

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

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

<sup>1</sup> Penn State University, PA, USA {chiehyang, txh357, clg20, txh710}@psu.edu

<sup>2</sup> Adobe Research, CA, USA {ryrossi, nenkova, sukim, ychan, eunyee}@adobe.com

\*Equal Contribution



**Crowd-AI Lab**  
crowdailab.net

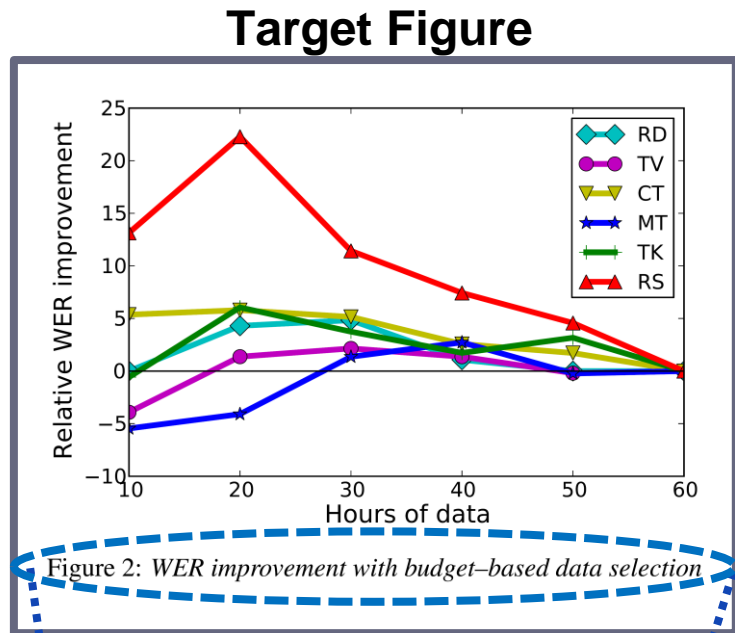


**Adobe**

# 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-world data are poorly written.
  - What do a “high-quality” caption need?

➔ **Any other information we can use?**



## Author-Written Caption

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

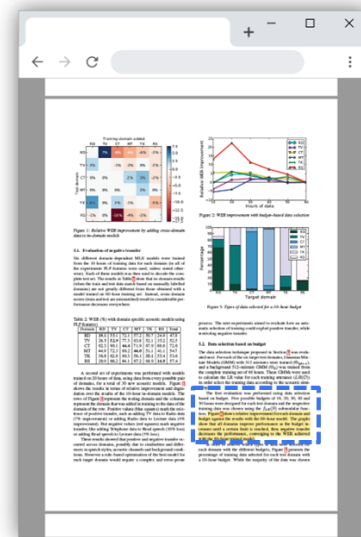
# Introduction

• **YES!**

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



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



“...function. **Figure 2** shows relative improvement for each domain and...”

“...as the results in **Figure 2** suggest, the optimal budget varies across different...”

“...budget, which was the best fixed budget from **Figure 2**. The results...”

“...the patterns of positive and negative transfer observed in **Figure 2**.”

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

# Introduction

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

**Target Figure**

Hours of data	RD	TV	CT	MT	TK	RS
10	0	0	5	-2	0	15
20	5	2	5	2	0	22
30	5	2	5	2	0	12
40	3	2	5	2	0	8
50	3	2	5	2	0	5
60	0	0	0	0	0	0

**Scientific Paper**

**Author-Written Caption**

Figure 2: WER improvement with budget-based data selection

## Extracting Mention(s) of the Figure

"...function. **Figure 2** shows relative improvement for each domain and..."

"...as the results in **Figure 2** suggest, the optimal budget varies across different..."

"...budget, which was the best fixed budget from **Figure 2**. The results..."

"...the patterns of positive and negative transfer observed in **Figure 2**."

Summarization

Figure 2: Performance of different target domains and budgets. The graphs show the improvement of the WER obtained with a fixed budget for each target domain as the budget increases, and negative transfer decreases the performance, converging to the performance achieved with the 60-hour training model.

# 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	<b>P</b>	14.0	<u>.374</u>	<u>2.067</u>	<u>.205</u>	<u>3.507</u>	<u>.334</u>	<u>2.201</u>	<u>.570</u>	<u>1.095</u>	<u>.682</u>	<u>1.196</u>
	<b>P+O</b>	14.0	<b>.381</b>	<b>2.106</b>	<b>.212</b>	<b>3.635</b>	<b>.340</b>	<b>2.242</b>	<b>.571</b>	<b>1.097</b>	<b>.685</b>	<b>1.202</b>
	<b>P+O+B</b>	38.3	.321	1.452	.154	1.916	.265	1.537	.546	1.044	.639	1.082
<b>TrOCR</b>	<b>Figure</b>	10.0	.220	1.464	.073	1.653	.195	1.502	.534	1.033	.610	1.096
<b>BEiT+GPT2</b>		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 Pegasus<sub>P+O+B</sub> 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		
			Peg <sub>P+O</sub>	Peg <sub>P+O+B</sub>	Caption
<b>TrOCR</b>	41	5.99	<.001***	.006**	.001**
<b>Peg<sub>P+O</sub></b>	20	4.54	-	.253	.973
<b>Peg<sub>P+O+B</sub></b>	24	4.93	-	-	.318
<b>Caption</b>	19	4.53	-	-	-

## Domain Expert Study on ranking “which one is the best”

n = 90	Avg. Ranking↓	T-Test on Avg. Ranking	
		Peg <sub>P+O+B</sub>	Caption
<b>Peg<sub>P+O</sub></b>	2.152	.016*	.015*
<b>Peg<sub>P+O+B</sub></b>	1.930	-	.923
<b>Caption</b>	1.919	-	-

**Pegasus<sub>P+O+B</sub>**: 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.

The background of the slide is white and decorated with various colorful, stylized illustrations of fruits and vegetables. These include a blue kiwi, a slice of orange, a green leafy vegetable, a green lime, a yellow lemon, a yellow banana, a green leafy vegetable, a yellow slice of orange, a red tomato, a yellow banana, and a whole orange. The illustrations are scattered around the edges of the slide.

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

**<https://arxiv.org/abs/2302.12324>**