Microsoft’s new image-captioning AI will help accessibility in Word, Outlook, and beyond

Microsoft has developed a new image-captioning algorithm that exceeds human accuracy in certain limited tests. The AI system has been used to update the company’s assistant app for the visually impaired, Seeing AI, and will soon be incorporated into other Microsoft products like Word, Outlook, and PowerPoint. There, it will be used for tasks like creating alt-text for images — a function that’s particularly important for increasing accessibility.

“Ideally, everyone would include alt text for all images in documents, on the web, in social media — as this enables people who are blind to access the content and participate in the conversation,” said Saqib Shaikh, a software engineering manager with Microsoft’s AI team in a press statement. “But, alas, people don’t. So, there are several apps that use image captioning as way to fill in alt text when it’s missing.”

The new algorithm is twice as good as its predecessor says Microsoft

These apps include Microsoft’s own Seeing AI, which the company first released in 2017. Seeing AI uses computer vision to describe the world as seen through a smartphone camera for the visually impaired. It can identify household items, read and scan text, describe scenes, and even identify friends. It can also be used to describe images in other apps, including email clients, social media apps, and messaging apps like WhatsApp.

Microsoft does not disclose user numbers for Seeing AI, but Eric Boyd, corporate vice president of Azure AI, told *The*

Verge the software is “one of the leading apps for people who are blind or have low vision.” Seeing AI has been voted best app or best assistive app three years in a row by AppleVis, a community of blind and low-vision iOS users.

Microsoft’s new image-captioning algorithm will improve the performance of Seeing AI significantly, as it’s able to not only identify objects but also more precisely describe the relationship between them. So, the algorithm can look at a picture and not just say what items and objects it contains (e.g., “a person, a chair, an accordion”) but how they are interacting (e.g., “a person is sitting on a chair and playing an accordion”). Microsoft says the algorithm is twice as good as its previous image-captioning system, in use since 2015.

The algorithm, which was described in a pre-print paper published in September, achieved the highest ever scores on an image-captioning benchmark known as “nocaps.” This is an industry-leading scoreboard for image captioning, though it has its own constraints.

The nocaps benchmark consists of more than 166,000 human-generated captions describing some 15,100 images taken from the Open Images Dataset. These images span a range of scenarios, from sports to holiday snaps to food photography and more. (You can get an idea of the mixture of images and captions by exploring the nocaps dataset here or looking at the gallery below.) Algorithms are tested on their ability to create captions for these pictures that match those from humans.

Grid View



It’s important to note, though, that the nocaps benchmarks capture only a tiny sliver of the complexity of image captioning as a general task. Although Microsoft claims in a press release that its new algorithm “describes images as well as people do,” this is only true insomuch as it applies to a very small subset of images contained within nocaps.

“Surpassing human performance on nocaps is not an indicator that image captioning is a solved problem”
As Harsh Agrawal, one of the creators of the benchmark, told *The Verge* over email: “Surpassing human performance on nocaps is not an indicator that image captioning is a solved

problem.” Argawal noted that the metrics used to evaluate performance on nocaps “only roughly correlate with human preferences” and that the benchmark itself “only covers a small percentage of all the possible visual concepts.”

“As with most benchmarks, [the] nocaps benchmark is only a rough indicator of the models’ performance on the task,” said Argawal. “Surpassing human performance on nocaps by no means indicates that AI systems surpass humans on image comprehension.”

This problem – assuming that performance on a specific benchmark can be extrapolated as performance on the underlying task more generally – is a common one when it comes to exaggerating the ability of AI. Indeed, Microsoft has been criticized by researchers in the past for making similar claims about its algorithms’ ability to comprehend the written word.

Nevertheless, image captioning is a task that has seen huge improvements in recent years thanks to artificial intelligence, and Microsoft’s algorithms are certainly state-of-the-art. In addition to being integrated into Word, Outlook, and PowerPoint, the image-captioning AI will also be available as a standalone model via Microsoft’s cloud and AI platform Azure.

https://www.theverge.com/2020/10/14/21514405/image-captioning-seeing-ai-microsoft-algorithm-word-powerpoint-outlook