE's dataset addresses the generalization concerns by generating hard negatives in different forms, as detailed above, with minimal difference from the positive captions (to ensure that the hard negatives don't have blatantly obvious chances). As stated earlier, the concerns about bias are addressed through adversarial refinement.

2.5 Baselines

This section covers the baselines achieved by existing vision-languge models on Sugarcrepe. We will first introduce these methods in brief and further illustrate their performance on SugaCrepe. (Hsieh et al., 2023)

1. ViT-B/32 based CLIP (Contrastive Language–Image Pre-training) as introduced in (Radford et al., 2021) is designed to understand and generate connections between textual descriptions and visual content. It does this by learning to map images and their corresponding textual descriptions into a shared embedding space, effectively allowing the model to predict the alignment between unseen images and text. This multimodal interaction enables the model to perform tasks such as zero-shot classification, where the model can classify images into categories it has never seen before during training, by using textual descriptions of those categories.

2. NEGCLIP Finetuned ViT-B/32 as introduced in (Yuksekgonul et al., 2023) is a variation of CLIP. In the NEGCLIP Finetuned ViT-B/32 models, multimodal interactions between image and text are enhanced through the use of hard negatives in the training process. NEGCLIP extends CLIP's original approach by introducing hard negatives—manipulated or alternative text descriptions that closely resemble the original descriptions but are incorrect. This method aims to further refine the model's ability to accurately associate images with their correct textual descriptions, thereby improving its performance on tasks requiring a nuanced understanding of both visual and textual content.

3. Pretrained ViT-B/32 as introduced in (Dosovitskiy et al., 2021) uses a standard pretrained model without any specific modifications for hard negative handling. The Pretrained ViT-B/32 model, in its original form, does not directly facilitate multimodal interactions between image and text in the way models specifically designed for multimodal tasks do, such as CLIP. The Vision Transformer is primarily focused on processing images

Model	Training	Hard Negative Used	ARO+CREPE				SUGARCREPE		
			REPLACE	SWAP	NEGATE	SHUFFLE	REPLACE	SWAP	ADD
Human			95.33	100	99.33	96.00	98.67	99.50	99.00
ViT-B/32	Pretrained	N/A	75.71	71.58	76.89	72.06	80.76	63.27	75.09
	CLIP finetuned	N/A	77.06	68.81	61.19	63.04	84.76	70.83	85.58
	NEGCLIP finetuned	REPLACE	94.51	90.04	85.06	88.15	88.27	74.89	90.16
		SWAP	82.88	94.48	77.57	87.00	85.54	76.21	86.56
		NEGATE	77.24	68.91	99.54	64.28	84.97	70.29	85.84
		Released in [43]	85.72	94.35	83.51	90.45	85.36	75.33	87.29
RN50	CLIP from scratch	N/A	69.93	59.96	55.36	68.78	69.54	60.33	67.63
	NEGCLIP from scratch	REPLACE	89.04	66.51	60.90	75.23	74.32	62.65	72.92
		SWAP	72.33	92.29	64.51	84.84	73.31	68.35	71.93
		NEGATE	70.09	60.29	99.45	69.03	72.74	60.89	70.47
		REP + SW + NEG	86.30	88.60	99.34	82.93	75.26	67.69	73.08

Table 1: Evaluating Pre-existing methods on ARO+CREPE and SUGARCREPE benchmarks based on Accuracy. Source: (Hsieh et al., 2023)

* [43] refers to (Yuksekgonul et al., 2023)

through its architecture, which is adapted from transformers used in natural language processing. However, it does not inherently include mechanisms for integrating or interpreting textual data alongside visual data without further modifications or specific training regimes designed to enable such multimodal interactions.

4. RN50 Trained with CLIP, as introduced in (He et al., 2015) uses a contrastive learning framework to match images with their corresponding text descriptions, thereby learning a shared representation space for both modalities. The training involves directly learning from raw images and textual descriptions without relying on pre-existing embeddings or features, fostering a deep understanding of both modalities.

5. RN50 Trained with NegCLIP, as introduced in (He et al., 2015) introduces an additional complexity by incorporating hard negatives—carefully selected or generated image-text pairs that are similar but not correct matches. This approach aims to refine the model's ability to discern more nuanced differences between images and their associated text, enhancing its comprehension of both modalities.

Refer to Table 1 for a comparison of model performances across different training methods and hard-negative forms. We will be focusing on the SUGARCREPE benchmark.

3 Team member contributions

Geerisha Jain contributed to finding suitable datasets, researching baselines and model performances and led the Exploratory Data Analysis.

Madhura Deshpande contributed to finding suitable datasets and baselines.

Shaurya Singh contributed to creating and formatting the report, formulating the problem definition and researching on methods to generate hard negatives.

Shrey Madeyanda contributed to formulating the problem definition, some Exploratory Data Analysis, metrics and researching methods to achieve strong attribute binding.

References

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020. Cops-ref: A new dataset and task on compositional referring expression comprehension.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe:

Fixing hackable benchmarks for vision-language compositionality.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Theo M.V. Janssen and Barbara H. Partee. 1997. Chapter 7 - compositionality. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 417–473. North-Holland, Amsterdam.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally?

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

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it?