Novel Visual-Textual Attention Fusion Models for Context-Aware Image Alt Text Generation

Stanford CS224N Custom Project

Benjamin B. Yan
Department of Computer Science
Stanford University
bbyan@stanford.edu

Abstract

Images are indispensable to web communication but are often inaccessible to individuals with blind or low vision. Screen readers use descriptions embedded in the HTML alt tag, which is often missing in images; while captions are more plentiful, they are written to supplement the image with context rather than replace it—and are typically devoid of visual details requisite for accessibility. Furthermore, existing description generation systems are often generic and agnostic to contextual implicature. To address this issue, we introduce an image description model using a novel multi-modal architecture that fuses visual embeddings from an image input and text embeddings from a caption context input. The description is generated using a GRU recurrent neural network (RNN) decoder, with Bahdanau attention units applied over tensors from both spatial image patches and caption tokens to encourage the RNN to extract from both visual and textual knowledge. We also develop a novel beam search variant with a brevity penalty to generate thorough descriptions. We demonstrate that our multi-modal method outperforms unimodal models with state-of-the-art architectures that do not use context input in BLEU score on the Wikipedia-based Concadia dataset. Our work represents a contribution toward enriching visual accessibility through context-aware NLP systems.

Keywords: Image description model, multi-modal, recurrent neural network, context-aware NLP systems

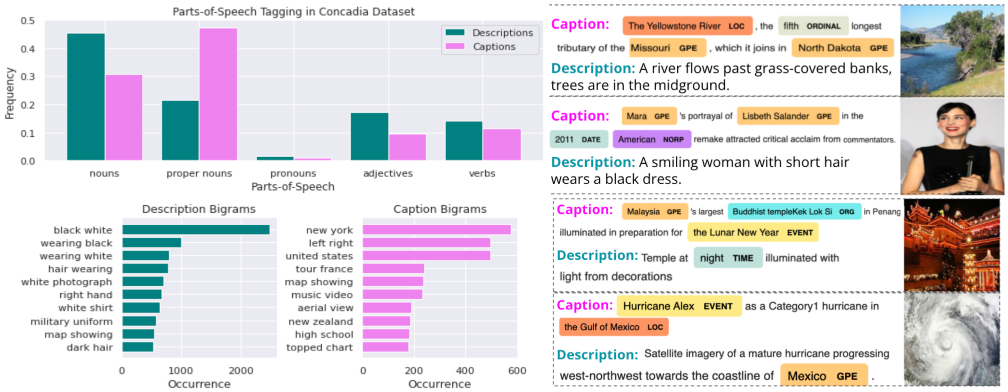
1 Key Information to include

- Mentor: Vincent Li
- External Collaborators (if you have any): N/A
- External Mentors: Elisa Kreiss and Prof. Christopher Potts
- Sharing project: N/A

2 Introduction

Generating image descriptions is a challenging task that mandates astute visual and natural language reasoning over complex and diverse scenes, including recognition and relational knowledge of people, objects, and terrain—and the ability to convey that semantic information in coherent sentences [1]. A key motivation for this task is that images are not directly accessible to blind or low vision individuals. They rely on screen readers that read alt text descriptions of the image, which are often missing, especially on social media where coverage drops to 0.1% [2]. While captions are more profuse, they are generally written to supplement the image with context and non-visual knowledge [2], making them an unreliable standalone proxy. This discrepancy in accessibility creates a need for automated systems that can substitute images with salient, coherent descriptions.

Current methods and datasets—such as MS-COCO and Flickr30K—often map images with generic reference captions or descriptions. Studies illustrate that the text they produce suffers from being rigid, one-size-fits-all, and contextually agnostic [11,12]. One particular challenge is discerning what content from the image is relevant to describe, and the lasting question of what makes a quality description. For instance, given a photograph of a husky sitting on a plush sofa, there would reasonably be different descriptions depending on whether the image appears on a furniture website to sell the sofa, or if the image appears on a dog adoption website. Existing unimodal systems such as Google Show-And-Tell [3] cannot accommodate additional text-based context, which leads us to propose a multi-modal system that generates descriptions through attention over both deep visual embeddings of the image and vector embeddings of the context.



Caption: The Yellowstone River LOC , the fifth ORDINAL longest tributary of the Missouri GPE , which it joins in North Dakota GPE

Description: A river flows past grass-covered banks, trees are in the midground.

Caption: Mara GPE 's portrayal of Lisbeth Salander GPE in the 2011 DATE American NORP remake attracted critical acclaim from commentators.

Description: A smiling woman with short hair wears a black dress.

Caption: Malaysia GPE 's largest Buddhist templeKek Lok Si ORG in Penang illuminated in preparation for the Lunar New Year EVENT

Description: Temple at night TIME illuminated with light from decorations

Caption: Hurricane Alex EVENT as a Category1 hurricane in the Gulf of Mexico LOC

Description: Satellite imagery of a mature hurricane progressing west-northwest towards the coastline of Mexico GPE .

Figure 1: (Left) Parts-of-speech tagging and most frequent bigrams of the Wiki-based Concadia dataset. (Right) Examples from the dataset with Named Entity Recognition (NER) parsing of the captions and descriptions.

The data backbone of our system is Stanford NLP's Concadia, a corpus engineered by Kreiss. et. al [2] by mining Wikipedia images along with their alt-text (*descriptions*) and *captions*. While captions and descriptions are often perceived synonymously, this work argues they fulfill distinct purposes—*descriptions* are created to replace the image, while *captions* are created to support and contextualize the image. As visualized in **Figure 1**, we performed a linguistic analysis of Concadia using tools from SpaCy and NLTK. We find that proper nouns and named entities are substantially more abundant in *captions* than *descriptions*, which aligns with their contextual mission, while adjectives and verbs are more abundant in *descriptions*.

For accessibility, the goal is for the generated text to supplant the Wikipedia images, so this research focuses on decoding images into *descriptions*, with novel multimodal attention and embedding fusion architectures to robustly incorporate the *captions* as joint input. Our findings demonstrate that context embeddings strengthen the quality and detail of synthesized descriptions. We also contribute to ongoing neural machine translation (NMT) research an original variation of beam search decoding with a brevity penalty on traversal length to penalize excessively frugal and un-detailed descriptions.

3 Related Work

The image-to-text field has recently blossomed with a wellspring of generative models, vast datasets with annotated images, and human-computer interaction (HCI) research into visual context.

- **CAPWAP (Captioning with a Purpose):** To tackle the issue of generic reference descriptions, Fisch et. al [9] used large visual question-answering datasets with reinforcement learning to optimize generated descriptions to answer specific questions. They demonstrated with human raters that their generated descriptions tended to be more informative, less generic, and semantically relevant than traditional image-to-text baselines.
- **HCI Approach to Image Descriptions:** Stangl et. al [8] interviewed 28 blind or low vision (BLV) individuals to learn how image scenarios impact the relevant visual content that a description should have. They find that across different scenarios from news to e-commerce to travel planning, sub-themes for content that BLV users wanted can substantially differ,

providing a strong argument for a shift away from one-size-fits-all generic descriptions and toward improving the quality and responsiveness of descriptions to context.

- **Google Show and Tell:** Vinyals et. al [3] at Google Brain trained a ML system named Show and Tell to perform the *image* → *description* task with state-of-the-art results, using deep Inception CNNs to initialize the image encoder and a LSTM model trained to maximize the Bayesian likelihood of the *target description* sequence given the input *image*.

Our work is an organic extension of the aforementioned Concadia dataset paper [2], and incorporates a contextual framework toward description generation along the vein of HCI research from Stangl et. al. While CAPWAP works to anticipate and satisfy latent and *implicit* user needs for image descriptions, we work with the case where *explicit* captions are provided but not descriptions, which is common on social media and image-sharing platforms like Instagram [11]. Finally, we employ at a high-level the encoder-decoder paradigm for description synthesis from Vinyals et. al, with added functionality for the inclusion of context knowledge in the new task *{image, caption}*→ *description*.

4 Approach

We devise a new encoder architecture in Python that combines trailblazing advances in the computer vision and NLP fields, namely the VGG-16 CNN model [4] that won the 2014 ImageNet and a widely popular GloVe model [5] pre-trained over the Wikipedia corpus. We truncate the VGG-16's final classification layer so that it generates an embedding of 81 vectors (dimension $d = 512$) from an image corresponding to 9 by 9 spatial patches. Image preprocessing was done using the Tensorflow tutorial [2] for image captioning *linked here*. We obtain the word sequence from the caption using a NLTK regex tokenizer, truncate it to length 19 if necessary, and tag each word with its corresponding vector ($d = 300$) if it lies in the GloVe vocabularly and a padding vector otherwise.

We concatenate the vectors and map each one injectively to a 512-dimensional hyperspace, forming 19 vectors ($d = 512$). This leads to a fused embedding tensor of 100 vectors ($d = 512$), 81 and 19 from the image and context, respectively. We pass each vector through a *trainable* Tensorflow linear layer with dropout for fine-tuning with the aim of extracting features relevant for descriptions. The final 100 by 256 tensor is delivered to the decoder as input, with 100 Bahdanau additive attention weights so that the model can take a weighted sum of the information in *both* the 81 image and 19 caption vector embeddings for each generated word.

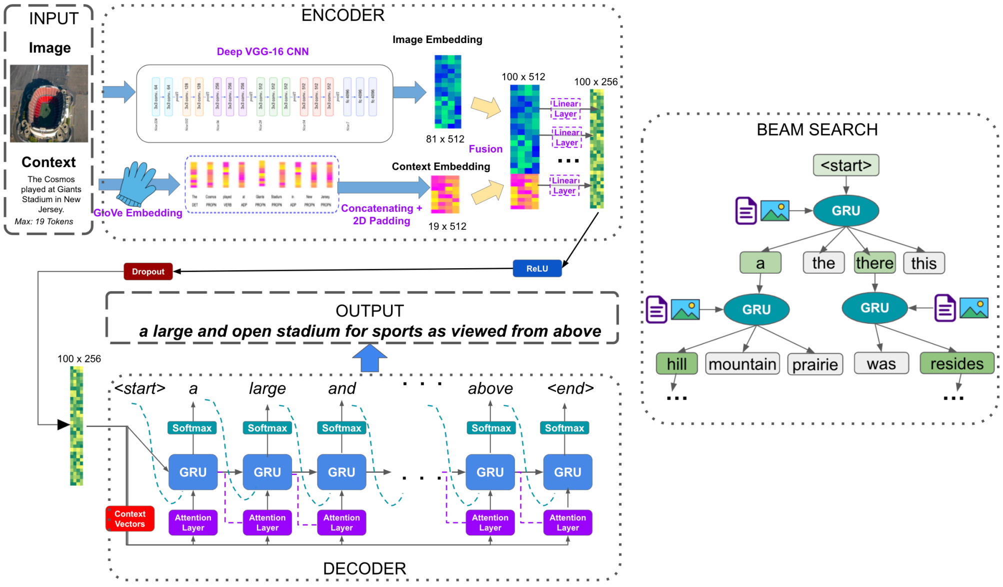


Figure 2: (Left) Schematic of our encoder-decoder model from *{image,context = caption}* to *description*. (Right) High-level diagram of post-training beam search heuristic that explores promising description sequence branches based on conditional softmax probability from the RNN.

The decoder—adapted from the Tensorflow tutorial [3] to apply joint attention over the image and caption tensors—is a recurrent neural network with Gated Recurrent Units (GRU) cells trained to maximize the joint likelihood / log-likelihood $p(T)$ of the target description T given its input tensor I—combination of the image and context embeddings—from the encoder. Given the target word sequence $T_1, \cdots, T_n$, we are searching for the optimal parameters $\theta*$ with [3]

$$\begin{aligned}
\theta^* &= \arg\max_\theta p(T; I, \theta) \\
&= \arg\max_\theta \log(p(T_1, T_2, \cdots, T_n; I, \theta)) \\
&= \arg\max_\theta \log\left[\prod_{i=1}^{n} p(T_i \mid I, \theta, T_1, \cdots, T_{i-1})\right] \\
&= \arg\max_\theta \sum_{i=1}^{n} \log(p(T_i \mid I, \theta, T_1, \cdots, T_{i-1}))
\end{aligned}$$

As the above equation describes a sum of log-likelihoods of individual words $T_j\{1 \leq j \leq n\}$ in the target sequence, and the GRU at time $i \geq 1$ generates a probability distribution P_{i+1} to predict the next $(i+1)th$ word (Figure 2)—with T_1 being reserved as the **<start>** token—we can consider the negative log-likelihood of T_{i+1} in the distribution, or $-\log(P_{i+1}(T_{i+1}))$, to be the cost at the particular word T_{i+1}. Summing over non-start target words T_2 to T_n, we get a total loss of $\ell = -\sum_{j=2}^{n} \log(P_j(T_j))$ for attempting to predict target description T, which is used for training.

For post-training, we use a *beam search* (Figure 2) of width $w = 3$, which at iteration $k \geq 1$, considers the top (most probable) w best sequences of length k by joint log-likelihood of their words, generates new sequences of length $k+1$ from them using the GRU cell, and prunes all but the best w new sequences for the next iteration. However, we found that in our generated descriptions, the vanilla beam search was often innately favoring shorter and non-detailed descriptions. We hypothesize this is due to ranking candidate sequences by joint likelihood of their words, since extending an existing sequence $S_1, \cdots, S_m$ with a new word S_{m+1} can never increase joint log-likelihood LL as

$$\begin{aligned}
\log(P(S_1, S_2, \cdots, S_m, S_{m+1})) &= \log(P(S_{m+1} \mid S_1, S_2, \cdots, S_m)) + \log(P(S_1, S_2, \cdots, S_m)) \\
&\leq \log(P(S_1, S_2, \cdots, S_m)).
\end{aligned}$$

which disadvantages longer descriptions. Consequently, we introduce a novel beam search variant inspired by BLEU scoring that imposes a multiplicative *brevity penalty* on each of the sequences in the last set of candidates. For a candidate sequence $S_1, \cdots, S_m$ of length m, the modified log-likelihood score $LL^*(S_1, \cdots, S_m)$ would be

$$LL^*(S_1, \cdots, S_m) = LL(S_1, \cdots, S_m) * \underbrace{\max(1, e^{1-\frac{m}{\Gamma}})}_{\text{brevity penalty}},$$

for tunable parameter Γ, which is the minimum description word length that leads to no penalty. A graph of $\max(1, e^{1-\frac{m}{\Gamma}})$ over m is provided below; since LL is non-positive for discrete probabilities, a higher $\max(1, e^{1-\frac{m}{\Gamma}}) \geq 1$ due to shorter length decreases the modified LL^* of a candidate.

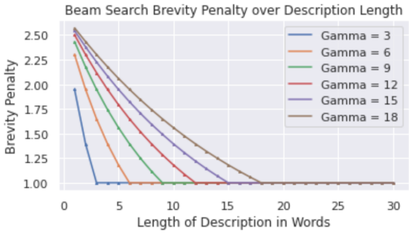


Figure 3: Graph of the brevity penalty for different values of parameter Γ.

Furthermore, we note that a low Γ does little to prevent favoritism of shorter candidate branches, while a too large Γ can encourage ranting behavior in the RNN to avoid the brevity penalty, as shown below. A good balance we have found has been $\Gamma = 10$.

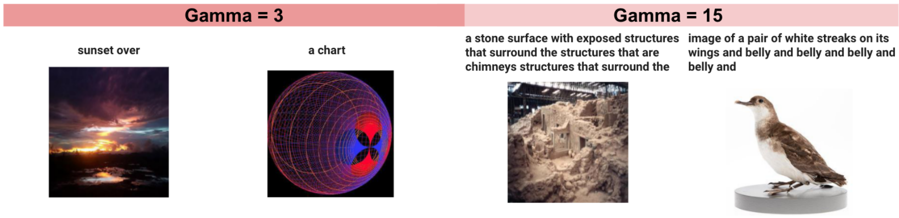

Gamma = 3
Gamma = 15
sunset over
a chart
a stone surface with exposed structures that surround the structures that are chimneys structures that surround the
image of a pair of white streaks on its wings and belly and belly and belly and belly and

Figure 4: A sample of description results with sub-optimal selections of parameter Γ.

5 Experiments

5.1 Data

We use Concadia as the primary dataset, which is a corpus from Kreiss et. al composed of 96,918 images with corresponding alt-text descriptions and captions mined across Wikipedia. We also use the Microsoft COCO (MS-COCO) [7] 2015 dataset, a large-scale computer vision trove of 82,000 images, each with multiple descriptions annotated through crowdsourcing. The models that we pre-train on MS-COCO are deployed on Concadia as baselines—apart of a transfer learning evaluation of how description systems perform outside of their trained image domains. The specific tasks for each dataset are delineated in the table below.

Task	**Training Dataset**	**Prediction Mapping**	**Relevant Datasets**
Image Description Generation	{🖼, **reference description**}	{🖼} → **predicted description**	MS-COCO, Concadia
Image Description Generation with Caption Context	{🖼, **reference caption**, **reference description**}	{🖼, **reference caption**} → **predicted description**	Concadia

Table 1: We run the uni-modal task of *image* → *description* on both MS-COCO and Concadia, while we run the multi-modal task of *{image, caption}* → *description* on just Concadia.

For training, we used 70K {image,caption,description} samples from Concadia and 60K {image,text} pairs from MS-COCO, as well as 9K samples from each dataset for validation.

5.2 Evaluation method

For evaluation, we used cumulative BLEU scores up to 4-grams—considered the contemporary state-of-the-art metric for assessing neural machine translation (NMT)—between the reference descriptions and the model-synthesized descriptions. For the three models we pre-train on MS-COCO, we report their BLEU scores on our MS-COCO validation set; for all models, we report their BLEU scores on the Concadia validation set.

5.3 Experimental details

The models were compiled in Tensorflow and training for 40 epochs was dispatched over Google Cloud GPUs and TPUs with $\geq$ 12 GB of RAM. For the uni-modal baseline models trained to perform *image* → *description* instead of *{image, caption}* → *description* instead, the model encoder consists exclusively of the VGG-16 CNN and a linear layer over the image embedding, with the caption embedding sub-system removed. Each model used an Adam optimizer with a learning rate of 0.001 to perform mini-batch gradient descent with a batch size of 64. We run models with encoder dropout rates of $0\%, 25\%, 50\%$ for hyper-parameter tuning and determining how to prevent regularization in the model.

Teacher forcing was used in the decoder during training to help the model stabilize and converge by supplying the ground truth at time t as input to the GRU at time $t+1$ instead of the GRU's earlier prediction at time t. For beam search in the post-training, we used a beam width of 3 and a brevity penalty parameter of $\Gamma = 10$.

5.4 Results

In the graph below, all of the uni-modal models—which perform *image* $\rightarrow$ *description*—serve as baselines for the evaluation of the multi-modal model trained on the Concadia dataset, which performs *{image, caption}* $\rightarrow$ *description*. Additionally, the reason that we do not evaluate the Concadia-trained models on MS-COCO is that they operate in the domain where caption context is available to produce descriptions, which is not the case in MS-COCO.

We find that on the Concadia dataset (pink bars), the multi-modal model outperforms all four of the baselines with a BLEU score of 31.9. This is compared with a BLEU score of 26.5 from the uni-modal model trained directly on Concadia, which suggests that captions—while they are not designed to replace the image—can provide useful embedding information alongside the image for the model to produce better descriptions.

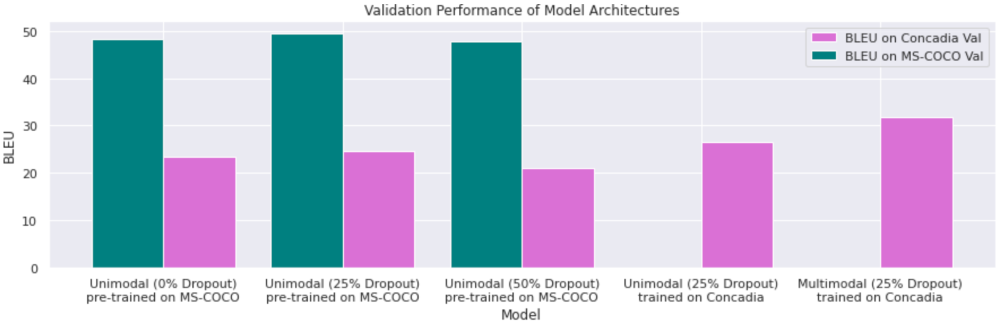


Figure 5: BLEU validation scores (scaled 0-100) for different model architectures and hyperparameters.

For hyperparameter tuning, we also found 25% dropout rate on the encoder to be slightly more optimal than the other dropout rates of 0% and 50% on the validation sets. We note that the unimodal models that were pre-trained on MS-COCO perform quite well on MS-COCO validation (teal bars), with BLEU scores in the $45 - 49$ range. One reason is that while the MS-COCO dataset has at least 5 descriptions annotated for each image, the Concadia dataset has a one-to-one pairing of images and their captions and descriptions. Although their BLEU scores degrade considerably when evaluated on the Concadia dataset, they all score at least 20, which suggests that there is a promising path of transfer learning—likely with some fine-tuning on the new image domain—toward image description systems being effective on new image datasets, which can be crucial due to the massive global diversity of image domains on the Web that is growing at unparalleled rates.

6 Analysis

A sample of images from the Concadia validation dataset are provided below, with the model's generated sequences—using the beam search variant with width 3—on top of the image and the input caption with NER tagging provided below.

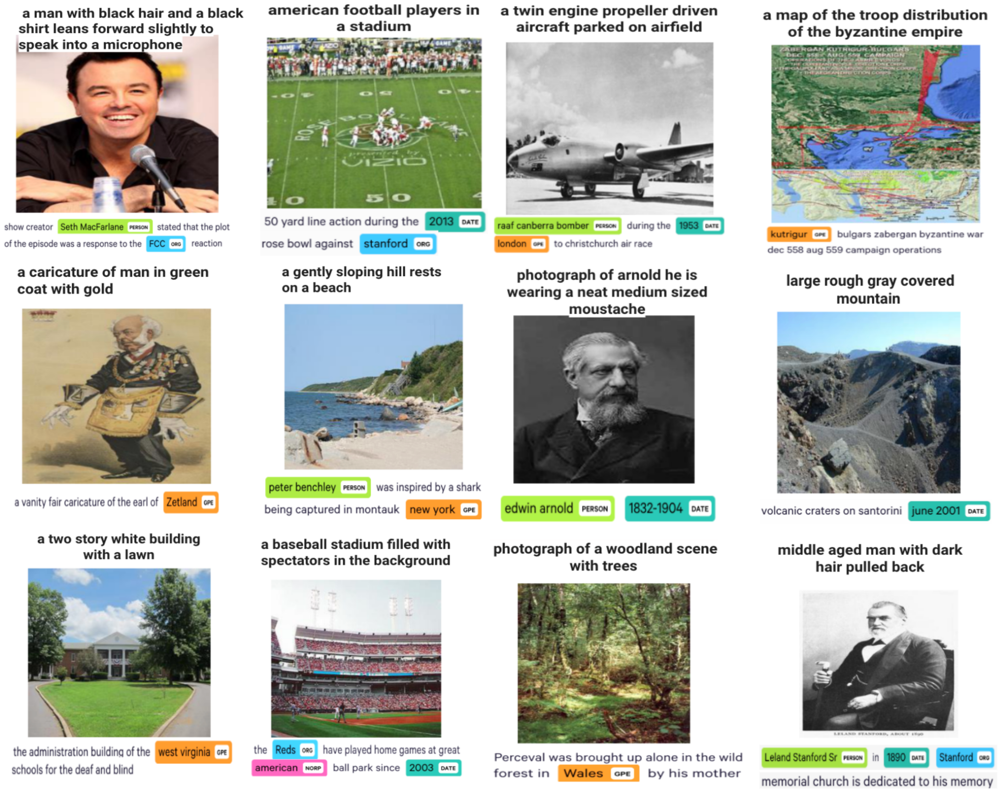

a man with black hair and a black shirt leans forward slightly to speak into a microphone
show creator Seth MacFarlane PERSON stated that the plot of the episode was a response to the FCC ORG reaction
american football players in a stadium
50 yard line action during the 2013 DATE rose bowl against stanford ORG
a twin engine propeller driven aircraft parked on airfield
raaf canberra bomber PERSON during the 1953 DATE london GPE to christchurch air race
a map of the troop distribution of the byzantine empire
kutrigur GPE bulgars zabergan byzantine war dec 558 aug 559 campaign operations
a caricature of man in green coat with gold
a vanity fair caricature of the earl of Zetland GPE
a gently sloping hill rests on a beach
peter benchley PERSON was inspired by a shark being captured in montauk new york GPE
photograph of arnold he is wearing a neat medium sized moustache
edwin arnold PERSON 1832-1904 DATE
large rough gray covered mountain
volcanic craters on santorini june 2001 DATE
a two story white building with a lawn
the administration building of the west virginia GPE schools for the deaf and blind
a baseball stadium filled with spectators in the background
the Reds ORG have played home games at great american NORP ball park since 2003 DATE
photograph of a woodland scene with trees
Perceval was brought up alone in the wild forest in Wales GPE by his mother
middle aged man with dark hair pulled back
Leland Stanford Sr PERSON in 1890 DATE Stanford ORG memorial church is dedicated to his memory

Figure 6: Synthesized descriptions alongside the image and input captions in the Concadia val dataset.

We find that generally, the model is able to synthesize informative descriptions that adequately capture the people, objects, and terrain, as well as their relational dynamics, present in the image. We also notice that the generated descriptions contain visual information that is not present in the caption, suggesting that the model is able to cogently extract and write about attributes on the image. On the top right example above, we observe that the model is able to incorporate the "byzantine" information available in the context into the synthesized description. We highlight some errors with visual reasoning or poor language generation in the images below.

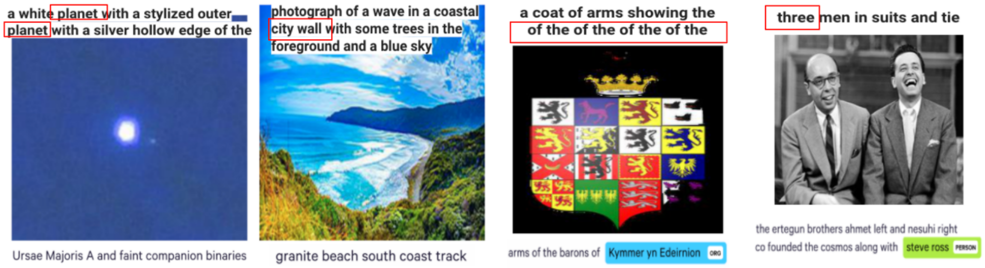

a white planet with a stylized outer planet with a silver hollow edge of the
Ursae Majoris A and faint companion binaries
photograph of a wave in a coastal city wall with some trees in the foreground and a blue sky
granite beach south coast track
a coat of arms showing the of the of the of the of the
arms of the barons of Kymmer yn Edeirnion ORG
three men in suits and tie
the ertegun brothers ahmet left and nesuhi right co founded the cosmos along with steve ross PERSON

Figure 7: Synthesized descriptions with red error boxes showing visual reasoning or language mistakes.

We see that in the leftmost image, it mis-identifies the two binary stars in the image as planets, while in the second left image, it manufactures a "city wall" that is not present in the image although the other details are descriptive and correct, which can be due to regurgitating descriptions from the training set. On the second to right image, we see an example of decoder breakdown where the model continually repeats "of the", which, which can be attributed to the RNN attending over "of" and "then" several times; a potential future alleviation strategy is to penalize repetitive words or bigrams in the decoder. The rightmost image makes a visual error of miscounting the number of individuals.

To help provide transparency and understanding into why the decoder makes its word choices at each step of the description generation process, we provide a new program that extracts the 100 attention weights from each time t the GRU cell runs and builds both visual and textual heatmaps over the image and caption tokens, respectively. The first 81 weights represent attention over 9 x 9 spatial patches of the image, while the last 19 weights represent attention over the word embeddings in the caption. We demonstrate the use of this program on an image of a beach that inspired the movie *Jaws*.

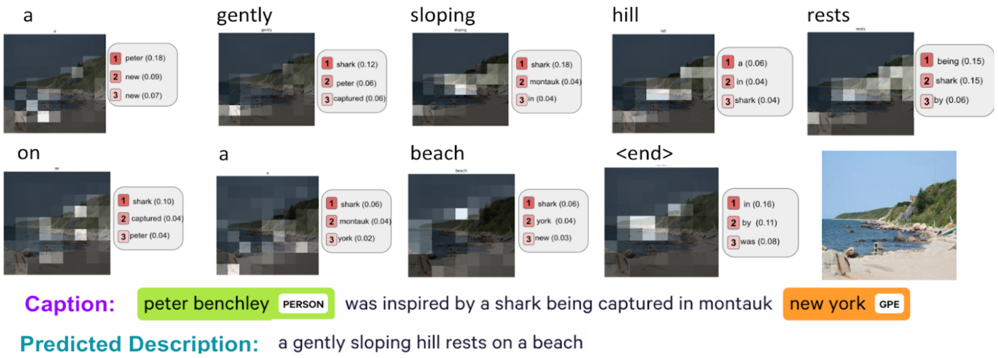


Figure 8: Brighter patches on the image had higher attention when the GRU cell generated the corresponding word. To the right of each image heatmap are the top three caption tokens ranked by attention weight.

From the heatmap, we see that when the GRU generates "sloping" and "hill", the brightest areas on the image are appropriately the patches containing the grassy hill with an ascending band of bright spatial patches. We also see that when the GRU generates "beach", the highest patches of attention on the image are generally the water and sand areas, and the most salient word is intriguingly "shark". While the attention heatmaps do not fully capture why the RNN makes each word choice, it can help provide a degree of transparent insight that can help the effort toward understanding how context and image features specifically impact the word descriptions generated by deep learning algorithms.

7 Conclusion

We find that image descriptions benefit from the multi-modal inclusion of context caption embeddings when they are available, which can provide salient signal information to the decoder. On the Wikipedia-based Concadia dataset, we found that our multimodal image description system that map from *{image,caption}* → *description* outperforms uni-modal baselines in BLEU score with the state-of-the-art Show and Tell architecture. Our results suggest a promising and auspicious deep learning approach toward improving visual accessibility in domains where captions are significantly more prevalent than alt-text descriptions, such as social media or e-commerce sites. While the author has come to the conclusion, alongside many others, that automatic systems remain far from fully imitating human quality descriptions, the author believes that building context-sensitive models represent a worthwhile contribution toward advancing visual accessibility.

One major limitation of our multi-modal model is its dependence on the availability and quality of caption context, while uni-modal models require just an image to produce descriptions. Furthermore, while the multi-modal model was trained on the diverse corpus of Wikipedia images, captions, and descriptions, we have not empirically assessed its performance in other image domains like Instagram or Facebook. Angles for future work include experimenting with different embedding strategies for the caption, such as using *Bidirectional Encoder Representations from Transformers* (BERT) or word2vec representations. Also, we would want to potentially assess human judgement and evaluation on the quality and features of the synthetic descriptions—in comparison with the ground truth descriptions—through human participant trials.

References

[1] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. (2018). Caption Crawler: Enabling Reusable Alternative Text Descriptions using Reverse Image Search. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, pages 1–11, Montreal QC, Canada.* ACM Press.

[2] Elisa Kreiss, Noah D. Goodman, Christopher Potts. (2022). Concadia: Tackling Image Accessibility with Descriptive Texts and Context.

[3] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015). Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3156-3164.*

[4] Karen Simonyan, Andrew Zisserman. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing* (EMNLP 2014), 12:1532–1543.

[6] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. (2016). Neural Machine Translation by Jointly Learning to Align and Translate.

[7] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. ECCV.

[8] Stangl, A., Verma, N., Fleischmann, K.R., Morris, M.R., Gurari, D. (2021). Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility.*

[9] Fisch, A., Lee, K., Chang, M., Clark, J., Barzilay, R. (2020). CapWAP: Captioning with a Purpose. ArXiv, abs/2011.04264.

[10] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. *In The World Wide Web Conference on* - WWW '19, pages 549–559, San Francisco, CA, USA. ACM Press.

[11] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2020. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. arXiv:2012.11696 [cs].

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alttext Dataset For Automatic Image Captioning. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.