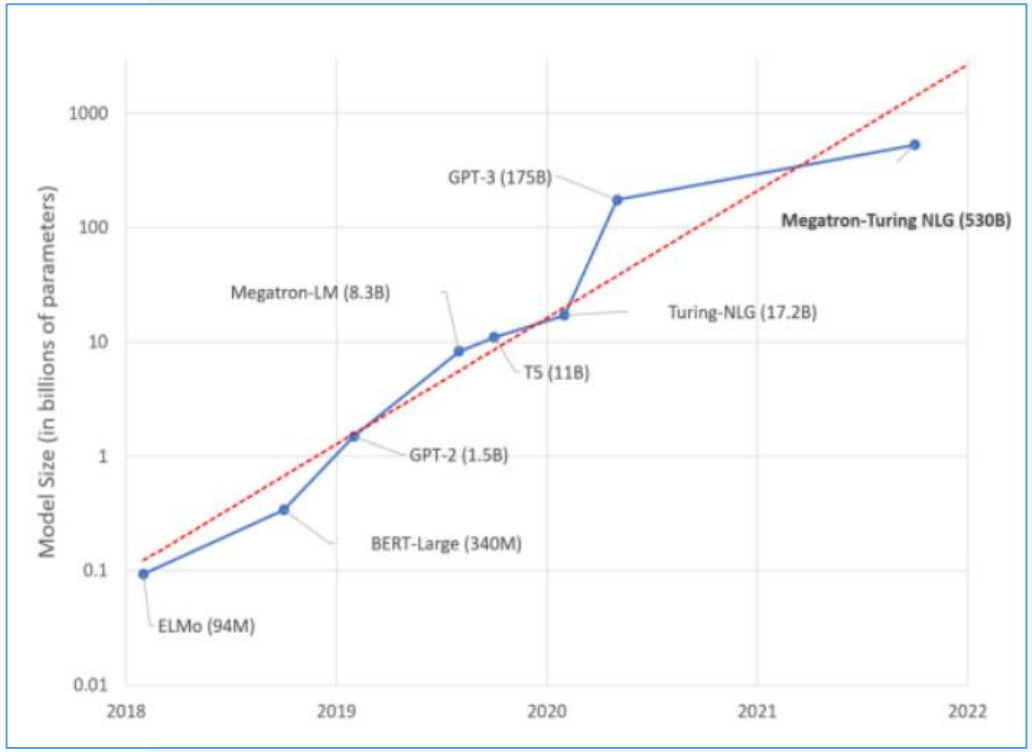
Ch1. Model memory requirement

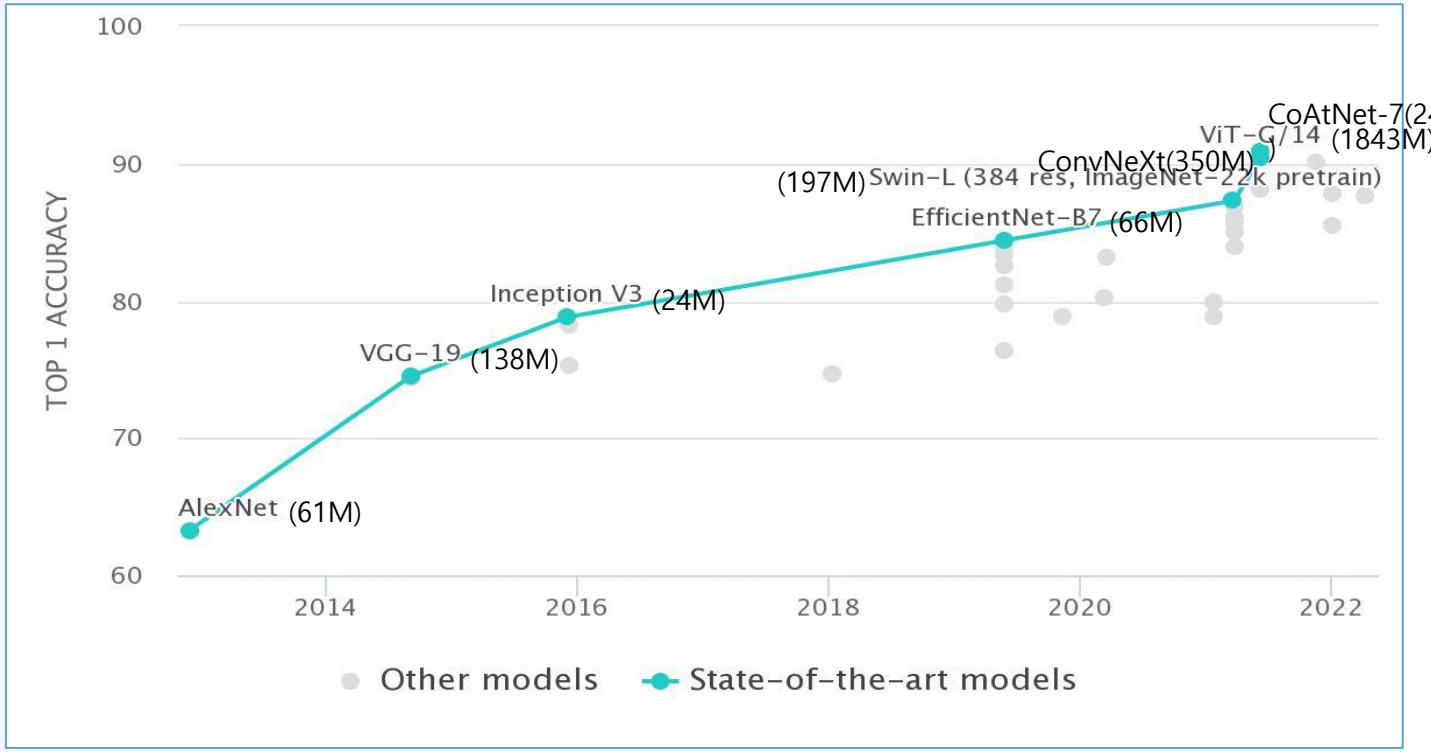
ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning

AI 모델 크기 변화

- 언어 모델

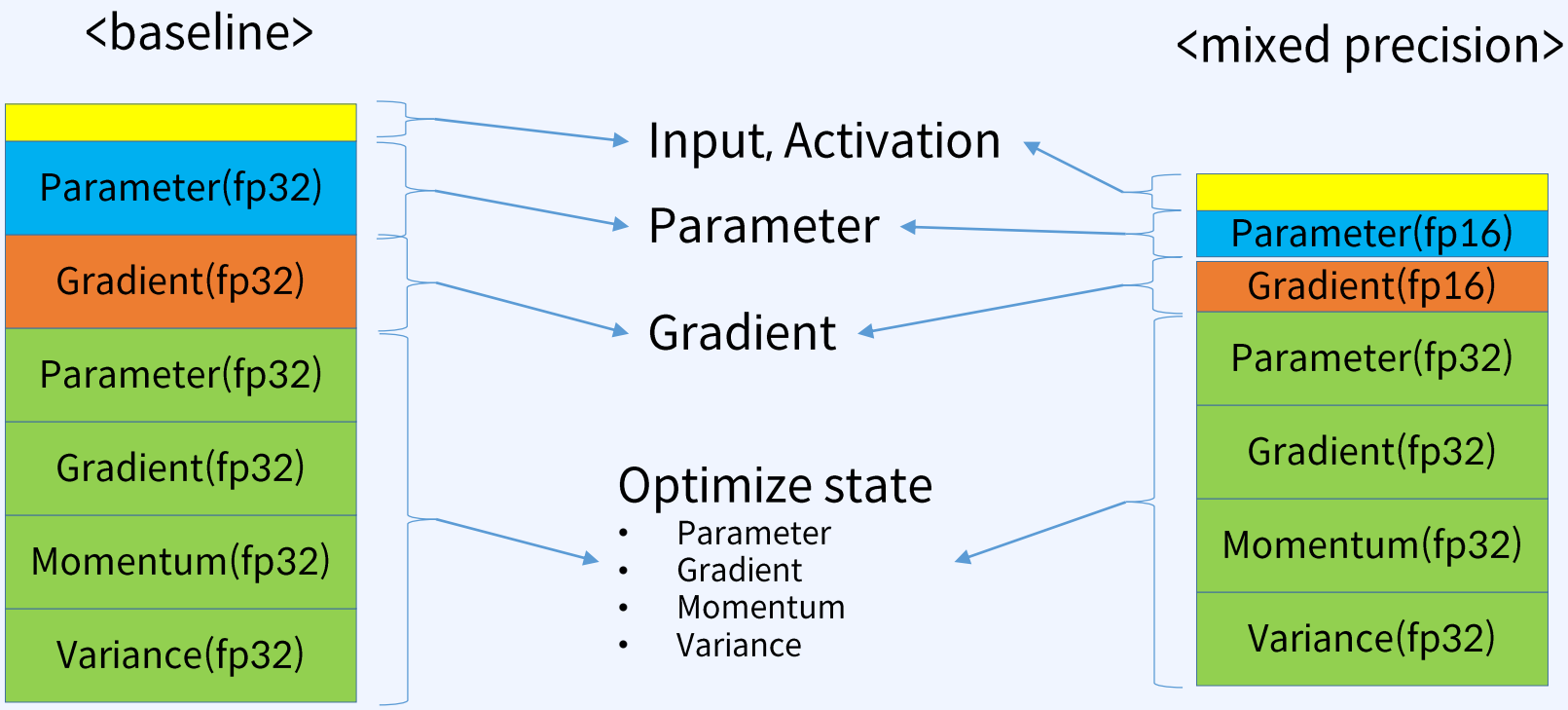


- Vision 모델



AI 모델 GPU 메모리 최소 요구량

- AI 모델 GPU 메모리 사용 구성 요소(with adam optm.)



Ex. AlexNet 최소 메모리 요구량

Layer name	Activation shape	Activation size	Kernels	Biases	Parameter
Input image	227x227x3	154,587	0	0	0
Conv-1	55x55x96 (k=11, s=4, p=0)	290,400	34,848	96	34,944
MaxPool-1	27x27x96 (k=3, s=2)	69,984	0	-	0
Conv-2	27x27x256 (k=5, s=1, p=2)	186,624	614,400	256	614,656
MaxPool-2	13x13x256 (k=3, s=2)	43,264	0	-	0
Conv-3	13x13x384 (k=3, s=1, p=1)	64,896	884,736	384	885,120
Conv-4	13x13x384 (k=3, s=1, p=1)	64,896	1,327,104	384	1,327,488
Conv-5	13x13x256 (k=3, s=1, p=1)	43,264	884,736	256	884,992
MaxPool-3	6x6x256 (k=3, s=2)	9,216	0	-	0
FC-1	4096x1	4,096	37,748,736	4,096	37,752,832
FC-2	4096x1	4,096	16,777,216	4,096	16,781,312
FC-3	1000x1	1,000	4,096,000	1,000	4,097,000
Total		936,323			62,378,344

<mixed precision 기준>

<Parameter> : 62,378,344 x 2byte
= 124,756,688byte

<gradient> : 62,378,344 x 2byte
= 124,756,688byte

<Optimize state>

Parameter : 62,378,344 x 4byte
= 249,513,376byte
Gradient : 62,378,344 x 4byte
= 249,513,376byte
Momentum : 62,378,344 x 4byte
= 249,513,376byte
Variance : 62,378,344 x 4byte
= 249,513,376byte

<activation (1batch)>: 936,323 x 4byte
= 1,209,312byte

<total>: 1,248,776,192byte = 1.25GB
Memory requirement = 1.25GB + α

AI 모델의 최소 메모리 요구량 계산식

- Parameter의 메모리 할당 크기 + Optimize state 메모리 할당 크기 + Activation 메모리 할당 크기 + α
  - Parameter의 크기 = gradient의 크기 = momentum의 크기 = variance의 크기
  - Optimize state의 크기 = Parameter의 크기 x 4
- Parameter의 크기 x 2byte + gradient의 크기 x 2byte + optimize state의 크기 x 4byte + activation의 크기 x 4byte + α

➔ parameter의 크기 x 20byte + activation의 크기 x 4 byte + α

Ex. AI 모델 최소 메모리 요구량

- AlexNet

Overall, AlexNet has about 660K units, **61M parameters**, and over 600M connections. Notice: the convolutional layers comprise most of the units and connections, but the fully connected layers are responsible for most of the weights. 2018. 2. 12.

https://www.cs.toronto.edu › tutorials › tut6_slides PDF

A Closer Look at AlexNet

61M x 20 + α = 1.22GB + α

- CoAtNet-7

https://medium.com › coatnets-6608442da4d2

CoAtNets. Brief notes on a class of… | by m0nads - Medium

(2021). CoAtNet models (pronounced “coat” net) for computer vision emerge as a… ... by a CoAtNet model (**CoAtNet-7**, Top Accuracy: 90.88%, 2440M **parameters**, ...

2240M x 20 + α = 44.8GB + α

- GPT3

GPT-3's full version has a capacity of **175 billion machine learning parameters**. GPT-3, which was introduced in May 2020, and was in beta testing as of July 2020, is part of a trend in natural language processing (NLP) systems of pre-trained language representations.

https://en.wikipedia.org › wiki › GPT-3

GPT-3 - Wikipedia

175B x 20 + α = 3.5TB + α

Summary

- AI 모델의 최소 메모리 요구량 계산 방법
- AI 모델의 최소 메모리 요구량 계산식
- 대략적인 AI 모델 최소 메모리 요구량 계산