Enhancing Generative Commonsense Reasoning Using Image Cues

Hwanjun (Michael) Yi[1], Soumya Sanyal[2], Xiang Ren[2]

[1]Portola High School, [2]University of Southern California

23yimichael@gmail.com, soumyasa@usc.edu, xiangren@usc.edu

Generative Commonsense Task

Commongen Task [1]: Generate coherent sentences given their respective keywords (concepts) and a corresponding image.

dog, frisbee, catch, throw

A dog leaps to catch a thrown frisbee.

In this work, we compare the current Commongen model with a visual question answering approach with a fine-tuned VL-T5 baseline.

Motivation To Use Images

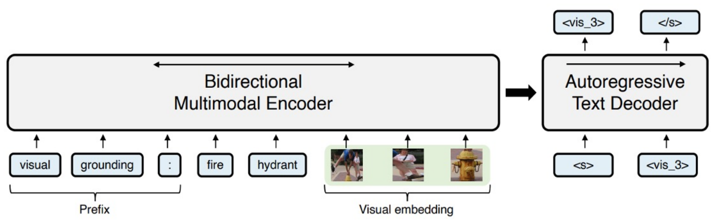
- Mimic how humans would approach the Commongen task:
  - **Develop a scene** using the concepts
  - **Make sentence** adapted from the scene
- Visual information gained from the image simulates the scene development process and thus **enhances commonsense reasoning**

Methods

VL-T5 Fine-Tuning

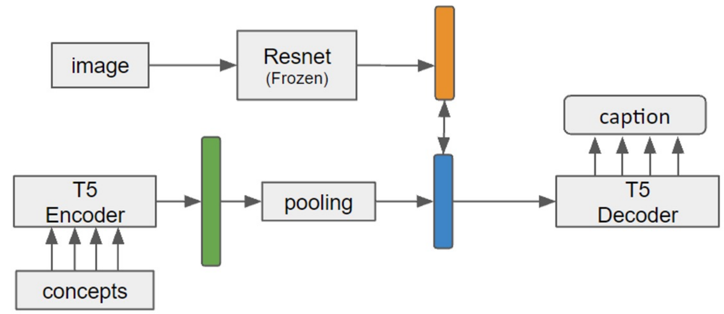
1. **Extract image features** using Detectron2, an object detection algorithm **[2]**.
2. **Visual question answering**: ask question → model answers. Example: “vqa: what is the image caption using concepts: dog, frisbee, catch, throw?”
3. The question and the image features are **processed** by VL-T5’s bidirectional multimodal **encoder [3]**.
4. The vector that the encoder generates is sent to the model’s autoregressive text **decoder**, thus **generating** our desired **sentence [3]**.

Created by Wu et. al. 2019 at Facebook AI Research, the figure illustrates all the features extracted from an image of a bike race using Detectron2 **[2]**.



Created by Cho et. al. 2021 at University of North Carolina at Chapel Hill, the figure illustrates how both text and visuals are processed to generate a text output within the VL-T5 architecture **[3]**.

Commongen Model



1. **Process** the concepts by using an **encoder** adapted from T5, a text-to-text baseline model developed by Google, and a **pooling** layer to generate the **concept embeddings**.
2. **Process** the image by using the **ResNet** deep residual neural network to generate the **image embeddings**.
3. **Calculate the contrastive loss $J_t(\theta)$** between the newly generated image and concept embeddings to inject visual knowledge.
4. Use the T5 model’s **decoder** to generate the desired sentence using the vision-injected vector (shown in blue).

$$J_t(\theta) = \log \sigma\left(u_o^T v_c\right) + \sum_{j \sim P(w)} \left[\log \sigma\left(-u_j^T v_c\right)\right]$$

Results

Model Name	Structure	BLEU	SPICE
T5-base	Concepts → Sentence	31.96	28.86
iCommongen-mean	Concepts + Image → Sentence	33.27	29.447
I&V (T5-base)	Concepts + Scene Graphs → Sentence	**40.16**	**30.57**
VisCTG (T5-base)	Concepts + Caption → Sentence	34.722	28.808

- Performance metrics used:
  - **BLEU** assesses the quality of text relative to **human translation**
  - **SPICE** evaluates the quality of captions relative to their respective **image**
- Using images is helpful w.r.t. T5-base, indicating that visual information **enhances commonsense reasoning.**
- Model underperforms compared to baselines that use **scene graphs or image captions** instead of images, showing that the image information is likely suboptimal.
- **Future Direction**: Use pre-trained vision-language models (CLIP) to better encode the vision and the concepts, instead of ResNet.

Acknowledgements

I’d like to thank Prof. Xiang Ren, my mentor Soumya Sanyal, the INK Lab researchers, Dr. Katie Mills, my centermentor Michelle Emelle, and my family for providing this opportunity to conduct research at a college level. Doing so as a high school student has been nothing short of a privilege, one that all of these people have provided for me this summer at SHINE.

References

[1] Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., & Ren, X. (2020, November 30). Commongen: A constrained text generation challenge for Generative Commonsense reasoning. arXiv.org. Retrieved July 22, 2022, from https://arxiv.org/abs/1911.03705

[2] Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. GitHub. Retrieved July 22, 2022, from https://github.com/facebookresearch/detectron2

[3] Cho, J. (2021). Unifying vision-and-language tasks via text generation. arXiv. Retrieved July 22, 2022, from https://arxiv.org/pdf/2102.02779.pdf