Figuring Out Figures: Using Textual References to Caption Scientific Figures

Stanley Cao, Kevin Liu
CS224N Final Project, Winter 2022

The Problem & Background

Figures are essential channels for densely communicating complex ideas in papers. Automatic captioning could:

1. Reduce the burden for authors to write high-quality captions;
2. Improve accessibility to impaired readers;
3. Extract useful information as text from figures that could be used in later pipelines.

However, captioning scientific figures is harder than normal images:

- State of the art models rely on region/object extraction (unavailable for figures).
- Even after trying more advanced architectures, best performance is currently achieved with a CNN+LSTM.

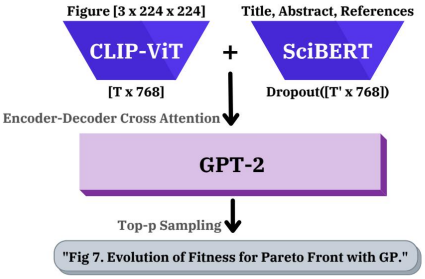
Data

- **SciCap dataset**, curated by Hsu et al. as a starting point, which contains **figures and captions** for CS papers from 2010-2020.
  - **133K** figures without subfig, split 80/10/10 train/val/test
  - Extracts first sentence of caption
- arXiv metadata to pull **titles and abstracts** for each paper
- Custom data pipeline to search the full text of each paper for **references**. Original verbatim caption found through Striped Smith-Waterman algorithm and masked out.

Methods

- Modeled as a sequence to sequence problem
- Fully transformer architecture
- **Vision Encoder Decoder architecture**: (figure, title, abstract, references) → caption
  - **Encoder**: CLIP ViT-B/32 + SciBERT-base (text features)
  - CLIP used as model trained on general Internet imagery
  - **Decoder**: GPT-2 base model or DistilGPT2
- **We used the original captions from SciCap**, rather than the normalized ones, to encourage equation/number prediction
- To generate the caption, we use top-p sampling with $p = 0.9$. A diagram is shown to the right.

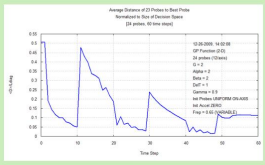
Model Architecture



Data Example — Figure 1/13.3K (test set)

INPUTS

SciCap:



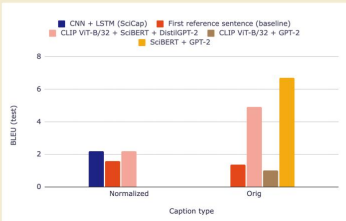
Title: Pseudorandomness in Central Force Optimization
Abstract: Central Force Optimization is a deterministic metaheuristic for...
References: ...Fig. 7 plots CFO's " Davg curve" for GP. Davg is the normalized average distance between the probe with th...

OUTPUTS

Ground truth caption: Fig. 7. Evolution of GP Function avgD
Predicted: Fig. 7. Evolution of Fitness for Pareto Front with GP.

Experiments

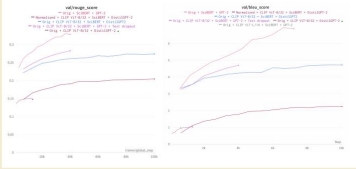
- Ablations over image/text inputs performed
- Extensive model/hyperparameter search: DeiT, GPT-2, CLIP ViT/L-14, lr scheduling, text dropout to encourage image usage



Analysis

- Reference data improves performance over pure image input
- However, creates balance issues between modalities (text-only outperforms image + text models)
- Potential benefits from scale, text dropout (improvements from ViT-L)

Image?	Text?	Caption type	Model	BLEU (test)	ROUGE-L F1 (test)
✓		Normalized	CNN + LSTM (SciCap)	2.19	—
	✓	Normalized	First reference sentence (baseline)	1.59	0.09
✓	✓	Normalized	CLIP ViT-B/32 + SciBERT + DistilGPT-2	2.21	0.18
	✓	Orig	First reference sentence (baseline)	1.38	0.10
✓	✓	Orig	CLIP ViT-B/32 + SciBERT + DistilGPT-2	4.92	0.26
✓		Orig	CLIP ViT-B/32 + GPT-2	1.02	0.13
✓	✓	Orig	SciBERT + GPT-2	6.71	0.30



Conclusion

- Transformer model architecture achieves a better performance than the CNN+LSTM model when it is given textual metadata (e.g., references to the figure, the title, and the abstract).
- However, because our model learns more from textual references, experimentation should be done with more expressive image encoders, or perhaps improving the image encoder architecture.
- Pre-processing the image a valuable path forward (e.g., vectorizing, LaTeX), because the vision encoder could leverage the patterns found in the consistency from image pre-processing.

References

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3258–3264, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. 2021b. doi: 10.48550/ARXIV.2103.00020. URL https://arxiv.org/abs/2103.00020.