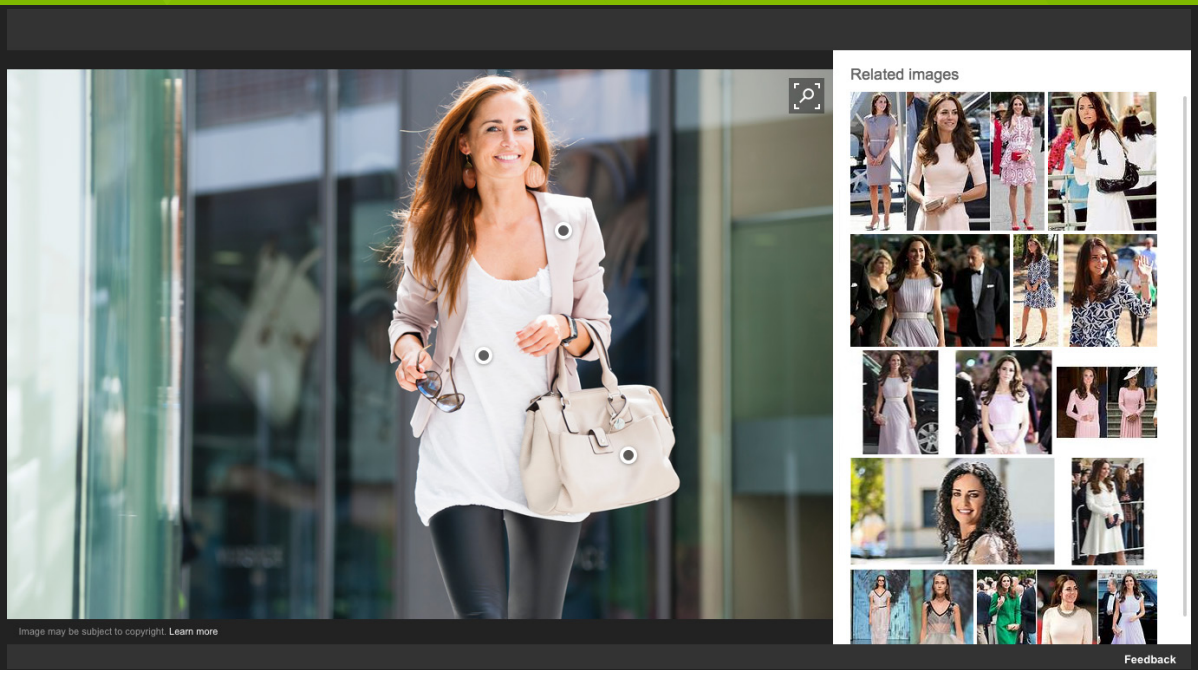
SUCCESS STORY | BING, SEARCH FROM MICROSOFT

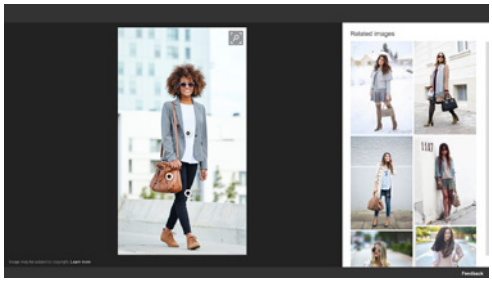
GPU-POWERED SMARTER, FASTER VISUAL SEARCH



NVIDIA

Bing deploys NVIDIA technology to speed up object detection and deliver pertinent results in real time.

VISUAL SEARCH: A WORTHY CAUSE



Bing: Group Object Detection

In Visual search is seen as the next great search frontier, and Microsoft's Bing has tapped the power of NVIDIA GPUs to make it a reality. At the same time, they've leveraged the NVIDIA® CUDA® profiling toolchain and cuDNN to make the system more cost-effective.

But visual search at scale is no easy matter: Instantly delivering pertinent results when users mouse over objects within photos requires massive computations by algorithms trained to classify, detect, and match the images within images.

It's also well worth the effort.

"A picture is worth more than a thousand words," said Yan Wang, senior engineer with Bing. "When you have a picture, you're that much closer to what you're looking for."

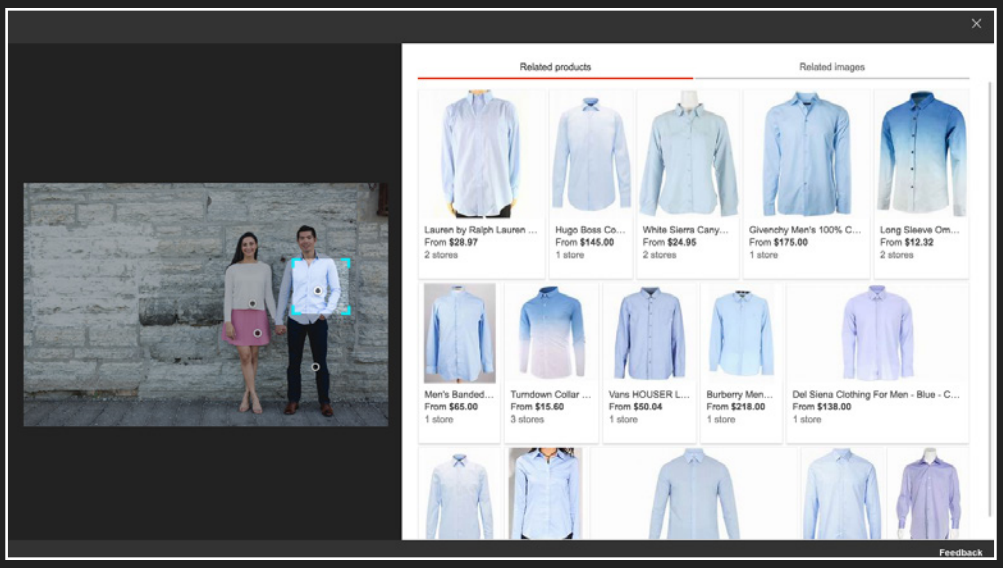
Before now, however, it was a lengthy wait for what you were looking for. In 2015, Bing introduced image-search capabilities that enabled users to draw boxes around sub-images or click on boxes of sub-images already detected by the platform; they could then use those images as the basis of a new search.

Bing sought a solution that was fast enough to keep up with user expectations. They transitioned their object-detection platform from CPUs to Azure NV-series virtual machines running NVIDIA Tesla® M60 GPU accelerators. In doing so, Bing slashed their object-detection latency from 2.5 seconds on the CPU to 200 milliseconds. Further optimizations with NVIDIA cuDNN lowered that to 40 milliseconds, well under the threshold for an excellent user experience on most applications.

CUSTOMER PROFILE

Bing

Organization	Industry	Location	Organization Size	Website
Bing team, Microsoft	Technology	Seattle, Washington	124,000 employees	www.microsoft.com



Bing:Dress Shirt Search

PRODUCTS

NVIDIA Tesla M60 GPUs on Azure

NVIDIA Tesla K40s

NVIDIA cuDNN

CREATING A SUPERIOR EXPERIENCE

The payoff for the move to NVIDIA GPUs was instantaneous, with inference latency reduced immediately by 10X. But Bing's engineers weren't about to stop there.

They incorporated the NVIDIA cuDNN GPU-accelerated deep learning library into their code and updated their driver mode from the Windows Display Driver Model to the Tesla Compute Cluster, dropping latency to 40 milliseconds for a total performance improvement of 60X. To detect more object categories on an image, they moved from a fast R-CNN two-stage process to a one-stage "single shot detection" process. This sped up the feature 10X and enables detection of over 80 image categories.

The Bing team also leverages a filter triggering model and Microsoft's ObjectStore key-value store to limit the amount of data they need to process and cache results for future use. This helps them save over 90 percent of their costs, making it more economically feasible to service the volume of requests they receive daily.

The user experience offered by Bing Visual Search reflects these extra efforts. From the Bing search page, a user can select "image search," type in text or upload a picture, and then either select hotspots automatically detected on the picture or draw a box on the parts of interest to trigger near-instantaneous search results. Drawing the box over, say, a purse, generates numerous purse-buying opportunities, complete with pricing.

On the development and deployment side, switching to NVIDIA GPUs has empowered the Bing team to be more agile and increase their rate of learning and innovation. With CPUs, it would take months to run updated models on the entire dataset of billions of images after every significant change. With GPUs, this process is now instantaneous, making it practical to update the models frequently and provide more features for Bing's users.

"[A speedier model-update process] drastically cuts down our innovation and production cycle from over a month on each update to almost instantaneous."

Yan Wang,
Senior Engineer, Bing

GROUNDBREAKING MOMENT FOR VISUAL SEARCH

Real-time object detection and visual search are now possible, making Bing Visual Search a groundbreaking moment. With the ability to process deeper and more complex models, Bing Visual Search can support more categories of detectable objects. And speedier updates for back-end models frees Bing to up the ante on the development front.

"It drastically cuts down our innovation and production cycle from over a month on each update to almost instantaneous," said Wang.

The potential impact of Bing Visual Search could be transformational for online retailers, who will be able to take their products directly to consumers' searches rather than waiting for the searches to find them. But it doesn't take a lot to imagine what Bing Visual Search can do for other industries as well, such as travel and education.

For instance, a user swept away by a picture of a beach could immediately match that photo to a real location and book a vacation. Or an art student could take a photo of a painting at a museum and instantly identify other paintings that might have influenced or been influenced by the painting in question. The possibilities are endless.

www.nvidia.com

© 2018 NVIDIA Corporation. All rights reserved.

NVIDIA