Why is AI Different?

What Can AI Do For Me?

Dis Prez Be Tots huMan Made!!

The Black Hole (1979) Disney

Elysium (2013) Neil Blomkamp

What Is AI Really?

https://stablediffusionweb.com/#demo
Prompt: Artificial Intelligence

- **Artificial Intelligence**:

A **Search Engine,** With Matrix Compression,
That Returns A Single Next-Most-Likely **Token**

A Compute Intensive Probabilistic Squirt Gun!!

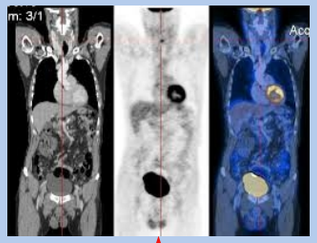


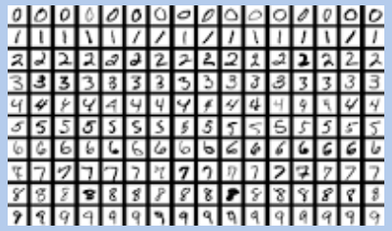

https://stablediffusionweb.com/#demo
Prompt: A computer connected to a squirt gun shooting 0's and 1's.

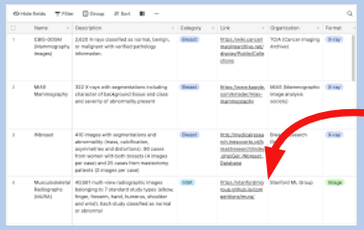
https://stablediffusionweb.com/#demo
Prompt: A Crowd of Regular People
Around a Robotic Parrot.

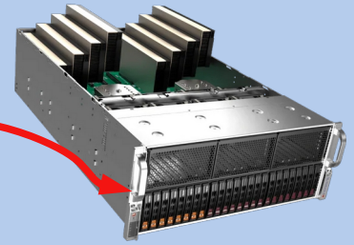
Artificial Intelligence – Just The Interface

Machine Learning – Complex Data



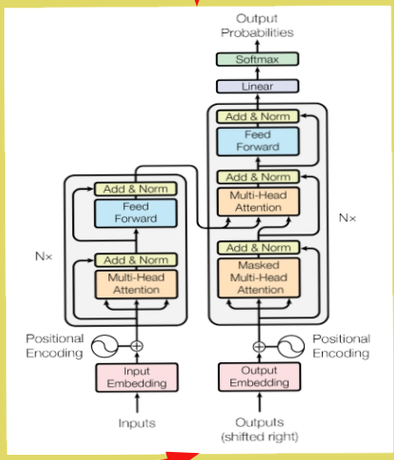






https://stablediffusionweb.com/#demo
Prompt: Albert Einstein as a Robot.

Transformer Models

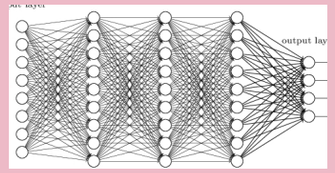


Vector Database

142.2, 3.22, 16520.2, 5.0, 5223.19...
5.62, 8.46, 550.2, 5.0, 5223.15, 12...
76.62, 12420.2, 925.0, 11623.19, 1...
622.22, 43520.2, 3673.0, 2366.163...

Update Matrices
With New Numbers (Info)

Neural Nets



Non-linear
Neural Nets

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{array}$$

Linear Algebra

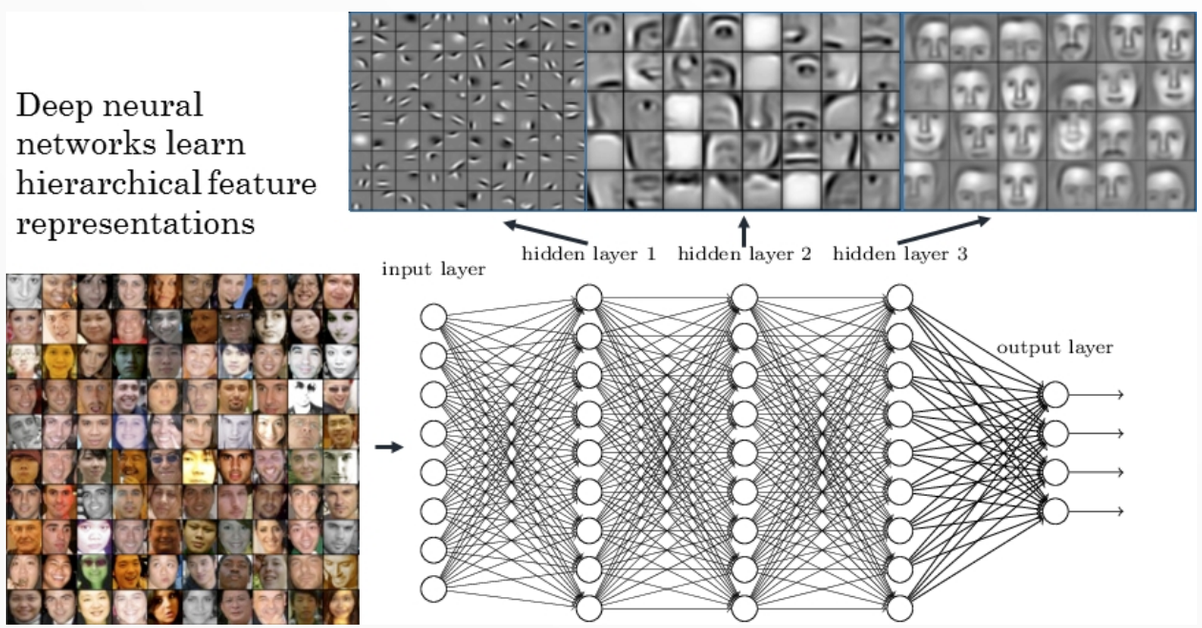
Why ML? Why You?

- Public ML is **1 Year Old**
  - Exponential Adoption: 100M Consumers, 3 Months
  - 7nm GPU + COVID Pushed ML Roll-out into 2022
- Street-Level Talent Needed **Everywhere**
  - Lead Non-Adopter Neighbors, Get 100 Grad Students
  - COVID Created Mass SW Pool, Lab→Pub→Github→**You**
  - Data Conv, Activity Stabilization, GAN/Synth Training
  - Sec/IT Adm, Teachers, Marketing, Dbg/QA, Plant Mgt

https://cdn.openart.ai/stable_diffusion/141d409516352ce1ad2b5ffb93f924d3b12247c5_2000x2000.webp

Neural Networks

- **Sentiment**:
  - Sequential Model (Matrices)
  - Functions In-Between



Matrix: 1x8→ [8x9→9x9→9x9→9x4] →4x1

https://medium.com/@fenjiro/face-id-deep-learning-for-face-recognition-324b50d916d1

stored as fat or glycogen f
metabolism begins in the

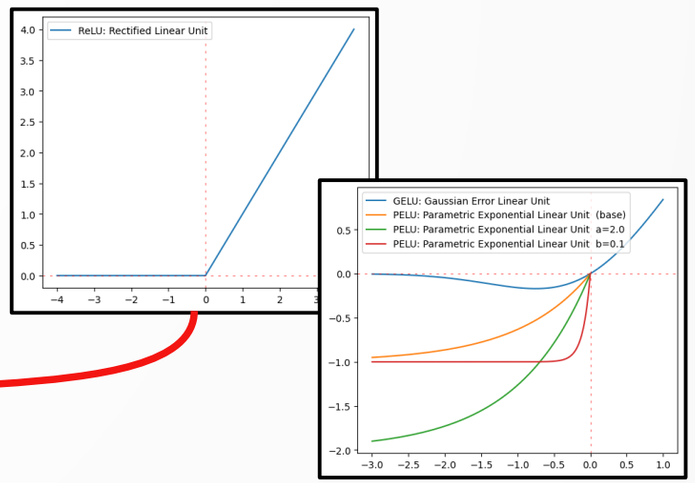

```
{
  "value": "stored",
  "confidence": 0.9965261220932007,
  "geometry": [
    [
      0.1145450367647059,
      0.07421875
    ],
    [
      0.16256893382352944,
      0.0888671875
    ]
  ]
},
```

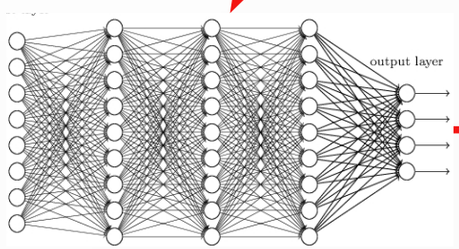

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & w_1 & x_1 \\ 1 & w_2 & x_2 \\ 1 & w_3 & x_3 \\ 1 & w_4 & x_4 \\ 1 & w_5 & x_5 \\ 1 & w_6 & x_6 \\ 1 & w_7 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

https://en.wikipedia.org/wiki/Design_matrix

Activation Func - Neuron?

- **ReLU($[y_i]$)** = $[y_1, \ldots, 0, y_5, 0, y_9]$:
  - Nothing Good to Say? Say Nothing!!
  - Train on **Positive Correlation** Only





- **SoftMax($[y_i]$)** = $[p_i]$:
  - Tight Correlation to Wide Percent
  - **e^{1003}** vs **e^{1000}** = **95%** vs 5%
    - Normally 1003 vs 1000 is 50.15% vs 49.85%

$$P(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{K} e^{y_j}}$$

e = 2.718...

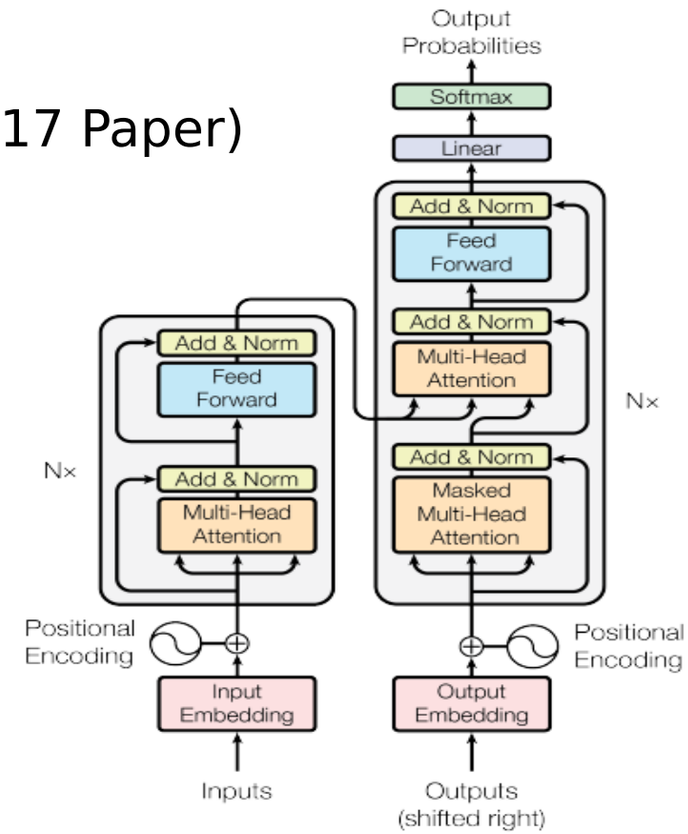
Solving Big Problems!!

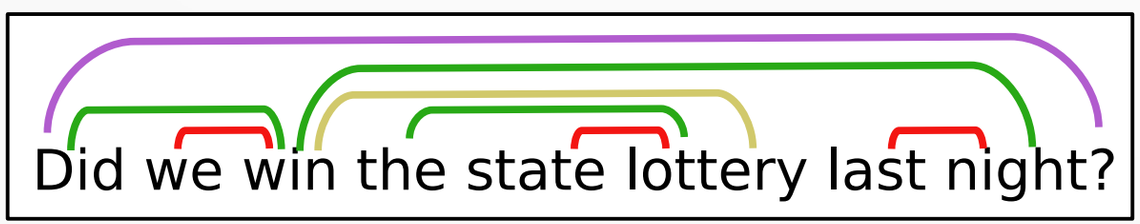
- **Solving Language, Med-Image, … Models!!**
  - Raise Lots of Investor Cash
  - Buy All The Datasets You Can
  - Play With Expensive Computers
  - Give High-Profile Interviews

Intermission
Brief Q&A
Refresh Your Beer

Attention Transformer

- **Attention Is All You Need**: (2017 Paper)
  - Unifies All NN Fields Into One!!
- **Transformer** Model
  - Input Context:
    - **In-Context** of Tokens Into “Relations”
  - Attention & Attention **Head**:
    - **River Bank** vs **Savings Bank** vs **West Bank**
    - **My Bank** vs **Jos. A. Bank** vs **Money in the Bank**





Input & Output Vocab’s

- **In**: **Attention** & **Out**: **SoftMax Choice(s)**:
  - Chat: Input: Unicode Output: Unicode
  - Text To Speech: Input: Unicode Output: Cosines/Wavelets
  - Speech-to-Text: Input: Cosines/Wavelets Output: Unicode
  - Noise Suppression: Input: Cosines/Wavelets Output: Cosines/Wavelets
  - Image OCR: Input: [X,Y: RGB] Packets Output: Unicode
  - Movie Creation: Input: Unicode Output: [X,Y: RGB] Frames
  - Code Creation: Input: Unicode Output: Semantic Code
  - Med Diagnostics: Input: MRI/CT Images Output: Disease Diagnosis
  - **Agentic AI: Input: Audio Speech Output: Call Other AI’s**

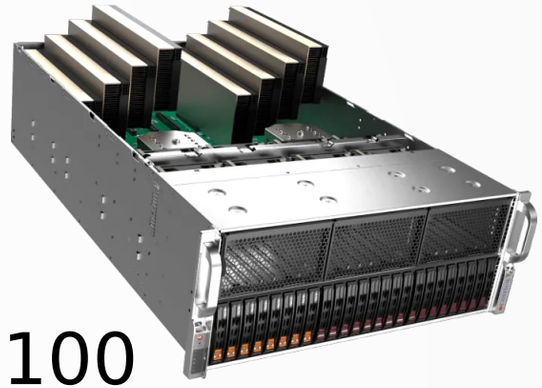
Custom Input → Input Relations → Trained Model Purpose → Output Relations → **Custom Output**

Generic Trainable Core

https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/nvidia-a100-pcie-3qtr-top-left-2c50-d.jpg

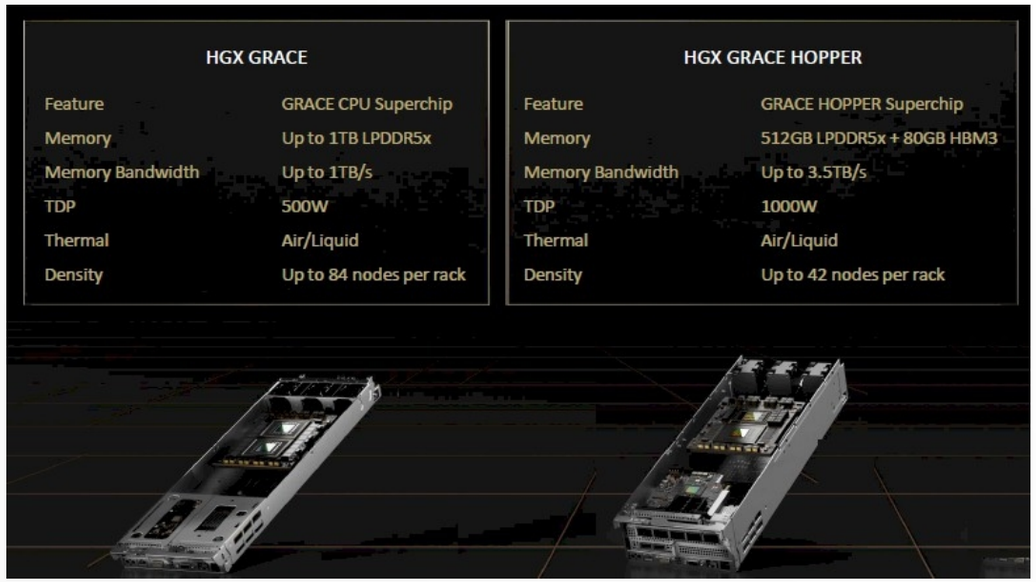
<7nm, >50B Transistor Chips:

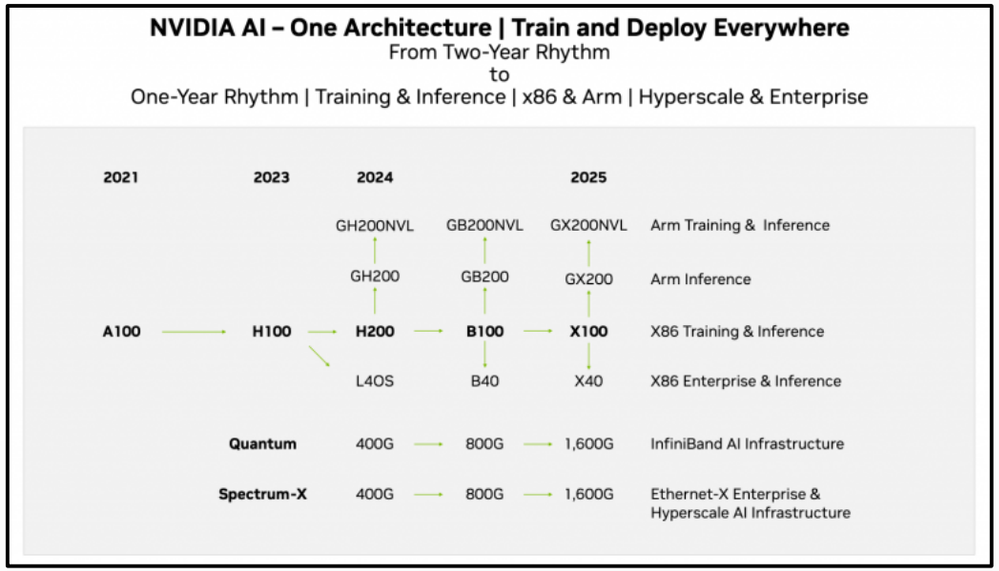
- Hardware Side of Exponential Era



- Peta/Tera Flops To Choose 1 Token
- Tensor Cores Expedite Model Training
- Nvidia >10X: A100, H100, H200, B100, X100

https://www.supermicro.com/en/accelerators/nvidia





https://www.hpcwire.com/2023/10/16/annual-gpu-upgrades-nvidias-plan-for-faster-chips/

Best Opportunity is NOW!

- **Desktop and Small Biz ML is Brand New**:
  - Accounting Firms Now Spending $300M/yr on ML
- **Covid => Training Materials and Code**:
  - Free SW Resources are Amazing, Google Colab Too
- **Hardware is Available and Inexpensive**:
  - Like Patching SW, Any Model Can be **Retrained**
  - From Scratch or **Model Upgraded** on Your Data

So What Does This All Mean For Me!!!

The Future is Bright

- **Personal AI Operating Systems**:
  - Genetic-Health, Micro-Nutrition, Precise Robotic Surgery
  - Automated Maintenance, Self-Flying Shuttles, Education
  - Instant Movies, V-BestFriends, e-Psychiatrist, Assistants
- **Logistics Optimization Everywhere**:
  - Precision Agriculture, Power Grid, International Travel
  - RT Lang Translation, Auto-Construction, Internet Routing
  - Hyper-Real Gaming, Material Science, Robotic Assembly

Adapt or Retire?

- **Opportunities in Disruptive Tech**:
  - Every Industry Is/Will Adopt ML As A New Tech Base
  - ML Data-Conv/Automation In High Demand NOW!!
  - Stability AI: Base Models Open, Hire or Roll Your Own
  - **Hugging Face**: 1000’s of Free Models to Adapt

- **Rules-Based Jobs Disappear**:
  - Lawyer, Med-Tech, Accountant, Author Gone 3 Years?
  - Programmer, Architect, Pilot?, Teacher Gone 5 Years?

So Long And Thanks For All The Fish!!!

Nvidia Desktop GPU’s

- **Quadro**:
  - A6000 48GB, A5000 24GB, A2000 12GB
  - **Engineering** Workloads, Lower Power, Highly Reliable
  - Prev Gen: RTX 5000, RTX 4000 **no “A”** (much weaker)

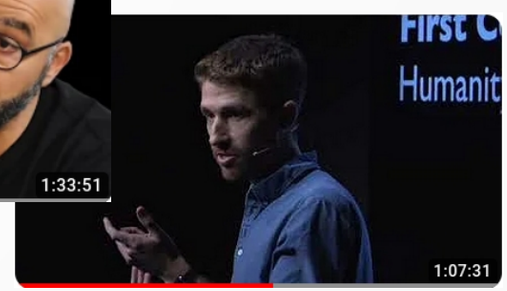
- **GeForce**:
  - RTX 4090 24 GB (A Beast), RTX 4070 12GB
  - **Gaming** DLSS AI Texture Generation InCard
  - Consumes Massive Power, Needs Special Tower/MBoards

Under Control?

https://www.youtube.com/watch?v=Se91Pn3xxSs

- **Regulatory Controls Regime**
  - Who Benefits Financially?
  - Who Controls Context-Bias?
  - Monitoring? No Cameras Needed
  - AI Weapons Treaties?

https://www.youtube.com/watch?v=nltlE4wqv3g

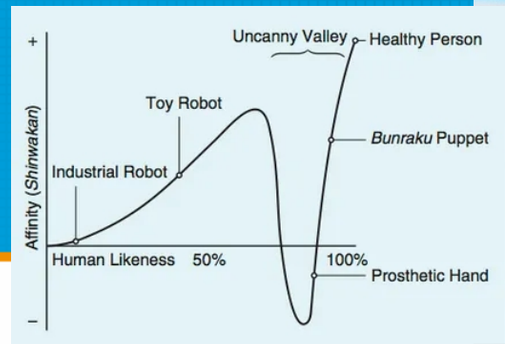


https://www.youtube.com/watch?v=fDHvUviV8nk

https://www.youtube.com/watch?v=xoVJKj8lcNQ

- **ML Only Practical Since 2022 from COVID?**
  - ML Software Around for Decades, Stable, Open Source
  - 2020 <7nm Hardware Makes ML A Commodity Now!!

Uncanny Valley



http://2045.com/news/33979.html

- **Japan's Aging Population**:
  - Urbanization means few babies into fewer workers
  - Automation and Robotics sustain Aging Population

- **Masahiro Mori's Paper**: (2012)
  - AI: Useful to **Repulsive** to Amazing

- **Ghost In The Shell**: (2017)
  - ScarJo's Face on Killer Cyborg (Cool or Frightening)?

OpenAI GPT-3

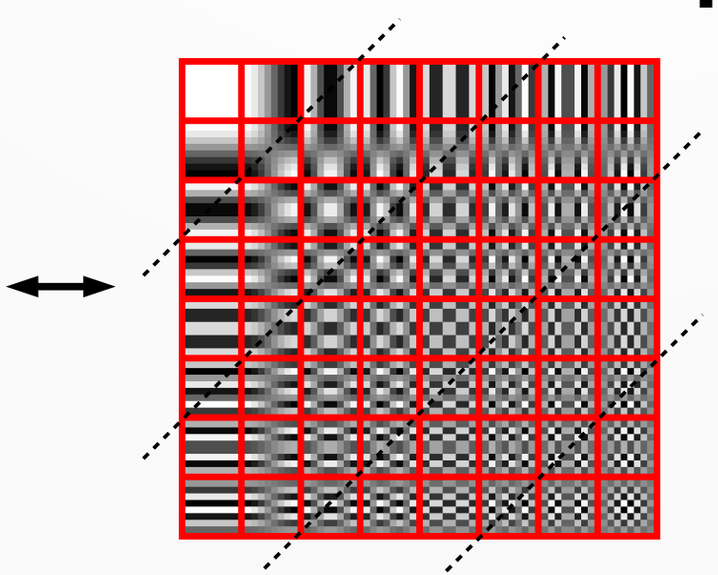
- **DataCenter Models**:
  - Nodes: 125M, 350M, 760M, 1.3B, 7B, 13B, 175B
  - VRam: FP-32: 700GB of GPU Ram
  - Attn: 175B: 96 A-Layers, 96x128 Tokens Each
  - Context: 2048 Tokens of User Input
  - Batch: 125M: 0.5M Tokens, 175B: 3.2M Tokens
  - Cost: 355 GPU Years, ~$4.6M / Training Run Min
  - Power: Approx 30 MWatts / Training

https://lambdalabs.com/blog/demystifying-gpt-3

Diffusion Training

- **DCT Transform** (no -er) **(JPEG)**
  - Input: High-Rez **Digital** Image
  - Weights: Discrete **Cosine** Coefficients
  - Output: Floating Point Numbers
  - Transform: Pixel **Ints** to **Freqs** and **Back**

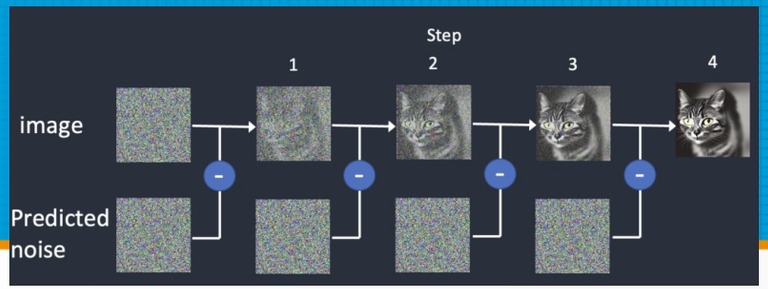
Trivial Image Decomposition Via Basic Math



$$G = \begin{bmatrix} -415.38 & -30.19 & -61.20 & 27.24 & 56.12 & -20.10 & -2.39 & 0.46 \\ 4.47 & -21.86 & -60.76 & 10.25 & 13.15 & -7.09 & -8.54 & 4.88 \\ -46.83 & 7.37 & 77.13 & -24.56 & -28.91 & 9.93 & 5.42 & -5.65 \\ -48.53 & 12.07 & 34.10 & -14.76 & -10.24 & 6.30 & 1.83 & 1.95 \\ 12.12 & -6.55 & -13.20 & -3.95 & -1.87 & 1.75 & -2.79 & 3.14 \\ -7.73 & 2.91 & 2.38 & -5.94 & -2.38 & 0.94 & 4.30 & 1.85 \\ -1.03 & 0.18 & 0.42 & -2.42 & -0.88 & -3.02 & 4.12 & -0.66 \\ -0.17 & 0.14 & -1.07 & -4.19 & -1.17 & -0.10 & 0.50 & 1.68 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \downarrow v. \end{matrix}$$

$u \rightarrow$

Diffusion Training



https://stable-diffusion-art.com/how-stable-diffusion-work/

- **Frabbit or Bunrog?**: (2D Color Cross-Attention Along X,Y Dims)
  - Dissolve Many Rabbit and Frog Images To Noise Panel
  - Dissolve Rabbit Only on Sharp Background
  - Dissolve Frog Only on Sharp Background
  - Dissolve Backgrounds, Keep Animal Sharp
  - **Attention** Map Words to X,Y,RGB Probs
  - Retrain Thousands Of Times, Min Error
  - Ask Stable Diffusion:
    - “photo of a hybrid between a rabbit and a frog”

Model Bias?

https://cdn.openart.ai/stable_diffusion/141d409516352ce1ad2b5ffb93f924d3b12247c5_2000x2000.webp

Hard Work Already Done

```python
from keras import layers
from keras.models import Model
from mltu.model_utils import residual_block

def train_model(input_dim, output_dim, activation='leaky_relu', dropout=0.2):

    inputs = layers.Input(shape=input_dim, name="input")
    input = layers.Lambda(lambda x: x / 255)(inputs)

    x1 = residual_block(input, 16, activation=activation,
                        skip_conv=True, strides=1, dropout=dropout)
    x2 = residual_block(x1, 16, activation=activation,
                        skip_conv=True, strides=2, dropout=dropout)
    x3 = residual_block(x2, 16, activation=activation,
                        skip_conv=False, strides=1, dropout=dropout)
    x4 = residual_block(x3, 32, activation=activation,
                        skip_conv=True, strides=2, dropout=dropout)
    x5 = residual_block(x4, 32, activation=activation,
                        skip_conv=False, strides=1, dropout=dropout)
    x6 = residual_block(x5, 64, activation=activation,
                        skip_conv=True, strides=1, dropout=dropout)
    x7 = residual_block(x6, 64, activation=activation,
                        skip_conv=False, strides=1, dropout=dropout)

    squeezed = layers.Reshape((x7.shape[-3] * x7.shape[-2], x7.shape[-1]))(x7)
    blstm = layers.Bidirectional(layers.LSTM(64, return_sequences=True))(squeezed)

    output = layers.Dense(output_dim + 1,
                          activation='softmax', name="output")(blstm)
    model = Model(inputs=inputs, outputs=output)
    return model
```

https://pylessons.com/ctc-text-recognition

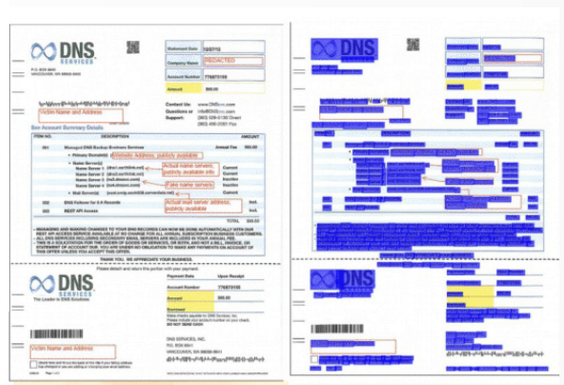
```python
def residual_block(x: tf.Tensor, filter_num: int, ...):
    x_skip = x
    x = layers.Conv2D(filter_num, kernel_size,
                      padding = padding, strides = strides,
                      kernel_initializer=kernel_initializer)(x)
    x = layers.BatchNormalization()(x)
    x = activation_layer(x, activation=activation)

    x = layers.Conv2D(filter_num, kernel_size,
                      padding = padding,
                      kernel_initializer=kernel_initializer)(x)
    x = layers.BatchNormalization()(x)

    if skip_conv:
        x_skip = layers.Conv2D(filter_num, 1,
                               padding = padding,
                               strides = strides,
                               kernel_initializer
                                  =kernel_initializer)(x_skip)
    x = layers.Add()([x, x_skip])
    x = activation_layer(x, activation=activation)
    if dropout:
        x = layers.Dropout(dropout)(x)
    return x
```

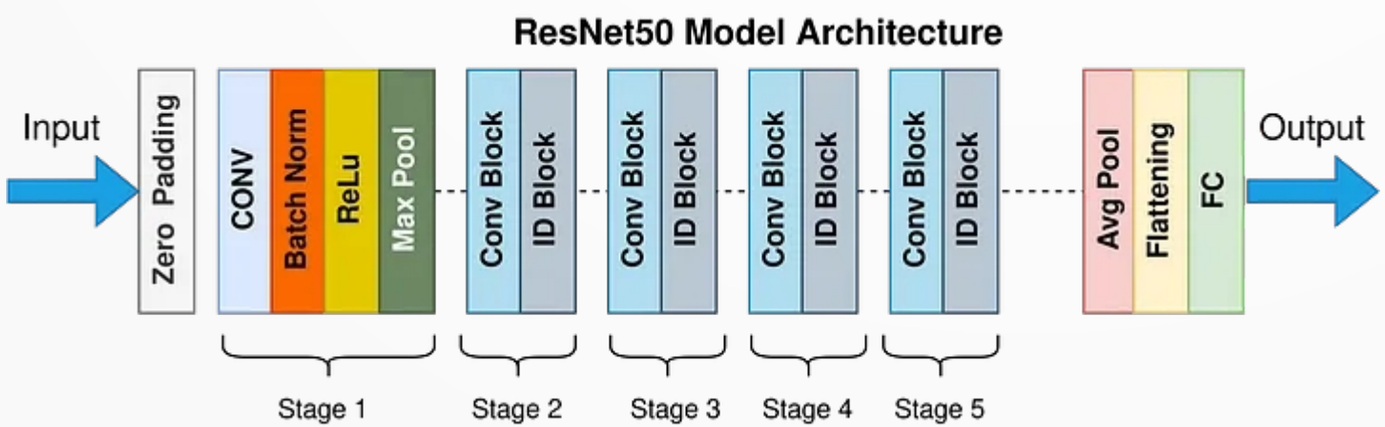
https://github.com/pythonlessons/mltu/blob/main/mltu/tensorflow/model_utils.py

Mindee DocTr & ResNet50

https://github.com/mindee/doctr/blob/main/docs/images/ocr.png

- DocTr Receipt/Invoice/... OCR
  - Fully Implemented ResNet50
  - OpenSource And Models Public
  - Retrain on Private/Custom Image Docs






```
tf.keras.applications.resnet50.ResNet50(
    include_top=True,
    weights='imagenet',
    input_tensor=None,
    input_shape=None,
    pooling=None,
    classes=1000,
    **kwargs
)
```

https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

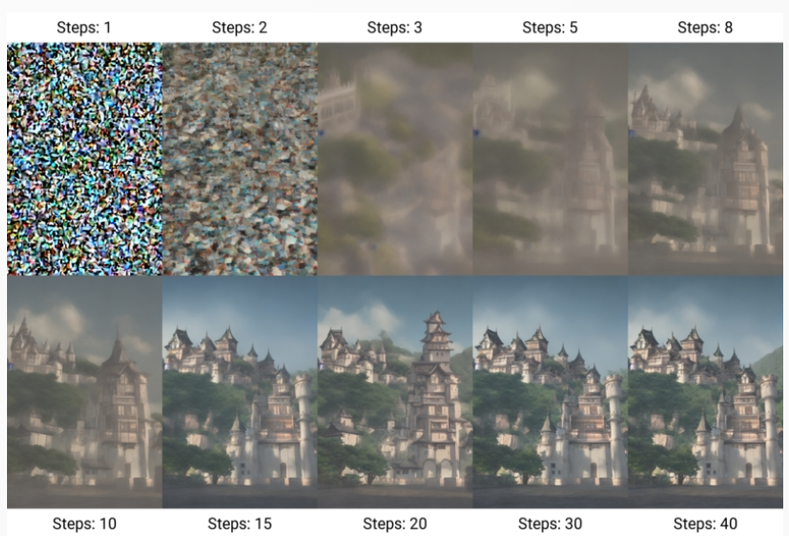
General DC Model Info

- **Chat Model Parameters (Nodes) and Sizes:**
  - GPT-4 (1.76**T** Nodes), 120 Layers, >175K Vocab, 32K Context
  - LLaMA 2 (70**B** Nodes), 32 Layers, 32K Vocab, 4K Context

Attention

- **Cross-Attention:**
  - Input and Output Radically Different
  - EnglishText-to-FrenchAudio, Text-to-Video

- **Image Diffusion Model**:
  - Image Blurring To Pure Noise and Back
  - Pixel Attention in Multiple Dimensions



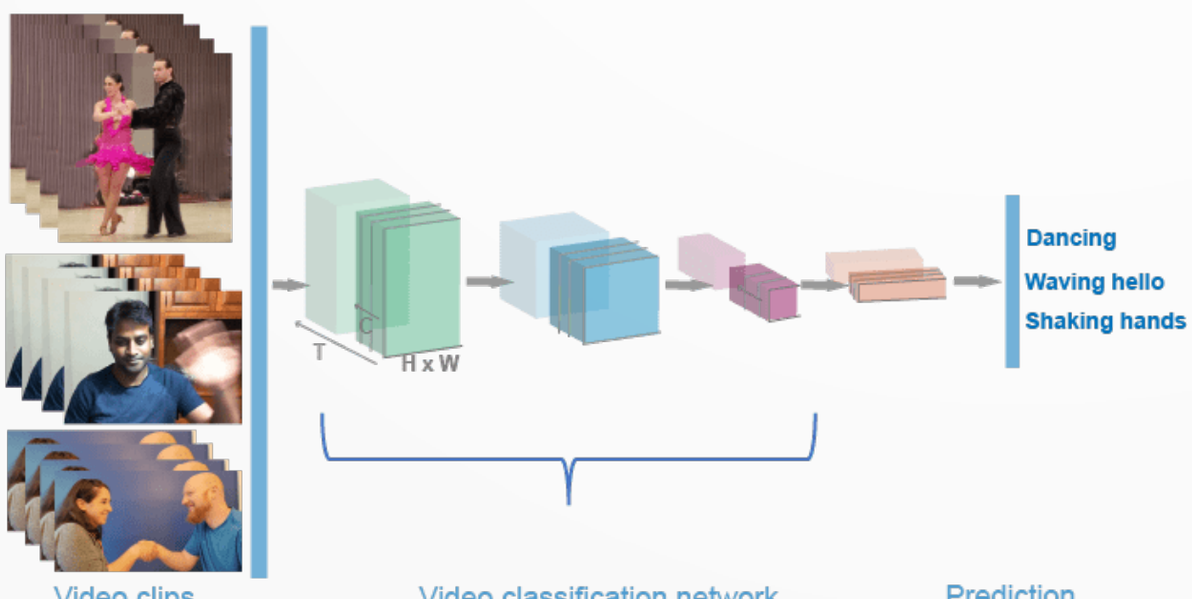
https://en.m.wikipedia.org/wiki/File:X-Y_plot_of_algorithmically-generated_AI_art_of_European-style_castle_in_Japan_demonstrating_DDIM_diffusion_steps.png

How AI Models Work

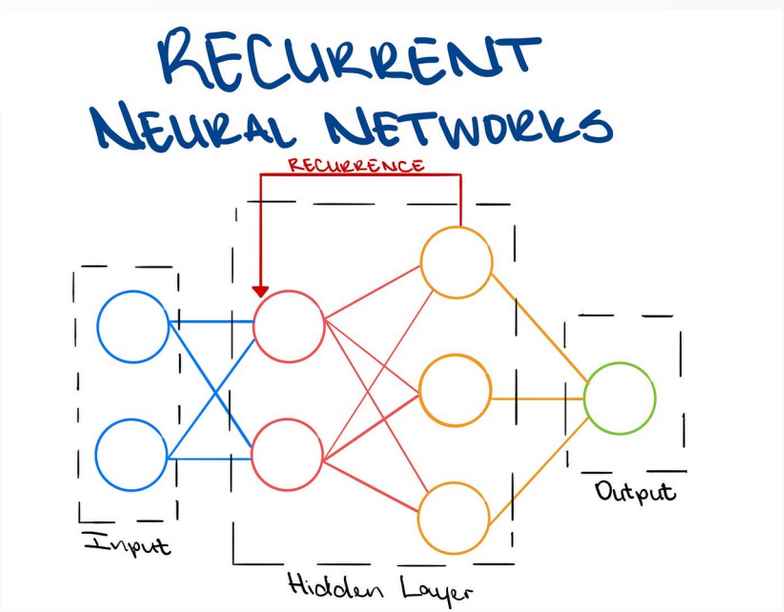
- **Model 8-Bit Quantization**
  - Crop Nodes From Floating Point 16 to 8 (or 4 bit)
  - Model Perf Mostly Unchanged at ½ GPU Memory
- **Low Rank Adaptation (LoRA)**
  - Second Matrix To Downsize/Specialize Large Model
- **Rank One Model Editing (ROME Paper)**
  - Tweak One Node Number And Change “Knowledge”
  - Find Node That Relocates Eiffel Tower to ROME Italy.

Recurrent Neural Net

- **Feedback For Video Classification**
  - **Previous Output** Influences Decisions
  - Notion of Context or Short Term Memory
  - Recognition of Human Activity In Video

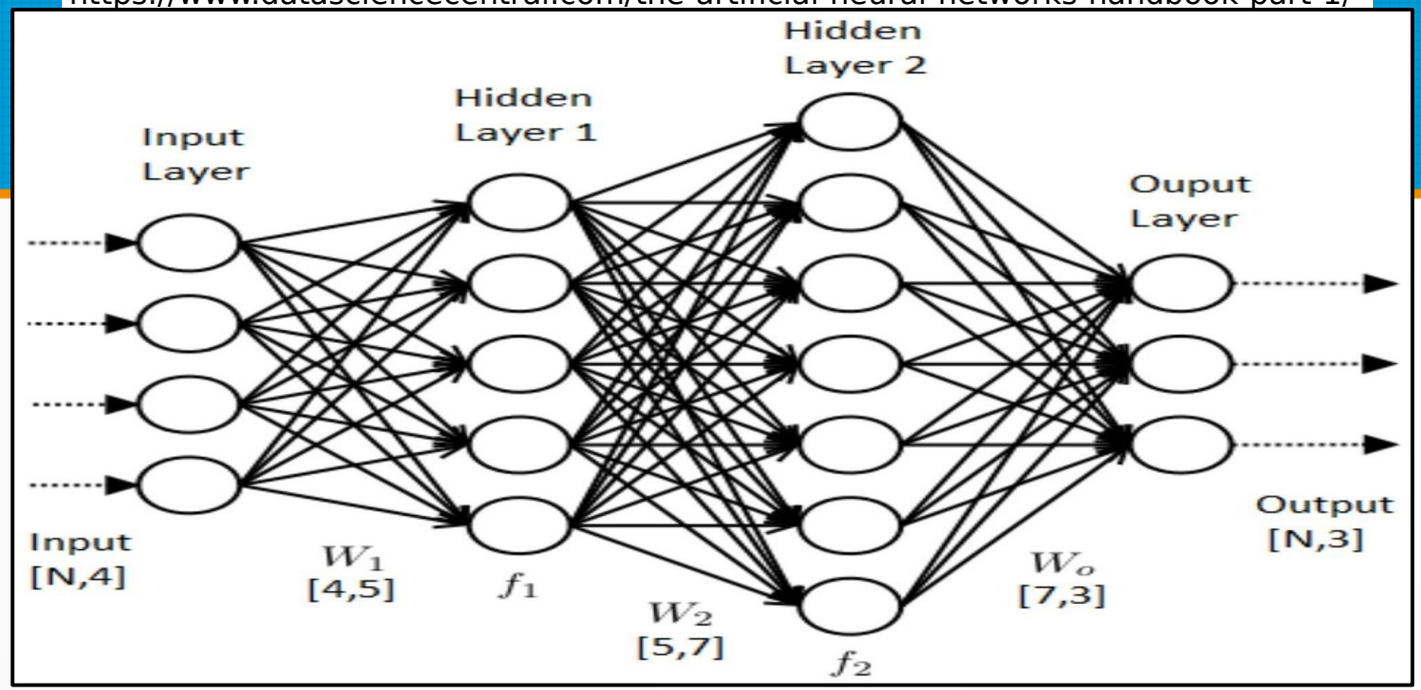


https://www.mathworks.com/help/vision/ug/video-classification-using-deep-learning.html



https://towardsdatascience.com/introducing-recurrent-neural-networks-f359653d7020

https://www.datasciencecentral.com/the-artificial-neural-networks-handbook-part-1/



INPUT

$$X \begin{bmatrix} & 1 & 2 & \dots & n \\ 1 & a_{11} & a_{12} & \dots & a_{1n} \\ 2 & a_{21} & a_{22} & \dots & a_{2n} \\ 3 & a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m & a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

RELU()

RELU()

SOFTMAX()

Probable Output:

Yes or No,
Set Choice,
Next Pixel,
Next Freq.,
Next Letter

Multi-Head & MultiModal

- **Multi-Head Attention**:
  - 1 Input: Each **Head** Identifies A Property of Input
  - En/Fr/Sp, Adult/Child, US/EU/BR, Angry/Joking/Formal

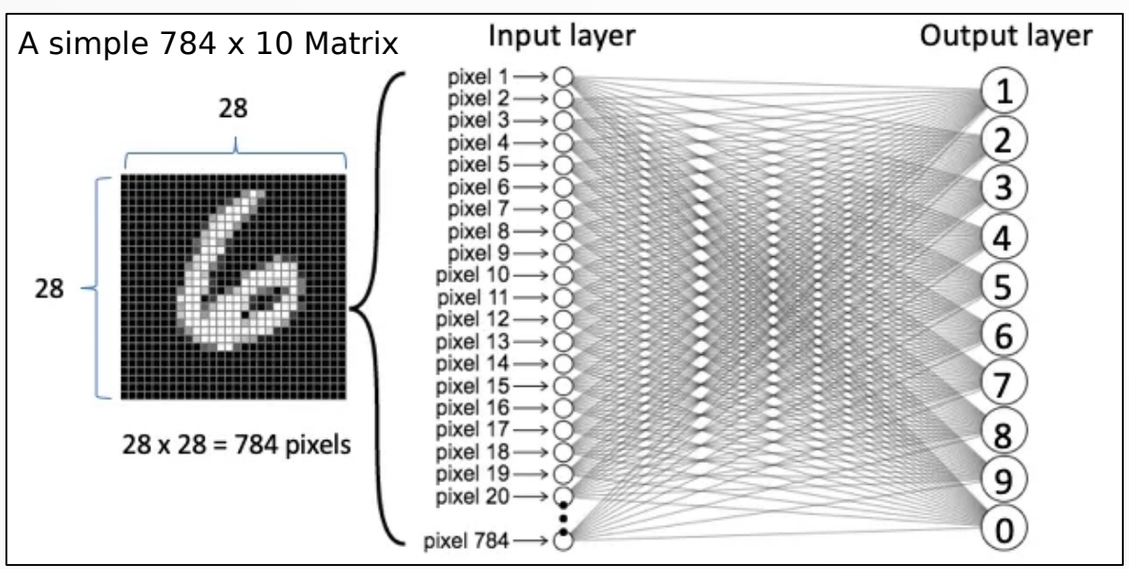
- **MultiModal Models**:
  - **Youtube**: Video + Audio + Captions + Genre Tags + Comment Stream
  - Synchronized Streams Can be Co-Modeled as One
  - Describe A Custom Genre → Model Creates Whole Movie

Linear Neural Net

- **Classifier / Cluster**:
  - Trivial Input **Vocab** to Output **Vocab**
  - Matrix **Trained** Until In→Out Perfect
  - **Embedding** Packs **Filtering** Know-How

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & w_1 & x_1 \\ 1 & w_2 & x_2 \\ 1 & w_3 & x_3 \\ 1 & w_4 & x_4 \\ 1 & w_5 & x_5 \\ 1 & w_6 & x_6 \\ 1 & w_7 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

https://en.wikipedia.org/wiki/Design_matrix



https://medium.com/dataman-in-ai/module-6-image-recognition-for-insurance-claim-handling-part-i-a338d16c9de0

https://www.buzzfeed.com/adambvary/inside-the-all-plinko-episode-of-the-price-is-right

- **Attention Is All You Need**:
  - In and Out Can Be **Any Digitized Format(s)**
  - **Tensor** Training “Solves” For Unknown Patterns
- **Andrej Karpathy**: Let’s Build GPT
  - Excellent End-to-End Development and Training Example
  - Build An Infinite Shakespeare Model
- **Transformer** is Pervasive
  - Any In & Out Format Can Run on GPU
  - **All** Digitized Input and Output Works
  - Mathematically Elegant and Extensible

https://www.youtube.com/watch?v=kCc8FmEb1nY

PERL w/ Inline Python

- **Perl Imports Python Modules as Native**
  - Compile Python Shared lib Enabled - Python 3.11
    - pip3 install --extra-index-url https://developer.download.nvidia.com/compute/redist \
      tensorflow tensorboard keras python-doctr ...
  - Compile/Download PERL as Usual - PERL 5.38
    - cpan install Inline Inline::Python (Give it paths to Python EXE and lib/python3.11/libpython3.11.so)

The Future is Bright

- **Commodity ML is 1 Year Old, Use It NOW**:
  - Apply Principles to Your Existing Industry or Data
  - **ANY** Ugly-X to Simple-Y Conversion/Pattern/Filter
  - Retrain Github/**Hugface** to Your Needs... BINGO!!!
- **Get Books, Youtubes, Meetups and Dive In**:
  - COVID Created Free Books, Videos, Training Options
  - 4 Years Your Chance Will Be Over, Don’t Wait
  - “Exponential Growth” Tech So Start **ASA-F’n-P**