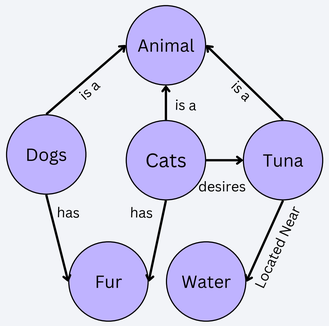
Enhancing Multimodal Learning with Knowledge Graphs

Ananya Kulshrestha, Ethan Harbaugh, Julianna Schneider

Motivation

Multimodal models hold great potential to address **tasks in accessibility**, where certain modalities may be inaccessible to users due to health conditions
→ We focus on image captioning as an exemplary case

Challenge
Capturing semantic information across complex object relationships seen in everyday scenes



Solution
Leverage the strong situational and relational inductive bias of Knowledge Graphs to guide learning

Related Work & System Design

We present a language model that uses CLIP image embeddings to generate image captions via recurrent NN.

Training Dataset
COCO [3]:
- 17.9K diverse, general purpose image-caption pairs

Images → CLIP Embeddings:
- Pre-compute CLIP [6] image embeddings for all COCO images

Captions → KG Embeddings
- Pre-compute word2vec style embeddings from ConceptNet5 [1] Numberbatch for all COCO captions

Evaluation Dataset
Viz Wiz:
- 5k image-caption pairs: images taken by blind people, captions written by seeing annotators

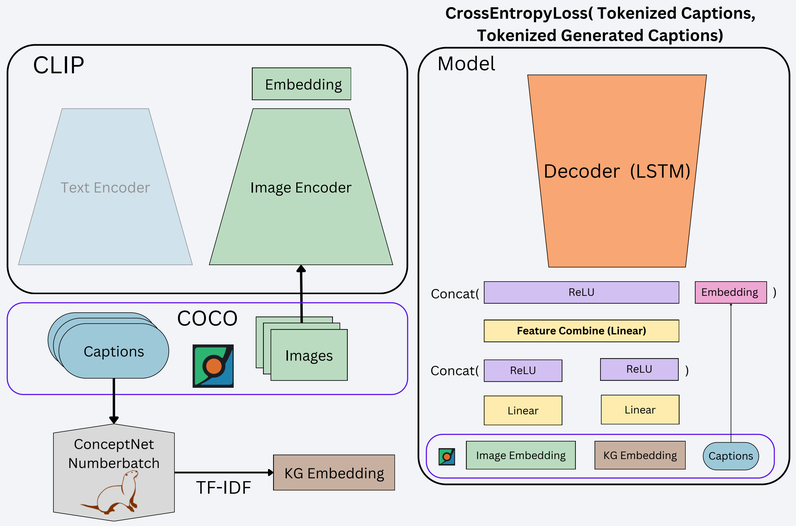
Images → CLIP Embeddings:
- Pre-compute CLIP [6] embeddings as in train

No KG information provided to avoid leaking caption

Dataset Integration
Vocabulary: union of words across train & eval captions

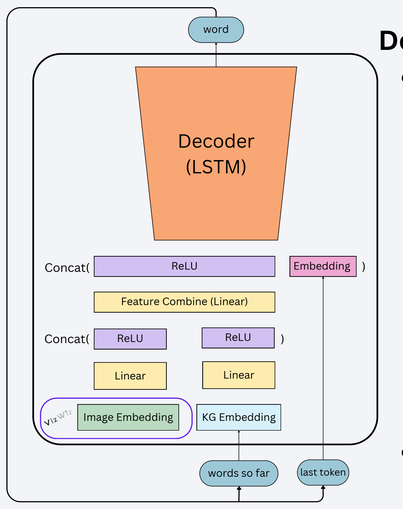
Methods

Training
- Predict next word using image embedding, target KG embedding, and trained embedding of last *target* word



$$E_{\text{caption}} = \frac{\sum_{w \in \text{caption}} \text{TF-IDF}(w) \cdot E_{\text{word}}^{(\text{Numberbatch})}}{\sum_{w' \in \text{caption}} \text{TF-IDF}(w')}$$

Inference
- Predict next word using image embedding, prediction KG embedding, and trained embedding of last *predicted* word.
- Apply trained decoder in autoregressive fashion to generate a caption of 20 words at most.



Decoding
- Modified beam search
  - Length normalization
    - Avoids bias towards shorter captions
  - Trigram blocking
    - Disallow repeated trigrams
  - Repetition penalty
    - Incentivize diverse word choice
- Maintain 3 beams at a time

Results & Discussion

	BLEU1	ROUGE-L	CIDEr	SPICE
Without KG	0.2288	0.1251	**0.0247**	0.0182
With KG	**0.2835**	**0.1438**	0.0158	**0.0284**

- Strengths:
  - KG inductive bias results in higher scores across three out of four standard image captioning benchmarks
  - Beam search modifications mitigate repetition issues in greedy decoding
- Limitations:
  - Low expressive power of our shallow architecture
  - CV challenge posed by low-quality images
  - Crowdsourced captions for challenging labeling task

Outlook
→ Proof of concept of utility of KGs in multimodal models
→ Room for improvement in strength of generative output

		Captures Relations	Correct Category	Accuracy	Diversity
No KG	a **black dog** lying next to pick up close shot of **orange juice carton** on the **blender with bat**	✗	✗	😐	✓
KG	a **person on an image** of someone is another man holding a **woman guy one boy** in black	✓	✓	😐	😐
Target	A view of a building with a picture of a man on it with a couple big trees in front of it.	✓	✓	✓	✓

Acknowledgements and References

Thank you to Professor Andreas, Dr. Tanner, Dr. Maune, and the entire 6.8611 course staff for their invaluable guidance and support!

[1] Robyn Speer, Joshua Chin, and Catherine Havasi (2017). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In proceedings of AAAI 2017.
[2] "Captioning Images Taken by People Who Are Blind." Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. European Conference on Computer Vision (ECCV), 2020.
[3] Lin et al. Coco: Common objects in context, 2017. https://cocodataset.org/home.
[4] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning." 2018.
[5] How to Generate Text – A Guide to Text Generation with Transformers. Hugging Face. https://huggingface.co/blog/how-to-generate.
[6] Radford et al. "Learning Transferable Visual Models From Natural Language Supervision." Proceedings of the 38th International Conference on Machine Learning (ICML). 2021.