Lecture 1: Introduction

COMP 5801H/4900A: Generative AI and LLMs

2026-01-06

Sriram Subramanian
Assistant Professor & Canada Research Chair, Carleton University
Faculty Affiliate, Vector Institute for Artificial Intelligence
Faculty Affiliate, Schwartz Reisman Institute for Technology and Society

Carleton University

VECTOR INSTITUTE

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

Carleton University

VECTOR INSTITUTE

SCHWARTZ REISMAN INSTITUTE FOR TECHNOLOGY AND SOCIETY

Research Topics

- **Reinforcement Learning**: Multi-agent, Distributional, Human-in-the-Loop, Offline, Model-based, Non-Markovian, Theory
- **Generative AI: Foundation Models, Transformers, Diffusion Models, Attention Mechanisms, AI Safety**
- **Game Theory**: Normal-form games, Extensive-form games, Stochastic games, Stackleberg games, Bayesian games, Mean-field games
- **Applications**: Geomatics, Material Design, Autonomous Driving, Fighting Wildland Fires, Portfolio Optimization, Stock Trading

Collaborations

Microsoft

esri

BMO

UNIVERSITY OF WATERLOO

RBC

DENSO

BOREALIS AI

Canada NRC·CNRC

UNIVERSITY OF ALBERTA

UNIVERSITY OF TORONTO

amii

Today’s Agenda

- Course Motivation & Goals
- Core Architectures (GANs, VAEs, etc.)
- Course Logistics & Policies

Discussion: Name one way GenAI has surprised you lately

The Paradigm Shift — From Analyzing to Creating

This image shows a Cat (Confidence: 98%)

Prompt: “Generate an image of a cat with an astronaut's suit”

How & Why Now?

Data-Model-Compute Triangle (Convergence Triangle)

Massive Datasets

Google Books

COMMON CRAWL

WIKIPEDIA
The Free Encyclopedia

Specialized Hardware

Architectural Breakthroughs

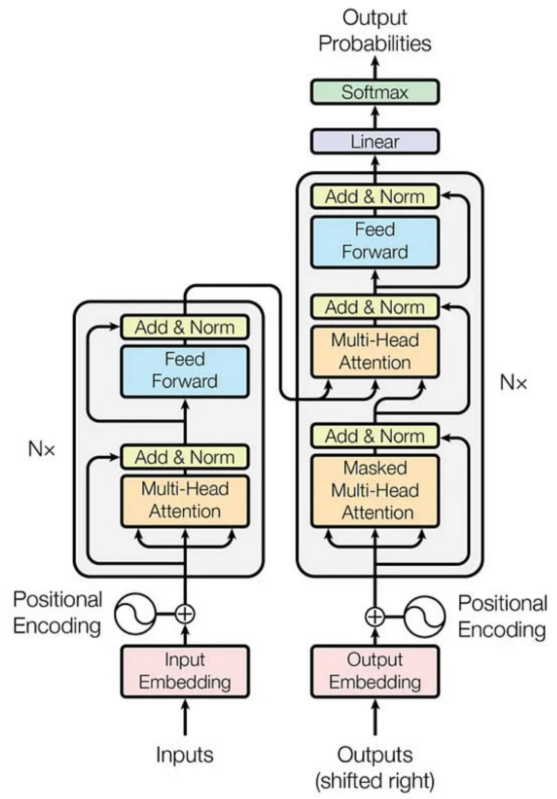


Figure 1: The Transformer - model architecture.

Credit: Vaswani et al.“Attention is all you need”

Generative AI

- Type of AI capable of creating **new content** (text, images, audio, video, code)
- Identifies patterns within massive datasets and uses the pattern to generate **original output**
- How Generative AI works?
    - Training: Uses vast amounts of **human-created content** to learn **underlying patterns (“joint probability”)**
    - Prompting: A human provides a “**prompt**", and the AI generates a response by predicting the **next logical element**

Discriminative AI

- Analyzes existing data to **categorize or distinguish** between different items (“decision boundary”)

- Does **not** create **new content**

- How Discriminative AI works?

  - Training: Uses vast datasets to identifies **the specific features** that most **effectively differentiate** the two (or more) groups (“**conditional probability**”)

  - **Ignores** the **overall “structure”** of the data and focuses only on the differences needed to make a decision

  - Prediction: When presented with **new, unseen data**, the model determines which side of the **learned boundary** the data falls on and assigns it a label or probability score

Course on Generative AI

- Discriminative AI: **Several other courses in CS**, COMP 3105, COMP 3106, COMP 4107, …
- COMP 5801H/4900A:
  - Only covers Generative AI
  - Focus on **creating new data** and **not classifying old data**
  - The shift from “**AI that classifies/predicts**” to “**AI that creates**”
  - Probability (The “What”) and Density estimation (The “How”)

Course Objectives

- Theoretical foundations of modern architectures
- Practical implementation skills
- Critical analysis of SOTA models

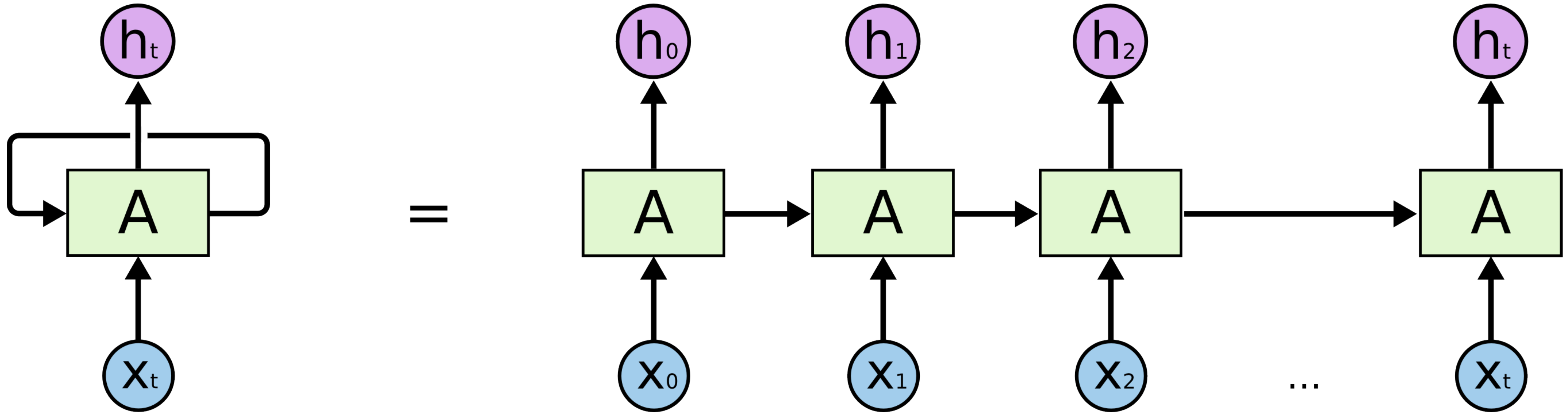
Why should you take this course?

- AI transitioning from a “niche research filed” to a “**general-purpose technology**”
- AI is becoming a **transformative technology**
- **Trillion-Dollar Opportunity**: McKinsey & Company estimates that Generative AI could add $2.6 trillion to $4.4 trillion annually to the global economy across 63 use cases (Citation: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier)
- The industry is shifting from building small, task-specific models (Discriminative AI) to deploying and fine-tuning **Foundation Models** (Generative AI)
- We are in a **Data-Model-Compute Triangle** explosion
- Generative AI provides a **Productivity Leap**

The Generative Toolkit

- Consists of four pillars
    - Generative Adversarial Networks (GANs)
    - Variational Autoencoders (VAEs)
    - Transformers
    - Diffusion Models
- Represents **foundational breakthroughs** of the last decade
- SOTA models **use a fusion** (for example, Stable Diffusion)

The Need for Persistence: Why We Need RNNs

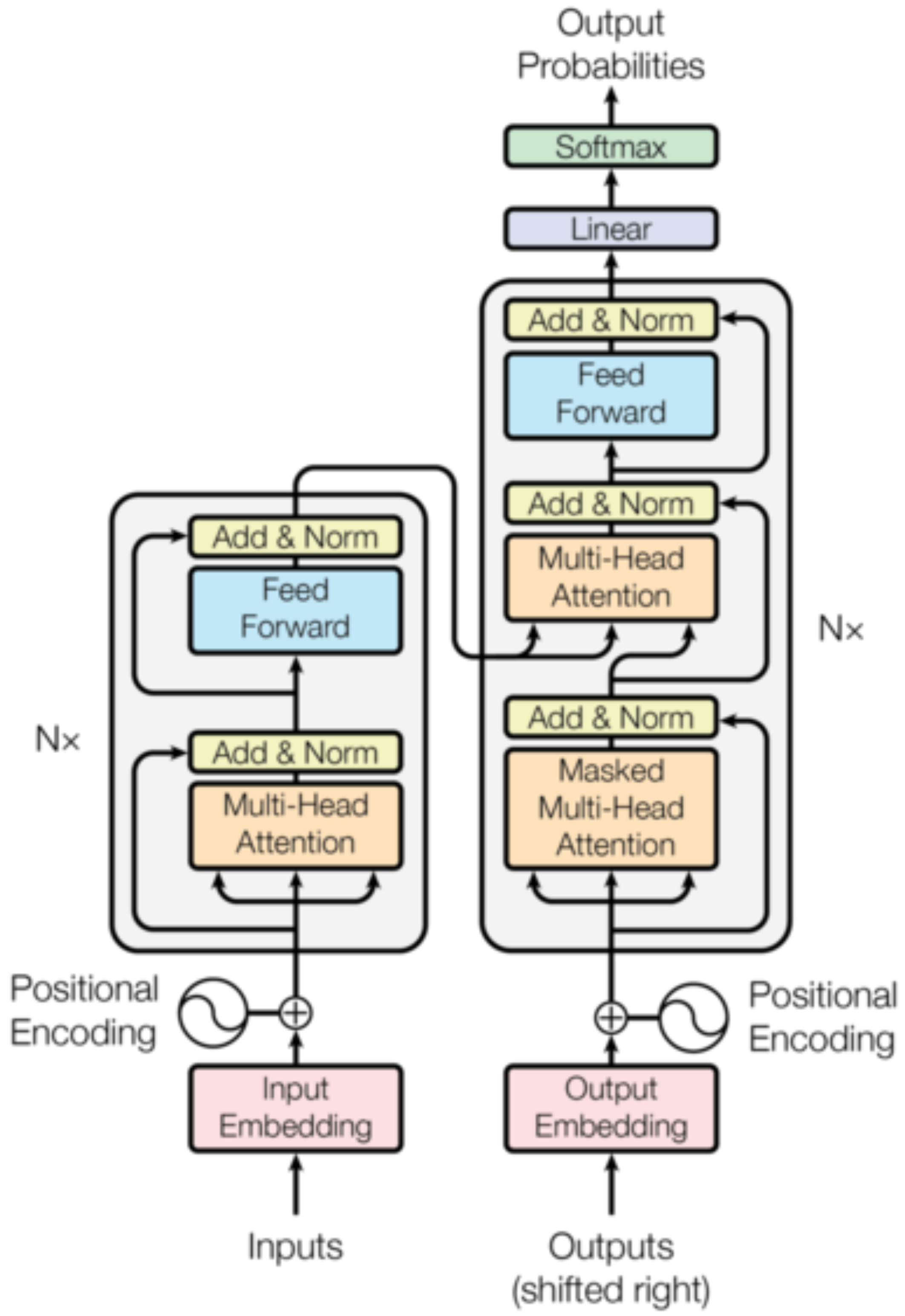


Unrolled Recurrent Neural Network

- Static Nets are **Problematic**
- Not useful for **sequential data** (text, speech, time series)
- RNNs introduce **loops**
- Limitations: **Vanishing gradients, Information bottleneck, Sequential Computation**

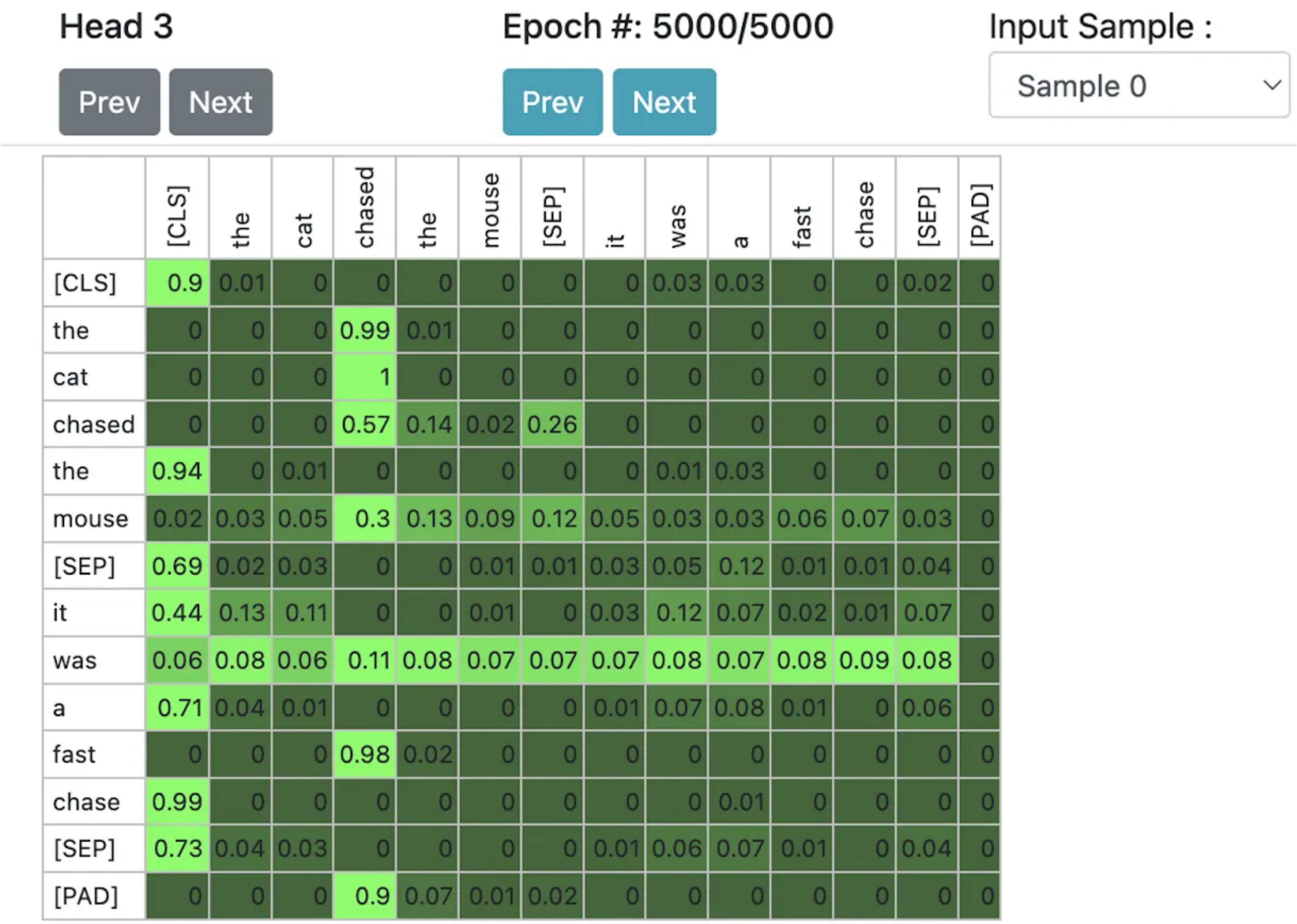
The Attention Revolution

- Important **Bottleneck** of RNNs: Loses detail over long distances
- Attention allows for **Parallelization**
- "**Attention is All You Need**": "Which other words in this sentence are most important to the current word?"
- **Global Context**: Unlike RNNs, the distance between words no longer matters
- **Transformers**: This single architectural breakthrough made Large Language Models like GPT **possible**



Credit: Vaswani et al."Attention is all you need"

The Attention Heatmap



	[CLS]	the	cat	chased	the	mouse	[SEP]	it	was	a	fast	chase	[SEP]	[PAD]
[CLS]	0.9	0.01	0	0	0	0	0	0	0.03	0.03	0	0	0.02	0
the	0	0	0	0.99	0.01	0	0	0	0	0	0	0	0	0
cat	0	0	0	1	0	0	0	0	0	0	0	0	0	0
chased	0	0	0	0.57	0.14	0.02	0.26	0	0	0	0	0	0	0
the	0.94	0	0.01	0	0	0	0	0	0.01	0.03	0	0	0	0
mouse	0.02	0.03	0.05	0.3	0.13	0.09	0.12	0.05	0.03	0.03	0.06	0.07	0.03	0
[SEP]	0.69	0.02	0.03	0	0	0.01	0.01	0.03	0.05	0.12	0.01	0.01	0.04	0
it	0.44	0.13	0.11	0	0	0.01	0	0.03	0.12	0.07	0.02	0.01	0.07	0
was	0.06	0.08	0.06	0.11	0.08	0.07	0.07	0.07	0.08	0.07	0.08	0.09	0.08	0
a	0.71	0.04	0.01	0	0	0	0	0.01	0.07	0.08	0.01	0	0.06	0
fast	0	0	0	0.98	0.02	0	0	0	0	0	0	0	0	0
chase	0.99	0	0	0	0	0	0	0	0	0.01	0	0	0	0
[SEP]	0.73	0.04	0.03	0	0	0	0	0.01	0.06	0.07	0.01	0	0.04	0
[PAD]	0	0	0	0.9	0.07	0.01	0.02	0	0	0	0	0	0	0

Credit: https://muneebsa.medium.com/deep-learning-101-lesson-30-understanding-text-with-attention-heatmaps-efe968a51bc2

Discussion: Transformers treat all words as having an equal ‘distance’ from each other. In doing so, have we lost the ‘flow’ of time that RNNs captured? Is ‘Order’ (Position) just a feature we add back in, or is it fundamental to how intelligence perceives the world?

Understanding vs. Generating: The BERT & GPT Divide

- **GPT (Generative Pre-trained Transformer)**:
  - Architecture: Decoder-only
  - Training Objective: Autoregressive (Causal Language Modelling) - It predicts the next token based only on preceding ones
  - Strength: Creative writing, coding, and open-ended conversation. It is a “storyteller”
- **BERT (Bidirectional Encoder Representations from Transformers)**:
  - Architecture: Encoder-only
  - Training Objective: Masked Language Modelling (MLM) - It sees the entire sentence at once and fills in “masked” blanks
  - Strength: Sentiment analysis, question answering, and named entity recognition. It is a “linguist”

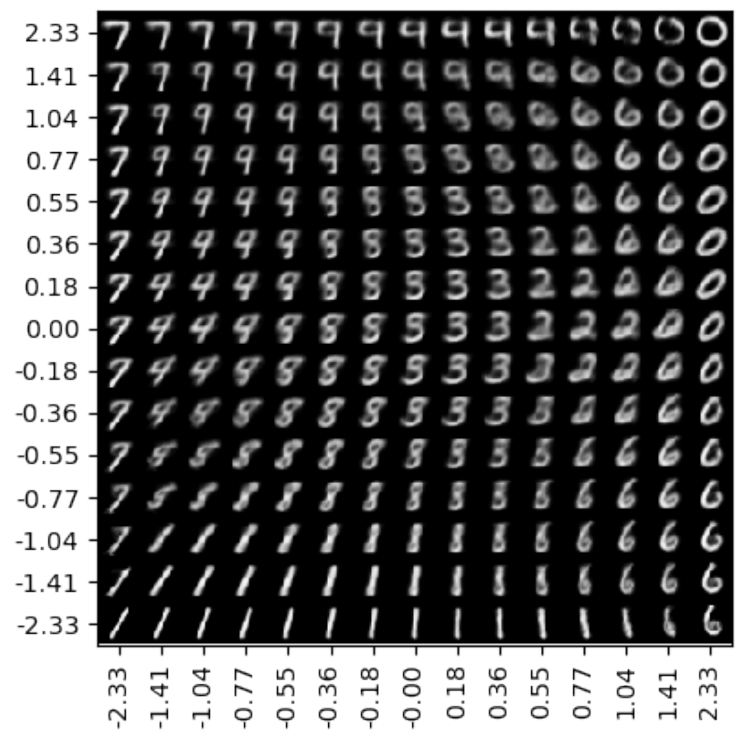
Discussion: GPT was trained simply to predict the next word. Is “Prediction” the same thing as “Understanding”? If you can predict the next word perfectly every time, do you eventually develop a world model, or are you just a “Stochastic Parrot”?

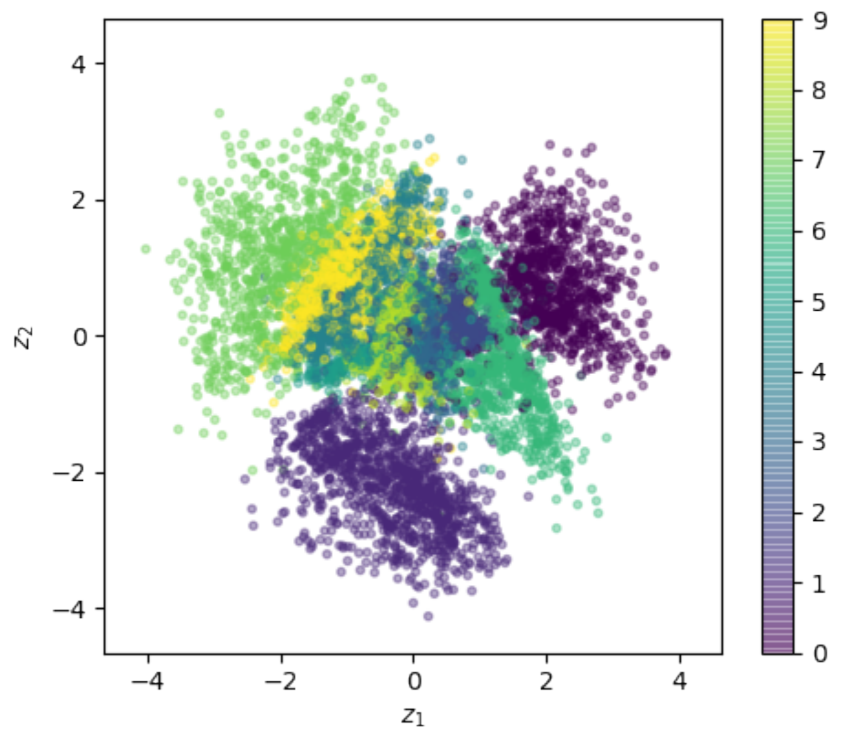
For additional reading: Read more about the Othello-GPT experiment - https://arxiv.org/pdf/2210.13382

RLHF: Aligning Raw Intelligence with Human Values

- The “**Base Model**” Problem
  - Model are trained on the internet
  - Could generate toxic content or follow instructions poorly
  - Could “hallucinate” facts because it is just imitating its training data
- Defining “**Helpfulness**”: Humans rank model outputs from best to worst
- **The Reward Model**: Train a second “critic" to learn human preferences, and teach LLMs
- **The Outcome** (Alignment): Safety guardrails, instruction following, and distinct “assistant” persona we see in ChatGPT or Gemini

VAEs & Latent Space: The Map of Creation





- From Complexity to Simplicity: VAEs compress high-dimensional inputs (e.g., 1024×1024 pixels) into a low-dimensional Latent Vector (z)

- This latent space organizes data by **features**

- The latent space is **continuous** and “**smooth**”

- Once trained, we can **throw away the original data**, pick a random point in the latent space, and the Decoder will create a “**realistic**” new image from that point

GANs & Adversarial Training: Learning Through Competition

- **The Two Players**:
  - **The Generator** (The Forger): Learns to create data that looks like the training set. Its goal is to maximize the Discriminator’s error rate.
  - **The Discriminator** (The Critic): Learns to distinguish between “Real” data and “Fake” data. Its goal is to minimize its own error rate.
- **Adversarial Training** (Zero-Sum Game): The networks improve simultaneously
- **High-Fidelity Outputs**: GANs are famous for producing sharp, high-resolution images
- **Convergence**: The training ends when the Generator is so good that the Discriminator can only guess with 50% accuracy (no better than a coin flip)

Diffusion Models: Generating Order from Chaos

- **The Diffusion Process**:
    - Forward (Destruction): Gradually adding **Gaussian noise** to an image until it is unrecognizable
    - Reverse (Creation): The model learns to **predict and subtract** that noise, step-by-step, to recover the original image
- **Stability** over GANs
- **Unrivalled Fidelity** - Current SOTA (e.g., Stable Diffusion, Midjourney, DALL-E 3)
- **Iterative Refinement**: “time” to fix errors and add intricate details

The Responsibility Gap: Ethics, Bias, & Governance

- **The Amplification of Bias**: If a training set contains historical gender or racial biases, the model doesn't just learn them—it concentrates them (e.g., generating only male doctors or female nurses)

- **The Problem of Hallucination & Truth**: Generative models prioritize “plausibility” over “truth.” In high-stakes fields like medicine or law, an ethical framework is required to prevent confident misinformation

- **Intellectual Property & Consent**: Who owns the output? Was the training data used legally? We must navigate the tension between innovation and the rights of original creators

Discussion: “If a model is trained on the entire internet, it reflects the internet’s majority views. If we ‘fix’ this by manually weighting minority data or forcing diverse outputs, are we making the model more accurate or less accurate to reality? Should AI represent the world as it is, or the world as we want it to be?”

Additional Topics (subject to time availability)

- Multimodal Models (Bridging text, image, and audio)
- RAG: Retrieval-Augmented Generation
- Model Efficiency
- Continual Learning
- Safety & Robustness
- System Design and Deployment

The Ultimate Goal: Understanding Intelligence

As you leave today, remember: we aren’t just learning how to code models. We are participating in the greatest intellectual experiment in history. Every time you tweak a weight or design a new architecture, you are asking the universe: “What is a thought?” and “What does it mean to know something?” The ultimate goal isn’t a better GPU or a larger dataset: it’s the answer to what makes intelligence possible.