

Generative AI

Module 10 of a course on *Ethical Issues in AI*

Prepared by

John Hooker

Emeritus Professor, Carnegie Mellon University

Chautauqua, June 2024

Topics

- How does generative AI work?
 - *Language models*
 - *Generative adversarial networks*
 - *ChatGPT etc.*
- Ethics of GPTs
 - *Creating document with GPTs.*
 - *Intellectual property issues (next module)*

Language models

- **Word associations**
 - *The basic tool.*
 - *What words tend to occur **near others** in a given type of document?*
- **Learning tools**
 - *Recurrent NNs, **GANs** (Generative Adversarial Networks)*
- **Large language models**
 - NNs may have **billions** of parameters

Language models

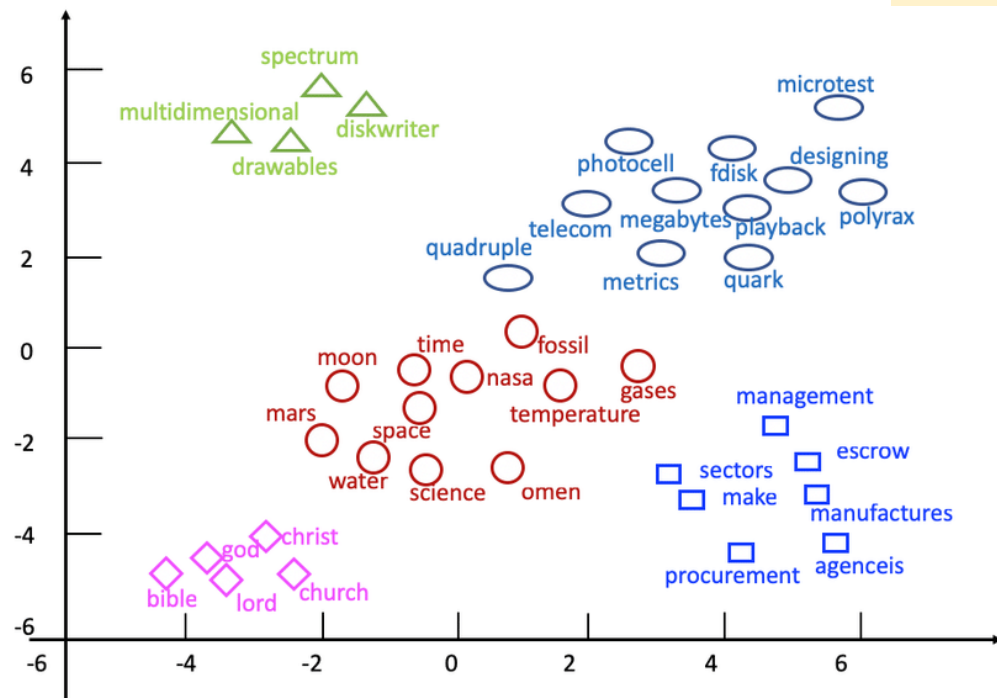
- **Spatial word embedding**

- *Word = point in higher-dimensional space.*

- Each coordinate is # of occurrences in a particular document.
- “Nearby” words tend to occur together.

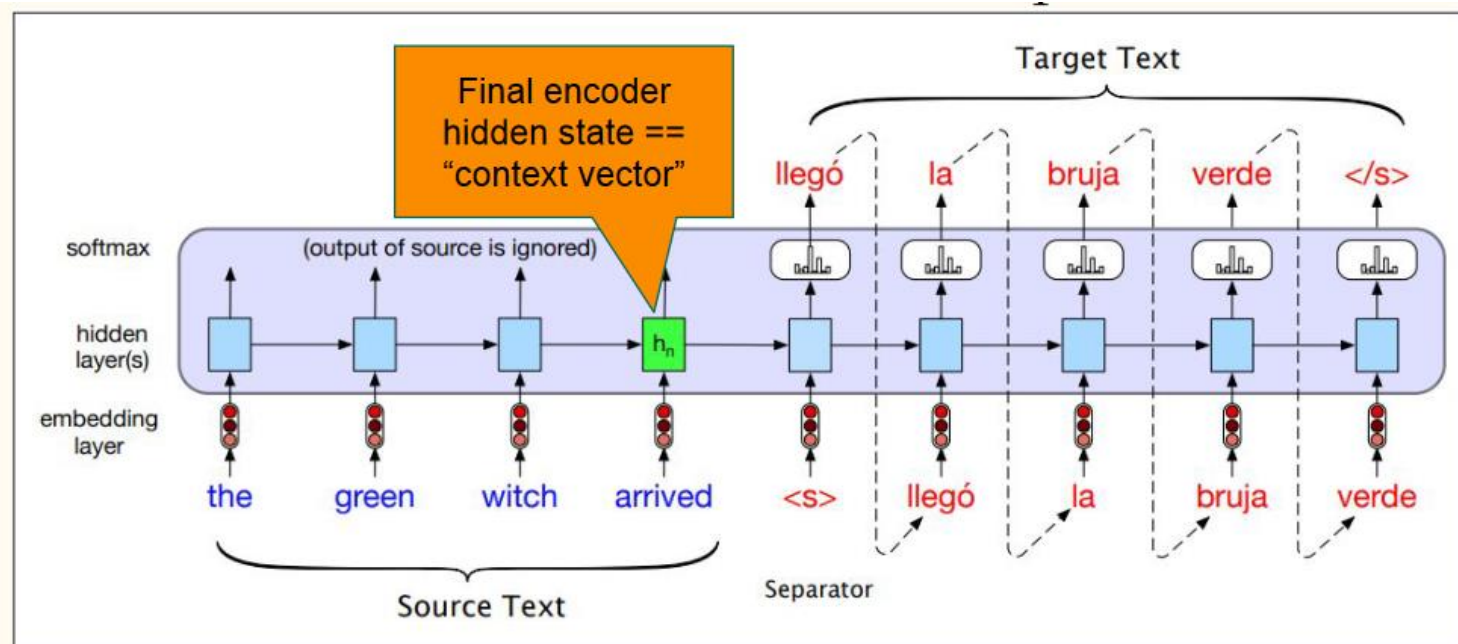
- *Dimension of space reduced from millions to hundreds.*

- Using technique similar to singular value decomposition.



Language models

- **Language translation.**
 - *Recurrent NN matches words in one language with those in another having similar spatial positions.*
 - Taking into account **context** and **order** of words.

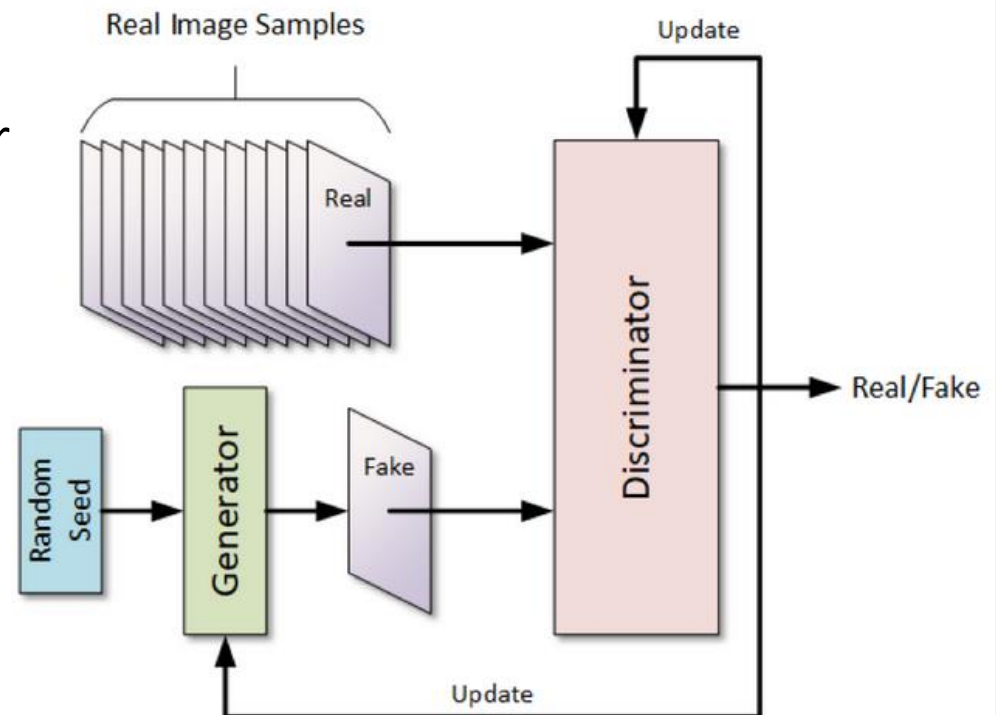


Generative Adversarial Networks

- **GANs** – Used for both **images** and **text**.

- *For example, “Draw a cat.”*

- **Generator** learns to create a fake image that the discriminator will classify as a cat.
- **Discriminator** learns to recognize fake images.
- This **adversarial** relationship generates a realistic but fictitious cat.



Generative Adversarial Networks



Cats that do not exist

Generative Adversarial Networks



People who do not exist

ChatGPT etc.

- **GPT** = Generative Pre-trained Transformer
 - Uses a **GAN** to **transform** existing material to **generated** text or images in response to a **prompt**.
 - It is **pre-trained** rather than learning on the fly.

Prompt: Create a Python script that calculates my age based on a birthday input

ChatGPT: Certainly! Here's a Python script that calculates your age based on a birthday input:

```
import datetime

def calculate_age(birthdate):
    today = datetime.date.today()
    age = today.year - birthdate.year [etc.]
```

ChatGPT etc.

- Chatbots date back to the 1960s (e.g., ELIZA)
 - *ChatGPT isn't really a chatbot.*
 - Unclear how to **train** a chatbot from online data.
 - *But it is very adept at finding text, images and voice that respond to a prompt.*
 - Due to co-occurrence of this material with verbiage similar to the prompt.
 - Based on vast training data and > **1 trillion** parameters.

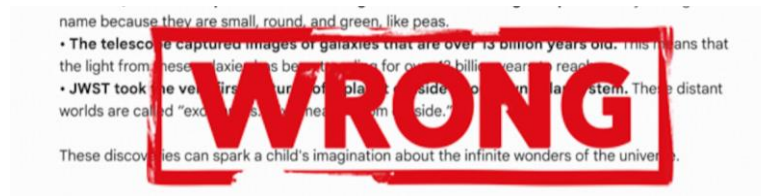


ChatGPT etc.

- Two key facts about GPTs
 - *They don't know what they are talking about.*
 - Language models **don't understand language**.
 - They would produce the same output if trained on **gibberish** that encodes text.

ChatGPT etc.

- Two key facts about GPTs
 - *They don't know what they are talking about.*
 - Language models **don't understand language**.
 - They would produce the same output if trained on **gibberish** that encodes text.
 - *They parrot what has **already been said**.*
 - With some modifications, cutting and pasting.
 - They **cannot reason** or construct an argument.
 - There is no ability (yet) to **distinguish true from false**.
 - Output can contain **mistakes** and **fabrications**.



Ethics of GPTs

- When is it ethical to **generate documents** with GPTs?
 - *We look at generalization and utilitarian arguments.*
- When do GPTs violate **intellectual property** rights?
 - *See the next course module.*

Note: The industry now euphemistically speaks of **document processing** rather than **document generation**

Ethics of GPTs

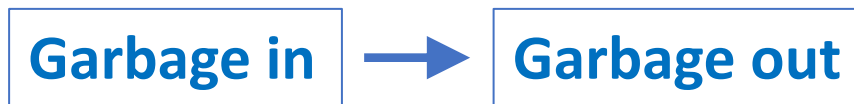
- **Generalization argument 1**

- *An essay or article carries an **implied promise**.*
 - The author has made a good faith effort to **say something that is true**.
 - The author has thought about the topic and describes **his/her own understanding** or argument.
 - In other words, the **author is, in fact, the author**.
- *Generating a document with a GPT **breaks this promise**.*
- *When done for convenience or profit, this is **not generalizable**.*

Ethics of GPTs

- **Generalization argument 2**

- *A new twist:*



Ethics of GPTs

- **Generalization argument 2**

- *A new twist:*



Ethics of GPTs

- **Generalization argument 2**

- *A new twist:*



- Meaningful material exists for training **only because** humans have created and researched **their own material**.
- GPT is **already** consuming its own output.
- *If everyone were to use GPTs...*
 - because they create usable material and are convenient...
 - ...then GPTs would **no longer create usable material**.
- *So, this practice is **not generalizable**.*

Ethics of GPTs

- **Utilitarian argument**

- *By laboring to create one's own documents...*
 - One hones **research and reasoning skills...**
 - ...which are in **short supply**.
 - One thereby makes a **greater positive contribution**.
- *Reliance on GPTs is therefore **not utilitarian**.*
 - This applies particularly to **students**.
 - In fact, one can argue that conscientious writing **creates literacy**.

