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

School of Computer Science

Benjamin Yan
bbyan@stanford.edu

CS224N Winter 2022

Background & Need

- **Generating image descriptions is a challenging task,** requiring machines to have astute visual, spatial, and language reasoning over complex scenes.
- **Images are not directly accessible to individuals with blind or low vision (BLV),** and alt-text is often missing for web images, **especially on social media where coverage is as low as 0.1%** [1].
- There is a pressing need for systems that can create automatic descriptions.
- Most existing models map from images to generic descriptions, and suffer from being rigid, one-size-fits-all, and contextually agnostic [2].

Sample image with its caption and description from the Wiki-based Concadia dataset.

- **Descriptions and captions fulfill separate purposes** [3]: descriptions are meant to replace the image, captions supplement and contextualize the image.
- We propose a novel **multimodal model** with a RNN decoder designed to utilize image-caption fused embeddings to generate descriptions.

- **Concadia** [3] is a corpus from Kreiss et. al composed of 96,918 images with corresponding alt-text descriptions and captions mined across Wikipedia.
- **Microsoft COCO 2015** is a large-scale computer vision dataset composed of ~82000 images, each with ≥ 5 human-annotated descriptions.

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: Unimodal (first task) and multimodal (second task) methods of generating image descriptions.

- A linguistic analysis of **Concadia** using NLTK and SpaCy tools reveals that *proper nouns* and *named entities*—often apart of context—are more abundant in captions than descriptions, while the reverse is true for *adjectives* and *verbs*.

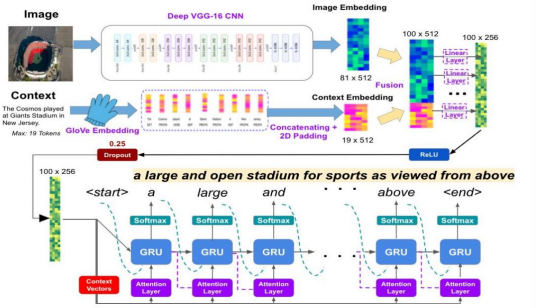
Named Entity	Descriptions	Captions
Person	28017	60282
Organization	19322	40665
Geopolitical Entity	16207	31404

Results from parts-of-speech and named entity recognition (NER) parsing of Concadia.

- For training, we use 70K *(image,caption,description)* samples from **Concadia** and 60K *(image,text)* pairs from **MS-COCO** for training, and 9K samples from each dataset for validation. Training was dispatched over Google Cloud GPUs.
- We conduct hyperparameter testing on dropout rate in the encoder, and evaluate BLEU performance between the uni-modal and multi-modal architectures.

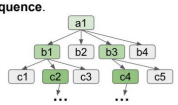
Multimodal Fusion Architecture

- A VGG-16 convolutional neural network (CNN) pre-trained on ImageNet is used to generate a Tensor embedding of the image with 81 512-dimensional vectors.
- The caption is tokenized: each word is tagged with a Wiki-GloVe vector.



Schematic diagram of description generation from image and caption

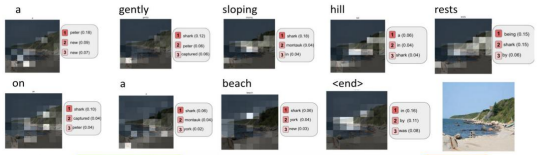
- We **form a joint embedding** of 100 512-dimensional vectors; each vector is passed through a shallow neural network with ReLU activation and dropout.
- The decoder is an RNN with GRU cells and Bahdanau additive **attention over both the visual and textual vectors to generate the description sequence**.
- Teacher forcing is used during training to help convergence by supplying the next recurrent cell with the reference token instead of predicted as input.
- We develop a novel **variant of beam search (width = 3)** with a custom brevity penalty to penalize excessively short and un-detailed description candidates.



Beam search (width = 2) diagram

Visual-Textual Attention Fusion Heatmap

- These heatmaps can help provide insight and transparency into each of the RNNs word choices, and in particular the areas of attention in the image and caption.



Caption: peter benchley PERSON was inspired by a shark being captured in montauk new york GPE

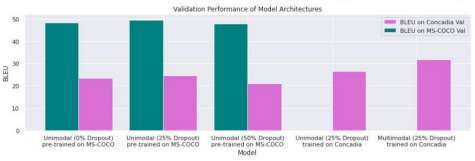
Predicted Description: a gently sloping hill rests on a beach

For each word generated by the decoder, the spatial attention map of the image and the top three tokens in the caption by attention weight are shown. The beach image inspired the movie Jaws.

Results

***Above:** On top of each image is the predicted description and below each image is the input context caption.*

***Above:** Image descriptions showing model visual / language reasoning errors or breakdowns.*



***Above:** BLEU validation scores (scaled 0-100) for different model architectures and hyperparameters after 40 epochs. Note that the models directly trained on **Concadia** were not evaluated on **MS-COCO**.*

Conclusions

- We find that **image description systems benefit from the multimodal inclusion of context caption embeddings**, which can provide salient information for the decoder.
- The high abundance of captions compared to alt-text on platforms such as social media suggest a promising avenue toward advancing visual accessibility within them using AI.
- While automatic systems are still far from fully imitating human image descriptions, context-sensitive models are a solid step forward toward less generic descriptions.
- **Future Work:** Applying other architectures such as transformers or BERT, as well as potentially assessing human judgement of the quality of the synthetic descriptions.

References

1. 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
2. Dognin, P.L., Melnyk, I.V., Mroueh, Y., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Young, R.A., & Belgodere, B.M. (2022). Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *J. Artif. Intell. Res., 73*, 437-459.
3. Elisa Kreiss, Noah D. Goodman, & Christopher Potts. (2022). Concadia: Tackling Image Accessibility with Descriptive Texts and Context.