# Image Captioning

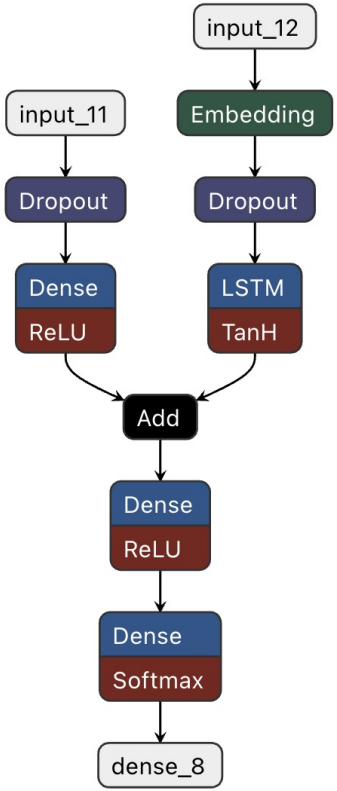
G. Aravind

# Proposed Approach

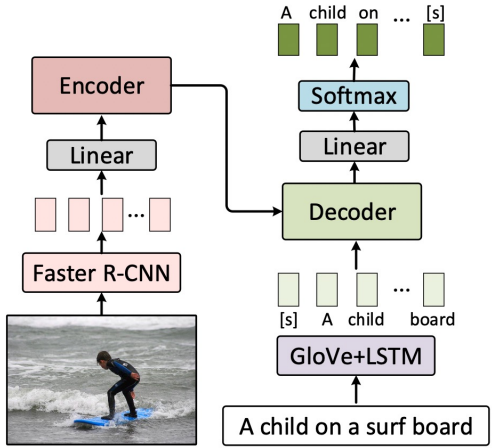
| Encoder-Decoder-I | Encoder-Decoder-II | Transformer based model |
|---|---|---|
| ✓ *Extract encoded features from the images using hid- den layers of Inception-V3 and EfficientNet-B0 models.* | ✓ *Instead of using Incpetion-V3 and EfficientNet-B0 models for detecting image features, implement a custom image-feature extractor using trainable CNN layers, and tune it to the dataset used for image captioning.* | ✓ *Used transformers for image captioning, in which the image encoder learns the deep im- age representation in a self-attention manner, then the caption decoder it to generate captions [1]* |
| ✓ *Create text features from captions using word embeddings. Also try Dynamic Meta-Embeddings [3] instead of ordinary embeddings here, followed by LSTMs.* | ✓ *Use Dynamic Meta-Embeddings [3] here as well and try to incorporate self-attention.* | ✓ *This model will capture both intra-modal and inter-modal interactions in a unified attention block* |
| ✓ *Decode the image and text features to predict the next word in the caption.* | | |

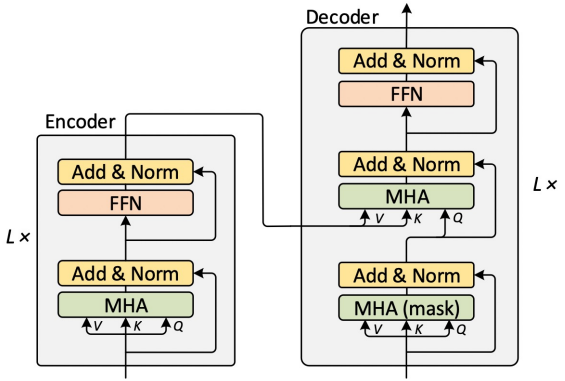
# Technical Details

## Basic Encoder-Decoder model




## Transformer based model



# *Contributions (Novelty)*

- ***Implemented** the non-transformer encoder-decoder models from scratch without using external code, and for transformer-based models [1], made key modifications to author's code [2] and architecture.*
- *Incorporated and implemented **Dynamic Meta-Embeddings** [3] myself, which enables the model to make use of multiple-word embeddings (Word2Vec, Glove and FastText) simultaneously using attention. This is a novel approach in image captioning, inspired by it's uses in other areas of NLP.*
- *Used **self-attention** at different layers of the architecture to see if the model performance can be improved. Also experimented with the architecture itself, while trying to improve the caption generation performance.*
- *Also tried to make use of Vision-Language pre-trained models like ViLBERT in the encoder-decoder architecture, but couldn't finish it in time as that proved to be too difficult to implement and generate results in such short time.*

# Results and Conclusion

a laptop computer sitting on top of a wooden table

| Architecture | BLEU Score [4] |
|---|---|
| Encoder-Decoder I | 0.23 |
| Encoder-Decoder II | 0.15 |
| Transformer Model | 0.29 |

a man riding a wave on top of a surfboard

a couple of giraffe standing next to each other

a cat that is sitting on top of a chair

a man riding a skateboard down a street

# References

[1] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang, “Multi- modal transformer with multi-view visual representation for image captioning,” 2019.

[2] Reference implementation for the third image-captioning architecture (transformer model): *https://github.com/MILVLG/mt-captioning*

[3] Douwe Kiela, Changhan Wang, and Kyunghyun Cho, “Dynamic meta-embeddings for improved sentence representations,” 2018.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, USA, 2002, ACL ’02, p. 311–318, Association for Com- putational Linguistics.