Background & Need

- Generating image descriptions is a challenging task, requiring machines to have a solid 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 0.71% [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].

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-Graph vector.

Results

![Schematic diagram of description generation from image and caption](image)

- 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.

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 all-text on platforms such as social media suggests a promising avenue toward advancing visual accessibility within them using AI.
- While automatic systems are still far from fully imitating human image descriptions, context-inclusive 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.

Visual-Textual Attention Fusion Heatmap

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

References