11-777 Report 2: Baselines and Model Proposal

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

1 Baseline Models and Metrics

For our baselines, we have 3 unimodal models, 2 simple multimodal models (with text and image modalities), and 4 competitive baselines, making use of large pretrained models.

The models will either generate outputs which will be compared against the positive and hard negative captions in the SugarCrepe, or will select one of those captions. It's performance will be determined by how often it selects the positive caption.

1.1 Unimodal Baselines

We are experimenting with three unimodal baselines, two using the text modality, and one using an image modality.

1.1.1 Multilayer Perceptron

We developed a model employing a multi-layer perceptron (MLP), using the PyTorch library, to classify captions as either positive or hard-negative. Our training data is sourced from the GenericsKB dataset (Bhakthavatsalam et al., 2020), a comprehensive knowledge base of generic sentences.

To prepare the training dataset, we assigned labels to sentences based on their score values: sentences with $scores \geq 0.5$ were labeled as positive (1), while the rest were labeled as hard-negative (0). To ensure balanced training, we sampled an equal number of positive and hard-negative instances from the dataset, resulting in a final dataset comprising 320,000 sentences, evenly split between positive and hard-negative instances.

After acquiring the training dataset, we employed the BERT Tokenizer to tokenize each sentence, resulting in a 24-token-wide vector to represent individual sentences.

Our MLP architecture comprises an input layer, succeeded by four hidden layers, and culminating in an output layer. This model outputs a score ranging from 0 to 1, serving as a measure of the commonsense correctness of the input caption.

The key insight from running this baseline is that even though MLPs offer a straightforward approach to text classification, their limitations become evident for projects such as ours which necessitates advanced compositional reasoning. We see that both positive and hard-negative captions are syntactically and semantically correct, and thus our model is not able to correctly distinguish between the two, resulting in an accuracy of around 50%.

1.1.2 VisualBERT-based model

We used a custom image captioning model, built upon the foundational principles of VisualBERT (Li et al., 2019) and pre-trained on the Flickr8k dataset. This model takes in only images for its input, and employs a Vision Transformer (ViT) for extracting intricate features from images and a GPT-2 tokenizer for nuanced text processing. The ViT, functioning as the image encoder, analyzes visual content to derive meaningful representations crucial for subsequent caption generation. Meanwhile, GPT-2, rooted in the Transformer architecture, captures intricate dependencies within sequential data. The tokenizer facilitates a seamless conversion of textual information into tokens, which is important for effective model processing. In both training and inference, the tokenizer plays a pivotal role in ensuring captions are appropriately formatted and tokenized. During pre-training, the model was fine-tuned with a beam search configuration, resulting in the elevation of the quality of generated captions.

The performance inconsistency of the model across the measures could be due to the absence of targeted fine-tuning, without which the model

*Everyone Contributed Equally – Alphabetical order

encounters difficulties in reconciling visual and textual information, impeding its ability to grasp the intricate relationships between the two modalities.

1.1.3 Vera

Vera is a general-purpose model that estimates the plausibility of declarative statements based on commonsense knowledge (Liu et al., 2023). It only makes use of the text modality. It has been trained on 7M commonsense statements created from 19 QA datasets and two large-scale knowledge bases. Vera is selecting the caption it considers more plausible, without looking at the image. We observe that it identifies the positive captions at a rate close to chance when run against the Sugarcrepe benchmark. This would indicate that the benchmark achieves its aim of generating hard negatives which are semantically correct and have subtle variations from the positive captions, which nonetheless generate a description different from what is in the image. On benchmarks like ARO+CREPE, language models have been able to outperform state-of-the-art vision-language models.

1.2 Simple Multimodal Baselines

We include 2 simple multimodal baselines, both making use of CNN as encoders and LSTM as decoders. One of the models also makes use of soft attention (making use of the approach detailed in Show, Attention and Tell).

1.2.1 Simple CNN-LSTM model

This model makes use of an early fusion of the image and text representations, and is implemented using the Keras library. For the CNN encocder, we make use of the Xception architecture provided by the library. This is used to extract the feature representations for the input images. We are using its default image size of 299x299, and resizing the input images are required.

For the captions associated with the images, during training, a tokenizer is used to convert the text to a sequence of integers, to be fed to the model.

The decoder is an LSTM layer of 256 units. This layer will take as its input a combined representation of the image and text representations.

For training data, we make use of 25,000 images (sampled from a total of 118K images), and their associated captions, from the COCO 2017 dataset. This curated subset was obtained by making use of coco-minitrain (). The authors mention that the performance of a model trained on this subset is positively correlated with the performance of a model trained on the full set.

The model performs rather poorly across the measures for the SugarCrepe benchmark, indicating that it has poor compositional reasoning. In addition, the captions produced by this model, while somewhat accurate, were not detailed in capturing all relevant details in the image. This indicates that for any model we choose to implement, which improves compositional understanding of an image, we need some alignment between the image and text representations, to provide a mapping between the object-attributes in the image and the text.

1.2.2 Enhanced CNN-LSTM with Visual Attention Model

This model inspired by the Show, Attend and Tell approach, introduces an enhanced attention mechanism to the task of image captioning, and is implemented using the PyTorch library.

For the encoder part of the model, we make use of the VGG19 architecture, pre-trained on ImageNet, to extract feature representations from input images. Images are resized to the default input size of VGG19 (224x224 pixels) before being processed. The VGG19 encoder transforms each input image into a set of feature maps that serve as the basis for attention-driven caption generation.

For the decoder part of the model, we use an LSTM network with 256 units, designed to generate captions based on the features extracted by VGG19 and the context provided by the attention mechanism.

We implement a soft attention mechanism, as described in Show, Attend and Tell, which allows it to learn to focus on different parts of the image at different steps of the caption generation process. This approach enhances the model's ability to describe images accurately by paying selective attention to details relevant to the caption being generated.

The model is trained on the COCO 2017 dataset, using a subset of 25,000 images and their associated captions. This curated subset was obtained by making use of coco-minitrain (Samet et al., 2020). Similar to the simple CNN-LSTM model, images are resized to fit the input requirements of the VGG19 encoder, and captions are tokenized into sequences of integers. The training process involves optimizing the alignment between the visual features, attention weights, and the generated captions to improve caption quality. While this approach did not yield any improvement in most of

the measures, there was a significant improvement in accuracy. when measured against the ADD form of hard-negatives. It indicates that attention is effective in identifying new objects and attributes in the caption, which were previously not present.

1.3 Competitive Baselines

We run 4 publicly available competitive baselines. These include:

1. CLIP ViT-B/32 (Radford et al., 2021) uses a ViT-B/32 Transformer architecture as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss. The model is finally tested on Sugarcrepe by feeding it an image and corresponding positive and hard-negative caption. The model returns a similarity score between the image and both the captions. The accuracy is calculated as a ratio between the number of times the positive caption was more similar to the image as compared to the negative caption and the total number of testing instances. We see a decline in model performance in the Swap and Add category, indicating that there's no binding between the attributes and objects in the text. Changing relative positions of objects and attributes, and adding new attributes and objects results in a poor performance. The key insight that we can conclude based on the results is that keeping a track of which attributes are bound to which objects, and the relative positions of those objects and attributes in the caption is important for better performance.

2. RN50 (He et al., 2015) is a Residual Network consisting of 50 layers. It consists of residual blocks with skip connections that allow gradients to flow through the network more effectively during training. The key insight from running this baseline is that RN50 is very effective at identifying existing objects in the image, and mapping them to the captions i.e. if an existing object gets replaced in the caption, the model identifies that, and selects the positive caption. However, it struggles if new objects are added, or if the existing objects are swapped in the caption text. There is also degraded performance in the benchmark when it comes any changes made in the relations (for example, replacing "running" with "walking").

3. RoBERTa-ViT-B-32 integrates RoBERTa's (Liu et al., 2019) NLP capabilities with the Vision Transformer's (ViT-B-32) (Dosovitskiy et al., 2021) image analysis. It is trained on the LAION-2B dataset and scaled to 12 billion parameters. It is designed to perform well in both understanding complex textual content and performing intricate visual analyses. The key insight from running this baseline is that RoBERTa-Vit-B is very effective in incorporating new objects into captions and adjusting to changes in object relationships within the text, that is, it shows a significant leap in adaptability and understanding. However, the captions generated followed a common pattern that was seen during the training. In some cases, it lacked creativity or failed to capture more abstract or nuanced aspects of the images.

4. CLIP Enhance-FineGrained, based on (Zhang et al., 2023) uses two loss functions to enhance the compositional understanding ability for any contrastive vision-language models loss like CLIP. This method significantly improves compositional reasoning in Vision-Language Models by refining the image-text contrastive learning framework through intra-modal contrast and cross-modal rank objectives. Intra-modal contrast enhances image-text alignment within the same modality, while cross-modal rank improves the model's understanding by ranking correct pairs higher than incorrect ones across modalities. This enhancement results in notable performance gains on challenging benchmarks like SugarCrepe, where the model excels in tasks such as adding objects, adding attributes, replacing objects, and replacing attributes. The substantial accuracy improvements demonstrate enhanced visio-linguistic understanding and reasoning capabilities. These results underscore the efficacy of the proposed approach in advancing compositional reasoning in Vision-Language Models, particularly in tasks requiring nuanced semantic variations and fine-grained image-text alignment.

Table 1: Results Table (Accuracy %)

Methods	REPLACE			SWAP		ADD	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
Multilayer Perceptron (Text only)	49.16	49.25	50.07	46.09	49.63	50.27	48.79
VisualBERT-based model (Image only)	60.35	52.53	56.04	60.00	65.46	43.74	48.55
Vera (Liu et al., 2023)	51.45	50.89	51.06	50.20	51.20	51.06	51.44
Simple CNN-LSTM model	48.79	46.94	54.59	47.75	45.19	22.59	39.65
CNN-LSTM with Visual Attention (Xu et al., 2016)	49.69	18.65	49.93	35.99	18.77	63.92	83.53
CLIP ViT-B/32 (Radford et al., 2021)	90.74	80.33	69.42	61.22	64.11	77.01	69.51
RN50 (He et al., 2015)	91.76	80.58	69.91	62.04	68.47	74.49	69.80
RoBERTa-ViT-B-32 (Liu et al., 2019)	92.85	84.89	72.40	62.85	71.02	87.34	79.91
CLIP Enhance-FineGrained (Zhang et al., 2023)	93.09	88.83	79.01	73.06	77.02	92.38	93.35

2 Results

We measure the performance of our selected baselines against the benchmark provided by Sugarcrepe, to check their compositional reasoning. The results are illustrated in the Table 1 below.

For the models using both image and text modalities, we check the cosine similarity of the generated captions against the positive and hard negative captions provided for that image by Sugarcrepe. If the similarity is higher for the positive caption, it will be considered an accurate result. For text-only models, our objective is to assess their ability to classify captions as either positive or hard-negative based on commonsense reasoning.

We will be evaluating this across 3 forms of hard negatives: Replace, Swap and Add.

REPLACE Accuracy The hard negative captions will have objects, attributes and relations (OARs) from the positive caption replaced with different values. This measure will check the model's ability to be able to just identify the OARs correctly (since replacing these in the captions will remove the original entirely). We expect a high score in this measure for the model we implement when it comes to replacing objects, since existing benchmarks perform rather well in this regard. However, those baselines show a degraded performance when it comes to replacing attributes and relations, and it is expected our proposed model will show an improved performance in this regard.

SWAP Accuracy The hard negative captions will have objects and attributes (OAs) from the positive caption swapped with other OAs from the caption. Swapping the relations will not result in semantically coherent sentences in a lot of cases, so this will not be done here. This measure will provide a more robust estimation of the model's ability to recognize the object-attribute bindings in the image, as compared to replacing the values with new ones, as all the objects and attributes originally present in the image, will still be present in the hard negative caption, albeit presented in a different order/context. The baselines show a degraded performance when it comes to these SWAP measures, and our proposed approach should show an improvement in this measure, if it achieves improved compositional reasoning as expected.

ADD Accuracy The hard negative captions will have new objects and attributes (OAs) in addition to the existing OAs in the positive caption. Again, adding relations might not result in sensible captions, so it will not be done here. It will allow us to determine if the model can identify the presence of new (and incorrect) OAs, even if the existing ones in the image are still present in the caption. The baseline show a good performance when it comes to identifying added objects, and a somewhat degraded performance in identifying added attributes. Our proposed approach should show a similar performance in identifying added objects, and improved performance in identifying additional attributes, if it as achieved improved compositional reasoning.

3 Model Proposal

In this section we propose our model using which we aim to solve the given problem. Figure 1 illustrates a labeled flow chart for the model.

3.1 Overall model structure

In this project, we propose an approach for a visual-language model with improved compositional reasoning. The problem we are trying to solve is to give our model the capability to identify slight vari-

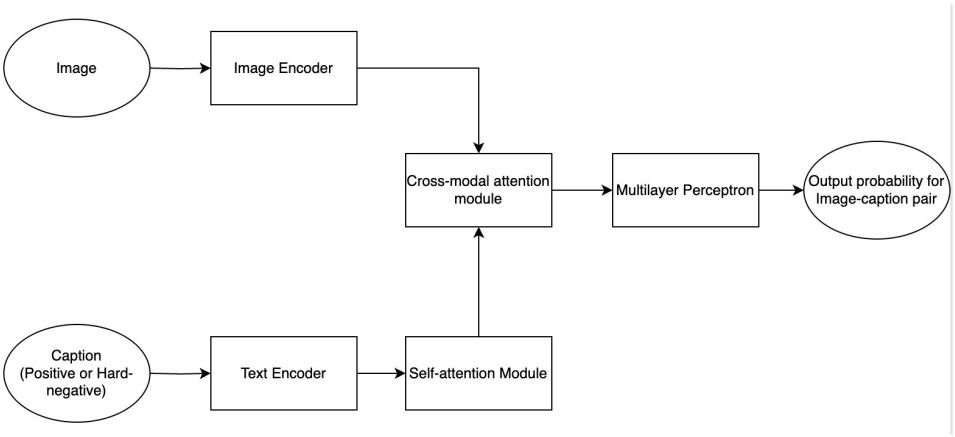

Figure 1: Model Proposal

ations in text associated with an image. Though these slight variations result in syntactically and semantically correct texts, we see that images corresponding to these texts would be fundamentally different. Current visual-language models are not able to reason well enough to identify these slight but significant variations to the captions.

The aim of our proposed model is to apply compositional reasoning to image-caption pairs, by creating a strong binding between the objects and attributes in the caption, and attempting to align those with the corresponding image representation. The idea is to learn a shared representation, which when passed through a classifier will allow it to select the positive captions and reject any associated hard-negative captions.

The model will consist of an image encoder to generate vector representations for the inputs images, and a text encoder to generate vector representations for the corresponding (positive and hard negative) captions for the image. Self-attention on the vector representation of the captions will be used to focus on the spatial relationships between the objects and attributes and create a stronger binding between them. For a caption, "The hat on the floor is red," we will get a strong binding between the object "hat" and the attribute "red."

The output from the self-attention module and image encoder will be fed as input to a cross-attention module to align the two representations. Finally, this shared representation will be the input for an MLP classifier, which will be used to predict the caption best representing the image (representation), from options containing positive and hard captions.

3.2 Encoders

3.2.1 Image Encoders

For image encoding, our primary choice is the Vision Transformer (ViT), as it splits images into patches and processes these through transformer blocks, enabling nuanced understanding of spatial hierarchies and object relationships within images. As an alternative, we consider ResNet-50, as it uses residual connections to enable deeper networks without the vanishing gradient problem, excelling in hierarchical feature extraction across various image types.

Comparison In terms of coverage, ViT offers exceptional coverage in identifying and interpreting complex visual patterns and relationships across the entire image due to its attention-based mechanism. It can analyze parts of an image in relation to the whole, capturing both minute details and global contexts. ResNet-50, although highly capable, primarily excels at recognizing patterns through its hierarchical feature extraction process, which might not capture the global context as effectively as ViT.

Efficiency In terms of efficiency, while ViT requires substantial computational power, its efficiency is also measured in its ability to rapidly adapt to new visual domains with minimal fine-tuning, thanks to its attention mechanism. ResNet-50, known for its structural efficiency, is optimized for fast feature extraction across various computing platforms, making it highly reliable and accessible for a wide range of applications.

3.2.2 Text Encoders

For text encoding, our primary selection is BERT (Bidirectional Encoder Representations from Trans-

formers), known for its deep understanding of context by processing text in both directions simultaneously, providing a comprehensive grasp of language nuances. As an alternative, we consider LSTM (Long Short-Term Memory) networks, which captures long-term dependencies in text sequences well, making them adept at maintaining narrative flow and context over extended passages.

Comparison In terms of coverage, BERT stands out for its ability to capture the full context of language, making it exceptionally effective in understanding the complex subtleties and variations in text. Its bidirectional nature allows it to grasp the meaning embedded in the entirety of a sentence or passage, far surpassing the capabilities of traditional unidirectional models. On the other hand, LSTM, while proficient in handling sequential data and capturing long-term dependencies, may not fully encapsulate the bidirectional context as comprehensively as BERT, potentially limiting its effectiveness in scenarios where understanding the immediate context is crucial.

Efficiency In terms of efficiency, BERT's architecture allows it to perform well in complex linguistic analysis without the necessity for extensive customization or model-specific adjustments, significantly reducing the need for extensive retraining while maintaining high levels of accuracy and adaptability. Meanwhile, LSTM's efficiency is notable for its simplicity and lower resource requirements, offering a viable alternative for completing our project with limited computational resources or in a situation where we have to prioritize sequential data processing over contextual depth.

3.3 Loss Functions

The model's final task will be multi-class classification, where the model will need to select the correct (positive) caption from a options that include hard-negatives. For this, we will make use of the cross-entropy loss function (our primary task loss), which is the standard for this task.

We will also need auxiliary loss functions for the sub-tasks that the model will need to perform. Firstly, for the self-attention module, we propose using the loss function from (Rassin et al., 2024), where they minimize that distance (maximize the overlap) over all pairs of modifiers (attributes) and their corresponding entity-nouns (objects).

For the cross-attention module, we propose using the Cross-modal rank loss (built on hinge loss) put forward in (Zhang et al., 2023), It employs a ranking loss with a threshold, which aims to keep a minimum semantic distance between true and hard negative image-text pairs. In addition, we can also consider using constrastive loss between the image-text pairs.

4 Team member contributions

Geerisha Jain contributed to running multimodal and competitive baselines, and wrote the related sections of the report. Contributed to the encoder section of the report.

Madhura Deshpande contributed to running the unimodal and competitive baselines and wrote the related sections of the report. Contributed to the loss function section of the report.

Shaurya Singh contributed to running the unimodal and competitive baselines, and wrote the related sections of the report. Contributed to the model proposal, and loss functions sections of the report.

Shrey Madeyanda contributed to running multimodal and competitive baselines, and wrote the related sections of the report. Contributed to the results metrics, model proposal, and loss functions sections of the report.

References

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements.

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.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

Nermin Samet, Samet Hicsonmez, and Emre Akbas. 2020. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision (ECCV)*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2023. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding.