CaptionWizard

Haoran Tang, Deepayan Sur,
Morteza Damghani Nouri, Jeremy Jang
CSCI – 567: Machine Learning

MOTIVATION

"Bridging the Gap Between AI and Human Engagement"

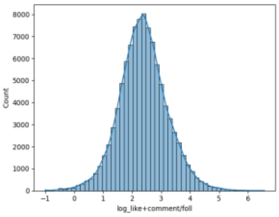
In the realm of social media, engagement is key. Current image captioning methods, while effective in describing content, lack the human touch that fosters connection and interaction. Our project is important because it addresses this gap, offering a solution that aligns with the way humans communicate and interact on social media. By generating captions that are not just accurate but also likable and engaging, we can significantly enhance the social media experience, both for individual users and brands looking to increase their online presence and engagement.

Caption generated by model:
a group of people with boxing mitts posing for a picture

Actual Caption:
ground: starting the weekend with a 🥊. thanks for having us

DATASET

- Our dataset moves beyond the factual, neutral tone of standard image captioning tasks like COCO and Flickr30k (e.g., "a man playing a guitar") to create more engaging captions for humans by incorporating qualitative aspects such as likes and comments.
- "**Multimodal Post Attentive Profiling for Influencer Marketing**," by Kim(ACM'20). selected images based on a likable threshold that is calculated by the number of likes and number of followers. Our Modified Scoring Formula: $\frac{log_{10}(Likes + Comments - 0.1 \times Followers)}{log_{10}(Followers)}$
- We have used only the top 75 percentile of this score, giving us a good uniform distribution of the data.

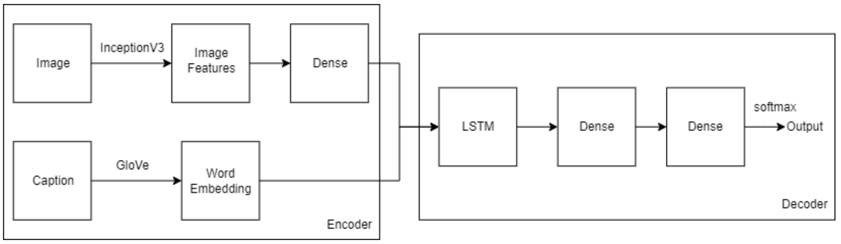


- We have concentrated on captions using English letters and words. To Remove non-English words we used a sentence-piece BiLSTM language detection model.
- We wanted to optimize engagement and enable us to set distinct thresholds for likability, tailored to various types of images, we are adopting a strategy that involves applying classification and clustering techniques. We experimented with the k-means algorithm to cluster our dataset.

MODEL ARCHITECTURE

Caption Preprocessing:

- Removing URLs, @mentions and #hashtags
- Removing non-English captions
- Removing Nan and empty captions from the dataset

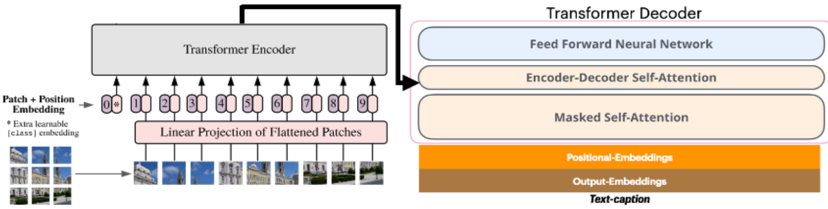


Model 1:

- Encoder-Decoder Architecture with InceptionV3 CNN encoder, and LSTM decoder
- Cross-Entropy loss function, ADAM optimizer
- Trained on 24k images, 20 epochs

Model 2:

- Base model: Vision Transformer (vit-image-captioning)
- Tokenization and Feature Extraction
- Finetuning the base model using our dataset and available resources


Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption

RESULTS

Base model caption: a woman in a pink shirt is holding a pink flower

CaptionWizard caption: a beautiful blonde haired girl in a pink dress with a pink bow

Base model caption: a woman with a black hair and a red tie

CaptionWizard caption: a beautiful blonde haired woman wearing a black dress and a black tie

Base model caption: a woman holding a cake with a smiley face

CaptionWizard caption: a little girl that is sitting at a table with a cake and cupcakes

Base model caption: a man and woman standing next to each other

CaptionWizard caption: a man and a woman standing in the grass by a lake

Future Works:

- Compare our model with other baseline models using common statistical metrics like METEOR, BLUE or ROUGE Score
- We are also working on a specialized metric for our use-case using a BERT embedding to compare with Instagram scraped data and the predicted outputs.