

The background features several large, overlapping, light blue watercolor-like shapes. On the right side, there are several thin, dark blue, curved lines that resemble a stylized 'C' or a series of concentric arcs. In the bottom left corner, there is a cluster of small, dark blue dots and splatters of varying sizes.



Image captioning to help blind people

Team: Youchen Zhang, and Hsuan Yu Lin



Project Goal

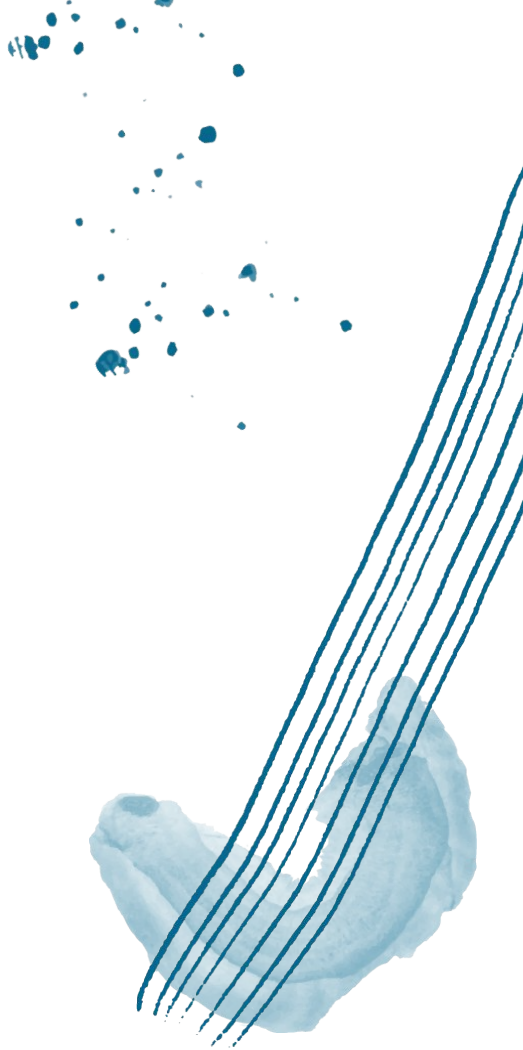
to help visually impaired population
by solving image captioning task
using VizWiz dataset



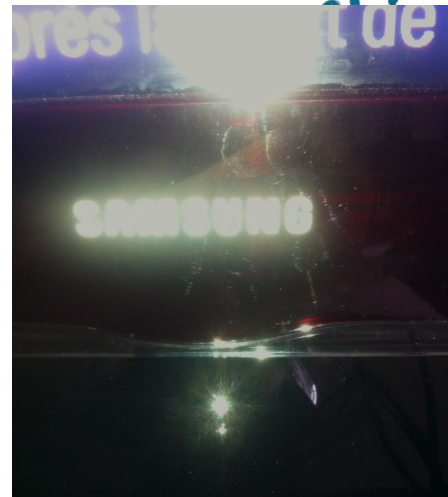


Dataset

The VizWiz-Captions dataset consists of approximately 39 thousand images each paired with 5 captions (17.5 GB). The images are captured by people who are blind in real world situations. Therefore, it exhibits different conditions than observed in the contrived environments of artificially created datasets, like MSCOCO and ImageNet.



Dataset





Methods

CNN-LSTM

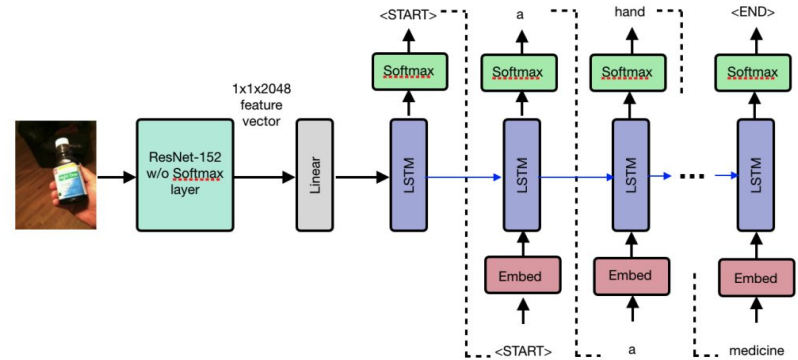
**CNN-LSTM with
attention**

CNN-Transformer



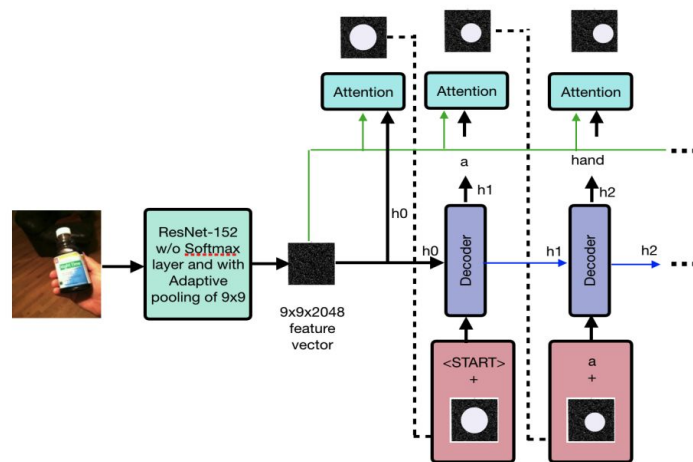
CNN-LSTM

The base model presented in this work is adopted from CNN-LSTM structure. The encoder is a deep convolutional neural network (CNN) without softmax layer. It produces embeddings of fixed length vectors from the input. ResNet-152 (pre-trained on ImageNet dataset) is used as encoder in this project images. The decoder is LSTM model. The first LSTM cell takes features produced from images and special start word S0 and predicts next word in the caption - S1. The second cell takes S1 and predicts second word in the caption S2. This step is repeated until the special stop word SN is produced.



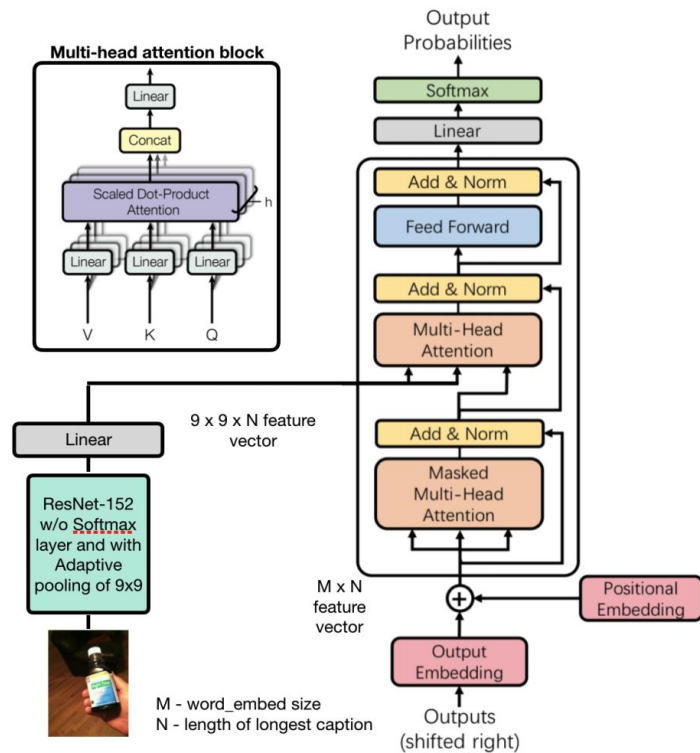
CNN-LSTM with attention

The attention model is built on top of the base model. The lower layers of encoder consist of ResNet-152 with deleted softmax layer. The last layer of encoder (Adaptive Pooling Layer) is changed to output 9x9 attention regions (instead of 1x1 in base model) to allow the model to attend to certain areas of an image. The decoder is LSTM model with attention. In contrast to the base model it looks at image features not only at the first step of decoding process, but at every step. This allows to learn to attend to specific locations in the image for each generated word.









CNN-transformer

The encoder is similar to encoder used in CNN-LSTM Attention Model with the only difference that CNN output is adapted to match word embedding size. It is necessary for multiplication in Multi-head attention layer. The decoder has more complicated structure in comparison to models above. However, as it processes the whole sentence rather than one word at a time, it is more computationally efficient. Transformer is not sequential, therefore to fully explain captions it is necessary to input not only words embeddings, but also positions of words in sentences. This is done by addition of positional embedding.



Analysis

VizWiz images	Ms COCO Images
 <p>GT: a DELL laptop computer screen showing window 7 home premium A computer screen with a windows dialogue box containing the login. Computer screen showing a Windows 7 home premium window with a Dell logo on it. A dell laptop with windows 7 home screen. window screen of dell desktop or laptop showed box of windows 7</p> <p>CNN-LSTM PM: a laptop computer sitting on top of a wooden table</p> <p>CNN-LSTM TFS: a computer screen with a blue background and white text</p> <p>Attention TFS: a computer screen with a message to restore a computer</p> <p>Transformer TFS: a computer screen with a message on it</p>	 <p>GT: The large brown bear has a black nose. A big burly grizzly bear is show with grass in the background. Closeup of a brown bear sitting in a grassy area. A large bear that is sitting on grass. A close up picture of a brown bear's face.</p> <p>CNN-LSTM PM: a person is holding a large, fluffy dog in the background.</p> <p>CNN-LSTM TFS: a brown bear is standing in the grass.</p> <p>Attention TFS: a white and black dog with a black collar</p> <p>Transformer TFS: a small dog is laying on the carpet</p>
 <p>GT: Quality issues are too severe to recognize visual content. Quality issues are too severe to recognize visual content. Quality issues are too severe to recognize visual content. Quality issues are too severe to recognize visual content. Quality issues are too severe to recognize visual content. Quality issues are too severe to recognize visual content.</p> <p>CNN-LSTM PM: a remote control sitting on a table</p> <p>CNN-LSTM TFS: quality issues are too severe to recognize visual content</p> <p>Attention TFS: quality issues are too severe to recognize visual content</p> <p>Transformer TFS: quality issues are too severe to recognize visual content</p>	 <p>GT: A woman stands in the dining area at the table. A room with chairs, a table, and a woman in it. A woman standing in a kitchen by a window A person standing at a table in a room. A living area with a television and a table</p> <p>CNN-LSTM PM: a picture of a living room with a tv on it</p> <p>CNN-LSTM TFS: a living room with a television and a television.</p> <p>Attention TFS: a room with a wooden table and a tv stand with a tv on it</p> <p>Transformer TFS: a room with a wooden floor and a table with a chair and a chair in the background</p>
 <p>GT: A person is holding a bottle that has medicine for the night time. A bottle of medication has a white twist top. night time medication bottle being held by someone a person holding a small black bottle of NIGHT TIME A bottle of what appears to be cough syrup held in hand.</p> <p>CNN-LSTM PM: a person holding a cell phone in front of a laptop</p> <p>CNN-LSTM TFS: a person is holding a bottle of medicine in their hand</p> <p>Attention TFS: a bottle of hand sanitizer with a white label</p> <p>Transformer TFS: a bottle of some kind of liquid that is being held by someone's hand</p>	 <p>GT: A man that is on a tennis court with a racquet. there is a male tennis player wearing a blue shirt playing on the court A person standing on a blue floor holding a tennis racket A tennis player is standing on the court. a person standing on a tennis court holding a racquet.</p> <p>CNN-LSTM PM: a woman holding a tennis racket in a tennis court</p> <p>CNN-LSTM TFS: a pair of blue and white tennis shoes with a white and blue collar</p> <p>Attention TFS: a person wearing a blue and white striped shirt with a white and blue striped shirt on the front</p> <p>Transformer TFS: a pair of blue and white unk unk unk unk is on a blue table</p>

Conclusion

This work shows that modern state-of-the-art algorithms pre-trained on other visual-linguistic datasets are not suitable to achieve top performance on VizWiz-Caption. This proves that dataset distribution is unique to the task and requires model fine-tuning or training from scratch on VizWiz data.

For all the models presented in this work solid performance is achieved on custom dataset split. The best performer is CNN-LSTM model with attention and beam search inference with beam = 3.

As addition of attention to the model shows to be beneficial to boost performance on VizWiz-Caption, exploring another architectures with different types of attention is suggested for the next step. Moreover, further fine-tuning of Transformer model can be done.

References

- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan Show and tell: A neural image caption generator CoRR, abs/1411.4555, 2014
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio Show, Attend and Tell: Neural Image Caption Generation with Visual Attention arXiv:1502.03044
- <https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>
- <https://github.com/salaniz/pycocoevalcap/blob/master/cider/cider.py>

Our Team



Youchen Zhang



Husan-Yu-Lin