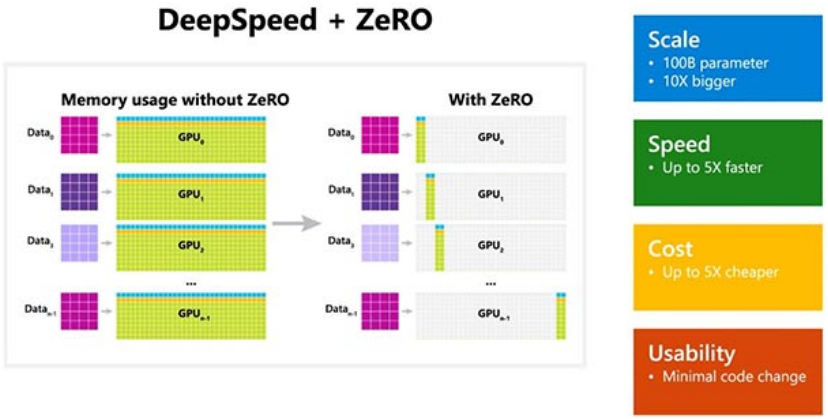
Microsoft announced DeepSpeed, a new deep learning library that can support the training of super-large scale AI models

Microsoft Research has recently sparked the rise of artificial intelligence (AI) researchers when it announced the successful development of DeepSpeed, a deep learning optimization library that can be used to train huge AI models with Scale up to 100 billion parameters.

Microsoft Research has recently sparked the rise of artificial intelligence (AI) researchers when it announced the successful development of DeepSpeed, a deep learning optimization library that can be used to train huge AI models with Scale up to 100 billion parameters.

In AI training, the bigger the natural language models you own, the greater the accuracy. However, the training of large natural language models takes a lot of time, and the costs involved are not small. DeepSpeed ??was born to overcome all of these difficulties: Improve speed, cost, training scale and usability.

In addition, Microsoft also mentioned that DeepSpeed ??also includes ZeRO (Zero Redundancy Optimizer), a parallel optimization technique that minimizes the amount of resources needed for models, while still helping to increase the amount of reference. Numbers can be trained. Using a combination of DeepSpeed ??and ZeRO, Microsoft researchers were able to successfully develop the new Turing Natural Language Generation (Turing-NLG) model - the largest language model available today with 17 billion parameters. .



Some highlights of DeepSpeed:

1. **Scale:** Large, advanced AI models such as OpenAI GPT-2, NVIDIA Megatron-LM and Google T5 are 1.5 billion, 8.3 billion and 11 billion parameters, respectively. ZeRO phase 1 in DeepSpeed ??can provide system support to run models with up to 100 billion parameters, which is 10 times larger than Google's largest model.

2. **Speed: The** throughput recorded will increase differently depending on the hardware configuration. On NVIDIA GPU clusters with low bandwidth connectivity (without NVIDIA NVLink or Infiniband), DeepSpeed ??achieves a throughput improvement of 3.75 times compared to using Megatron-LM only for standard GPT-2 models with 1.5 billion parameters. On the NVIDIA DGX-2 cluster with high bandwidth connection, for models with 20 to 80 billion parameters, DeepSpeed ??is 3 to 5 times faster.
3. **Cost:** From improvements in speed, training costs are also significantly optimized. For example, to train a model with 20 billion parameters, DeepSpeed ??requires 3 times less resources than usual.
4. **Availability:** Only a few minor code-related changes are needed to allow existing models to migrate to DeepSpeed ??and ZeRO. DeepSpeed ??does not require code redesign or refactoring the model.

Microsoft is open source for both DeepSpeed ??and ZeRO on GitHub, please refer.