CS 6384 Project Final Report

Kevin Tak Hay Cheung, Nithin Pingili, Jerry Xu
The University of Texas at Dallas
800 W. Campbell Road, Richardson, TX 75080
{KevinTakHay.Cheung, Nithin.Pingili, jerry.xu}@utdallas.edu

Abstract

In this project, we focus on the Referring Expression Comprehension (REC) task: Given an image and a set of textual expressions, we aim to return a bounding box surrounding the target described by the expressions. Solving the REC task is difficult because it involves leveraging techniques from both Computer Vision and Natural Language Processing. To solve this task, we built and trained our own models and fine-tuned a pre-trained model.

1. Introduction

Language and vision are two fundamental components of our daily life. We utilize referring expressions on a regular basis in our social and professional interactions. For instance, we might say something like, "Please bring me the tall yellow cup on the counter." Despite significant progress in bridging the gap between computer vision and natural language processing, comprehending referring expressions still poses a challenge. This challenge arises because it requires a thorough comprehension of complex language semantics and a variety of visual information, such as object relationships, attributes, and regions.

In referring expression comprehension tasks, objects in an image are often searched based on their category, attribute, and context. This search involves a deep understanding of both the language used in the referring expression and the visual information conveyed by the image. One of the most critical aspects of referring expression comprehension is ensuring that the expression is unambiguous. This is especially important when an image contains several examples of the same object category, and the referring expression must be precise enough to identify the exact instance being referred to.

Using natural language to encode information in referring expressions is incredibly valuable because it can convey a wealth of information that goes beyond simply identifying an object. These associations can help distinguish one object instance from another and provide context for the relationship between different objects in the image. Achieving this level of understanding requires the integration of natural language processing and computer vision techniques, which is a challenging but exciting area of research that has the potential to revolutionize the way we interact with visual information.

We explored three approaches to solving the REC task: building and training an Encoder-Decoder model, an Encoder-Only model, and fine-tuning a pre-trained model called FLAVA [13].

2. Related Work

Approaches that others have used to solve the REC task include:

2.1. CNN + LSTM

CNNs are able to generate rich image representations by embedding images into fixed-length vectors. These vectors can then be used for various visual tasks [12]. LSTMs are commonly used in text encoding and have shown good performance on various sequential modeling tasks [4].

2.2. Attention

Attention allows the model to focus only on certain parts of the input when processing a large amount of information. Attention is able to build element-wise connections between visual and textual information. Thus, the model utilize the information from some specific region in the image when encoding each word in the text and vice versa, leading to semantically-enriched visual and textual representations [9].

2.3. Modular Models

Modular models parse the referring expression into three phrase embeddings: subject, relationship, and location. Three visual modules then process these embeddings separately, and the model outputs a weighted sum of the module scores [14].

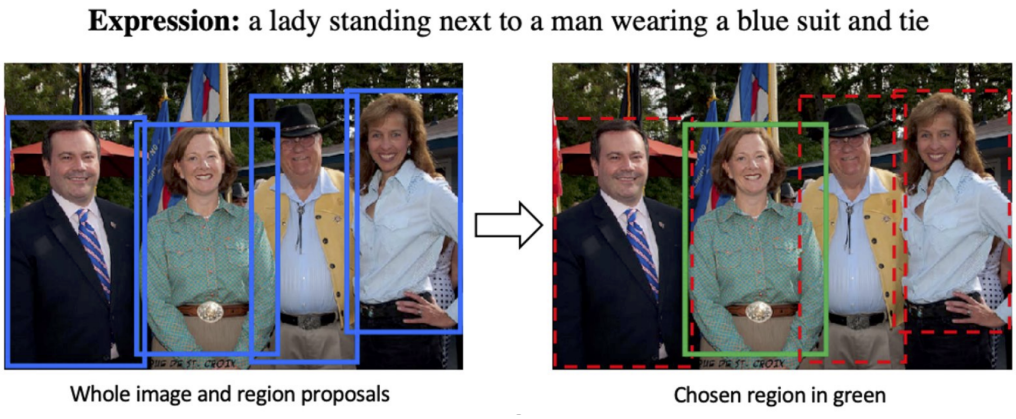

Figure 1. An example of the Referring Expression Comprehension task. Given an referring expression and an image, REC aims to localize the referential entity [11]. The blue boxes are candidates and the green box is the ground truth target.

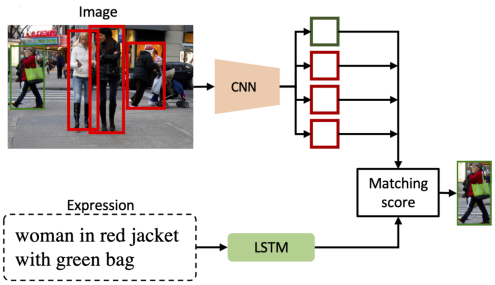

Figure 2. Solving the REC task using CNN + LSTM.

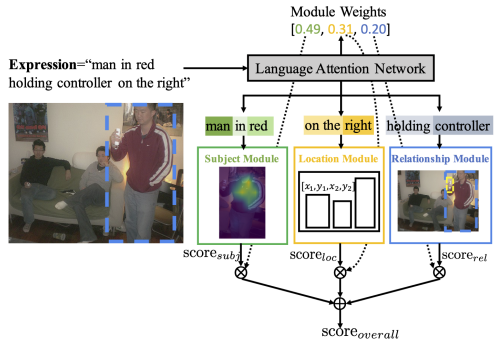

Figure 3. Solving the REC task using a modular model. Note that attention is used here.

2.4. Pre-Trained Models

The main approach here is to utilize separate pre-trained vision and language models. However, this approach has drawbacks, namely an inability to learn joint visual-language representations. This can result in poor generalization abilities. One approach to resolve this issue is ViLBERT (Vision and Language BERT), which extends BERT to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers [8].

3. Method

3.1. Encoder-Decoder Model

This model is a simplified version of the model presented the paper "Visual Grounding with Transformers" by Du *et al.* [3]. We used visual and textual backbones to preprocess the images and expressions. We used ResNet50 [5] as our visual backbone and uncased BERT [1] as our textual backbone. Next, these preprocessed inputs were used as inputs to visual and textual encoder branches respectively. The entire encoder structure can be stacked up to n times, where n is a user-chosen parameter.

In the Textual Encoder, the preprocessed text embeddings were passed through Multihead Attention layers, Add & Norm layers, and Linear layers. The outputs of the Textual Encoder were used as inputs to the Decoder and as inputs to the fusion module in the Visual Encoder.

In the Visual Encoder, the preprocessed visual embeddings were fused with the output of the Textual Encoder within the fusion module. The fusion module is a simple Multihead Attention mechanism that uses the visual embeddings as queries and textual embeddings as key-value pairs. The joint embeddings were then passed through Add & Norm layers and Linear layers. The new multimodal embedding was then used as input to the Decoder.

In the Decoder, the textual embeddings were passed through Multihead Attention layers and Add & Norm Layers. Next, the textual embeddings were jointly trained with the visual embeddings with a Multihead Attention layer. The new embeddings were then passed through more Linear layers and Add & Norm layers. Like the encoder, the decoder structure can be stacked up to n times, where n is

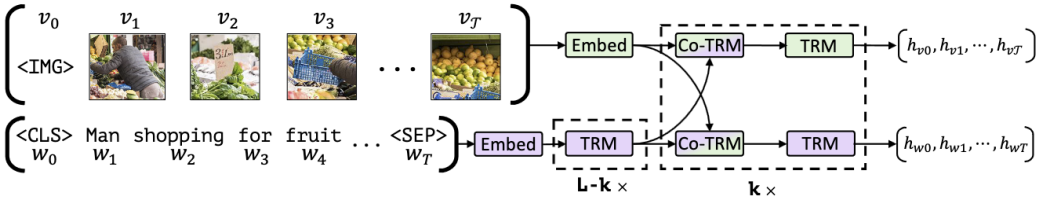

Figure 4. The ViLBERT architecture. It consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers [8].

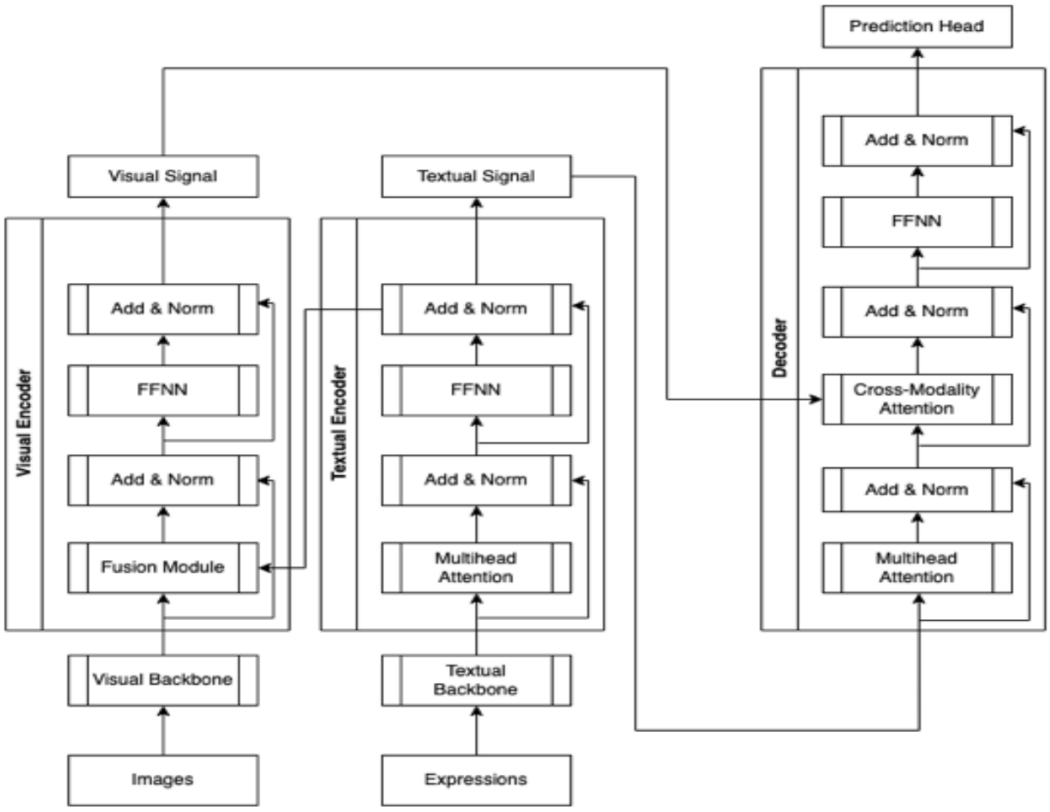

Figure 5. The architecture of the Encoder-Decoder model.

a user-chosen parameter.

The prediction head is a 3-layer neural network composed of linear layers. The first two layers consist of 512 units and use the ReLU activation function. The final layer consists of 4 units and uses the Sigmoid activation function. The prediction head predicts the four parameters of a bounding box, (x, y, w, h), where (x, y) are the coordinates of the top left corner of the bounding box and (w, h) are the dimensions of the bounding box.

3.2. Encoder-Only Model

This model is a simplified version of the Encoder-Decoder model. Here, we also passed the image and text into the visual and textual backbones and produced the visual and textual embeddings. Next, we concatenated the two embeddings into a joint embedding. This joint embedding was passed to the encoder and passed though Multihead Attention layers, Add & Norm layers, and Linear layers. Similar to the Encoder-Decoder model, we can stack the entire structure n times. Finally, the multimodal embedding was passed through a prediction head, which was described in the previous section.

3.3. FLAVA

FLAVA [13] (Foundational Language and Vision Alignment) is a model that aligns textual and visual information

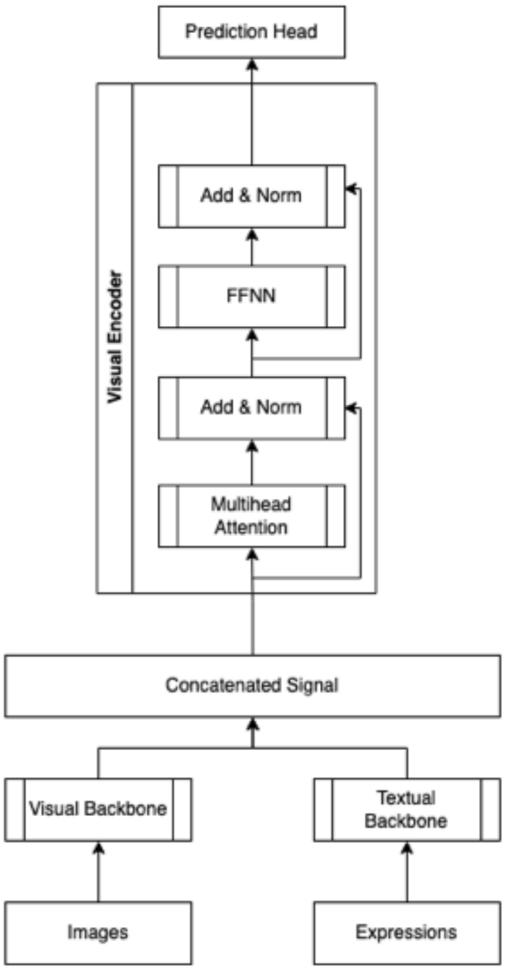

Figure 6. The architecture of the Encoder-Only model.

by jointly learning representations of text and image features. FLAVA uses ViT [2] to extract unimodal image representations and unimodal text representations.

FLAVA then uses ViT to align the text and image embeddings and create a combined representation of the input text and image. This joint representation is then used to predict a set of semantic labels, which describe the relationship between the input text and image.

FLAVA also makes use of a multi-level attention mechanism, which focuses on various granularities in the input text and image in order to enhance the alignment between text and image. Additionally, FLAVA is trained using a novel contrastive loss function that encourages the model to learn more discriminative representations of the input text and image.

FLAVA cannot take raw images and textual expressions as input. We first fed the images and expressions into a preprocessing module. The outputs of the preprocessing module were used as inputs to FLAVA. The raw outputs of FLAVA were then paseed through a prediction head. The prediction head is the same as one used in the Encoder-Decoder and Encoder-Only models. During training, we froze the layers in FLAVA and only allowed the parameters of the prediction head to update.

4. Experiments

4.1. Dataset

In this project, we used the RefCOCO family of datasets: RefCOCO [6], RefCOCO+ [6], and RefCOCOg [10]. These datasets are subsets of the MSCOCO (Microsoft Common Objects in Context) dataset [7].

RefCOCO and RefCOCO+ were collected in an interactive game-based setting (Amazon Mechanical Turk) while RefCOCOg was collected in an non-interactive setting. RefCOCO+ expressions are strictly appearance-based descriptions, meaning that "person to the right" is an invalid description. The average RefCOCOg expression has 8.4 words while the average RefCOCO and RefCOCO+ expressions have 3.5 words. Object categories in these datasets include person, vehicle, animal, accessory, sports, kitchen, food, furniture, electronics, etc.

For RefCOCO, the training set contains 40,000 referring expressions along with 19,213 images. The validation and test set both contain 5,000 referring expressions and 4,500 images.

For RefCOCO+, the training set contains 42,278 referring expressions along with 16,992 images. The validation set contains 3,805 referring expressions and 1,500 images. The test set contains 1,975 referring expressions and 750 images.

For RefCOCOg, the training set contains 42,226 referring expressions along with 21,899 images. The validation set contains 2,573 referring expressions and 1,300 images. The test set contains 5,023 referring expressions and 2,600 images.

An image in each of these three datasets may have multiple bounding boxes, and each bounding box may have multiple expressions. We only kept one expression per bounding box due to resource constraints. This resulted in about 180,000 image-text pairs. We selected a random subset of 50,000 image-text pairs as our dataset, again due to resource constraints. We used 40,000 image-text pairs for training and 10,000 image-text pairs for testing.

4.2. Evaluation Metric

We used Intersecion over Union (IoU) as our evaluation metric. In the field of object detection, IoU is a commonly used evaluation metric that helps assess the accuracy of object detection algorithms. It is defined as the area of overlap between the predicted bounding box and the ground truth bounding box divided by the area enclosed by the pre-

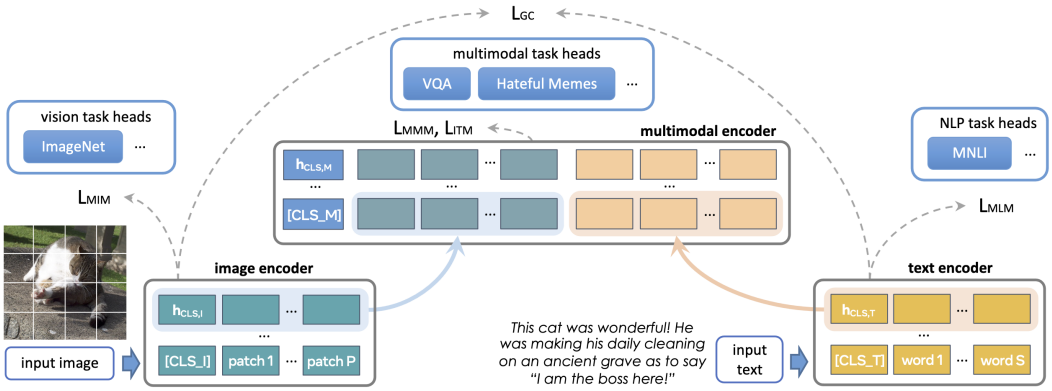

Figure 7. An overview of the FLAVA model, with an image encoder transformer to capture unimodal image representations, a text encoder transformer to process unimodal text information, and a multimodal encoder transformer that takes as input the encoded unimodal image and text and integrates their representations for multimodal reasoning.

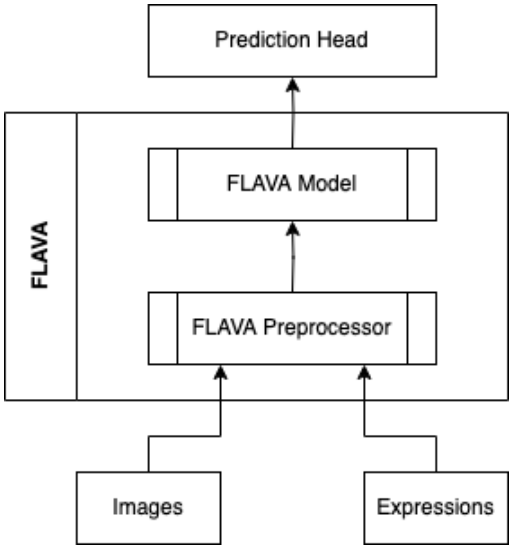

Figure 8. The architecture of the FLAVA model.

Figure 9. An image from the RefCOCO dataset. The expressions for the image were "biggest monitor," "front monitor," and "the monitor dead center apple logo." The green box represents the ground truth bounding box.

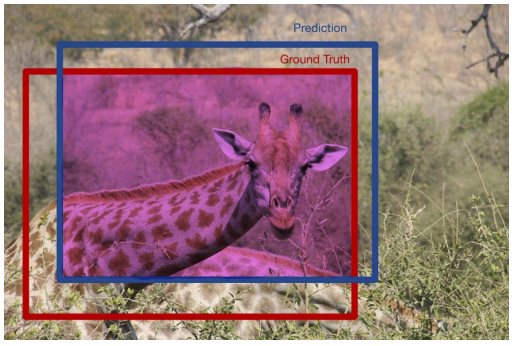

Figure 10. An illustration of the IoU metric.

dicted bounding box and the ground truth bounding box. In most cases, an IoU greater than 0.5 is considered a "good" prediction, as it indicates a significant degree of overlap between the predicted and ground truth bounding boxes. However, this threshold can vary depending on the specific application and context. For example, in certain scenarios where high precision is required (such as self-driving cars), a higher IoU threshold may be necessary to ensure accurate object detection.

4.3. Results

4.3.1 Training Results

We trained the Encoder-Decoder model for 10 epochs with the AdamW optimizer (learning rate = 1e-4 and weight decay = 1e-3) on Paperspace Gradient, the Encoder-Only model for 10 epochs with the AdamW optimizer (learning rate = 1e-5 and weight decay = 1e-2) on Paperspace Gradient, and the FLAVA model for 20 epochs with the AdamW optimizer (learning rate = 1e-4 and weight decay = 1e-2)

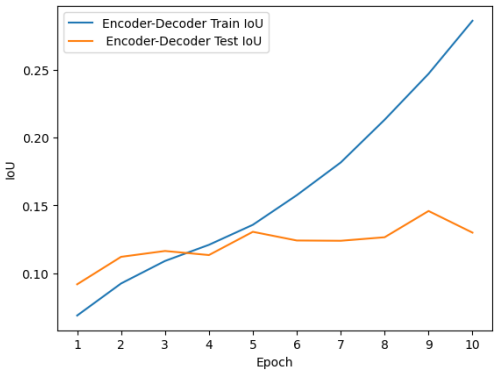

Figure 11. Encoder-Decoder training graph.

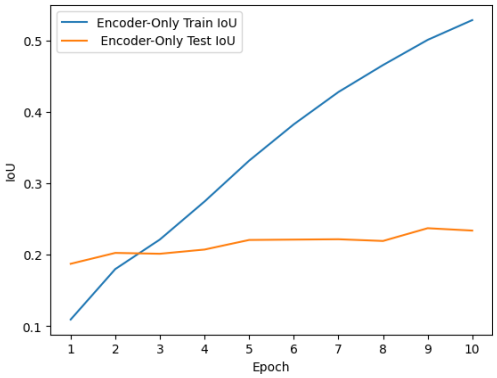

Figure 12. Encoder-Only training graph.

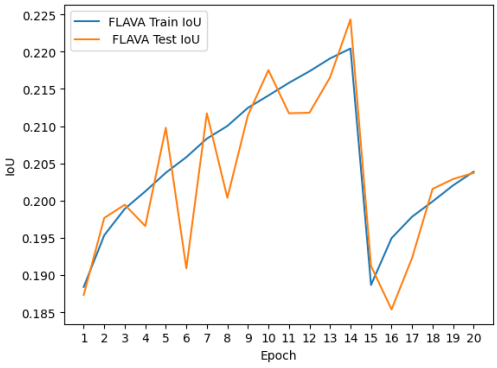

Figure 13. FLAVA training graph. The drop in IoUs at epoch 15 is due to us needing to re-load the model parameters following a Google Colab runtime disconnection.

on Google Colab. The training graphs are shown in Figures 11 to 13. From Figures 11 and 12, it is apparent that the Encoder-Decoder and Encoder-Only models are overfitting. We believe that this is due to the relatively small dataset that these models were trained on. 40,000 image-text pairs are nowhere near enough for the network to distinguish the objects in the images and spatial relationship

Model	Training IoU	Testing IoU
Encoder-Decoder	0.286079	0.129851
Encoder-Only	0.528473	0.233631
FLAVA	0.203883	0.203772

Table 1. Training and Testing IoUs at the end of training.

between the objects. From Figure 13, it is apparent that the FLAVA model is not overfitting. The Training and Testing IoUs are strongly correlated to each other. We believe that this is due to the fact that FLAVA was pre-trained on 70 million examples. Therefore, FLAVA contains much more knowledge than the other two models.

4.3.2 Qualitative Bounding Box Analysis

For the leftmost image in Figure 14, the expression is "man in back of surfboard." Here, the Encoder-Decoder model has learned the concept of "back of surfboard" but not the concepts of "man" or "woman."

For the middle image in Figure 14, the expression is "a man wearing black t shirt and holding a tennis ball in his hand." Here, the Encoder-Only model has learned the concept of "person," but not the concepts of "black t shirt" and "tennis ball."

For the rightmost image in Figure 14, the expression is "the whole pizza with fresh greens." Here, the FLAVA model has learned the concept of "pizza with fresh greens" but not the concept of "the whole pizza with fresh greens."

For us to have more accurate bounding boxes, all three models need to be trained for more epochs. Additionally, the Encoder-Decoder and Encoder-Only models also need to be trained on more data.

5. Conclusion

In this project, we have demonstrated the use of an Encoder-Decoder model, an Encoder-Only model, and a fine-tuned FLAVA model to localize a target object described by a referring expression phrased in natural language in an image. While the Encoder-Only model has the highest training and testing IoUs, it shows signs of overfitting. While the FLAVA model had the second-highest testing IoU, it shows no signs of overfitting so far. We believe that this is due to the fact that FLAVA was pretrained on a large dataset of 70 million examples. It was interesting to see that the Encoder-Only model performed better than the Encoder-Decoder model even though it is a simplified version of the Encoder-Decoder model. However, both the Encoder-Decoder model and the Encoder-Only model are overfitting. We believe that a lower learning rate can help here. Overall, all three models need to be trained for more epochs and more image-expression pairs to achieve better performance.

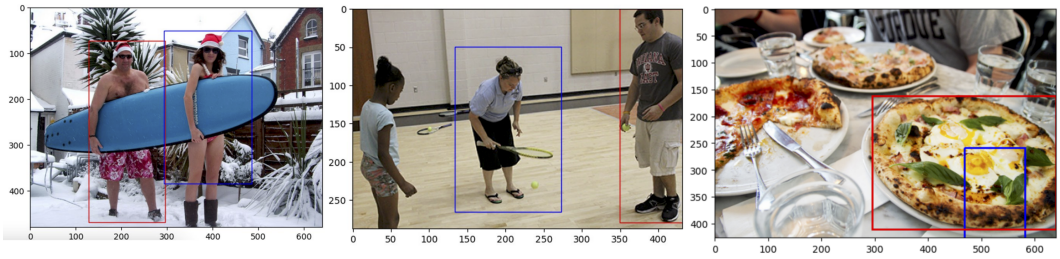

Figure 14. Sample results from the Encoder-Decoder model, the Encoder-Only model, and the FLAVA model respectively. The red bounding box is the ground truth bounding box, and the blue bounding box is the predicted bounding box.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[2] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[3] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. 2

[4] Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, oct 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016. 2

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 4

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. 4

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. 2, 3

[9] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation, 2020. 1

[10] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[11] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *CoRR*, abs/2007.09554, 2020. 2

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 1

[13] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 1, 3

[14] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension, 2018. 1