INLG 2023

# Summaries as Captions: Generating Figure Captions for Scientific Documents with Automated Text Summarization

Chieh-Yang Huang,[1*] Ting-Yao Hsu,[1*] Ryan Rossi,[2] Ani Nenkova,[2]
Sungchul Kim,[2] Gromit Yeuk-Yin Chan,[2] Eunyee Koh,[2]
Clyde Lee Giles,[1] Ting-Hao 'Kenneth' Huang[1]

[1] Penn State University, PA, USA {chiehyang, txh357, clg20, txh710}@psu.edu
[2] Adobe Research, CA, USA {ryrossi, nenkova, sukim, ychan, eunyee}@adobe.com
*Equal Contribution

Crowd-AI Lab
crowdailab.net

Adobe

# Introduction

- How to generate **high-quality captions for scientific figures**?
  - Existing vision-based approaches fail to generate reasonable captions.
  - A huge portion of the captions in real-word data are poorly written.
  - What do a "high-quality" caption need?

**Any other information we can use?**

**Target Figure**

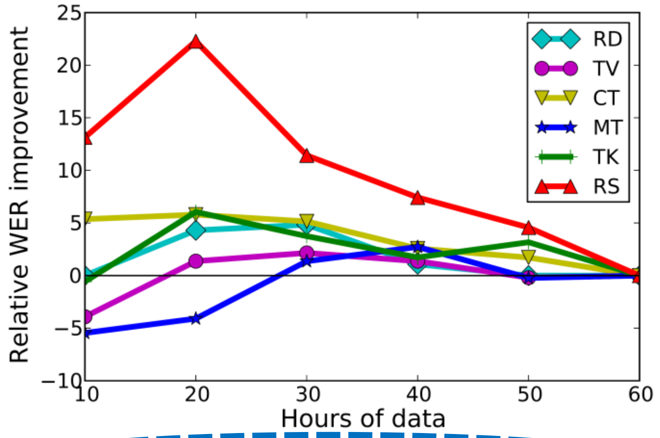



Figure 2: *WER improvement with budget–based data selection*

**Author-Written Caption**

Figure 2: **WER improvement with budget–based data selection**

# Introduction

- **YES!**

With **Awesome-Alignment**, we found that 75% of the information in the **caption** can be identified in the **figure-mentioning paragraphs**.

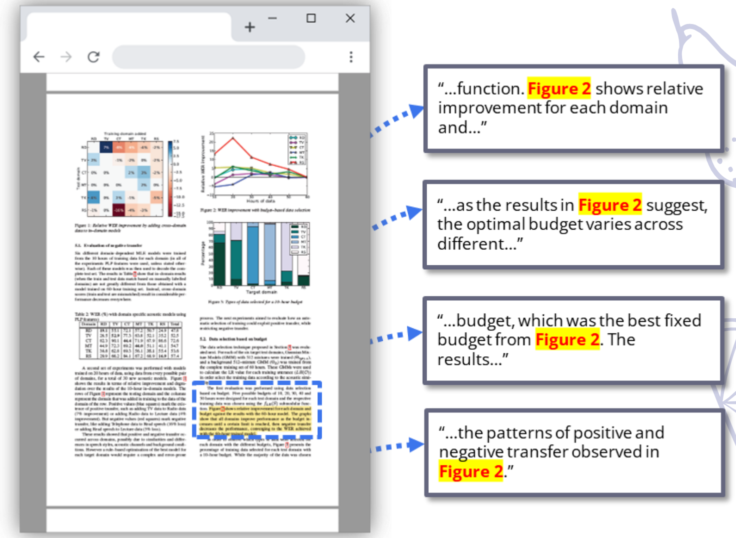
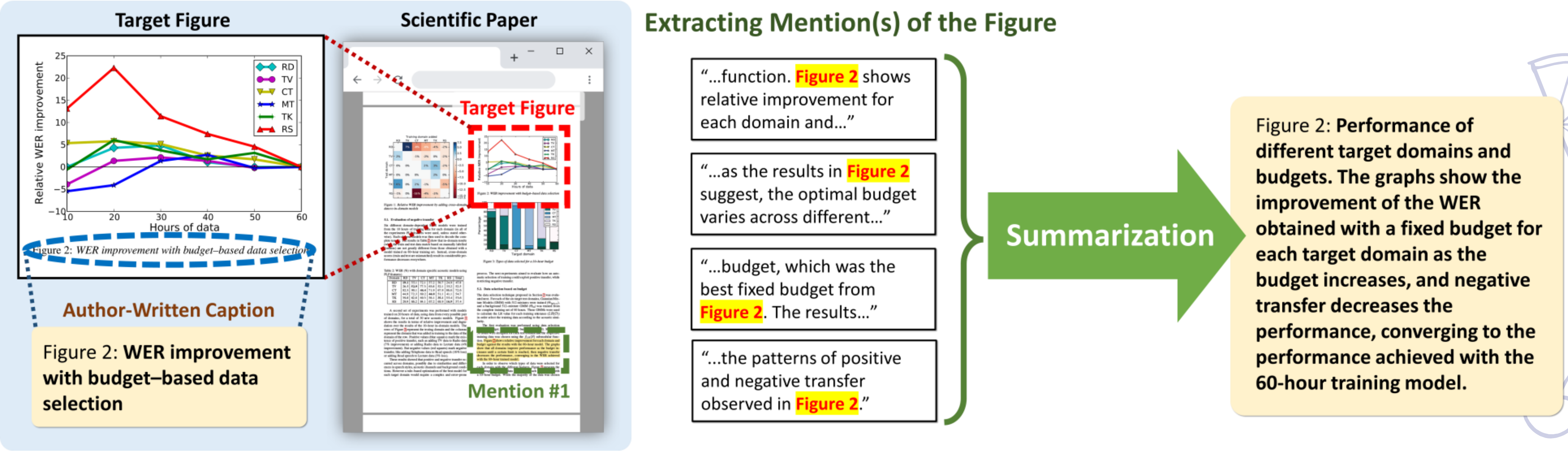



**Figure-mentioning texts contain 75% of the information needed to create a caption!**

➡ **How about solving it as a summarization task?**

# Introduction

- We formulate the scientific figure captioning task as a **summarization task**, and it works!



# Automatic Evaluation

- We trained a **Pegasus** model, taking figure-mentioning paragraphs as the input and generate the caption.
- All the experiments were conducted on **SciCap dataset**.
- Pegasus with Paragraph+OCR outperforms vision-based approaches!

| Model | Feature | Length | Rouge-1 (F1) | | Rouge-2 (F1) | | Rouge-L (F1) | | MoverScore | | BERTScore | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Score | Norm | Score | Norm | Score | Norm | Score | Norm | Score | Norm |
| Pegasus | P | 14.0 | .374 | 2.067 | .205 | 3.507 | .334 | 2.201 | .570 | 1.095 | .682 | 1.196 |
| | P+O | 14.0 | **.381** | **2.106** | **.212** | **3.635** | **.340** | **2.242** | **.571** | **1.097** | **.685** | **1.202** |
| | P+O+B | 38.3 | .321 | 1.452 | .154 | 1.916 | .265 | 1.537 | .546 | 1.044 | .639 | 1.082 |
| TrOCR | Figure | 10.0 | .220 | 1.464 | .073 | 1.653 | .195 | 1.502 | .534 | 1.033 | .610 | 1.096 |
| BEiT+GPT2 | Figure | 15.8 | .164 | 0.864 | .042 | 0.666 | .144 | 0.917 | .529 | 1.013 | .592 | 1.031 |

# How do human feel?

- The Mturk study indicates that vision-based model performs significantly **worse**.
- The domain expert study indicates $\text{Pegasus}_{P+O+B}$ is **ranked similarly** to ground-truth captions.

**Mturk Study on selecting "which one is the worst?"**

| n = 90 | #Maj. Votes↓ | Avg. Votes↓ | T-Test over Avg. Votes | | |
|---|---|---|---|---|---|
| | | | $\text{Peg}_{P+O}$ | $\text{Peg}_{P+O+B}$ | Caption |
| TrOCR | 41 | 5.99 | <.001*** | .006** | .001** |
| $\text{Peg}_{P+O}$ | 20 | 4.54 | - | .253 | .973 |
| $\text{Peg}_{P+O+B}$ | 24 | 4.93 | - | - | .318 |
| Caption | 19 | 4.53 | - | - | - |

**Domain Expert Study on ranking "which one is the best"**

| n = 90 | Avg. Ranking↓ | T-Test on Avg. Ranking | |
|---|---|---|---|
| | | $\text{Peg}_{P+O+B}$ | Caption |
| $\text{Peg}_{P+O}$ | 2.152 | .016* | .015* |
| $\text{Peg}_{P+O+B}$ | 1.930 | - | .923 |
| Caption | **1.919** | - | - |

**$\text{Pegasus}_{P+O+B}$**: Pegasus model but trained on caption with better quality (captions longer than 30 tokens).

# Conclusion

- Scientific figure captioning task can be solved via **text summarization.**
- Handling the **low-quality captions** in the dataset is challenging and will be something we should explore next.
- Filling the **missing 25% information** will probably still require the information from figures.

**Thanks! Please refer to our paper for more information.**

https://arxiv.org/abs/2302.12324