In 2011, **Google** with chipmaker **NVIDIA** found out that a computer vision algorithm it had trained on 2000 CPUs to differentiate cats from humans could achieve the same performance when trained on only 12 GPUs.

But GPUs aren't perfect for DL as well, mostly because of two things. First, they **can't work as a standalone chip** as they are limited by the kind of operations they perform. Second, **GPUs have very low cache memory**. That means the bulk of data is stored off-chip and must be retrieved when it is time for processing. This back-and-forth data flow ends up being a bottleneck for computation, capping the speed at which GPU can run a DL algorithm.

Now, Neural Magic comes up with a different methodology. **Instead of tinkering with the hardware, they modified the software**. It redesigned deep-learning algorithms to run more efficiently on a CPU by utilizing the chips' large available memory and complex cores. While the approach loses the speed achieved through a GPU's parallelization, it reportedly **gains back about the same amount of time by eliminating the need to ferry data on and off the chip**.

The algorithms can run on CPUs "at GPU speeds", the company says—but at a fraction of the cost.