Image Description

Presented By:

Muhammad Humam Choudhary - 210201080

Muhammad Rafay Qureshi - 210101035

Ariadne

Image Caption and Descriptor AI

Abstract

This study introduces an innovative image description model that uses diverse and complementary visual understanding datasets to create rich and precise captions. This model is trained on a mix of widely-used datasets including Flickr30k, Intrinsic Images in the Wild (IIW), ADE20K, MS COCO, and Visual Genome. Each dataset has its unique strength: Flickr30k specializes in human activities, IIW focuses on lighting and material properties, ADE20K offers fine-grained scene parsing, MS COCO provides object detection and segmentation, and Visual Genome includes dense visual relationships. This combination allows our model to capture a wide variety of visual elements and their interactions, leading to descriptions that are not only accurate but also detailed, covering objects, attributes, spatial relationships, and scene context. Experimental results show that our model surpasses existing methods in both quantitative metrics and human evaluation, proving the effectiveness of integrating diverse visual understanding datasets for image captioning.

Introduction

The ability to automatically generate accurate and descriptive captions for images is a fundamental challenge in computer vision, with wide-ranging applications in accessibility, content retrieval, and human-computer interaction. While significant progress has been made in image captioning over the past decade, current models often struggle to produce descriptions that are both semantically accurate and rich in detail. Many systems excel at identifying primary objects but fall short in capturing the hidden aspects of an image, such as human activities, scene context, and object relationships.

This limitation can be largely attributed to the narrow scope of training data. Most state-of-the-art models are trained on a single dataset, typically MS COCO or Flickr30k, which, despite their size and quality, have inherent biases and limitations in their visual coverage. For instance, MS COCO excels in object detection and segmentation but offers less insight into human activities or scene dynamics. Conversely, Flickr30k is rich in human activity descriptions but provides less information about object attributes or spatial relationships.

We posit that to generate truly comprehensive image descriptions, models must learn from a diverse array of visual understanding tasks. Each specialized dataset in the computer vision community offers unique insights: Intrinsic Images in the Wild (IIW) for understanding lighting and material properties, ADE20K for fine-grained scene parsing, and Visual Genome for dense visual relationships. By integrating these complementary datasets, we can train models to perceive and articulate a broader spectrum of visual elements.

In this study, we introduce a novel image description model that leverages the collective strengths of multiple, diverse datasets. Our approach moves beyond the conventional single-dataset paradigm, combining Flickr30k, IIW, ADE20K, MS COCO, and Visual Genome to create a more holistic understanding of visual scenes. We hypothesize that this multi-dataset strategy will enable our model to generate captions that are not only accurate in object identification but also rich in detail, capturing the full context of an image from object attributes to spatial relationships and scene dynamics.

Related Work

In this section, we provide relevant background on previous work in image caption generation and attention mechanisms. Recently, several methods have been proposed for generating image descriptions, many of which are based on recurrent neural networks (RNNs) and inspired by the successful use of sequence-to-sequence training in machine translation (Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; Kalchbrenner & Blunsom, 2013). This encoder-decoder framework is well-suited for image captioning, as it is analogous to "translating" an image to a sentence.

Kiros et al. (2014a) pioneered the use of neural networks for caption generation, using a multimodal log-bilinear model biased by image features. They later extended this work (2014b) to allow for both ranking and generation. Mao et al. (2014) adopted a similar approach but replaced a feedforward neural language model with a recurrent one. Vinyals et al. (2014) and Donahue et al. (2014) utilized RNNs based on long short-term memory (LSTM) units (Hochreiter & Schmidhuber, 1997). Notably, Vinyals et al. (2014) showed the image to the RNN only at the beginning, differing from Kiros et al. (2014a) and Mao et al. (2014) who presented the image at each time step.

Most of these works represent images as a single feature vector from a pre-trained convolutional network's top layer. In contrast, Karpathy & Li (2014) proposed learning a joint embedding space for ranking and generation, scoring sentence and image similarity based on R-CNN object detections and bidirectional RNN outputs. Fang et al. (2014) introduced a three-step pipeline incorporating object detections, applying a language model to detector outputs, followed by rescoring from a joint image-text embedding space.

Prior to neural network approaches, two main methods were dominant: generating caption templates filled in based on object detections and attribute discovery (Kulkarni et al., 2013; Li et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Elliott & Keller, 2013), and retrieving similar captioned images from a database, then modifying these captions to fit the query (Kuznetsova et al., 2012; 2014). These methods have since been overtaken by neural network approaches.

The use of attention in neural networks has a long history, with recent work sharing our spirit including Larochelle & Hinton (2010), Denil et al. (2012), Tang et al. (2014), and Gregor et al. (2015). Our work directly extends the attention mechanisms proposed by Bahdanau et al. (2014), Mnih et al. (2014), Ba et al. (2014), and Graves (2013).

One of the pioneering works in this field is the "Show and Tell" (Vinyals et al. 2015.) model. This model generates captions for images using Convolutional neural network for extracting features from images while the Long short-term memory network works on caption generation. While this proved its value by generating accurate captions but the generated captions lacked details.

Extending this the “Show, Attend and Tell” (Xu et al 2015) model used a attention mechanism that focuses on different parts of image at the time of generation which results in more accurate captions. However this model lacks in generating complex relation between different objects and their interactions

Another notable contribution is done by “BottomUp and Top Down Attention” (Anderson et al 2018) model. This model combines both bottom up and top down attention mechanism. While the bottom up selects relevent regions the top down weights the importance of this region. While this model generates more detailed captions however it struggles in object recognition.

Recently the transformer-based architecture was presented for this purpose showing promising results. The “Object Relational Transformer” (Herdade et al. 2019) model uses a self-attention mechanism to understand relations between objects, improving the generated caption. However this model struggles with large images or low end systems as it is computationally expensive hence in lower accuracy and slower output.

The model of “Hierarchical Question-Image Co-Attention” (Lu et al. (2016)) utilizes the Visual Question Answering (VOQ) to guide the model to where to look. This approch shows better result however it relies on the knowledge base it has.

Methodology

model V1

This intial model’s architechure was purposed and hand crafted however due to resource and time constraints was not able to implement or train. The model was divided into multiple subsection,

Image feature Extraction:

> The model used for this purpose was resnet 101, Due to its success in object recogniton this was chosen. Firstly this model will be trained on MS COCO dataset for object detection. Then it will be fine tune on Visual Genome Dataset first for object detection, then Object relation and at last object attribution.

Text Generation:

Transformer based model will be used for Text generation, Transformers were chosen due to their internal multiattention architechure. The Transformer will first be trained on LLAMA's dataset for learning sequnce generation of text.

Merging:

At last the model will contain a bidirectional LSTM for merging the Feature extraction and Text generation model. The Feature extracted will be concatenated with text generated from Transformer, and will generate a final output.

Model V2 - V4 are skipped due to their lower accuracy

Model V5:

This model was a cleaner version of v2 - v4, This model is divide into three parts, Encoder, Attention, And Decoder.

Encoder used was a pretrained resnet 50 model. Its main purpose was to extract features from a iimage.

Attention mechanisme was added as the transformer was removed. The purpose of attetion was to extract main defining objects and remove any noise or unrelated features coming from encoders output. And pass down the cleaned features to decoder

The Decoder is a simple BiLSTM, The main purpose of using bilstm was its accuracy and lower computation for generating text as it reads input sequnce in both directions forward and backwards.

Model v12:

Model V12 is the final model. This was extended from model V1. This model contains two parts. Encoder a resnet 18 pretrained on imagenet1k. And Decoder as Bert uncased and BiLSTM. The resnet101 from v1 was replaced by resnet 18 and transformer by bert uncased as bert uncased is made on top of a transformer and a pre trained LLM.

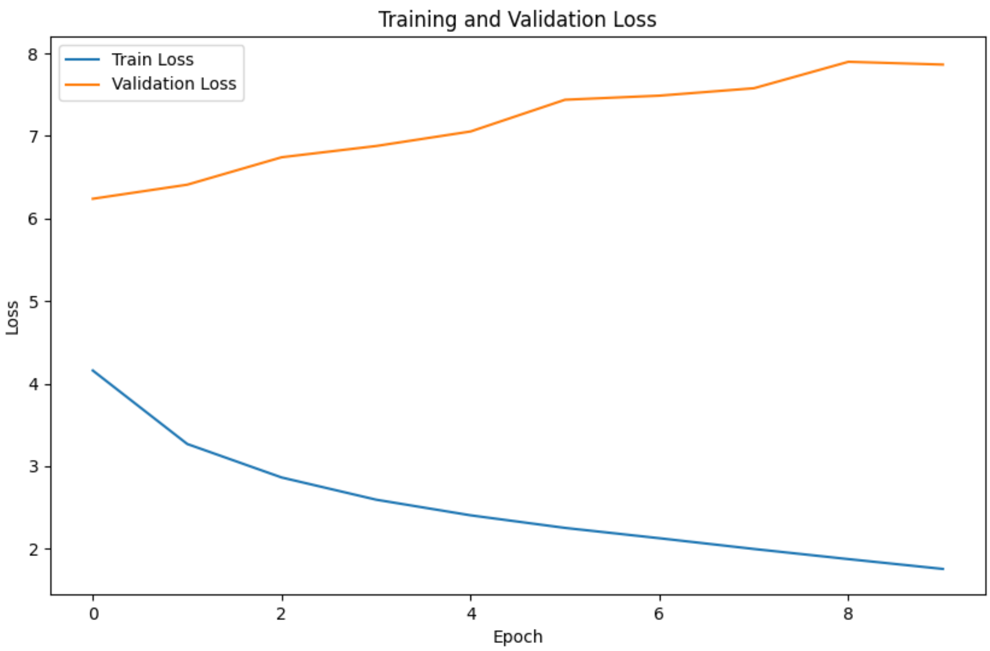
The learning rate was quite flactuating as the model was traing hard to find a global minima.

Results:

v5:


Training and Validation Loss
Train Loss
Validation Loss
280000
260000
240000
220000
200000
180000
Loss
0
2
4
6
8
Epoch

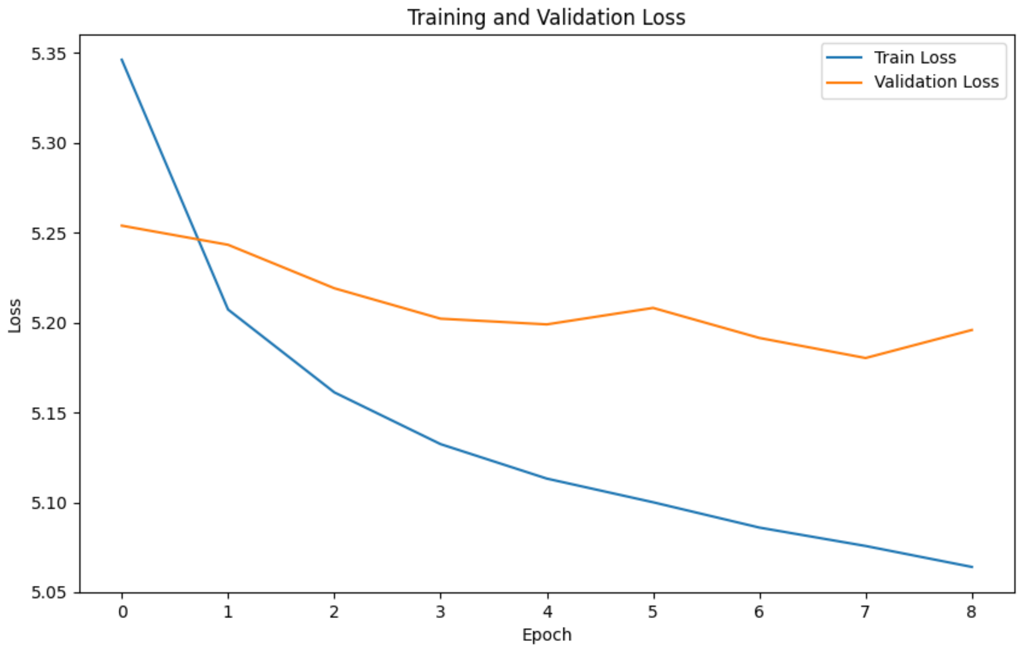
v6:


Training and Validation Loss
Train Loss
Validation Loss
8
7
6
5
4
3
2
Loss
0
2
4
6
8
Epoch

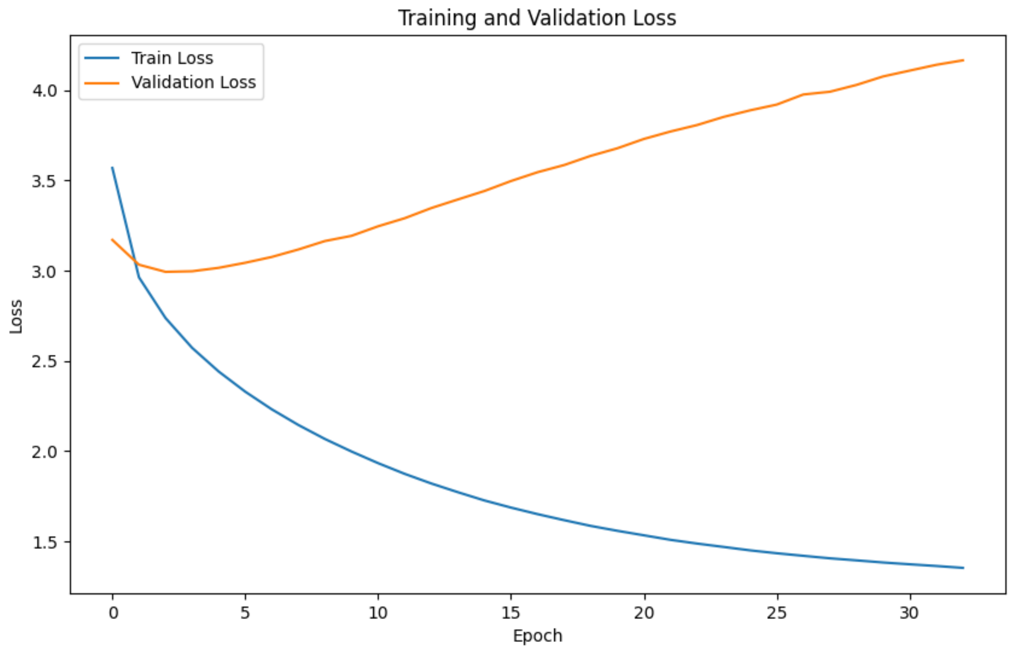
v7:


Training and Validation Loss
Train Loss
Validation Loss
Loss
Epoch
14
12
10
8
6
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

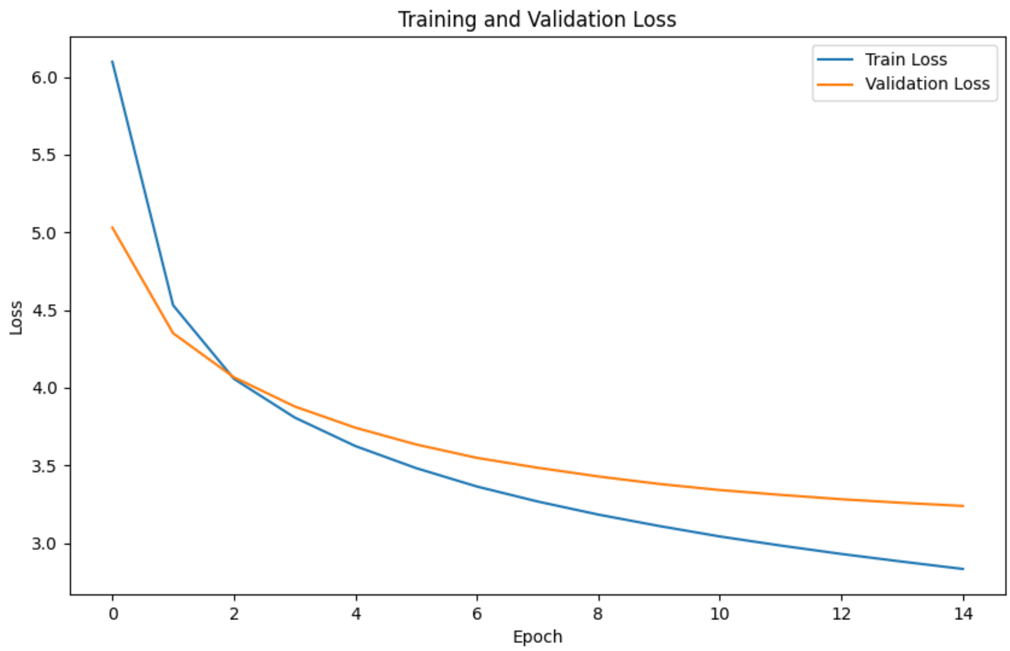
v8:


Training and Validation Loss
Train Loss
Validation Loss
Loss
Epoch
5.35
5.30
5.25
5.20
5.15
5.10
5.05
0
1
2
3
4
5
6
7
8

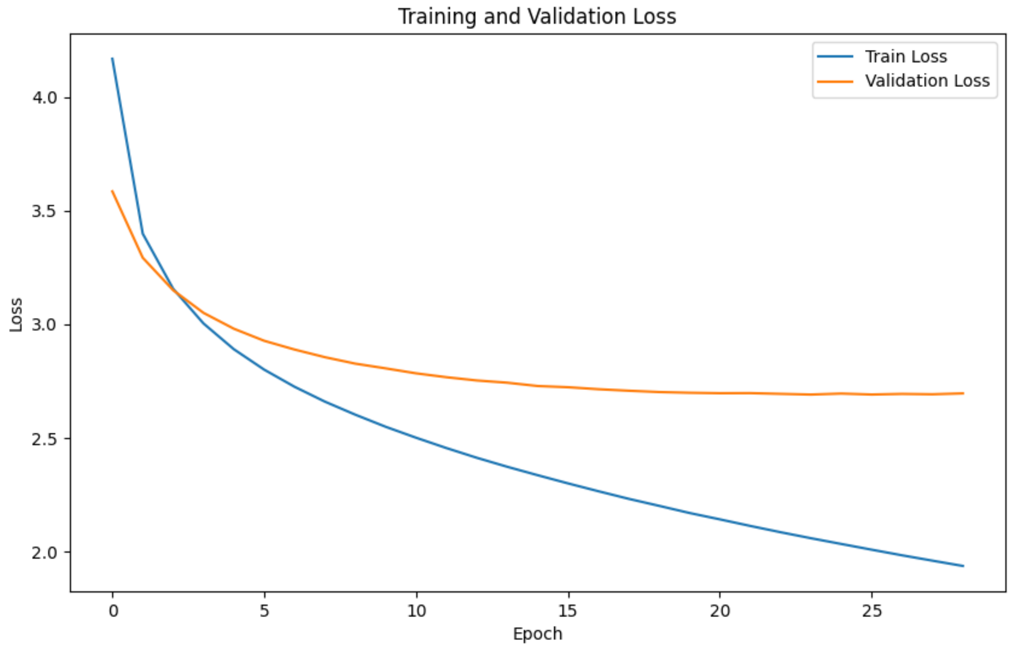
v10


Training and Validation Loss
Train Loss
Validation Loss
4.0
3.5
3.0
2.5
2.0
1.5
Loss
0
5
10
15
20
25
30
Epoch

v11:


Training and Validation Loss
Train Loss
Validation Loss
6.0
5.5
5.0
4.5
4.0
3.5
3.0
Loss
0
2
4
6
8
10
12
14
Epoch

v12:



Conclusion

Our research journey in image captioning led us through multiple architectural iterations, culminating in Model V12 as our best-performing version. Interestingly, our initial concept, Model V1, emerged as theoretically superior. Its meticulously crafted design, featuring ResNet-101 trained on MS COCO and Visual Genome for rich visual understanding, a Transformer pre-trained on LLAMA's dataset for eloquent text generation, and a bi-LSTM for seamless fusion, promised to capture both intricate visual elements and nuanced language. However, V1's substantial computational demands made it impractical within our time and resource constraints.

Instead, we adapted and simplified, arriving at V12, which uses ResNet-18 and BERT's uncased model. While V12 performs admirably, our V1 had the potential to set new benchmarks in caption accuracy and richness. This scenario poignantly illustrates the tension in AI research between theoretical ideals and practical limitations. Model V1 remains an aspirational blueprint, highlighting that today's constrained concepts may become tomorrow's breakthroughs as technology advances. Our journey underscores the importance of balancing ambition with pragmatism in driving the field forward.

References

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar,
Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua.
Learning phrase representations using RNN encoder-decoder
for statistical machine translation. In EMNLP, October 2014.

Kiros, Ryan, Salahutdinov, Ruslan, and Zemel, Richard. Multimodal neural language models.
In International Conference on
Machine Learning, pp. 595–603, 2014a.

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unifying visual-semantic
embeddings with multimodal neural language models. arXiv:1411.2539 [cs.LG], November
2014b.

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan.
Deep captioning with multimodal recurrent neural networks
(m-rnn). arXiv:1412.6632 [cs.CV], December 2014.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan,
Dumitru. Show and tell: A neural image caption generator.
arXiv:1411.4555 [cs.CV], November 2014.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A
neural image caption generator." In *Proceedings of the IEEE conference on computer vision
and pattern recognition*, pp. 3156-3164. 2015

Donahue, Jeff, Hendrikcs, Lisa Anne, Guadarrama, Segio, Rohrbach, Marcus, Venugopalan,
Subhashini, Saenko,
Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition
and description.
arXiv:1411.4389v2 [cs.CV], November 2014.

Hochreiter, S. and Schmidhuber, J. Long short-term memory.
Neural Computation, 9(8):1735–1780, 1997.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. arXiv:1412.2306 [cs.CV], December 2014.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest, Srivastava, Rupesh, Deng, Li, Dollar, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, ´ Margaret, Platt, John, et al. From captions to visual concepts and back. arXiv:1411.4952 [cs.CV], November 2014.

Kulkarni, Girish, Premraj, Visruth, Ordonez, Vicente, Dhar, Sagnik, Li, Siming, Choi, Yejin, Berg, Alexander C, and Berg, Tamara L. Babytalk: Understanding and generating simple image descriptions. PAMI, IEEE Transactions on, 35(12):2891–2903, 2013.

Yang, Yezhou, Teo, Ching Lik, Daume III, Hal, and Aloimonos, ´ Yiannis. Corpus-guided sentence generation of natural images. In EMNLP, pp. 444–454. Association for Computational Linguistics, 2011.

Kuznetsova, Polina, Ordonez, Vicente, Berg, Alexander C, Berg, Tamara L, and Choi, Yejin. Collective generation of natural image descriptions. In Association for Computational Linguistics: Long Papers, pp. 359–368. Association for Computational Linguistics, 2012.

Kuznetsova, Polina, Ordonez, Vicente, Berg, Tamara L, and Choi, Yejin. Treetalk: Composition and compression of trees for image descriptions. TACL, 2(10):351–362, 2014.

Larochelle, Hugo and Hinton, Geoffrey E. Learning to combine foveal glimpses with a third-order boltzmann machine. In NIPS, pp. 1243–1251, 2010.

Denil, Misha, Bazzani, Loris, Larochelle, Hugo, and de Freitas, Nando. Learning where to attend with deep architectures for image tracking. Neural Computation, 2012.

Tang, Yichuan, Srivastava, Nitish, and Salakhutdinov, Ruslan R. Learning generative models with visual attention. In NIPS, pp. 1808–1816, 2014.

Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623, 2015.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 [cs.CL], September 2014.

Mnih, Volodymyr, Hees, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. Recurrent models of visual attention. In NIPS, 2014.

Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. Multiple object recognition with visual

Graves, Alex. Generating sequences with recurrent neural networks. Technical report, arXiv preprint arXiv:1308.0850,
2013.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In *International conference on machine learning*, pp. 2048-2057. PMLR, 2015.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077-6086. 2018

Herdade, Simao, Armin Kappeler, Kofi Boakye, and Joao Soares. "Image captioning: Transforming objects into words." Advances in neural information processing systems 32 (2019).

Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Hierarchical question-image co-attention for visual question answering." Advances in neural information processing systems 29 (2016)