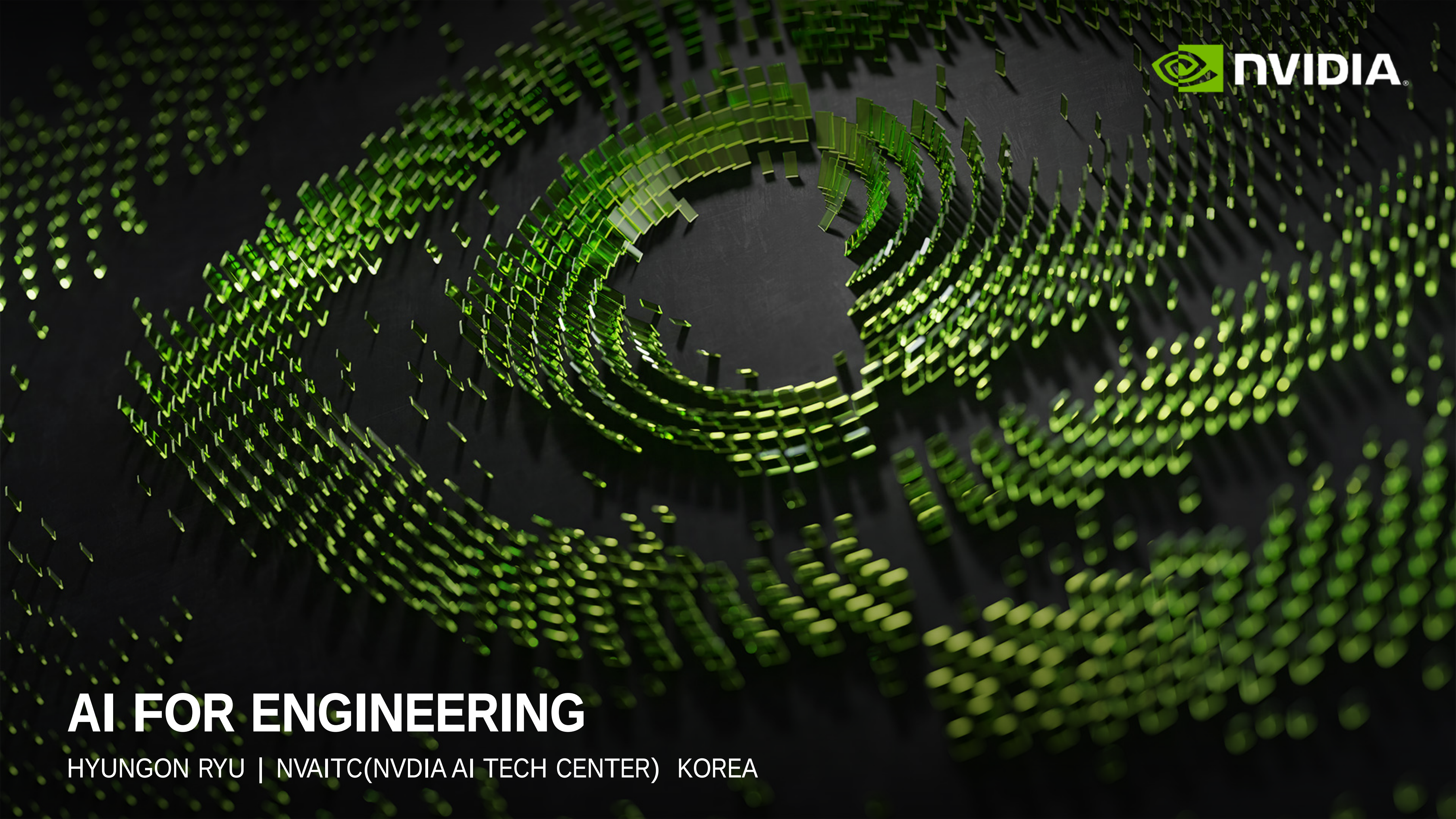


AI FOR ENGINEERING

HYUNGON RYU | NVAITC(NVDIA AI TECH CENTER) KOREA

KEYWORD FOR AI RESEARCH

DL Model	Demo only Paper only With sample With code With dataset With Checkpoint

application	Paperwithcode, github NEMO, RIVA, MONAI, Hugginface timm, einops,

Pair of (Input,Output)	(Image, Optical Flow) (text, image), (image, cls) (audio, text)
Data Loader	Dali, stream Augment, patch
preprocessing	Tokenizer, normalizer
Dataset	Image, WSI, X-ray/MRI, Lanauge(audio,text), video, 3D, stereo, Chemical, Protein, CFD

Technique	AMP, Data Parallel, Model Parallel, Quantization, hash, parameter sharing, checkpointing, ZeRO,
Train recipe	Learning rate schedule(Cosine, warm up), early stopping Optimizer(Adam), accumulation
Task	Multistage, multi modal, end2end, Pretrain/finetune, distill, quantization Regression, CLS, AE, GAN, Prompt, LM, AR, MLM, denoising, jigsaw, SuperRes
Objective	MSE, Cross Entropy, Dice, triplet, contrative
Model	Model : ResNet, EfficientNet, Unet, Hifi-GAN, transformer, BERT, BART, GPT-2, GPT-3 , NERF Module : Pool, Conv, LSTM, GRU, FCN, MLA, GNN, softmax, GeLU, ReLU, Residual, Skip Variation : Prenorm, postnorm,

DLFW	Pytorch, TF, Keras, DGL, PyG, JAX, pennylane, TorchANI WanDB, ignite, torchlightening,
DevOps	OS(Ubuntu,WSL2), PIP, Conda, Singularity, Docker, slurm/PBS/LSF, jupyter, NFS, Baremetal/Virtual, Ansible
Resource	GPU, TensorCore, multiGPU, MultiNode, IB,

EXAMPLES

Task : lung CT segmentation
Data pair : In:CT raw, Out : Segmentation
Dataset : COVID19-CT-Dataset
Augmentation : none
DataLoader : nefti reader(MONAI)

Task : 3D segmentation
Model : Unet(MONAI)
Optimizer : Adam
Recipe : train with warm up

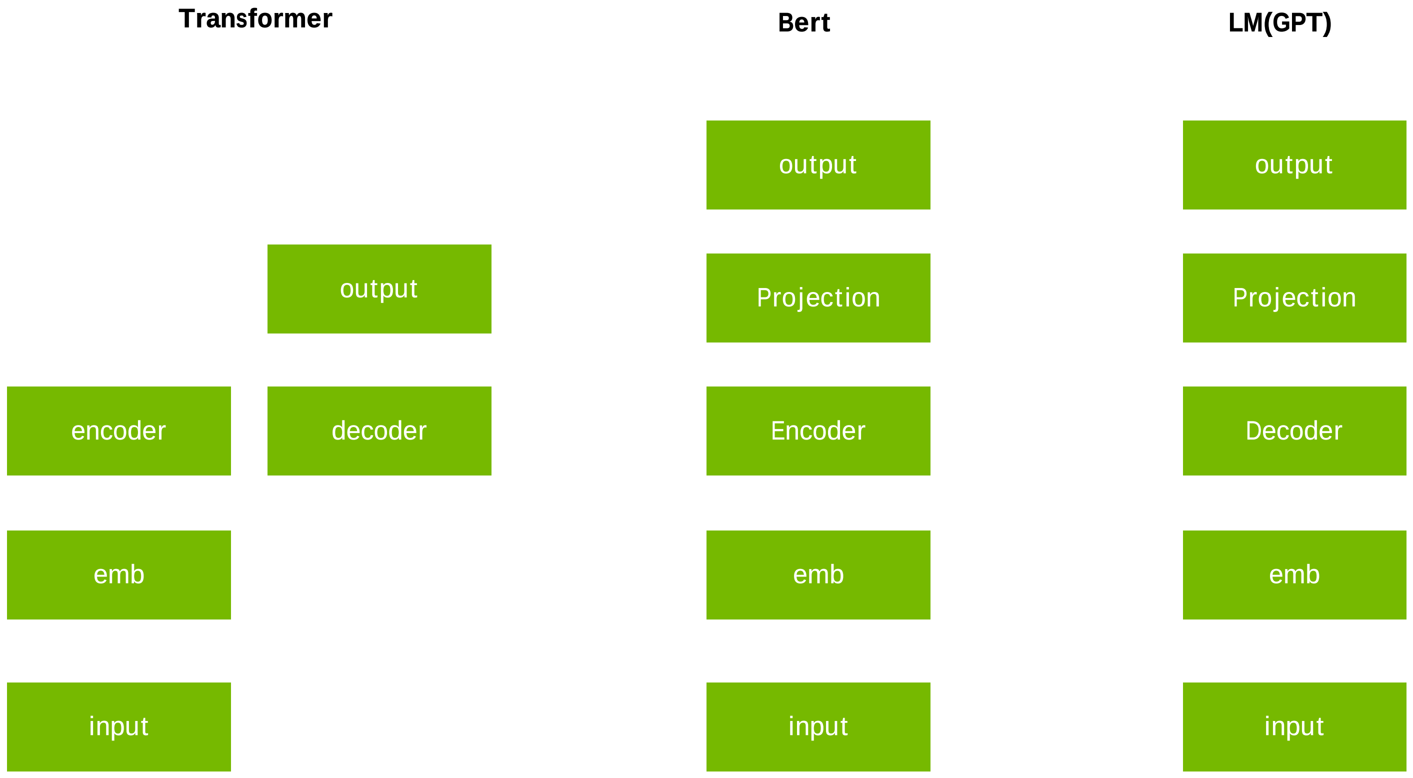
System : 1 node (2EA RTX8000 40GB)
OS : Ubuntu
DLFW : pytorch on NGC docker

Task : ASR
Data pair : In:audio, Out : text
Dataset : LibriLight
Augmentation : SpecAug
DataLoader : Nemo

Task : ASR
Model : ContextNet(Conv, SELayer)(NEMO)
Recipe : train with warm up

System : 2 node DGX-1 (8EA A100 80GB)
OS : Ubuntu
DLFW : pytorch on singularity, slurm

TRANSFORMERS



LLM(LARGE LANGUAGE MODEL)

420 node DGX-1(8EA A100)

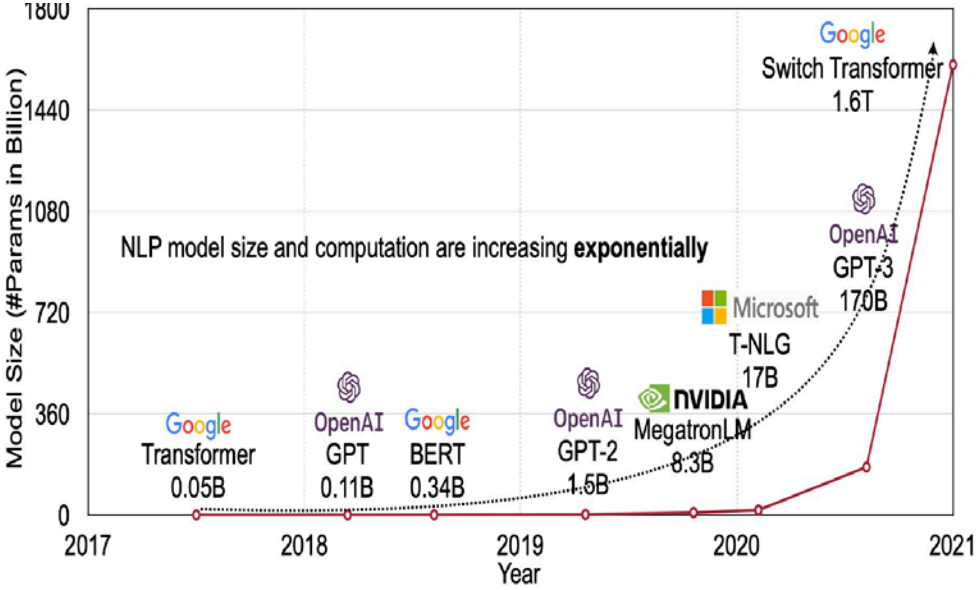


Image from https://hanlab.mit.edu/projects/efficientnlp_old/

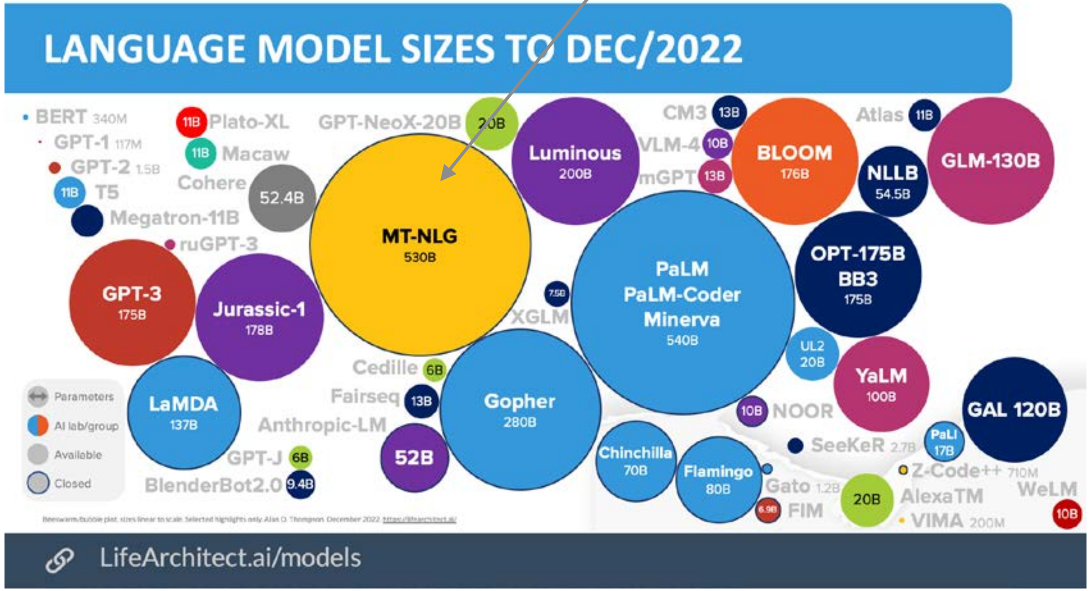
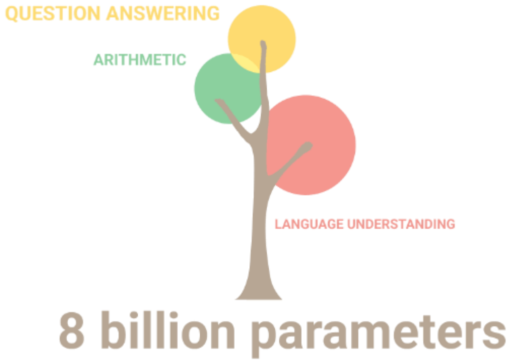
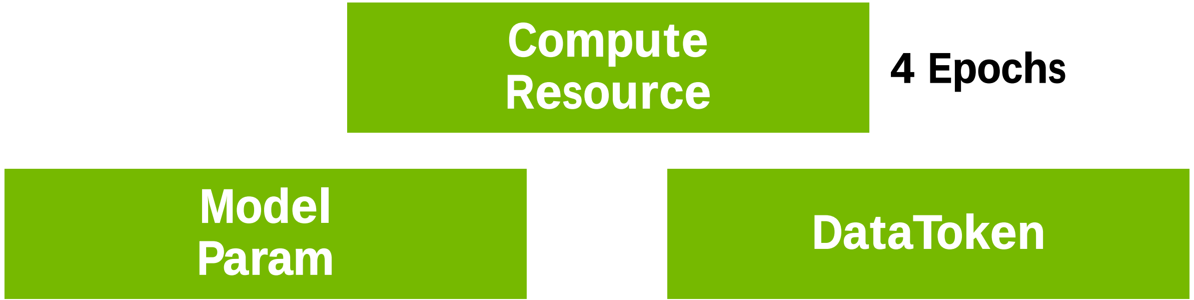


Image from https://lifearchitect.ai/models/

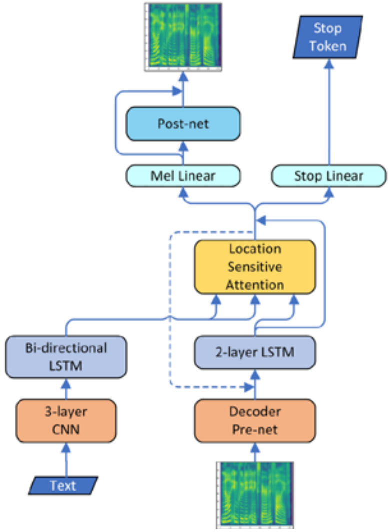
MODEL CAPABILITIES WITH SCALES



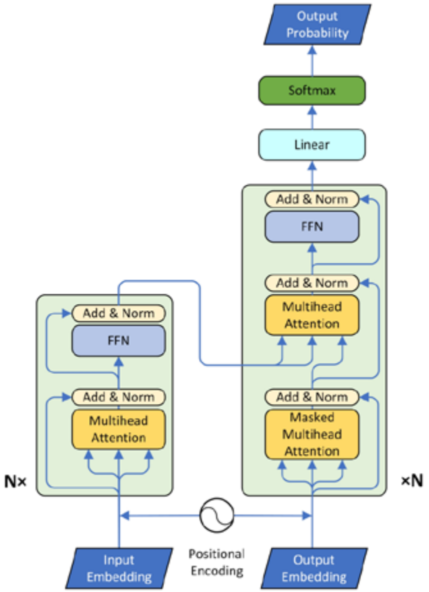


Transformer IN Various Domain

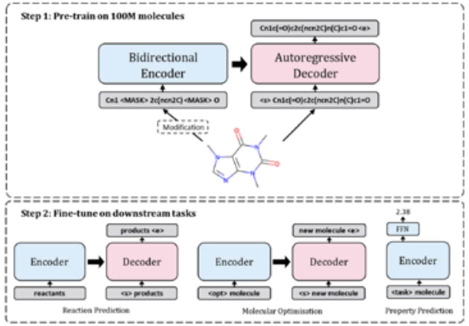
Neural Speech Synthesis with Transformer Network(2019)
https://arxiv.org/pdf/1809.08895.pdf



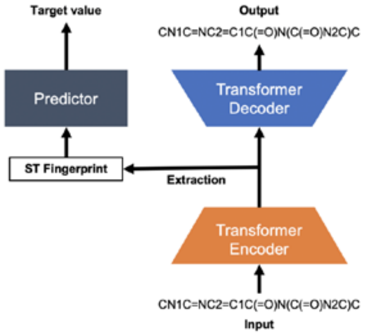
TTS(LSTM)



TTS(transformer)

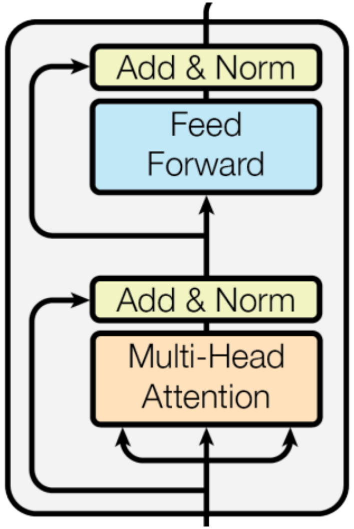


MolBART



Chemical(transformer)

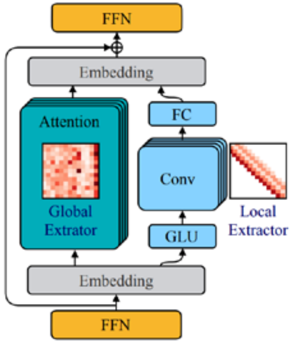
Various Transformer Layers



Replace
FF, MHA
Change order

Sparse Attention
Axial Attention
Graph Attention
Quaternion Transformer

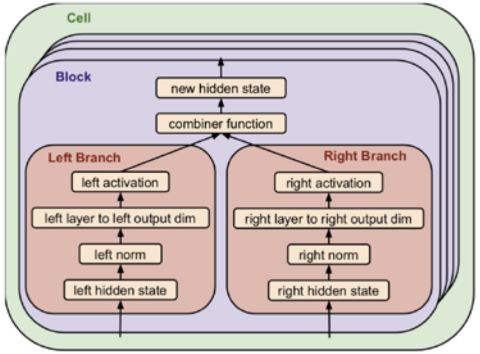
Lite Transformer



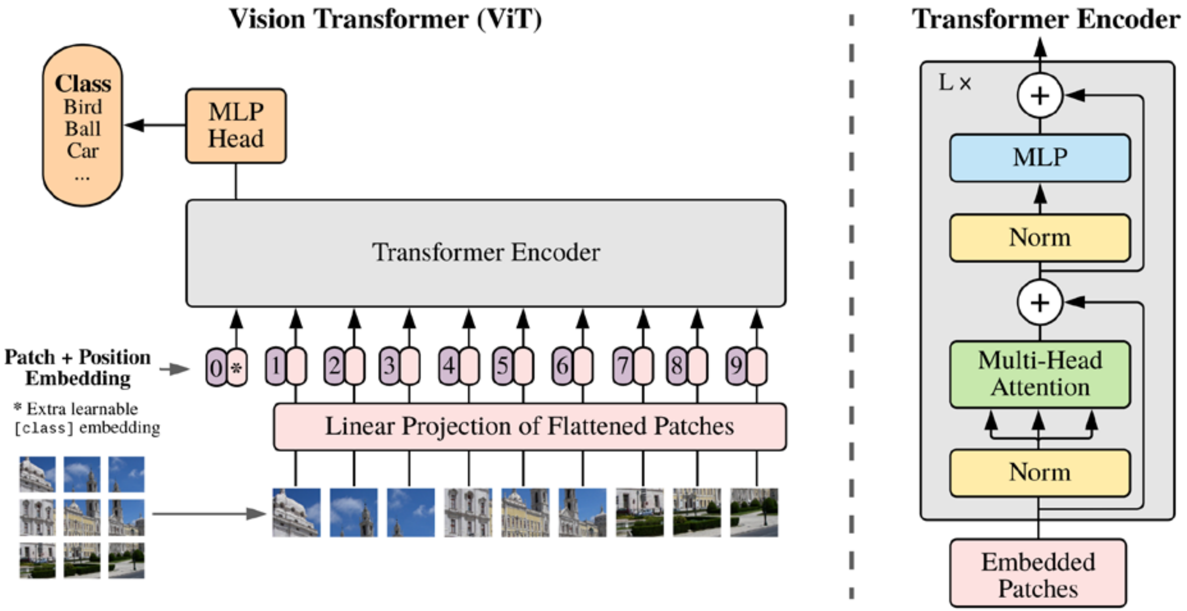
(a) Lite Transformer block

Longformer
Linformer
Reformer
Performer

Evolved Transformer(NAS)



Vision Transformer(ViT) ICLR2021



TRANSFORMERS

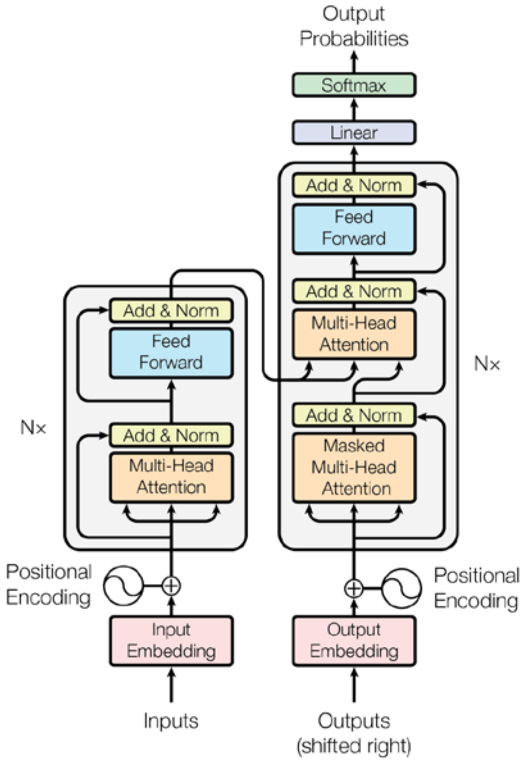
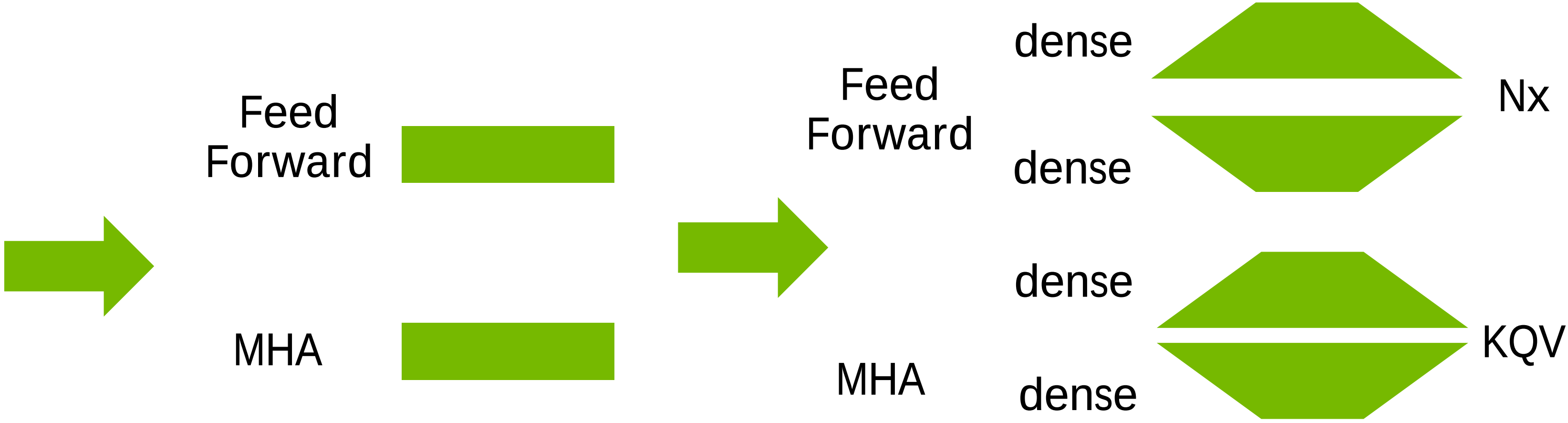


Figure 1: The Transformer - model architecture.

Attention Is All You Need



BERT BASE

BERT BASE

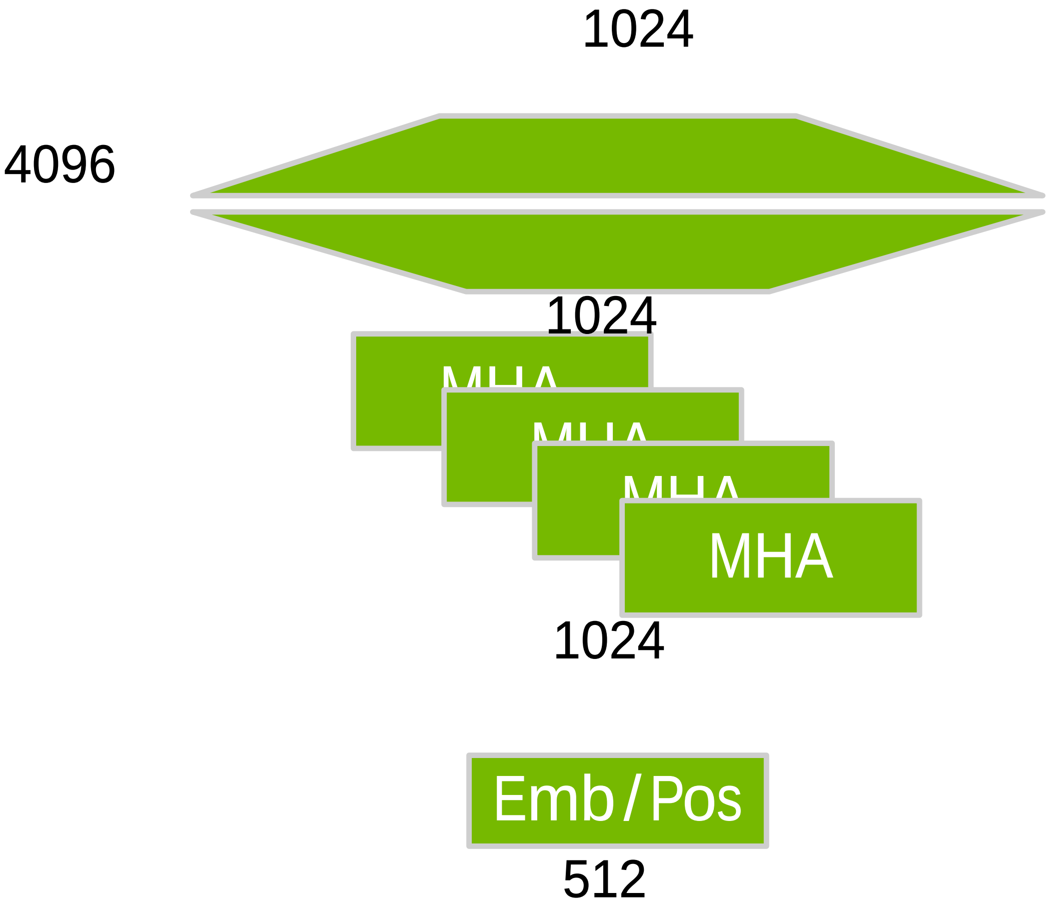
Pos : 512
numVOCA= 2^15

NumLayers: 12
dimModel : 768
dimHead :64
NumHeads : 12
Act : gelu
Dropout : 0.1
FF scale : 4
110M Param

BERT LARGE

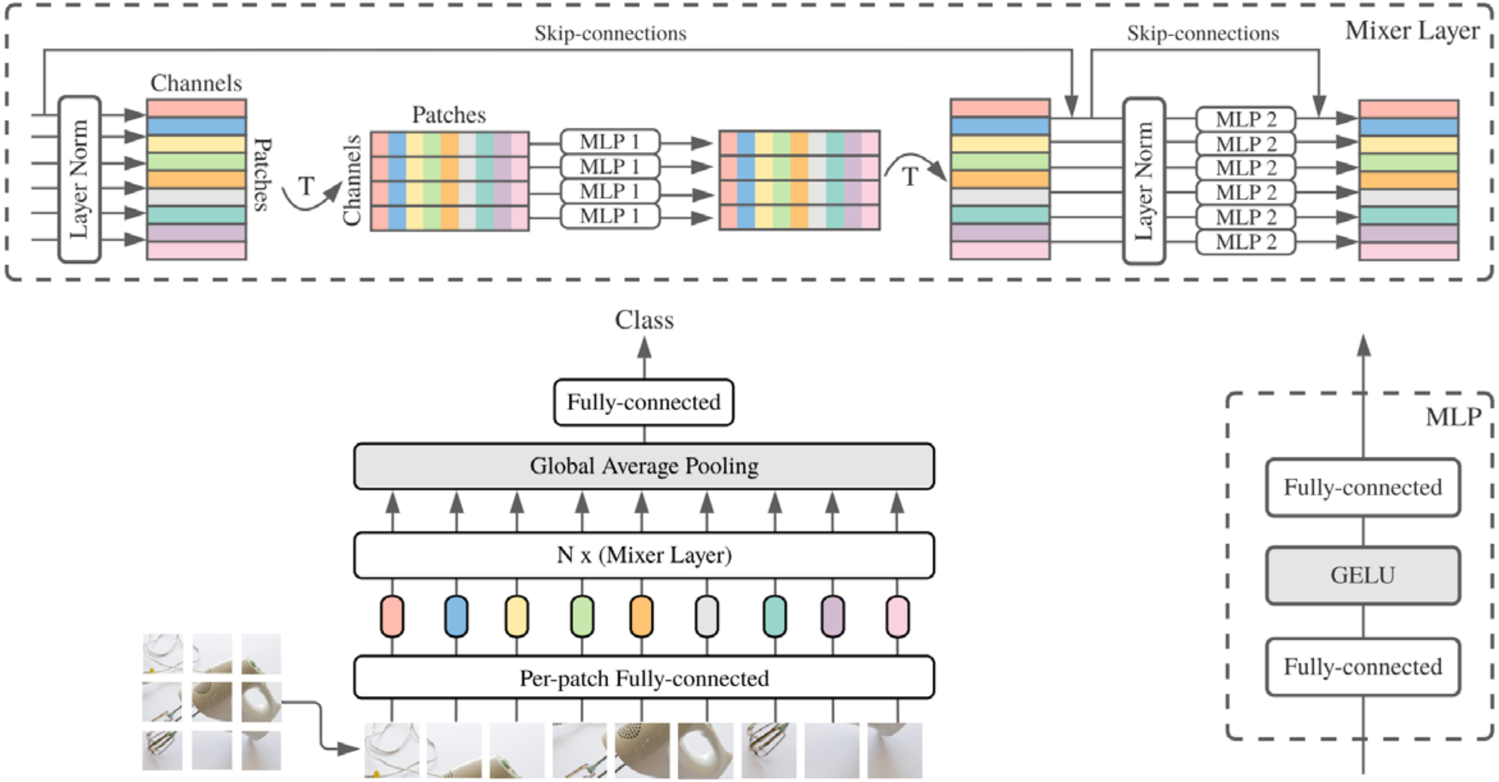
Pos : 512
numVOCA= 2^15

NumLayers: 24
dimModel 1024
dimHead :64
NumHeads : 16
Act : gelu
Dropout : 0.1
FF scale : 4
340M Param



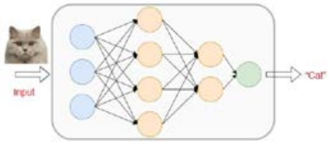
MLP-Mixer

MLP-Mixer: An all-MLP Architecture for Vision
https://arxiv.org/pdf/2105.01601.pdf



REVISIT MLP

MLP

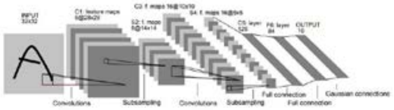


flatten
raw
1d input

MLP
sigmoid

softmax

CNN



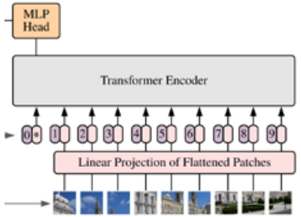
2d input

features
2d conv
relu

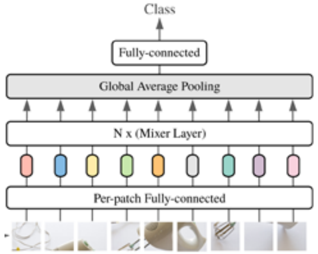
Onehot
encoding

Residual
SELayer

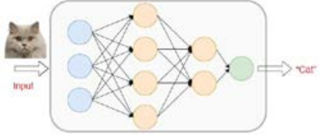
Transformer



MLP-Mixer



MLP(new)



encoded
1d input

repeat n
residual

relu/gelu

layernorm
dropout

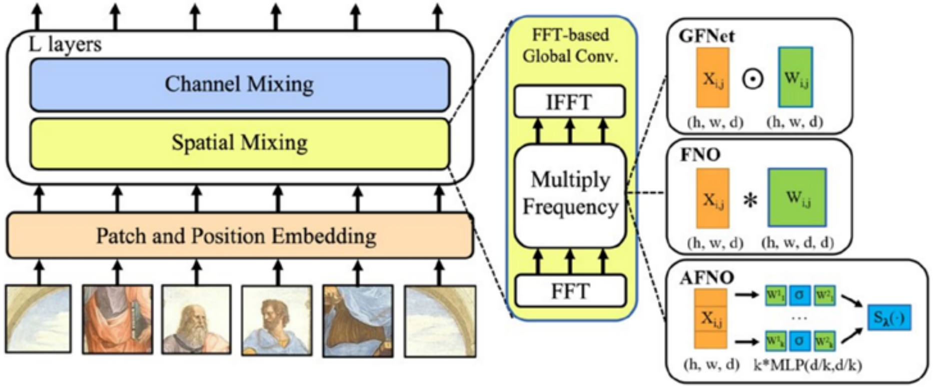
Softmax
Onehot
encoding

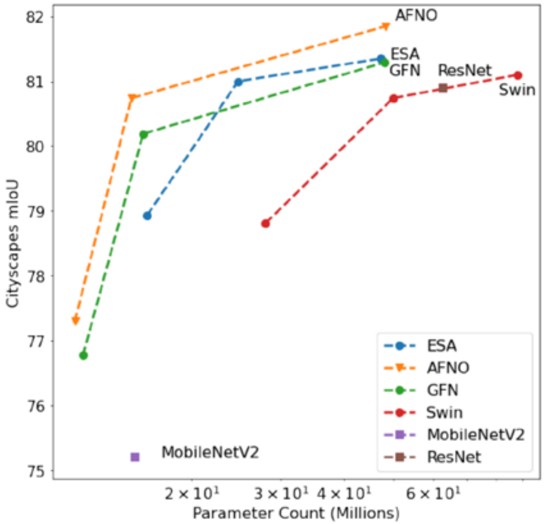
NVIDIA

AFNO (ICLR 2022)

Adaptive Fourier Neural Operators

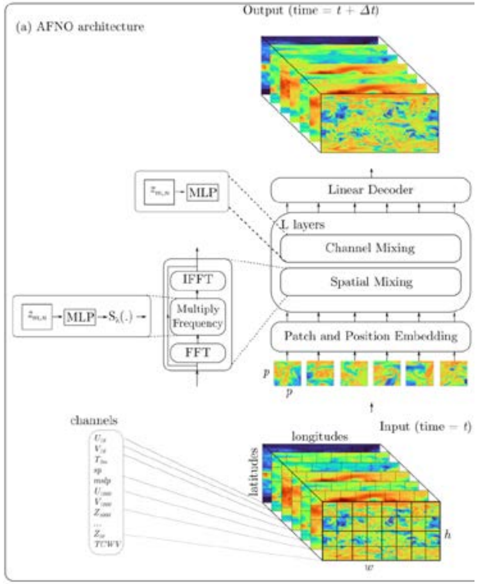
MLP-Mixer with FFT





FourCastNet

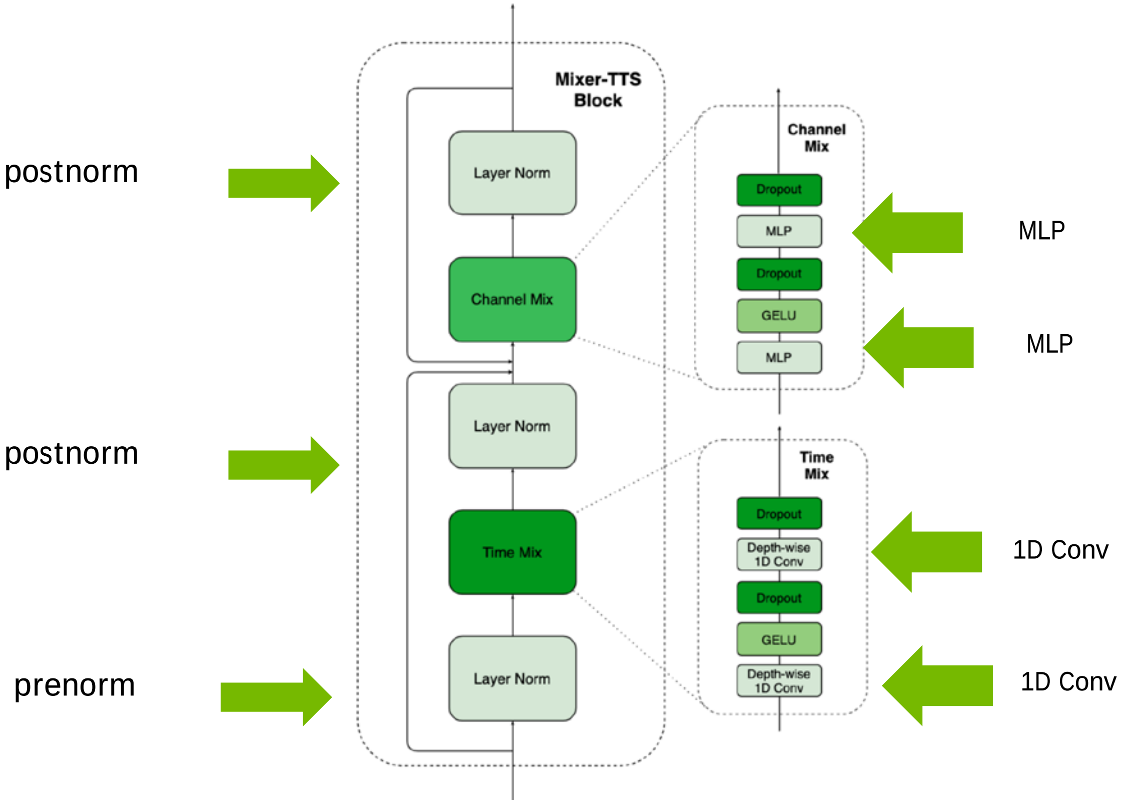
https://arxiv.org/pdf/2202.11214.pdf



Use AFNO for weather modeling(NWP)
FourCastNet generates a week-long forecast in less than 2 seconds
FourCastNet is about 45,000 times faster than traditional NWP models on a node-hour basis

MLP-MIXER VARIATION

Mixer-TTS



postnorm

postnorm

prenorm

MLP

MLP

1D Conv

1D Conv