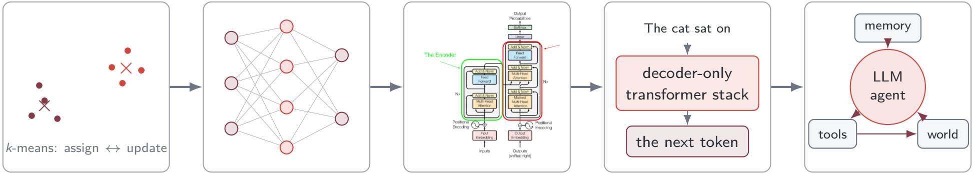
AI: A Brief Overview and Where We're Headed

Deep Learning Track — ATPESC 2026

Huihuo Zheng, Computer Scientist
Argonne Leadership Computing Facility (ALCF), Argonne National Laboratory
August 3, 2026
huihuo.zheng@anl.gov

Five overlapping waves shaping today's AI



Classical ML
1950s–

Deep learning
2012 inflection

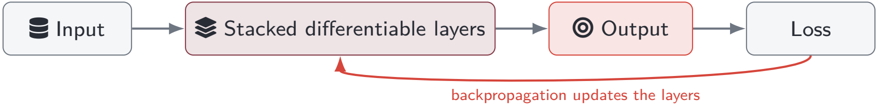
Transformer
2017

LLM
2018–

Agentic AI
2023 wave

These waves overlap rather than replace one another: each changes the learned representation, the architecture, or how a model interacts with the world.

Different structures call for different neural architectures



MLP / dense
fixed-size vectors

Dense nonlinear layers combine all input features.

CNN
spatial grids

Shared local filters scan images and spatial fields.

RNN / LSTM
ordered sequences

A recurrent state carries information through a sequence.

GNN
nodes and edges

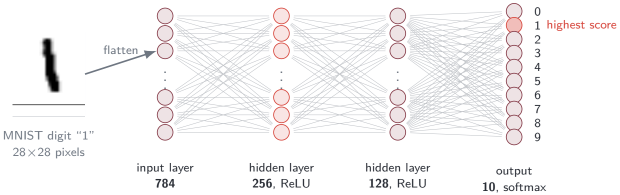
Message passing combines information along graph edges.

Transformer
tokens or elements

Attention directly connects relevant input positions.

All five learn representations with backpropagation. The architecture supplies an **inductive bias**: an assumption about which interactions matter. Next: CNNs, then transformers.

A concrete ANN: neurons form fully connected layers



One dense layer

$$\mathbf{h} = \mathrm{ReLU}(W\mathbf{x} + \mathbf{b})$$

For the first hidden layer, $W_1 \in \mathbb{R}^{256\times 784}$: each of 256 neurons has a separate weight for every input value.

Trainable parameters

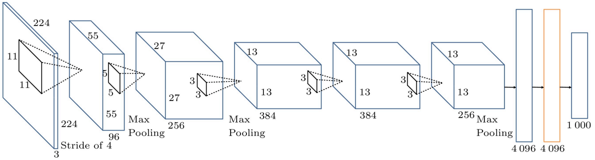
Layer	Output	Parameters
Dense + ReLU	256	200,960
Dense + ReLU	128	32,896
Dense + softmax	10	1,290
Total		**235,146**

Flattening retains every pixel value, but removes explicit 2D structure; the MLP must learn spatial relationships from data.

2012: AlexNet + ImageNet changed computer vision

🏆 AlexNet reached **15.3% top-5 test error** versus **26.2%** for the runner-up, validating deep CNNs trained on ImageNet with GPUs.

AlexNet: 5 conv + 3 fully connected layers



ImageNet at scale

1.2 million training images

1,000 object classes

Source: Krizhevsky, Sutskever, and Hinton, NeurIPS 2012.

2017: ImageNet top-5 error fell below 3%

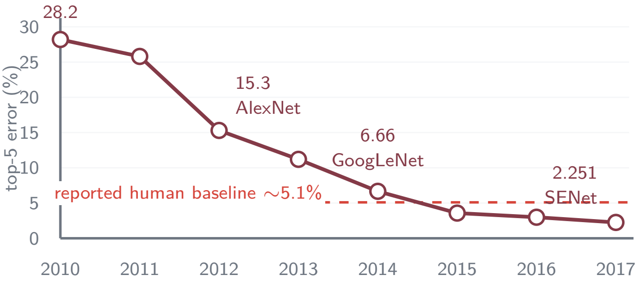
By the final ILSVRC in 2017, **29 of 38 teams exceeded 95% top-5 accuracy**. The winning SENet ensemble reached **2.251% top-5 error**, down from 28.2% in 2010.

Milestones

Year	Model family	Error
2010	Classical vision	28.2%
2011	Classical vision	25.8%
2012	AlexNet	15.3%
2014	GoogLeNet	6.66%
2015	ResNet	3.57%
2017	SENet	2.251%

Architectural progression: CNN → Inception → ResNet → channel attention.

Winning ILSVRC classification entries



Winning-entry values often use ensembles and are not directly comparable to single-model scores. Sources: ILSVRC results; ResNet and SENet papers.

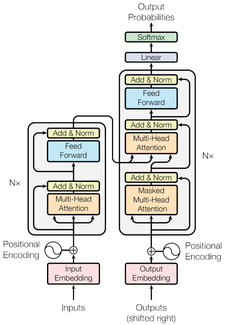
2017: The Transformer replaced recurrence with attention

The key change

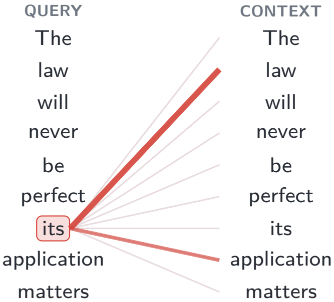
The Transformer modeled sequences with **attention** instead of recurrence.

During training, all token positions can be processed in parallel: better accelerator use and efficient scaling.

Transformer



Self-attention



"its" attends most strongly to "law"

Each token becomes a learned weighted combination of tokens in context.

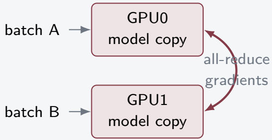
Attention Is All You Need — Vaswani et al., NeurIPS 2017. The original model was an encoder–decoder Transformer for machine translation; later LLMs adapted its attention-based design.

Scaling deep learning beyond one GPU

Three complementary strategies split a different dimension of the workload.

Split the batch

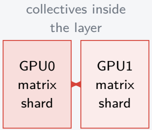
Data parallelism



More samples per step; synchronize gradients.

Split each layer

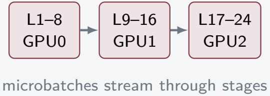
Tensor parallelism



A layer is distributed; communication is frequent.

Split the depth

Pipeline parallelism



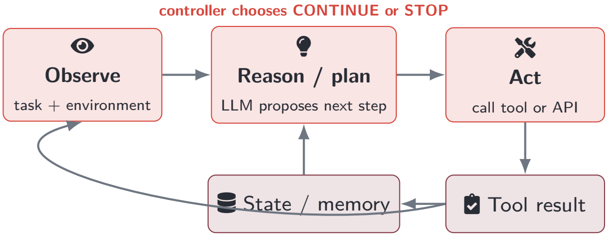
Layer groups form stages; activations move between them.

Modern training composes **data** × **tensor** × **pipeline parallelism**; FSDP/ZeRO can additionally shard model state.

From LLMs to agents: add tools, state, and a control loop

An **LLM** proposes text or decisions; an **agent runtime** executes actions, consumes feedback, maintains state, and decides whether to continue or stop.

A goal-directed feedback loop



Agent runtime

Tools affect external systems
State preserves progress
Controller loops and stops
Guardrails constrain actions

Example: diagnose a failed training run read logs → identify cause → edit configuration → relaunch → inspect results → stop.

argonne-lcf/ATPESC_MachineLearning

```
$ git clone https://github.com/argonne-lcf/ATPESC_MachineLearning.git
$ cd ATPESC_MachineLearning/01_intro_to_deep_learning/
```

ARGONNE
ATPESC2026
EXTREME-SCALE COMPUTING

ARGONNE TRAINING PROGRAM ON EXTREME-SCALE COMPUTING

Produced by Argonne National Laboratory, a U.S. Department of Energy Laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357.

Special thanks to the National Energy Research Scientific Computing Center (NERSC) and Oak Ridge Leadership Computing Facility (OLCF) for the use of their resources during the training event.