Case Study

Compression of Deep Learning-based MRI Reconstruction Model

- Project to compress a clients' MRI reconstruction model previously running on AWS high-tier instance(V100) to run on local PC w/Intel CPU

- Achieved 1,840x memory reduction, 160x inference acceleration, and 9x instance cost reduction without compromising accuracy

- Expected to facilitate purchase conversion by not transmitting private medical information to the outside

Accuracy(SSIM)	Model Size	Inference Time	Instance Costs
99.3% maintained	1,840x Reduced	160x Accelerated	9x Reduced
Client: 100.0%	Client: 643MB	Client: 9,430s	Client: $2,200/mth
ENERZAi: 99.3%	ENERZAi: 0.35MB	ENERZAi: 59s	ENERZAi: $247/mth

ENERZAi