ARIA
APPLIED RESEARCH IN ACTION

Image Caption Generation with Machine Learning for AR Wearable Devices

Elevate AR wearable devices with AI-generated image captions: Optimizing user focus, experience, and productivity.

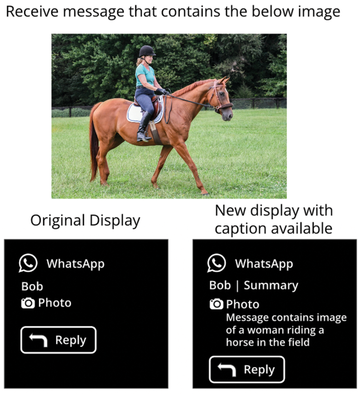
Yu-Chieh Wu

Anthony Bonner
ACADEMIC SUPERVISOR

Rashi Karanpuria, Shiyu Zhang
INDUSTRY SUPERVISORS

Model/Metrics	BLEU	ROUGE	CIDEr
PaLI	31.7%	49.6%	0.98
MUM	17.1%	36.7%	0.55

*PaLI: A Jointly-Scaled Multilingual Language-Image Model
*MUM: Multitask Unified Model - A new AI milestone for understanding information



PROJECT SUMMARY

AR wearable devices overlay digital information onto the user's real-world visual perception and create an interactive and enriched environment. Currently, in order to keep users focused on important hands-on tasks and only to notify them in the least disruptive manner, we do not display images in messaging notification on our AR wearable devices. In this project, we aim to research vision-language models that are capable of performing image captioning tasks, as well as selecting the one that best suits our need to integrate to our product. We evaluated and compared different internal and external models from different perspectives including performance, model size, latency, multilingual support, and compatibility with wearable devices. As a result, we were able to successfully identify and integrate a suitable vision-language model into our pipeline to support captioning for images in messaging notification, and keep users focused on their tasks by showing the summary of the image instead of the entire image.

Google

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing