Deep Learning for Vision & Language

Course Recap

RICE UNIVERSITY

Final Project

- PDF Project report (4 pages).
- Link to source code / github or google drive or dropbox links to code.
- 5 slides presenting your work -- ideally a video (optional) of you walking me through your project in case I have trouble running it or understanding your report
  [Motivation] [Problem Setup] [Model] [Experiments] [Results] / you can also submit a link for this part.

- Due: April 22nd

Grading Criteria

- **Originality:** Either in the idea itself, the application itself, or the experiments you present in your final report. Are you teaching me something new that is not obvious? The more clear this answer is yes, the more likely you get full points on this part.

  Technical soundness: Are you describing how your solution and the components that you used in your solution with good amount of details and correct technical accuracy? You should provide enough details to understand just by reading your report what you did. If I have to look at your code to understand what you did then technical soundness will not receive as high a score. If you re-used a component e.g. CLIP or something else, but from reading your report it seems obvious you're not understanding what this model does. Then, this can also lead to points deducted.

  Results: Does your report present results in a way that is easy to understand -- e.g. example of input outputs of your model -- and does your project provide quantitative empirical and/or statistical evidence of your solution -- e.g. plots/figures/tables/etc. Ideally most projects should have both types of "results".

  Presentation: Is your project report of top quality, (e.g. as shown in the template), or did you include figures that are just screenshots of some experiment you run on a notebook, e.g. your plots do not have clear labeling for what is being shown in the x-axis or what are the units, your images look too low resolution. Anything of those issues will get you points deducted automatically. Your presentation in your project report has to be scientific manuscript quality.

Common Mistakes that lead to Point Deductions

- Including figures you did not make yourself but just borrowed from somewhere else.
- Not exporting your figures properly leading to blurry fonts when zooming in
- Not making your text large enough for it to be legible without zooming in a lot into the document
- Not actually including any results of your method and discussion of results
- Not including references to original papers for components of your solution that you are clearly relying on
- Including screenshots of your command line or jupyter/colab notebook as quick way of showing results in your report.

Also be careful about the following

- Clearly indicate what you did – it should be clear what was done by you and what was not done by you by reading your report.
- Any code you submit will be by default assumed that was authored by you and your team mates unless indicated otherwise. It is okay to rely on other people's code or existing code but the Instructor should have an easy time judging what is yours and what is not.

Today

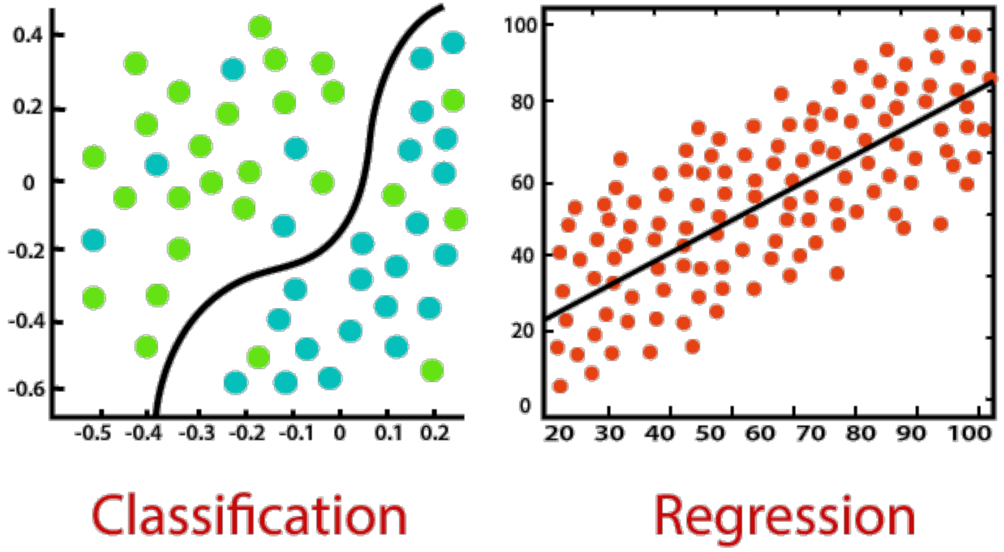
- Course Recap
- Future Directions

What you learned in this class?

- Machine Learning Basics (SGD, Losses, Evaluation)
- Computer Vision (CNNs, Detection, Segmentation, Vision Transform.)
- Natural Language Processing (RNNs, Transformers, GPTs)
- Vision and Language (RefExp, VQA, CLIP, cGANs, Diffusion)
- Self-supervised Representation Learning for Images, Text and Video
- Practical Implementation Aspects / Technical Skills

Machine Learning : Introduction

- Supervised Learning vs Unsupervised Learning (+Self-supervised)
- Classification vs Regression
- Least Squares Regression (Mean Squared Error MSE Loss – L1Loss)
- Simple Linear Classifiers e.g. Softmax Classifiers

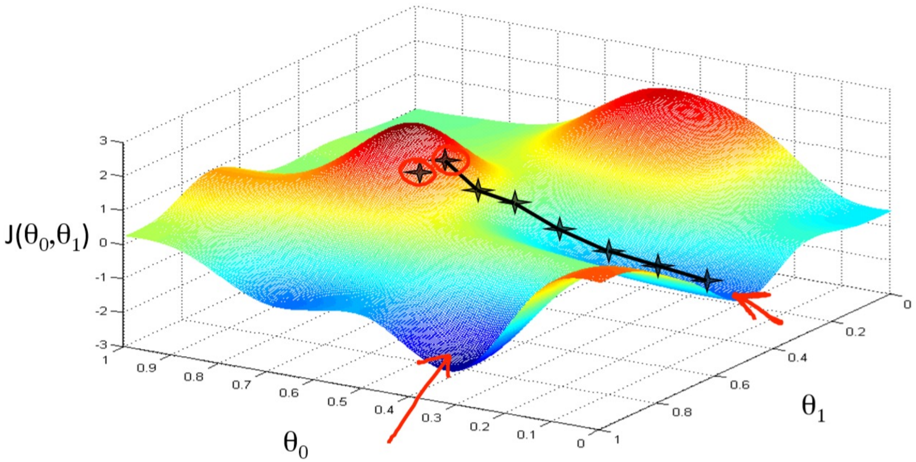


$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

https://static.javatpoint.com/tutorial/machine-learning/images/regression-vs-classification-in-machine-learning.png

Machine Learning: Optimization

- Gradient Descent (GD)
- (mini-batch) Stochastic Gradient Descent (SGD)
- Regularization, Momentum, Overfitting vs Underfitting
- Data Preprocessing and Data Augmentation
- Training / Validation / Testing

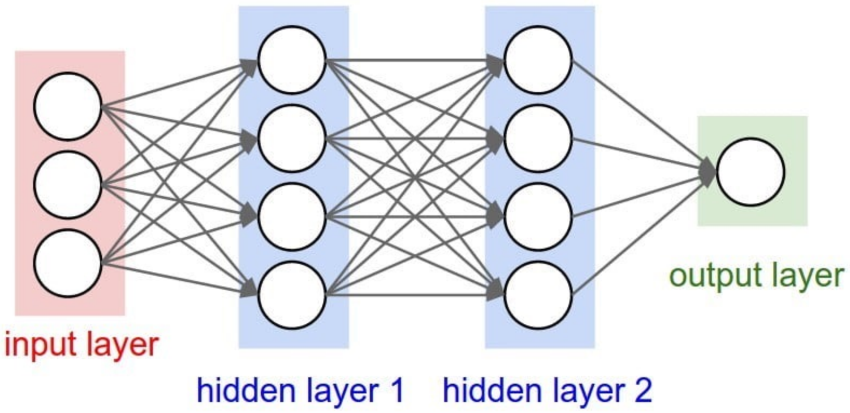


$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w_t}$$

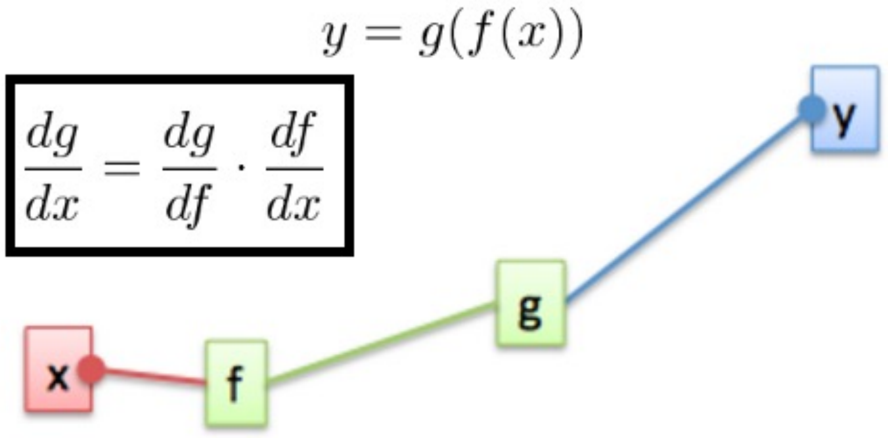
https://miro.medium.com/proxy/1*f9a162GhpMbiTVTAua_lLQ.png

Neural Networks: Backpropagation

- The Perceptron Model
- Multi-layer Perceptrons (Neural Networks of Linear Layers)
- Linear Layers and Non-linear Activations (ReLU, Sigmoid, Tanh)
- The backpropagation algorithm (Chain-rule) and SGD
- Pytorch’s automatic differentiation (loss.backward() and param.grad)



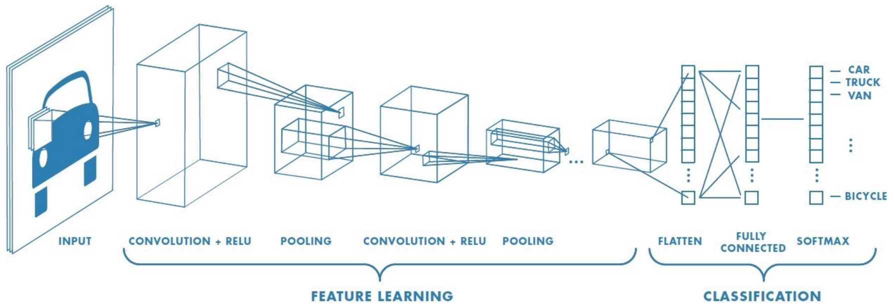
https://s7280.pcdn.co/wp-content/uploads/2020/07/Two-or-more-hidden-layers-comprise-a-Deep-Neural-Network.png



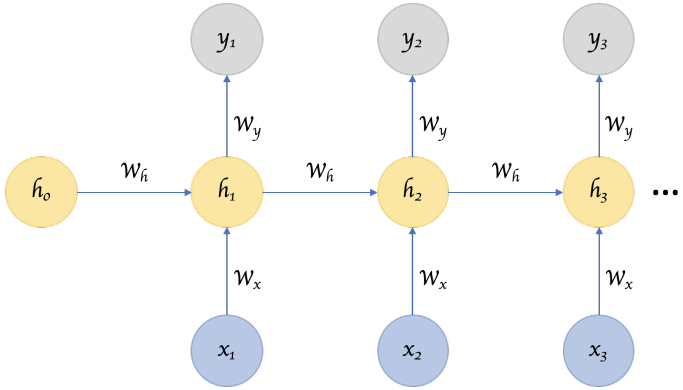
https://www.oreilly.com/library/view/hands-on-convolutional-neural/9781789130331/assets/85cb6c18-9cbc-44ff-9db6-c7d90d3c7cf0.png

Neural Networks: Models

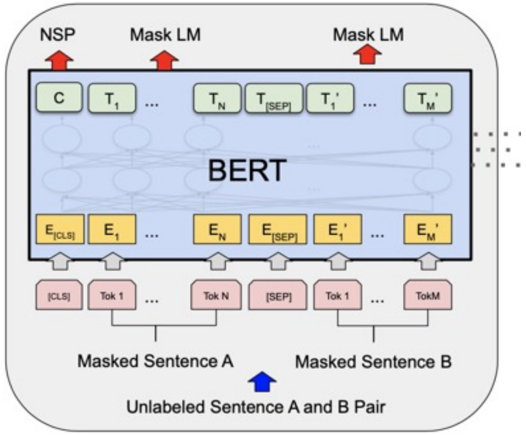
- Convolutional Neural Networks



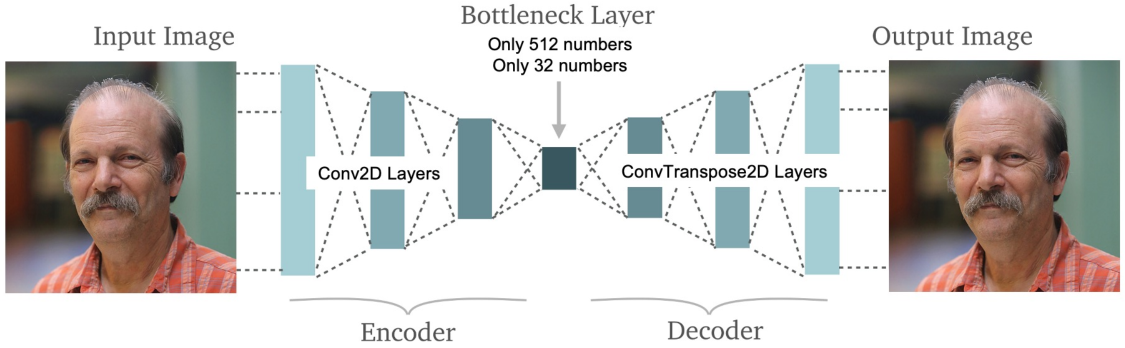
- Recurrent Neural Networks



- Transformer Networks



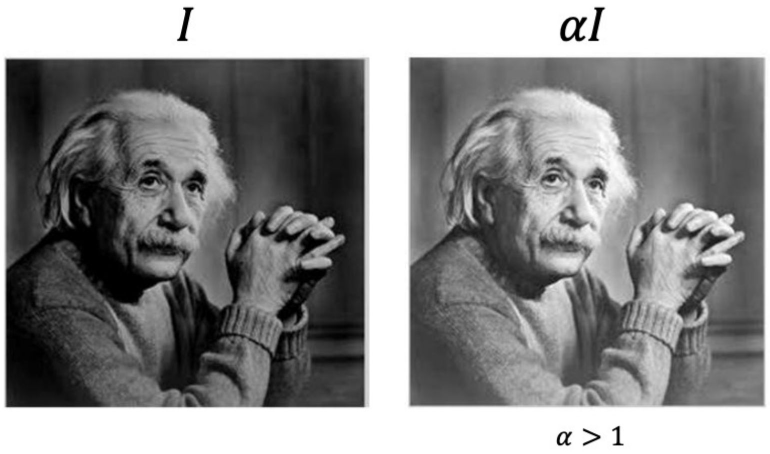
- Autoencoder Networks

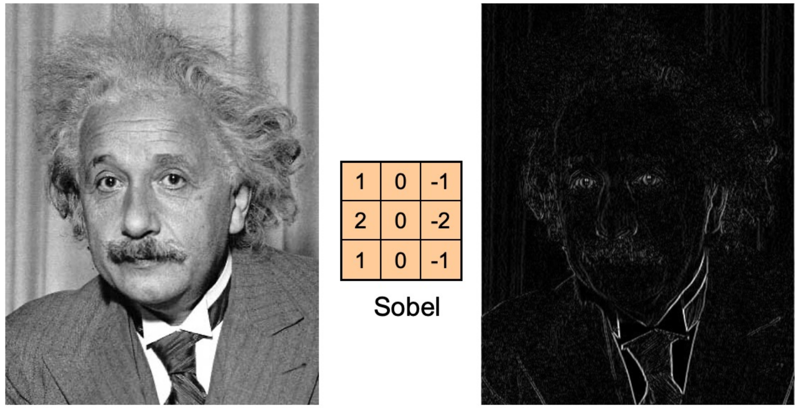


https://profiles.rice.edu/sites/g/files/bxs3881/files/2020-07/MosheVardi-500x500.jpg
https://www.edureka.co/blog/autoencoders-tutorial/

Intro to Computer Vision: Image Manipulations

- Image Processing and Manipulation
  (Brightness, Cropping, Normalizing, Resizing)
- Image Filtering and the Convolutional Operator
  (Box/Mean Filter, Gaussian Blur, Sobel Filtering)

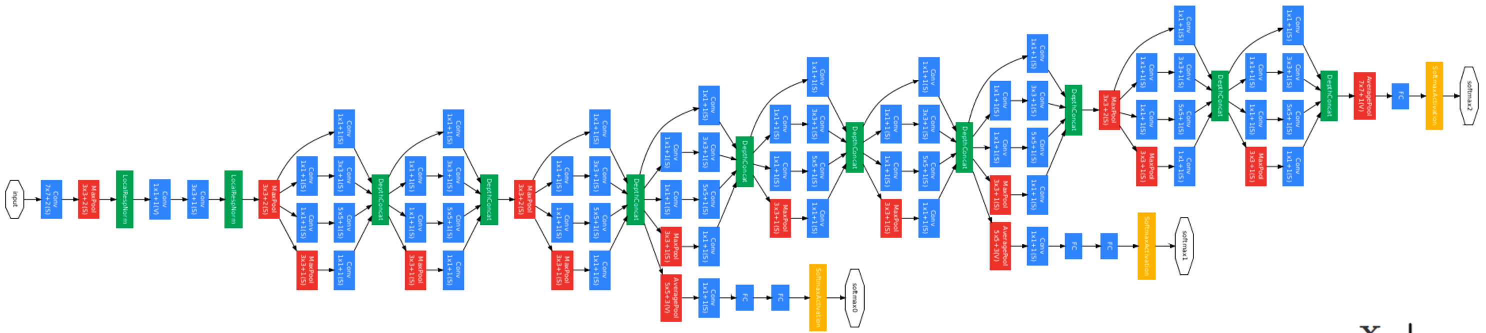


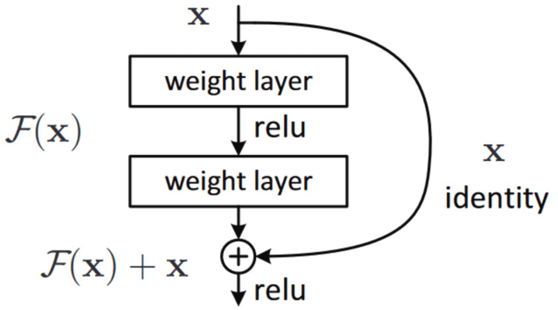


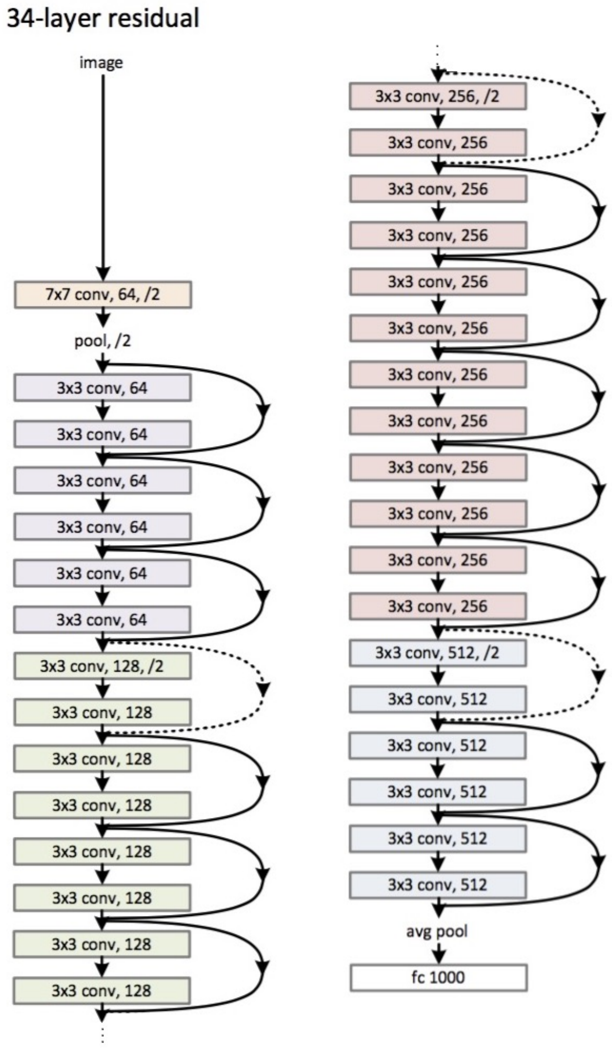
$$h[m,n] = \sum_{k,l} g[k,l]\, f[m+k, n+l]$$

Computer Vision: CNN Architectures

- Datasets: Imagenet (objects), SUN (scenes)
- Convolutional Neural Network Architectures for Images
  - AlexNet, VGGNet, GoogLeNet, ResNets, Densenet
- Layers: Dropout, Batch Normalization, Max Pooling





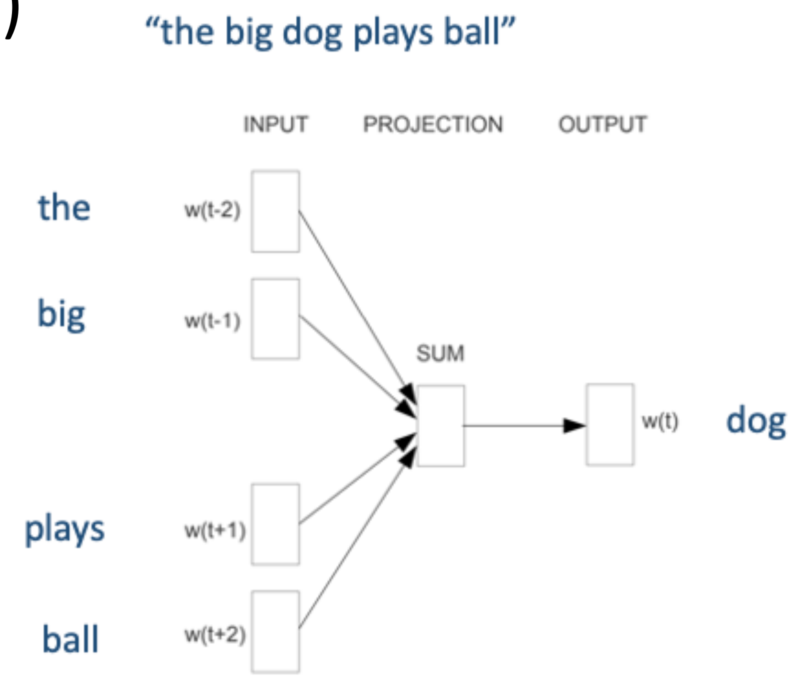


Intro to Natural Language Processing

- Representing text as Bag of Words
- Continuous Bag of Words (CBOW) -- i.e. Learned Word Embeddings
- Part-of-speech tagging, Text parsing, Entailment Resolution
- Tokenizers (including BytePairEncoding BPE)

bag-of-words representation

		dog	cat	person	holding	tree	computer	using			
person holding dog	{1, 3, 4}	[1	0	1	1	0	0	0	0	0	0]
person holding cat	{2, 3, 4}	[0	1	1	1	0	0	0	0	0	0]
person using computer	{3, 7, 6}	[0	0	1	0	0	1	1	0	0	0]

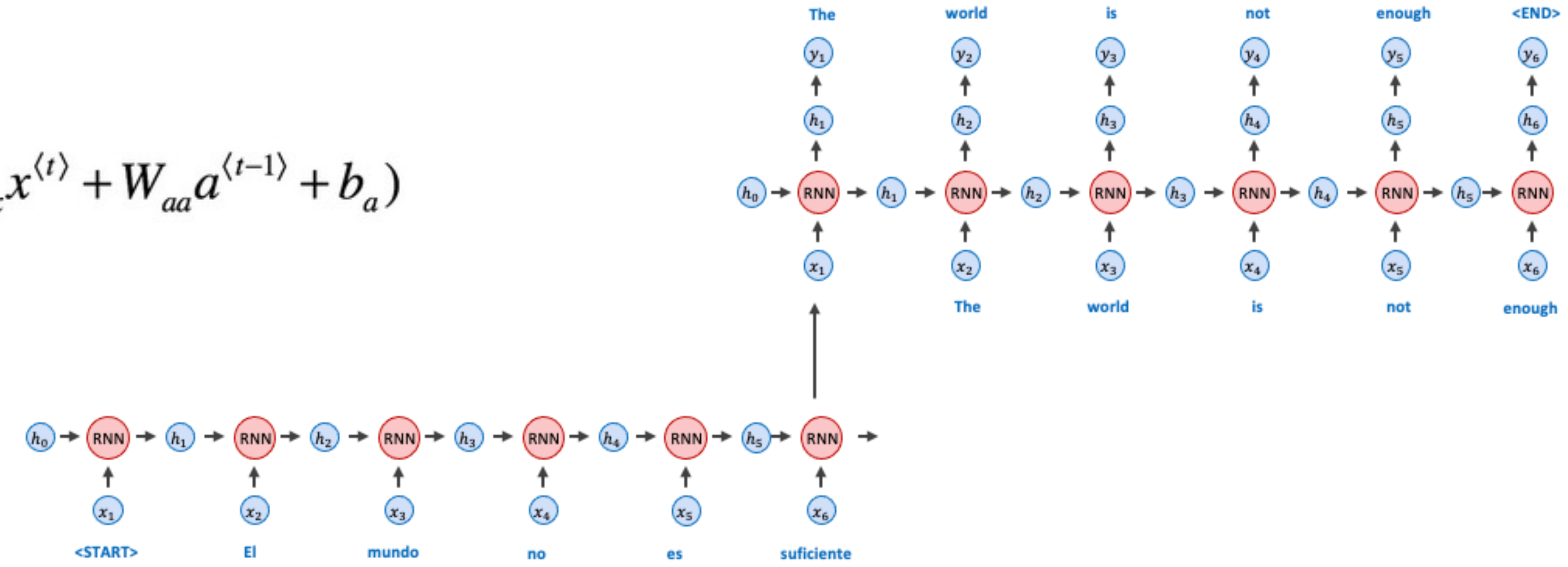


CBOW

Natural Language Processing: RNNs

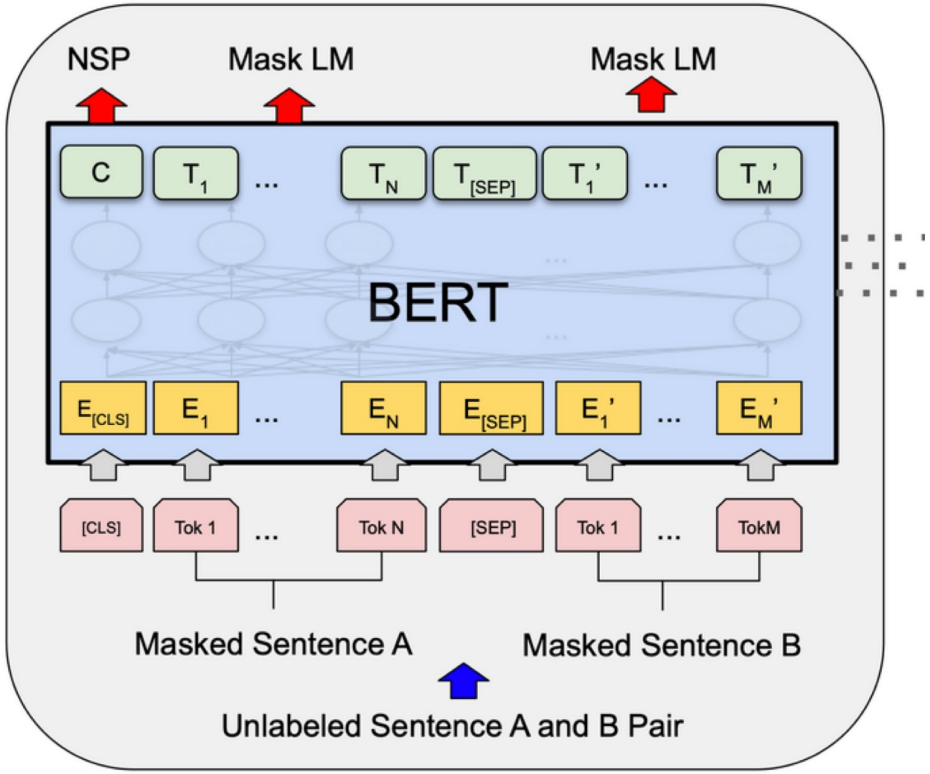
- Recurrent Neural Networks (RNNs)
  - Gated Recurrent Units (GRUs), Long-short Term Memory Networks (LSTMs)
  - Auto-regressive Models

$$a^{\langle t \rangle} = \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b_a)$$



Natural Language Processing: Transformers

- Transformer Models
  (Attention is all you need)
  - Single Head vs Multi-head Attention
  - Self-Attention and Masked Self-Attention
  - Positional Encodings
  - Masked Language Modeling (MLM)
  - The BERT Transformer Model
  - Other transformer models: GPT-2, GPT-3, T5
    BLOOM, OPT, LLAMA, BART



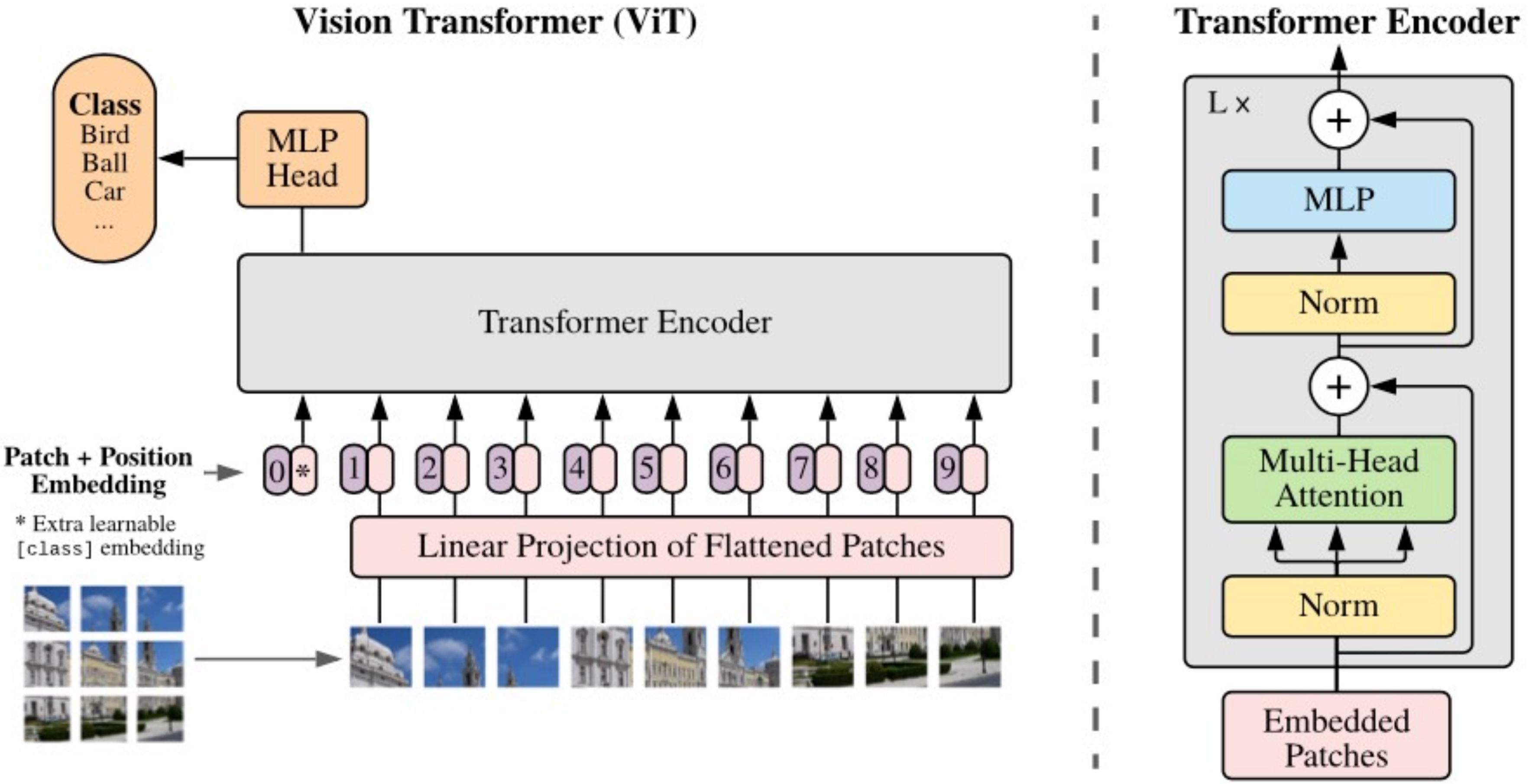
Pre-training

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$
$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

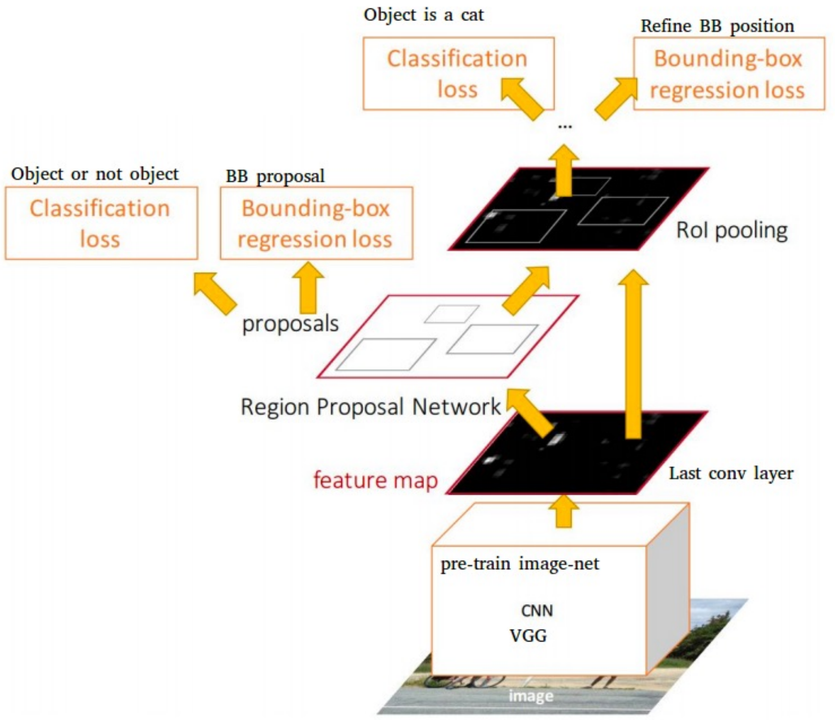
Computer Vision: Transformers

- Transformers for Images
  - The ViT Transformer

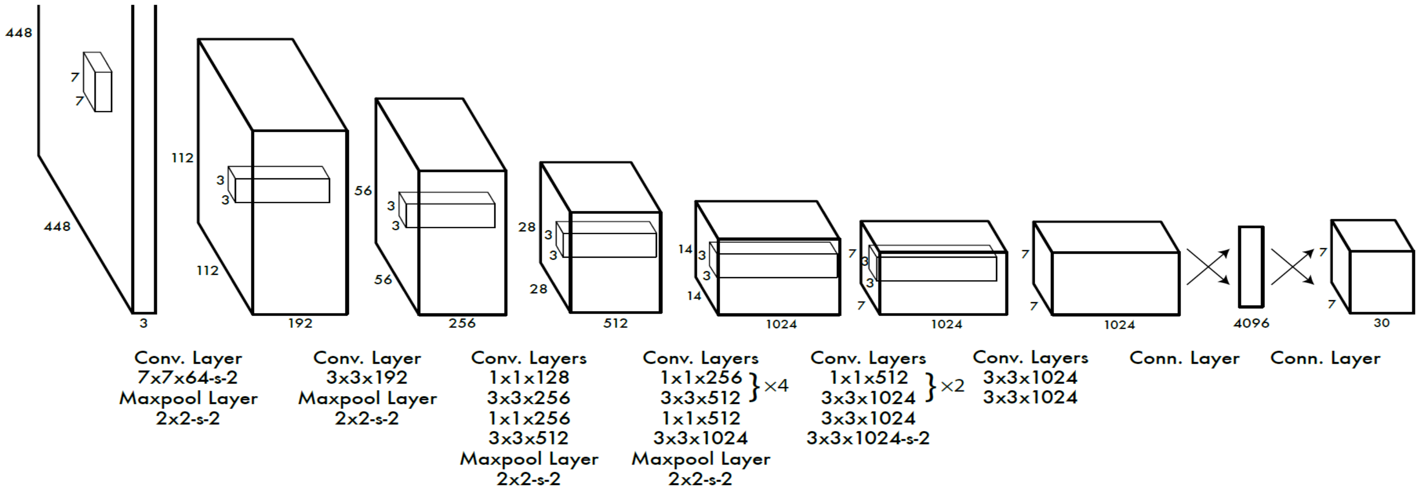


Computer Vision: Object Detection

- Convolutional Neural Networks for Object Detection
  - Two-Stage: RCNN, Fast-RCNN, Faster-RCNN
  - Single-Stage: You Only Look Once (YOLO), Single-Shot Detector (SSD)



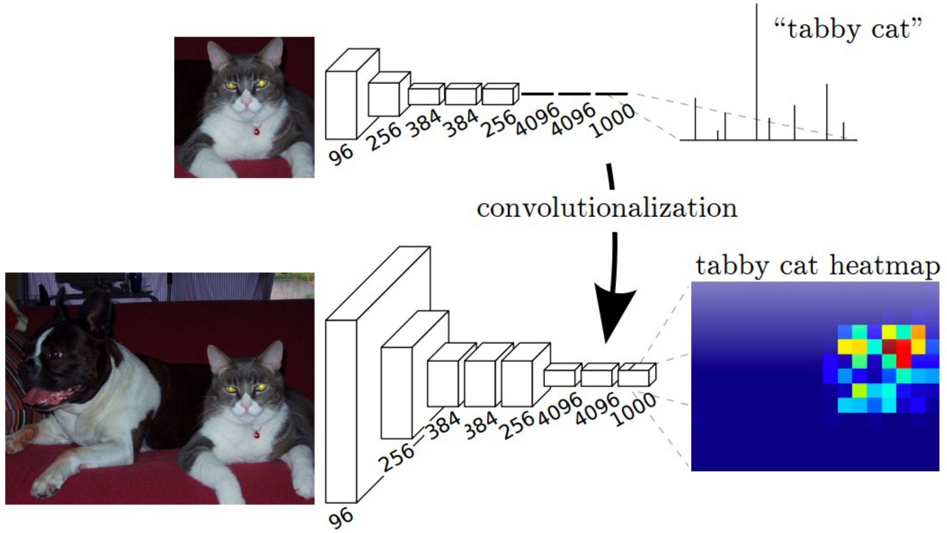
Faster-RCNN: Two-stage detector



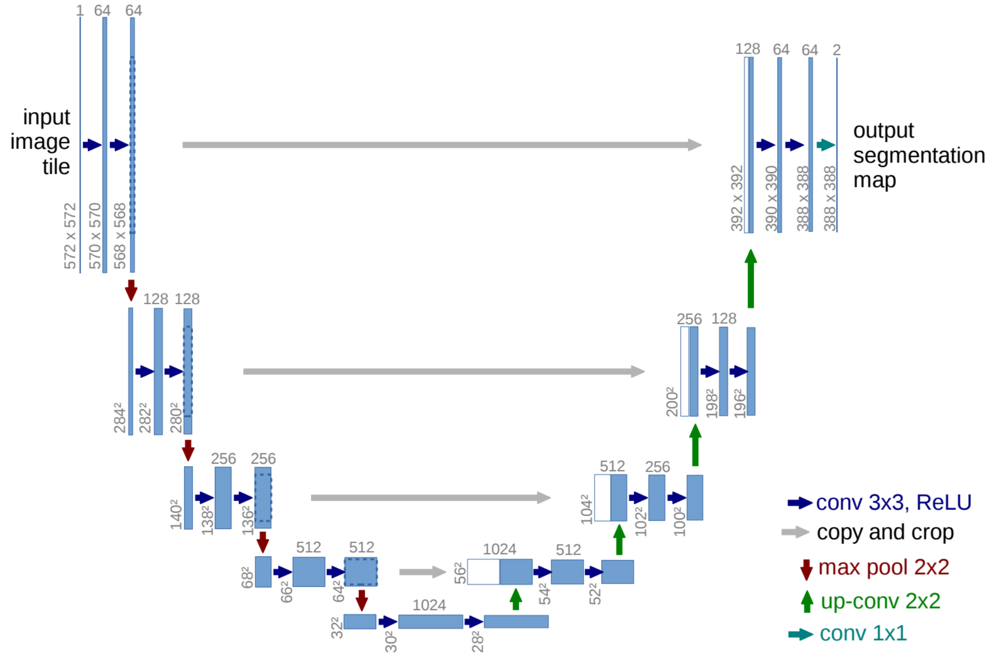
YOLO: Single-stage detector

Computer Vision: Segmentation

- Convolutional Neural Networks for Segmentation
  - Upsampling Convolutions, and Dilated Convolutions
  - **U-Nets**, Fully Convolutional Nets, and Mask-RCNN



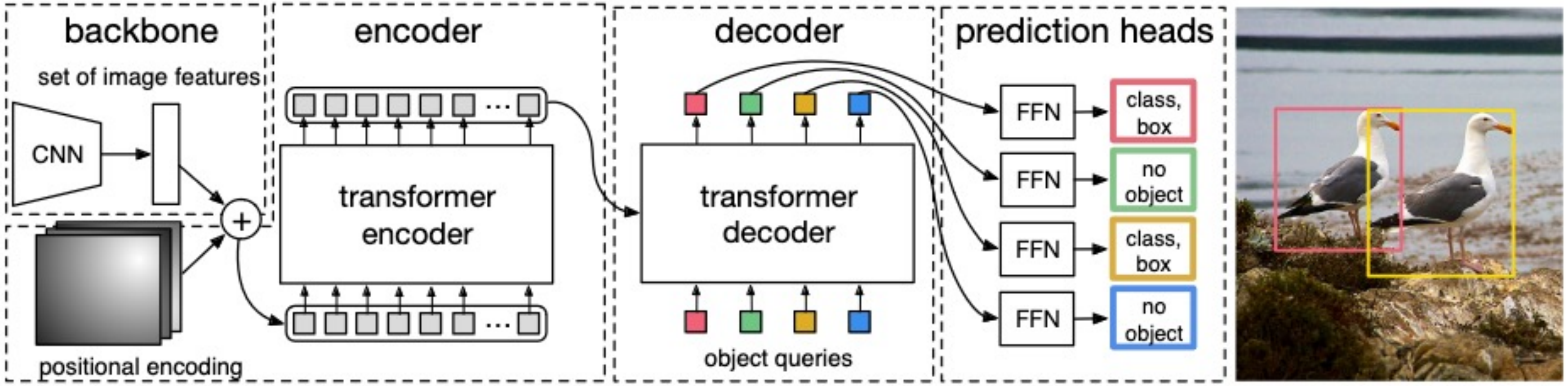
Fully Convolutional Networks



U-Net: Upsampling convolutions and skip connections

Computer Vision: Object Detection with Transformers

- Vision Transformers for Object Detection (DETR)
  - Hungarian Loss through Bipartite Matching

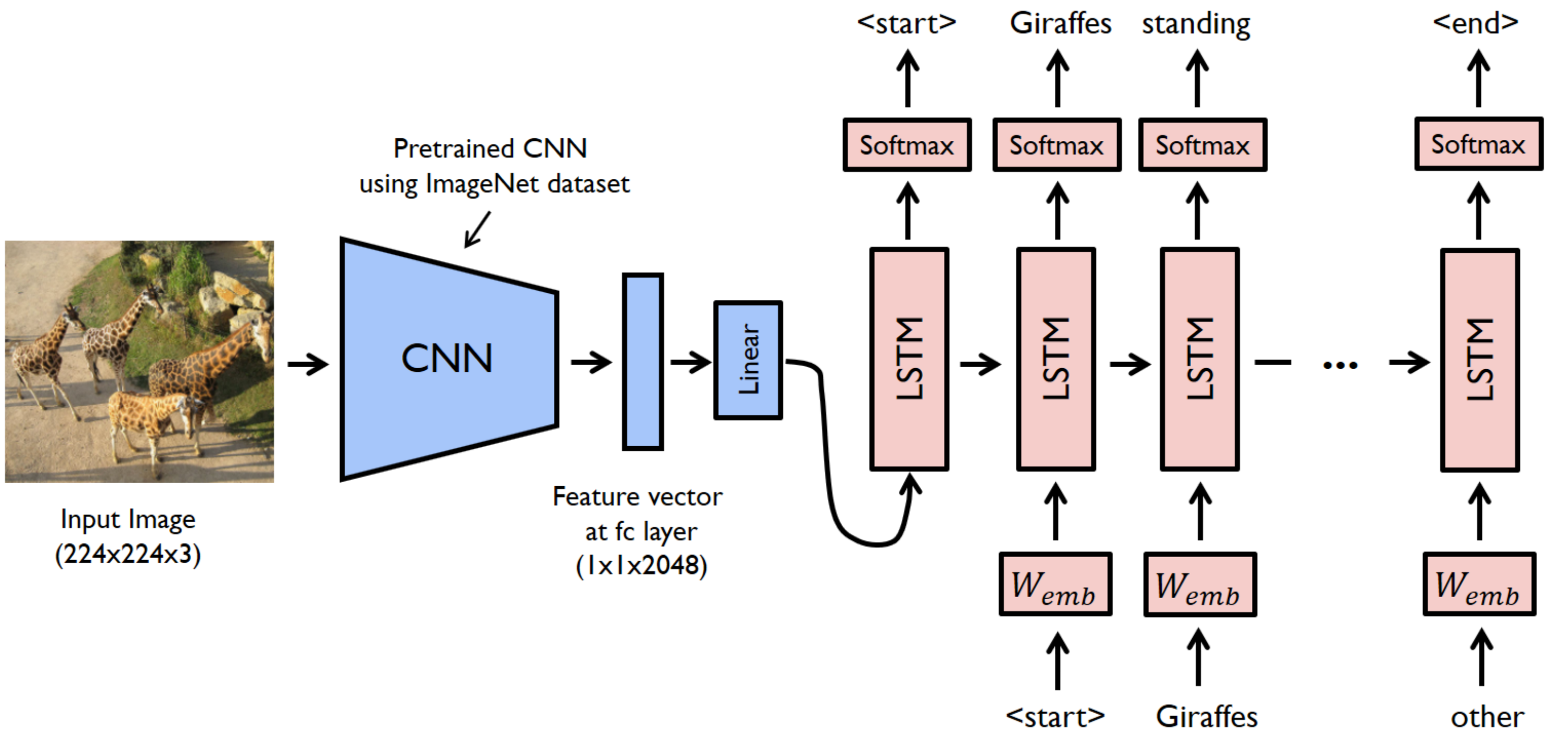


$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

Vision and Language

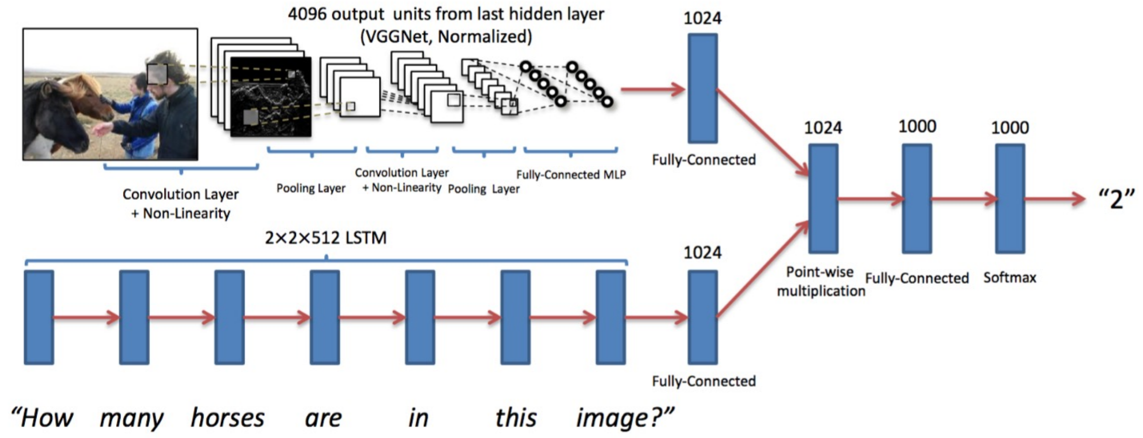
- Image Captioning (CNNs + RNNs): Autoregressive + Greedy decoding



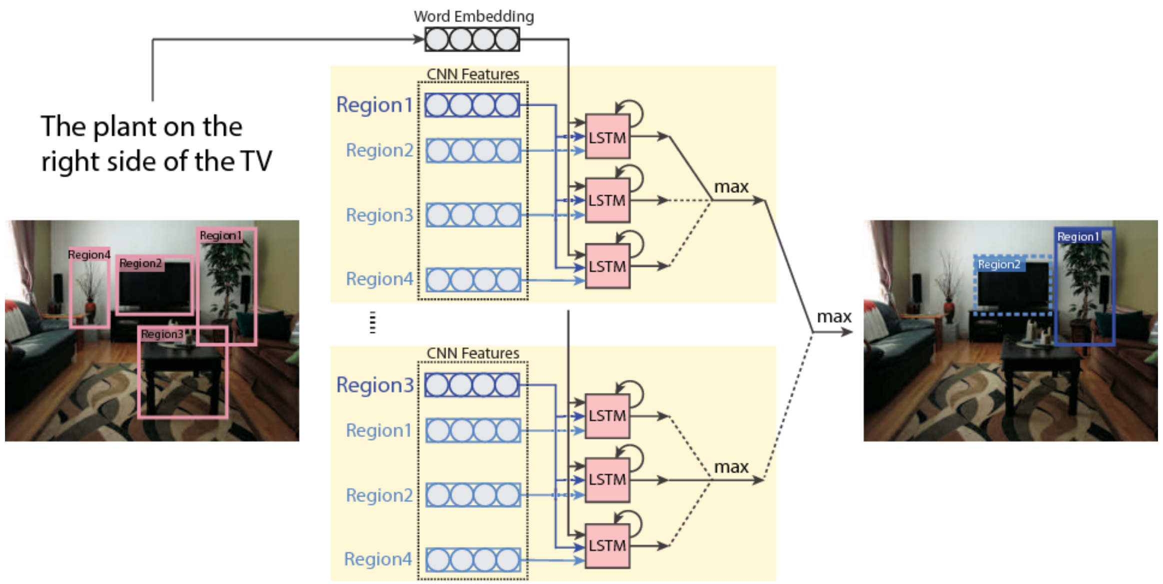
https://raw.githubusercontent.com/yunjey/pytorch-tutorial/master/tutorials/03-advanced/image_captioning/png/model.png

Vision and Language: VQA, RefExps

- Visual Question Answering (CNNs + RNNs + MLPs)
- Referring Expression Generation (Faster-RCNN + RNNs)
- Referring Expression Comprehension (Faster-RCNN + RNNs + MLPs)



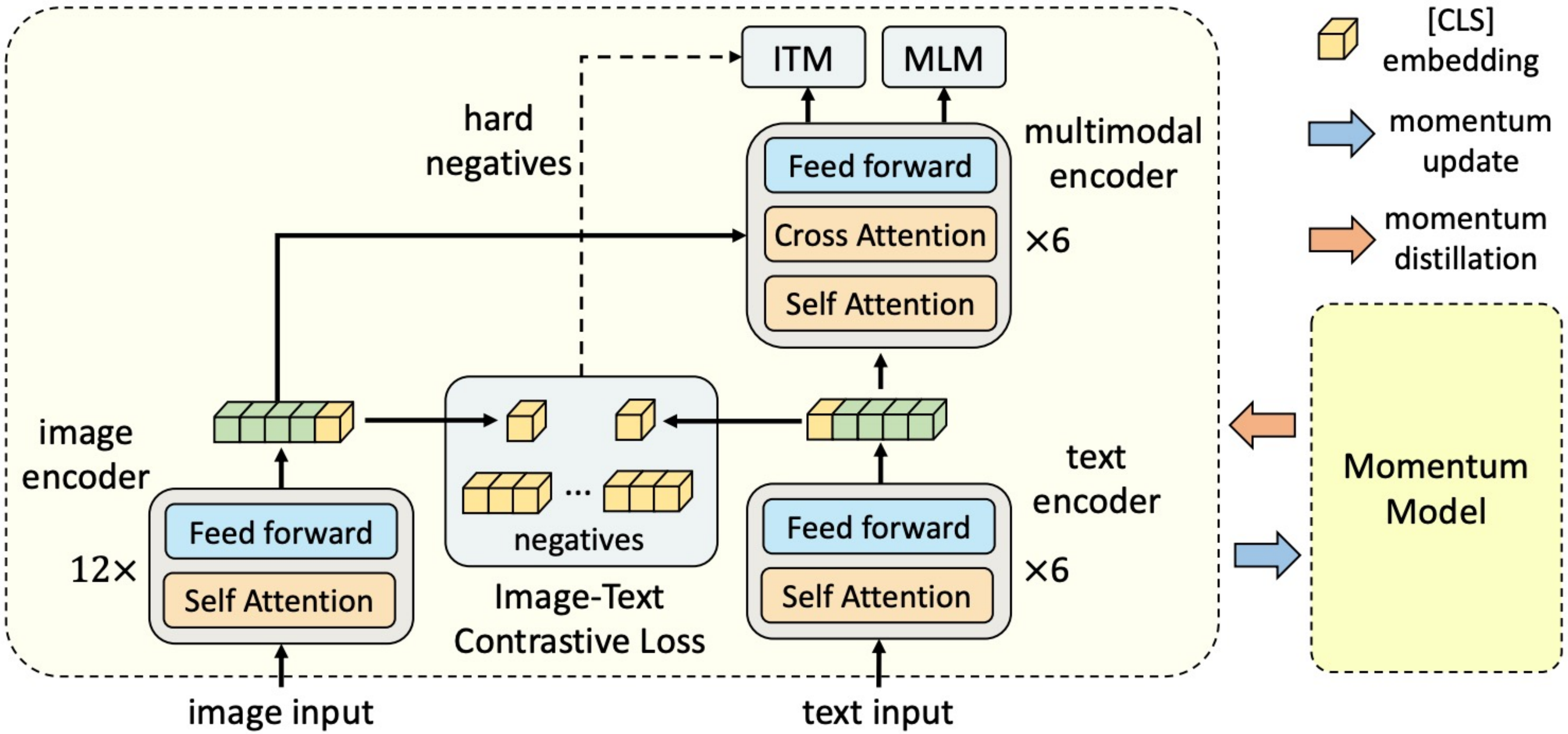
https://miro.medium.com/max/1400/1*QbWaFSNaO3GTgjQZOxhdDg.png



https://d3i71xaburhd42.cloudfront.net/f25b9aed37614aae007fc876f31eed0595ab9cd0/5-Figure2-1.png

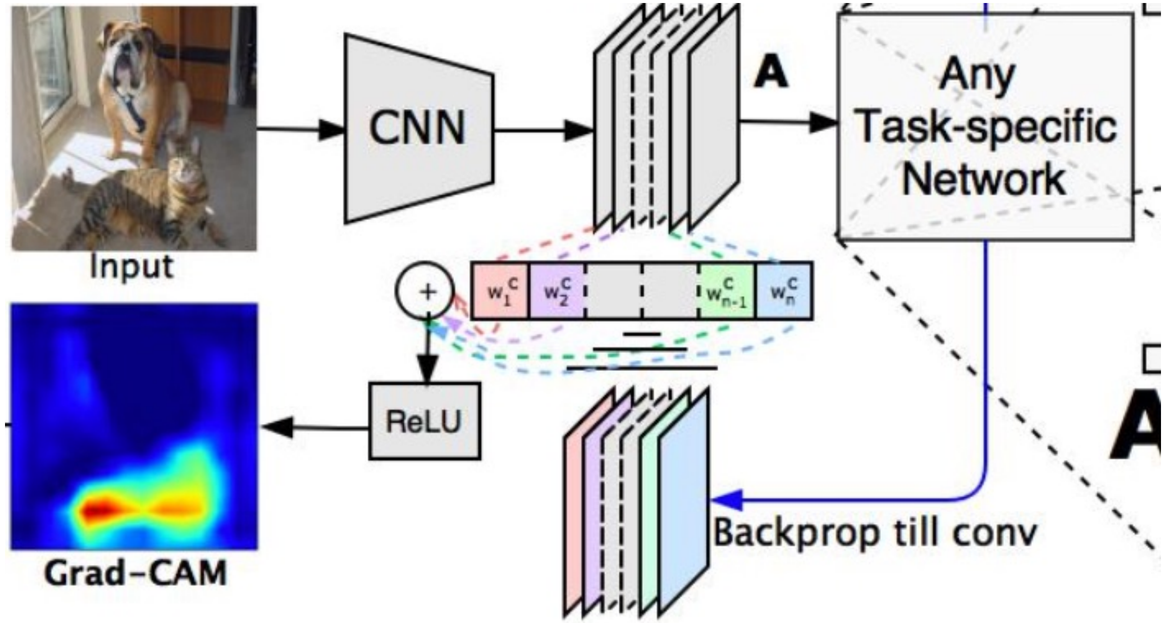
Vision and Language: Transformers

- Vision-Language Transformers (e.g. ALBEF)



Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

Vision and Language: Explanations

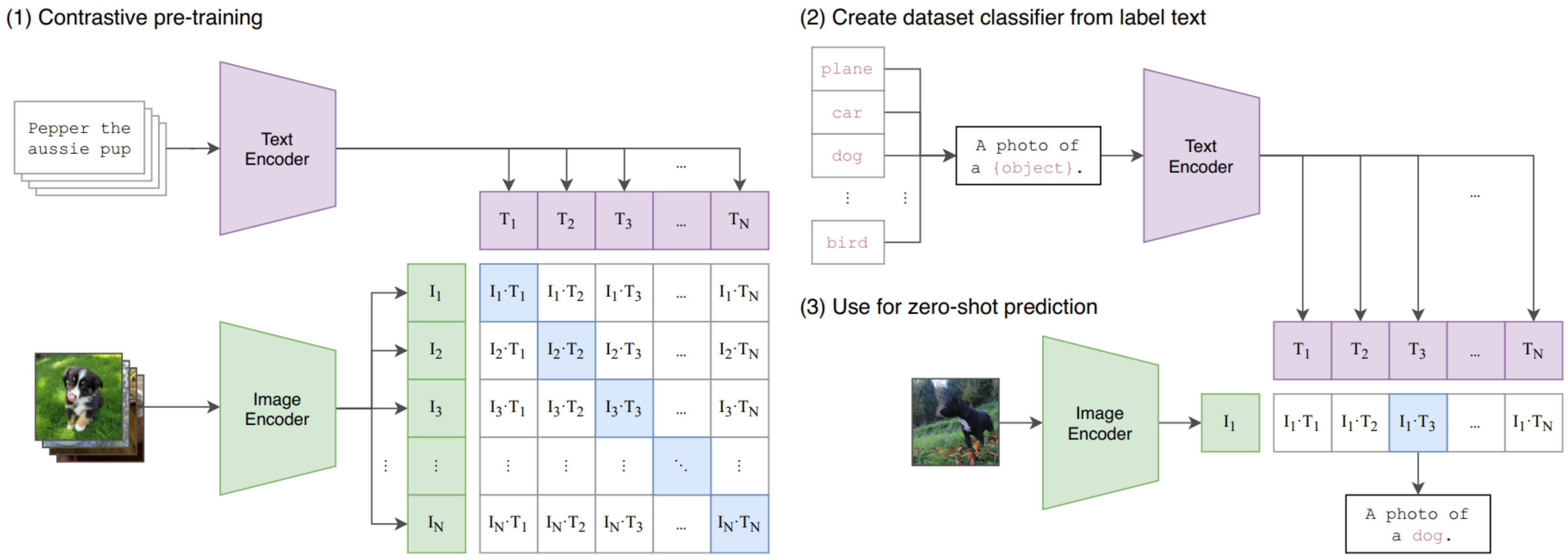


$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

Vision and Language: Contrastive Learning

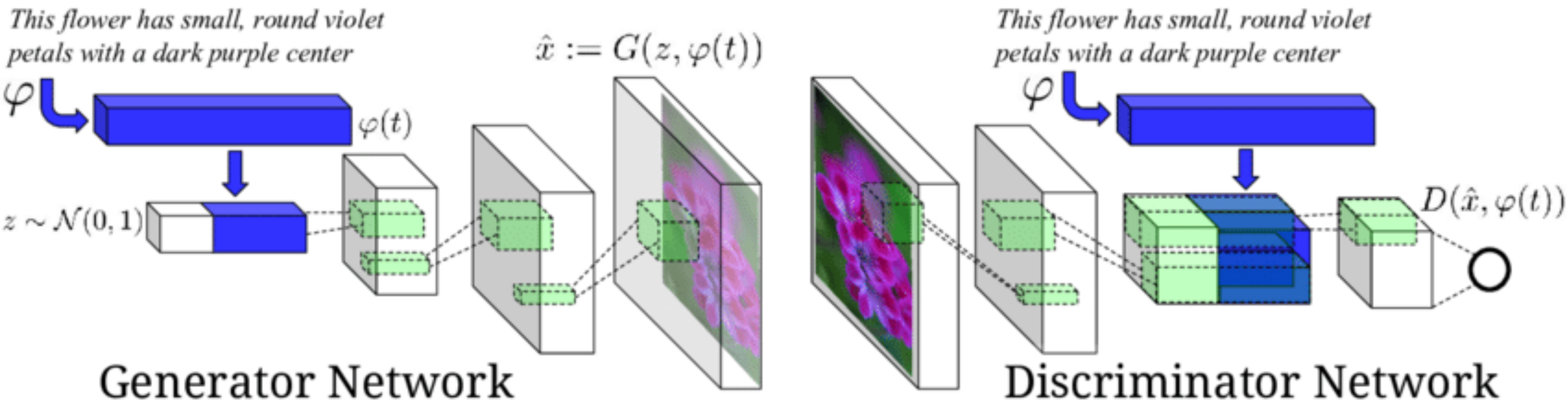
- Vision-Language Contrastive Learning (CLIP)
  - Zero-shot visual recognition through CLIP visual prompt engineering



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

Vision and Language: Text-to-Image

- Conditional GANs (Text-to-image synthesis)
- AutoEncoders + Transformers (DALL-E and DALL-E mini)



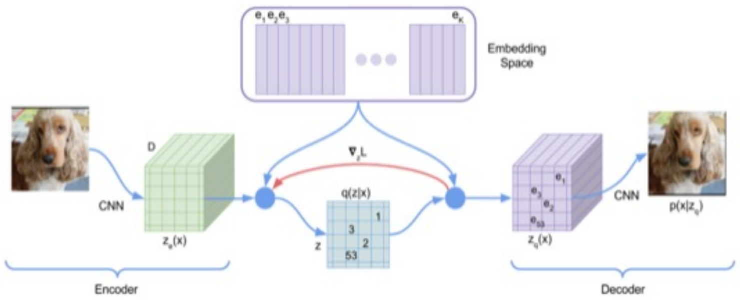
Reed, Scott & Akata, Zeynep & Yan, Xinchen & Logeswaran, Lajanugen & Schiele, Bernt & Lee, Honglak. (2016). Generative Adversarial Text to Image Synthesis.

Vision and Language: Text-to-Image

- Conditional GANs (Text-to-image synthesis)
- AutoEncoders + Transformers (DALL-E and DALL-E mini)

Step 1:

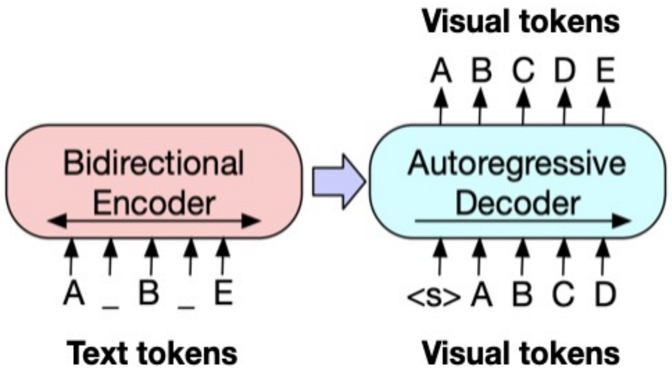
Learn Discrete Dictionary of Visual Tokens



VQVAE — Oord, Vinyals, Kavukcuoglu, 2017
VQGAN — Esser, Rombach, Ommer, 2021
dVAE - DALL-E — Ramesh et al 2021

Step 2:

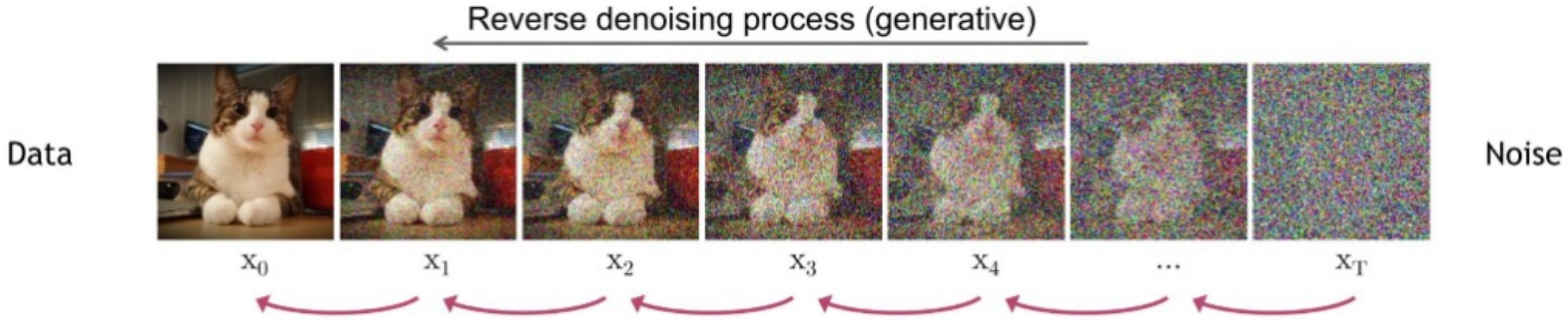
Build a scene as a composition of discrete visual tokens

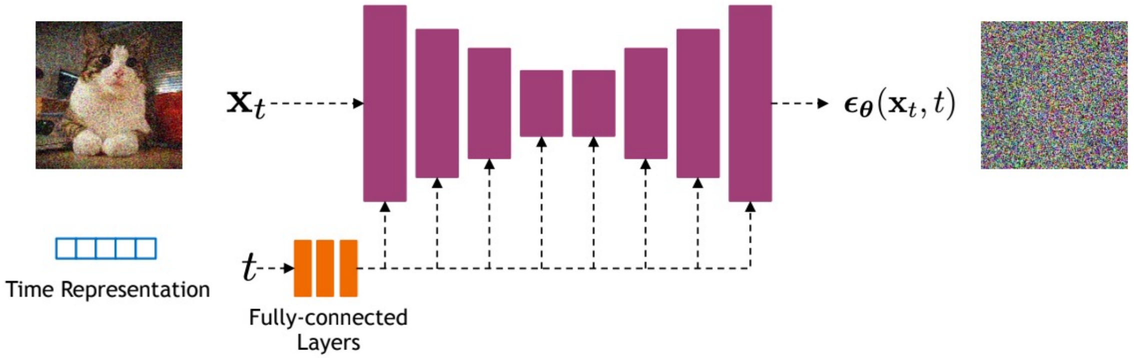


BART, GPT-3, etc

Vision and Language: Text to Image

- Reverse Diffusion Models (e.g. DALLE-2, StableDiffusion, Imagen)

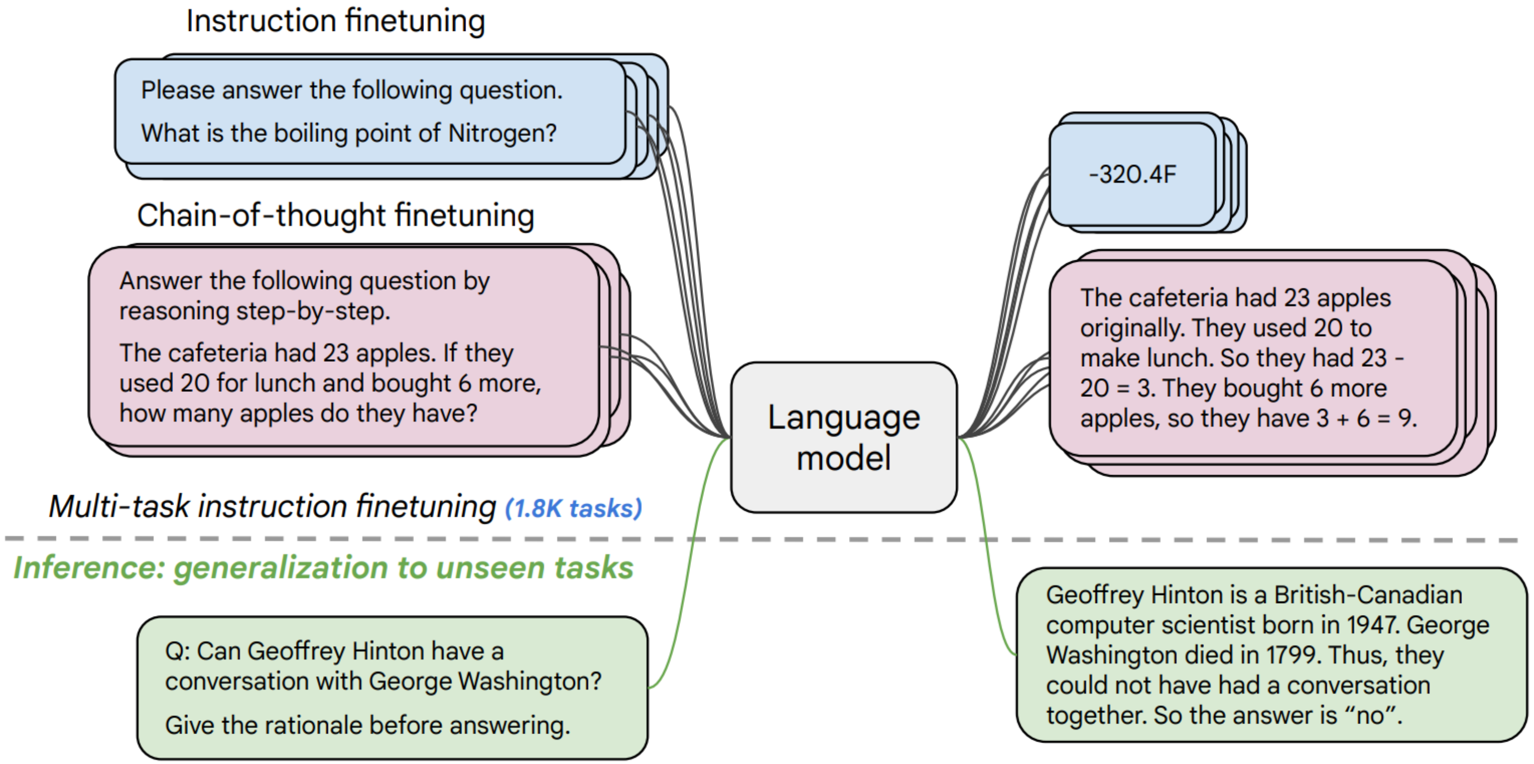




$$\mathbb{E}_{t,\epsilon}[||\epsilon_\theta(\mathbf{x}_t, c) - \epsilon||^2$$

Natural Language Processing: Instruction Following LLMs and Chatbot LLMs

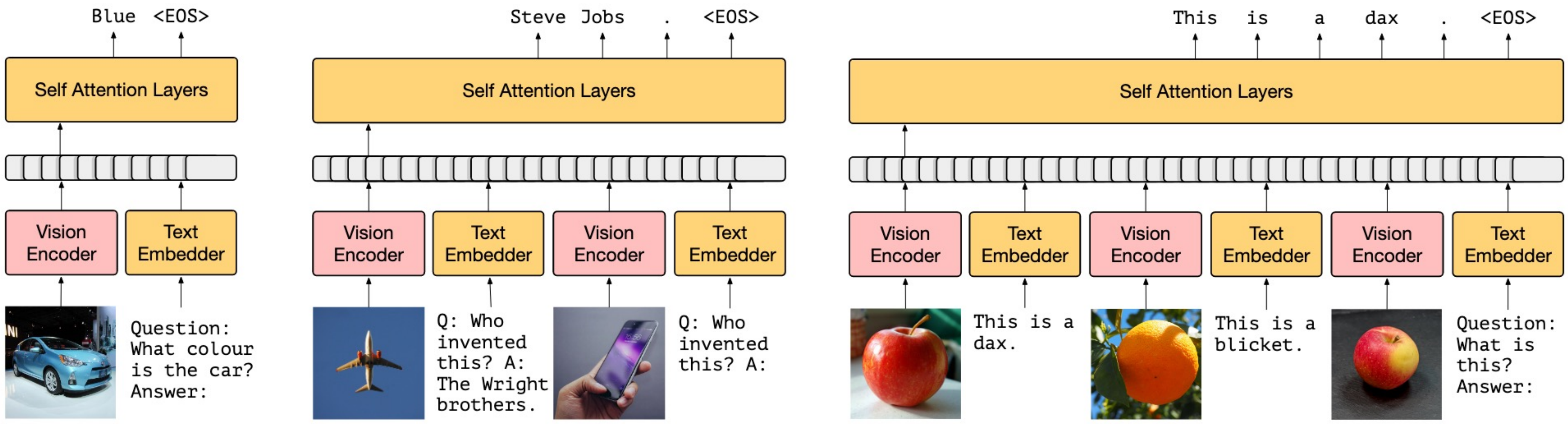
- Finetuned on Instructions: FLAN-T5, OPT-IML
- Tuned with Reinforcement Learning with Human Feedback: InstructGPT, ChatGPT



https://arxiv.org/pdf/2210.11416.pdf

Vision and Language: Advanced Multimodal Models

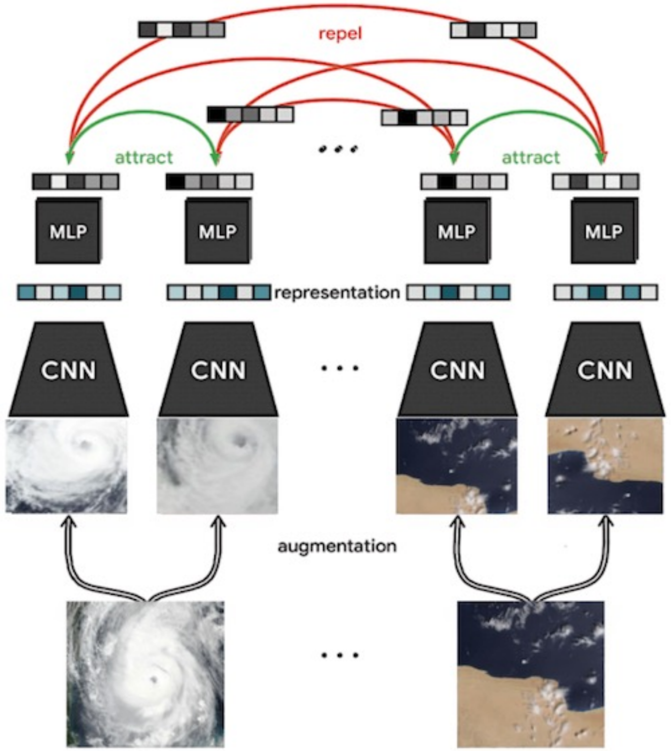
- Tuned LLMs with Image data:
  Frozen, Flamingo, GPT-4V, among others…

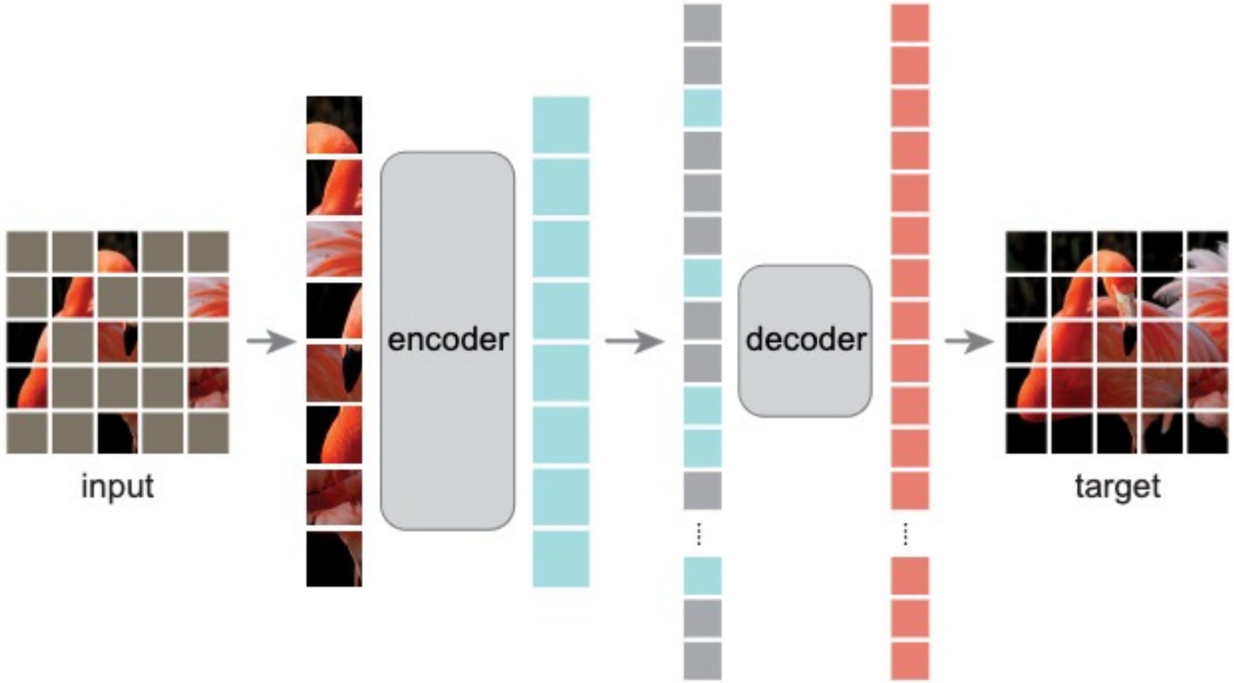


(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Computer Vision: Self-supervision

- Basic Pretext tasks: Colorization, context prediction, counting
- Contrastive Learning through Augmented Views: e.g. SimCLR
- Masked AutoEncoders (MAEs)





Practical Aspects

- Python + Pytorch + Automatic Differentiation + GPU
- Liveloss / Weights and Biases: For experiment Monitoring
- Matplotlib, LiveLossPlot, Torchvision, PIL (Python Imaging Library)
- Huggingface Transformers/Tokenizers Library

Practical Aspects

- Interactive Coding Tools
  - Google Colab Notebooks and Amazon SageMaker Lab
  - Powered by Project Jupyter

Practical Aspects

- Containers: Docker and Singularity
- Batch processing: SLURM and the Rice NOTS Cluster
- + Whatever you ended up needing for your course project

Practical Aspects: What else I recommend?

- **Pytorch's advanced features:**
  - Distributed training across multiple GPUs, and across multiple nodes with multiple GPUs. Torch.

```
from torch.nn.parallel import DistributedDataParallel as DDP
```

- **Cloud:** AWS, Google Cloud, Microsoft Azure, Oracle Cloud
  - On Demand vs Spot Instances
  - Submitting batch processing jobs through containers
  - Weights & Biases, Tensorboard, Comet
- **Other frameworks:**
  - Tensorflow, Apache MXNet, **JAX** (low level support for some optimized operations)
  - ONNX: Cross-platform compatible model checkpoint file.

Practical Aspects: User Interfaces/Demos Recommended

- Flask (or Django): For dynamic python-based server-side deployments
  - Jinja2 for templating your output HTML (if needed)
- JQuery, Bootstrap, React, ReactNative, Vue.js: For App Development

Things we didn't cover in the class

- Vision and Language Navigation
  - (e.g. see https://arxiv.org/abs/1711.07280)
- Visual Commonsense Reasoning
  - (e.g. see https://visualcommonsense.com/)
- Other topics:
  - Reinforcement Learning (Prof Vaibhav Unhelkar)
  - Graph Neural Networks and Graph ML (Prof. Arlei Silva)
  - Information Retrieval (Prof Xia Ben Hu)
  - Neural Radiance Fields (NERFs) (Prof Guha Balakrishnan)
  - 3D Computer Vision and Imaging (Prof. Ashok Veeraraghavan)
  - Robotics (Kaiyu Hang and Lydia Kavraki)

Rice University - Resources

lilie | liu idea lab for innovation & entrepreneurship

https://entrepreneurship.rice.edu/

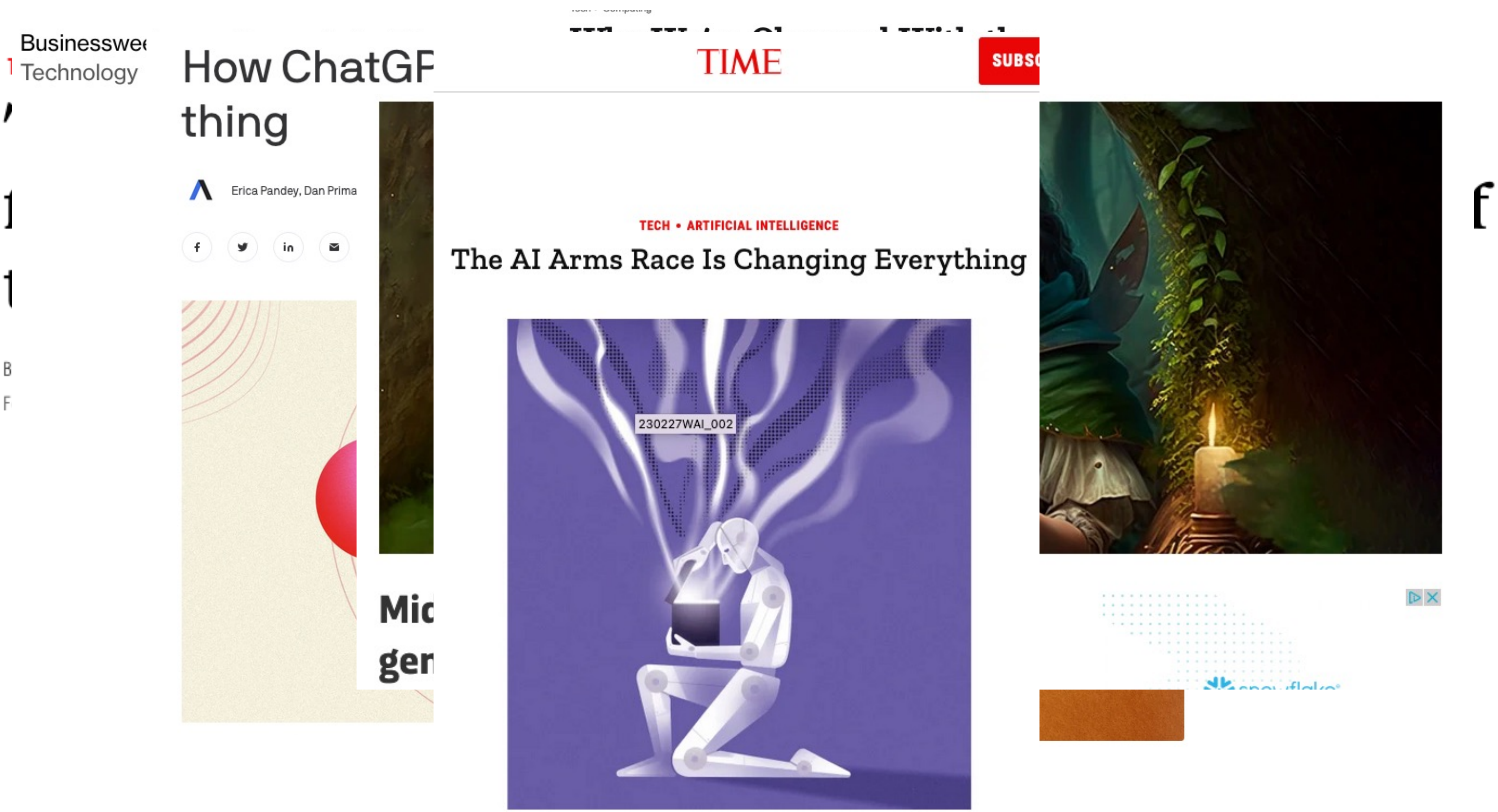
RICE VENTURES

Bolstering student entrepreneurs at Rice University

We are Rice University's student-led startup accelerator and entrepreneurship organization.

https://riceventures.org/

Other things to be aware about...



More recently

Home / Innovation / Artificial Intelligence

GPT-4 Turbo reclaims the 'best AI model' crown from Anthropic's Claude 3

OpenAI's latest update tops all 82 LLMs in the Chatbot Arena. Here's how to compare them for yourself.

Written by **Sabrina Ortiz,** Editor

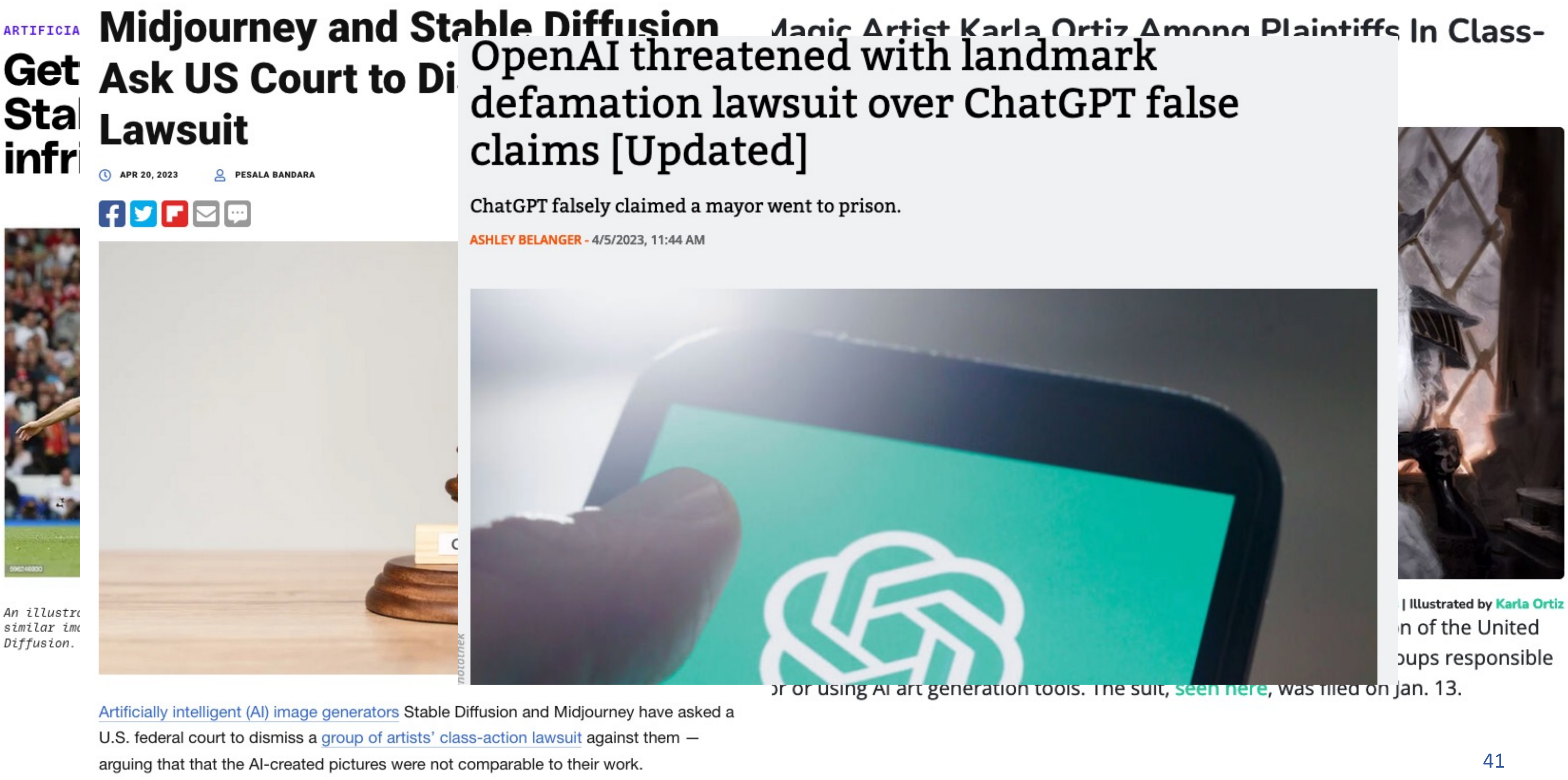
April 15, 2024 at 10:49 a.m. PT

Large Language Model

Introducing Meta Llama 3: The most capable openly available LLM to date

April 18, 2024

Other things to be aware about…



ARTIFICIA

Get Sta infri

An illustr similar im Diffusion.

Midjourney and Stable Diffusion Ask US Court to Di… Lawsuit

APR 20, 2023 PESALA BANDARA

Artificially intelligent (AI) image generators Stable Diffusion and Midjourney have asked a U.S. federal court to dismiss a group of artists' class-action lawsuit against them — arguing that that the AI-created pictures were not comparable to their work.

OpenAI threatened with landmark defamation lawsuit over ChatGPT false claims [Updated]

ChatGPT falsely claimed a mayor went to prison.

ASHLEY BELANGER - 4/5/2023, 11:44 AM

…agic Artist Karla Ortiz Among Plaintiffs In Class-

| Illustrated by Karla Ortiz

…n of the United …oups responsible …or using AI art generation tools. The suit, seen here, was filed on Jan. 13.

More recently

ARTIFICIAL INTELLIGENCE / TE

Getty laws
trial in the

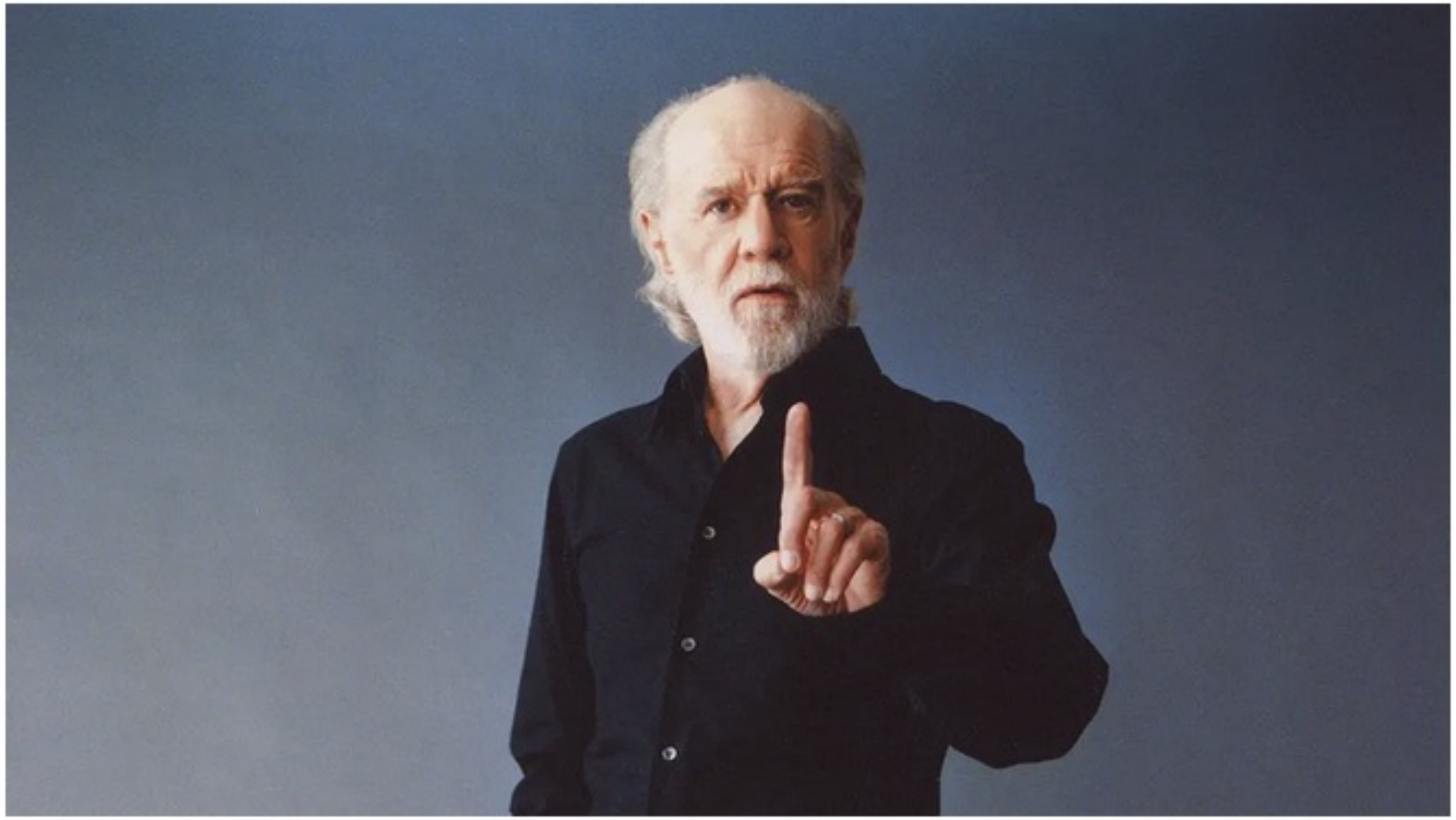
The Verge

HOME > BIZ > NEWS

Apr 2, 2024 5:11pm PT

George Carlin Estate Settles Lawsuit Over AI-Generated Comedy Special

By Gene Maddaus



HBO

Other things to be aware about…


Green Intelligence: Why Data And AI

NATURE AND ENVIRONMENT | GLOBAL ISSUES

How AI can help the environment

Natalie Muller | Neil King

04/19/2023

ChatGPT has created a buzz around artificial intelligence. From cleaning up polluting industries to disrupting deforestation, here are six AI innovations that can help the planet.

Other things to be aware about…

ALJAZEERA News Ukraine war

OPINION

Opinions | Labour Rights

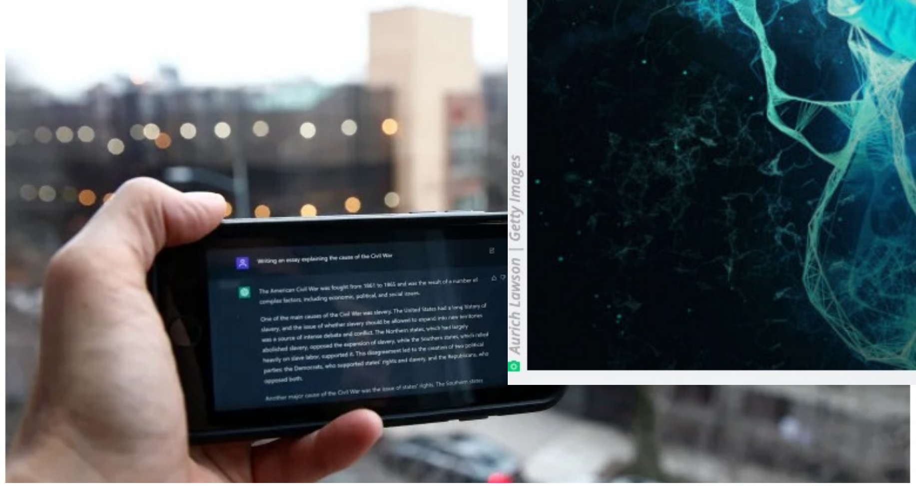
ChatGPT and the sweats
the digital age

The latest ChatGPT revelations are yet a
pervasive labour exploitation in digital

Nanjala Nyabola

Nanjala Nyabola is a political analyst and the author of "Di

23 Jan 2023



ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STOR

NOT SO FAST —

The mounting human and environmental costs of generative AI

Op-ed: Planetary impacts, escalating financial costs, and labor exploitation all factor.

SASHA LUCCIONI - 4/12/2023, 6:00 AM


Aurich Lawson | Getty Images

SUBSCRIBE

Published by the Berggruen Institute


Weerasekera for Noema Magazine

THE HUMAN

BY ADRIENNE WILLIAMS, MILAGROS MICELI AND TIMNIT GEBRU

OCTOBER 13, 2022

n order

Image g

BY BILLY PERRIGO

JANUARY 18, 2023 7:00 AM EST

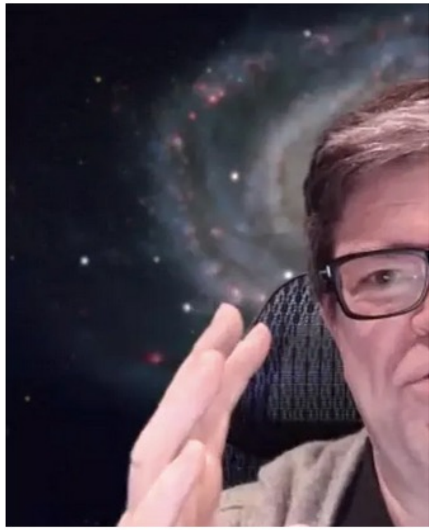
Other things to be aware about...

≡ WIRED BACKCHANNEL BUSINESS CULTURE MORE ⌄ SIGN IN SUBSCRIBE

ChatGPT is 'not parti
'nothing revolutionai
scientist

The public perceives OpenAI's ChatGPT being used and the same kind of work is learning pioneer.

Written by **Tiernan Ray,** Contributing W



Why hasn't the public seen programs like ChatGPT from a lot to lose by putting out systems that make stuff up," s

Collective[i] Forecast

MOTHERBOARD
TECH BY VICE

Everybody Please Calm Down About ChatGPT

The panic and hype around the surprisingly dumb chatbot is stopping us from talking about real issues with AI.

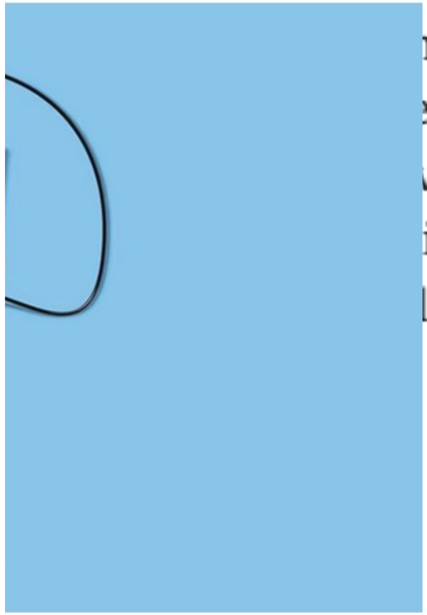
By Edward Ongweso Jr

December 16, 2022, 8:13am Share Tweet Snap

Listen to this article

Than You

ersity of



ned on
esent all
ve been well
ition, and
lóñez-Roman,

IARE ⌄ SAVED STORIES ↗ SAVE

Thanks Everyone

- Finish your course projects – I provided feedback to all progress reports and if I have not, let me know by email. At least in Canvas all my tickets are closed so far on progress report.

- Keep in touch – especially if you go on to do something great using computer vision / vision-language – always happy to hear back from students