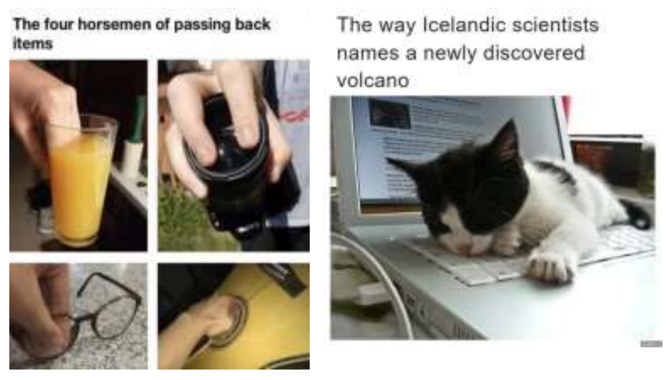
Live, Laugh, Learn: A Deep Learning Model for Meme Captioning

Presenters: Hisham Bhatti, Parum Misri, Samarjit Kaushik

CSE 493: Introduction to Deep Learning, Spring 2025

Introduction

Humor is **HARD**:
1. Need to understand the sociocultural context of an image
2. Humor is subjective and always evolving



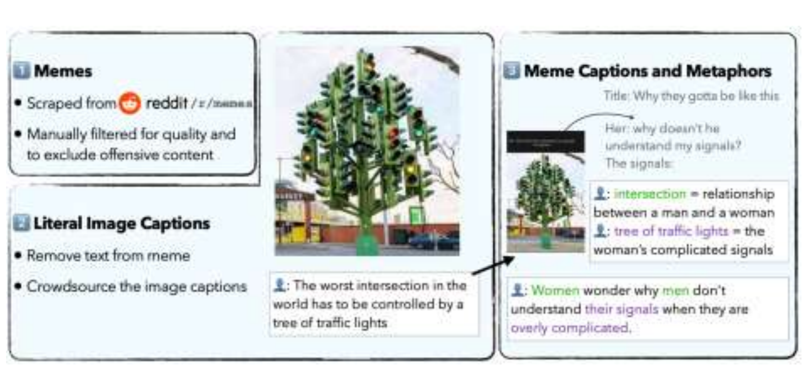
Existing Work

1. Lack of Diverse Data

4 Templates from MemeGenerator.net

2. Old/Generic Captions

3. Training Pipeline [4]



Goal

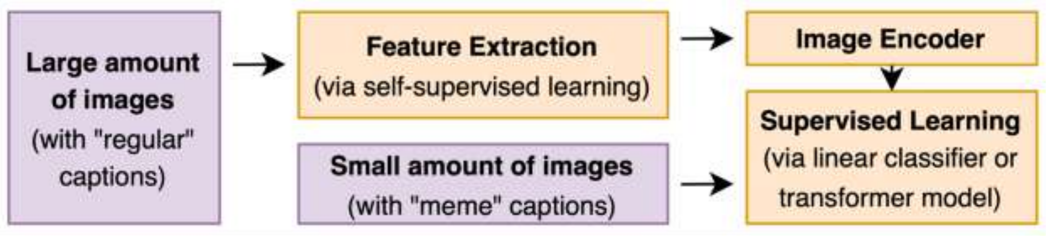
Develop a lightweight model that humorously captions diverse images using contextual understanding—rivaling larger LLMs with fewer parameters and less compute.

Conclusion

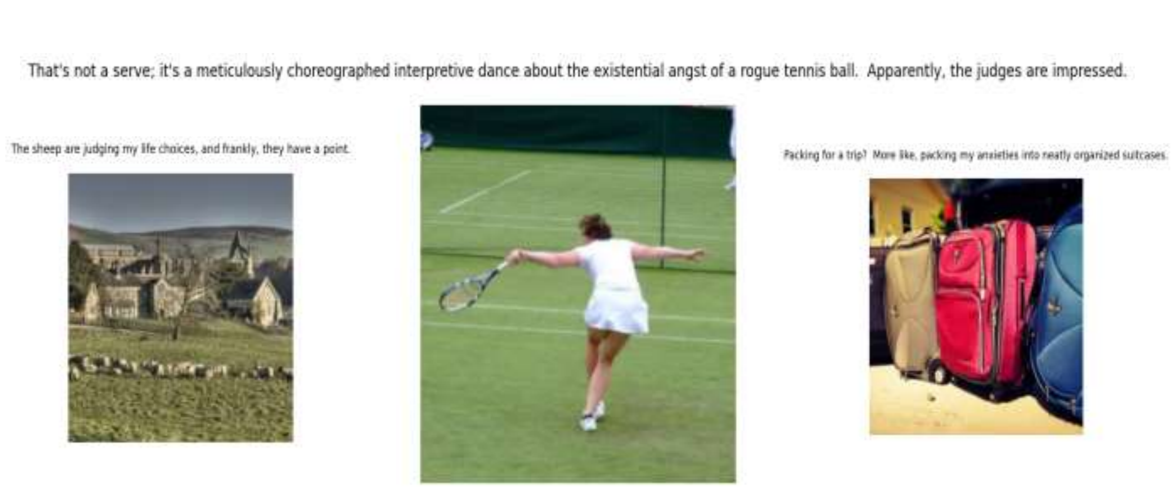
- Learning humor requires a diverse, massive, constantly-updating dataset
- Attributes unique to humor (diversity, inconsistent phrasing, subtle clues) make it especially hard to train

Future Work

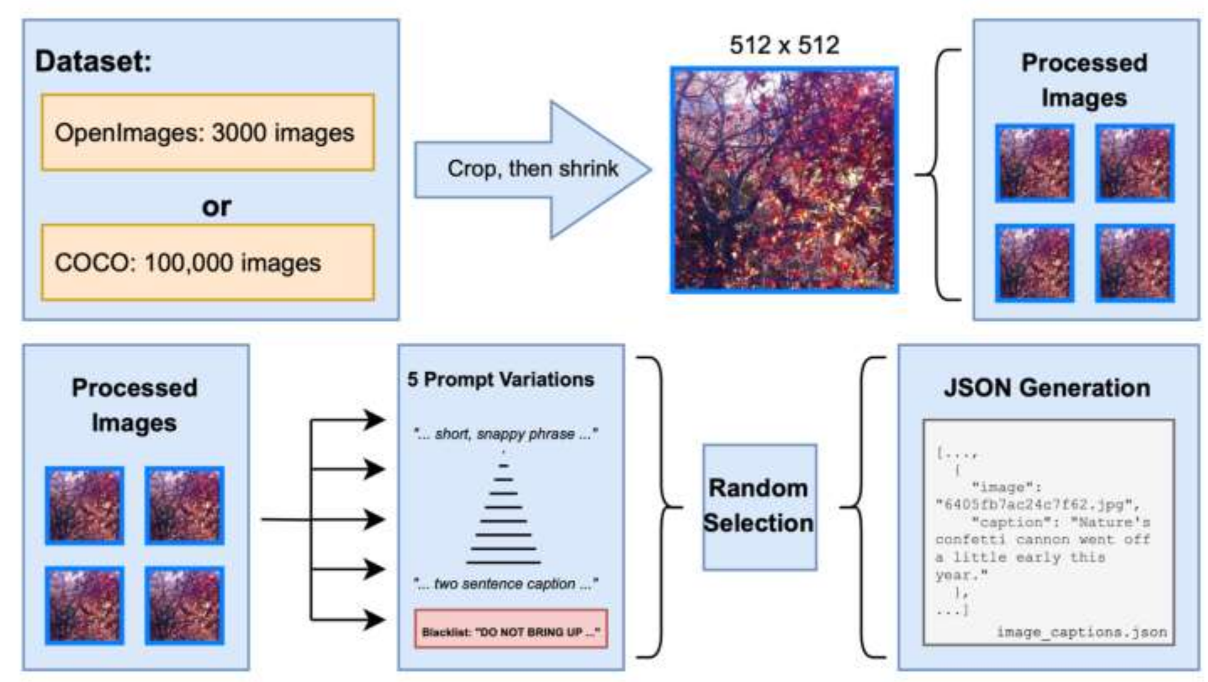
- Self-Supervised Learning
  - Pre-train on regular image captions, Post-train on meme captions



- Different architectures/data (e.g. a "meaning" along with caption)
- Increase the dataset size, quality (better prompts)



Data Collection



Q: *Why these datasets?*
A: Diverse dataset matches real-world images
Q: *Why do we generate our own captions?*
A: No meme caption dataset exists for diverse images
→ distillation via Gemini

You're a sharp, culturally-savvy wit tasked with crafting a humorous caption for an image. Keep it clever and dripping with wit.

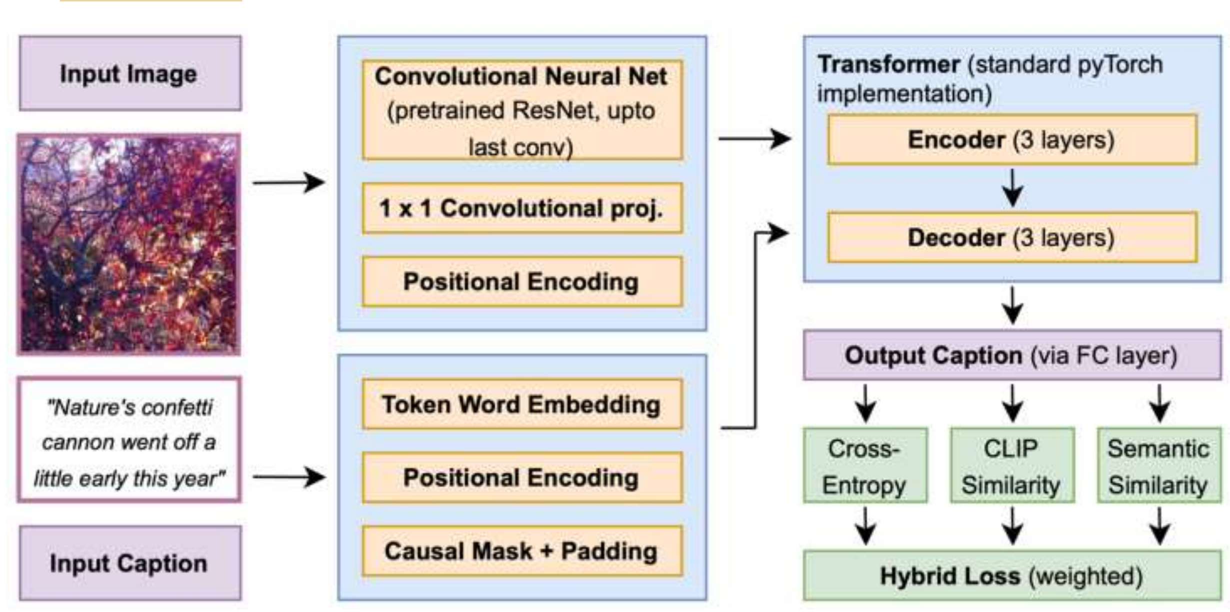
Instructions: {prompt_instruction} No Clichés Allowed: AVOID generic joke setups at ALL costs, like "when you...", "me: also me:", or "my therapist told me...". Stale won't cut it. Witty & Fresh: Go for unexpected twists, absurd spins, or sharp irony. Make it smirk-worthy, fast. Cultural Flavor: Capture universal themes—playful humor, cheeky wit, and absurd human quirks—that feel timeless and fresh, steering clear of specific trends or buzzwords. Image Hook: Don't narrate the image—grab its quirky essence and riff with wordplay or clever twists. Modern Vibe: Nail the bold, playful, and cheeky tone of today's viral posts. Output ONLY the caption.

WHEN GENERATING YOUR RESPONSE, AVOID THE FOLLOWING CLICHED SETUPS!

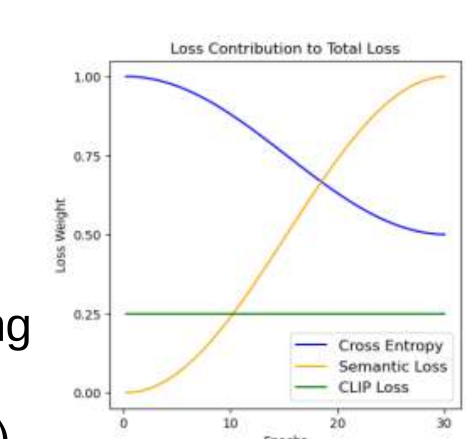
- DO NOT BRING UP THERAPISTS UNLESS THE IMAGE IS OF A THERAPIST OR IN A THERAPIST'S OFFICE.
- DO NOT USE THE SETUP "This [X] has more personality than [X]".
- DO NOT USE THE SETUP "My therapist said" or "My therapist told me".
- DO NOT SAY "I swear, [X]", and AVOID bringing up "my inner child".
- DO NOT MENTION "existential dread" or "existential crisis".
- DO NOT SAY "Seas the day".
- DO NOT SAY "This ain't my grandma's [X]".

Five different prompts and blacklist responses encourage diverse captions

Our Model

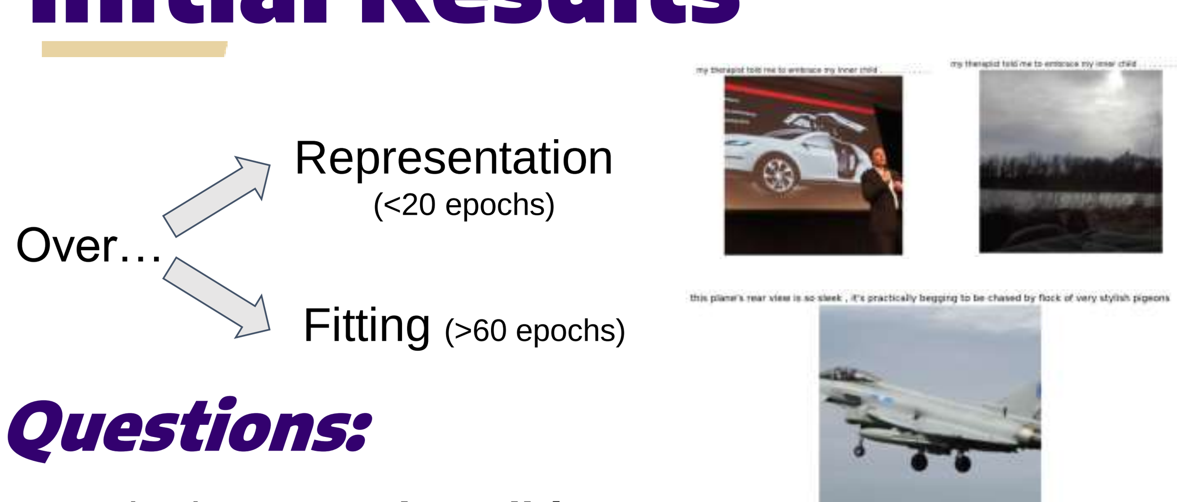


Loss weights change over the course of training (right):
- *Cross-Entropy:* text vs. text
- *Semantic similarity:* text embedding vs text embedding
- *CLIP Similarity:* image embedding vs text embedding (regularizer, doesn't backpropagate to model)



Initial Results

Over… Representation (<20 epochs)
Fitting (>60 epochs)



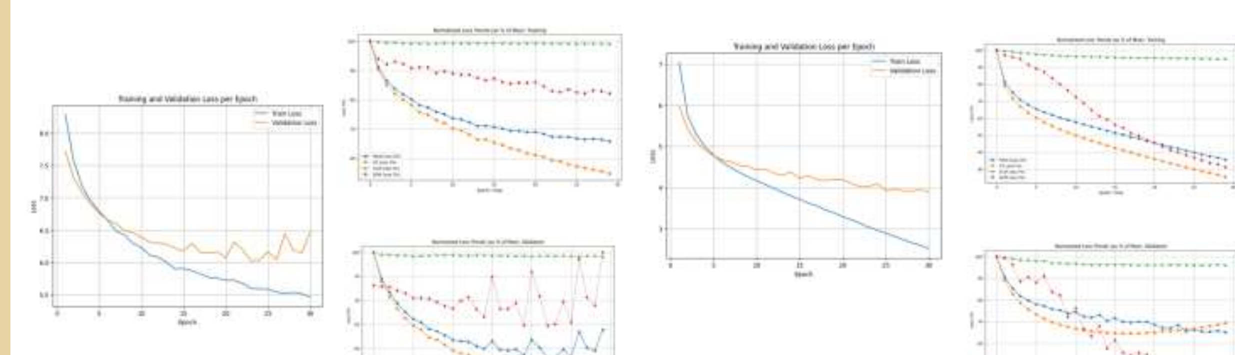
Questions:

1. Make it **comprehensible**?
2. How to balance underfitting/overfitting? (some sort of **semantic** score for early stopping)

1. Are we just training an **image classifier**, and then **overfitting** to image in training set?

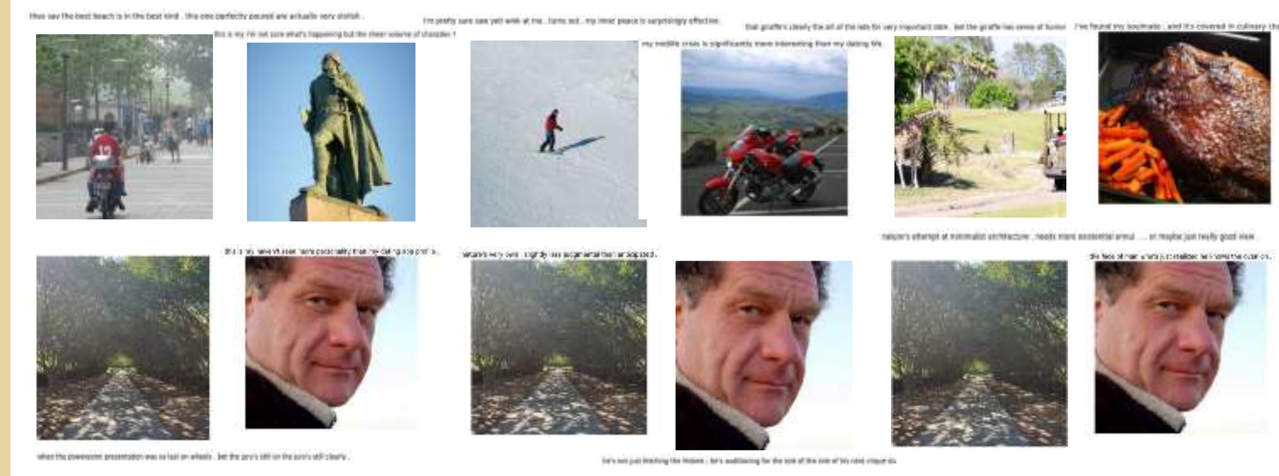
Experiments

1. More Training Data



3000 Images

60000 Images



3K Dataset

30K Dataset

60K Dataset

2. Byte-Pair Encoding (BPE)

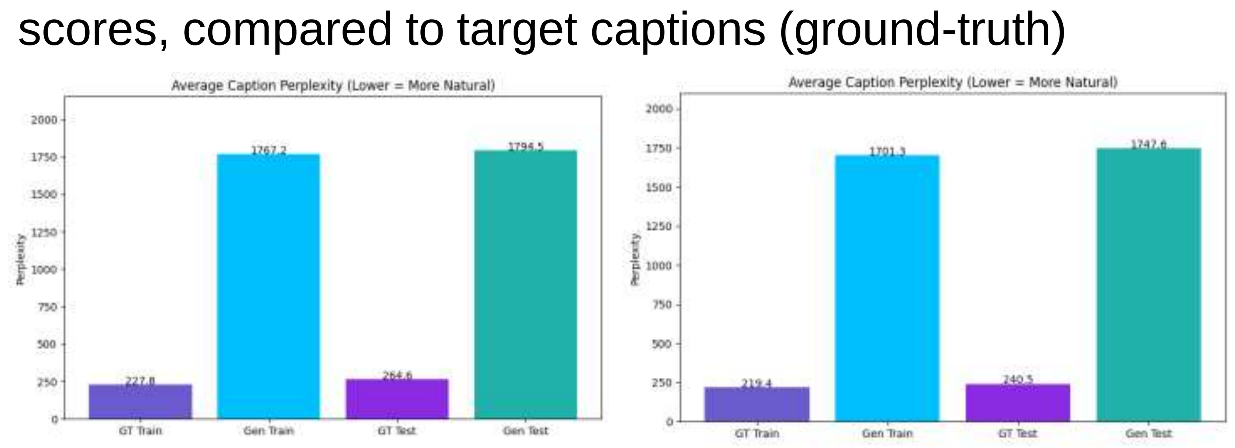
Vocab: 10K

Vocab: 20K

- Blocky with Inconsistent punctuation
- Hard to learn semantic structure w/o more data
- Reflects overuse of punctuation in meme captions

3. Semantics

Evaluating grammar and fluency via GPT2 perplexity scores, compared to target captions (ground-truth)

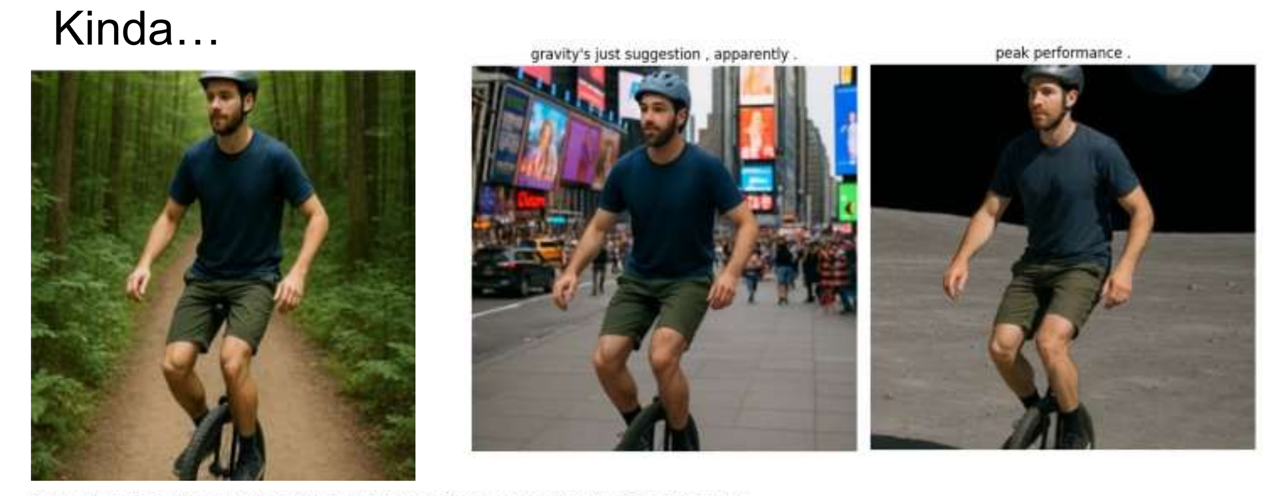


3000 Images

60000 Images

So … Does it Understand Context?

Kinda…



Is it funny?
…Sometimes :)

traffic lights nature's way of reminding us we're all just waiting for the traffic lights .

References

[1] Yuyan Chen, Songzhou Yan, Zhihong Zhu, Zhixu Li, and Yanghua Xiao. Xmecap: Meme caption generation with sub-image adaptability, 2024. 2
[2] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. ACM Computing Surveys, 56(3):1–39, Oct. 2023. 1, 2
[3] EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes, 2023. 2
[4] Abel L Peirson V and E Meltem Tolunay. Dank learning: Generating memes using deep neural networks, 2018
[5] Andreas Refsgaard and Frederik Lauenborg. MemeCam, 2023.

Acknowledgements: Prof. Ranjay Krishna and the entire Deep Learning Staff