行政院 專題演講

迎接生成式AI台灣需要的準備

(version r3)

August 21, 2023

孔祥重 (H. T. Kung)
Harvard University

Background

- Today, Taiwan is **widely recognized** by the world for its democracy, high-tech industry, and healthcare systems
- Sound **government policies** and **effective execution** were largely responsible for these successes
  - Specifically, Taiwan was able to capture the opportunity of the **PC revolution in the 1980s**. Since then, Taiwan has increased its wealth rapidly through the automation of industries, ODM/OEM in ICT (Information and Communication Technology) sectors, and, in recent years, the world-leading semiconductor industry
- **AI technology** is expected to be even **more transformative** than PC technology. For example, the human-like skills exhibited by **generative AI** in November 2022 have shocked the world

A Vision

- Taiwan has **unique strengths** in manufacturing, data in various domains (e.g., healthcare data), and government-industry and industry-industry collaboration
- Building on these strengths, Taiwan can aspire to be a global leader in (1) developing **AI applications** and (2) transforming AI computations into **AI chips**
- **A vision**: In the near future (let's say, 5 years from now), the world will look up to Taiwan to learn how to **apply AI** in industries and to **make chips for AI** computations.
- To realize this vision, the leadership of the Taiwanese government must acquire a broad and deep **understanding** of AI

Agenda for This Presentation

- Preamble: **AI is everywhere**
- An intuitive **explanation** of AI and generative AI
- **Generating images** from texts
- Technology **challenges**
- Conclusion

The perspectives presented here are based on my teaching, research, and consulting works with various organizations, and also discussions with many Taiwanese colleagues

Specifically, I will share with you what I learned about generative AI through studying and thinking (I have had a **busy 2023**!)

AI Is Everywhere

Example 1:
AI for Communication with
Low Earth Orbit (LEO) Satellite Constellations
(I taught a class at Harvard on this subject in Spring 2023)

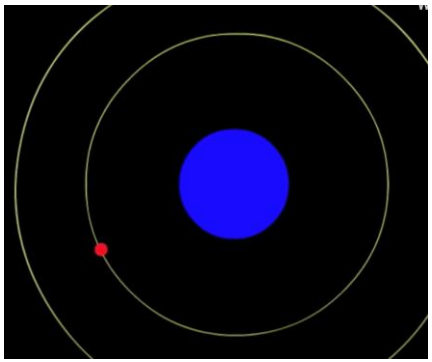
Example 2: 台中精機's AI 風火輪 Project
車床加工品質與刀具磨耗 AI 預診系統

Example 3: 台塑's AI efforts (e.g.,
minimizing energy consumption of 蒸餾塔)

LEO (Low Earth Orbit) Satellites: Emerging Global Networks

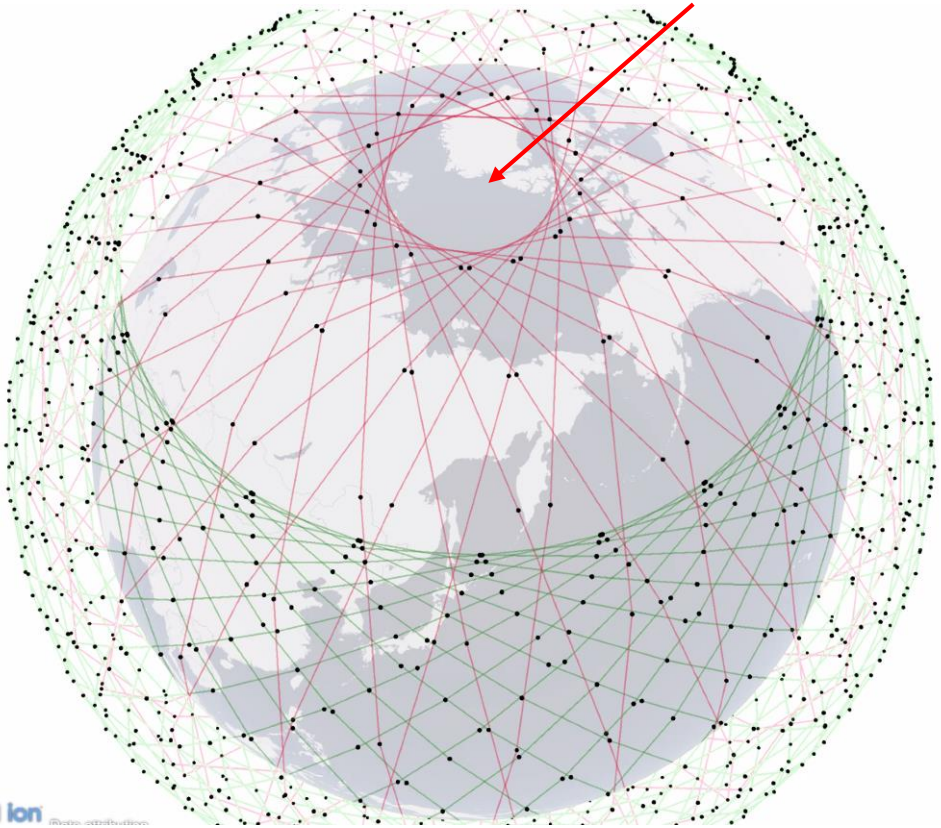
Satellite at an altitude

Multiple orbits/altitudes



Two constellations illustrated at different altitudes: red and green

Less populated area

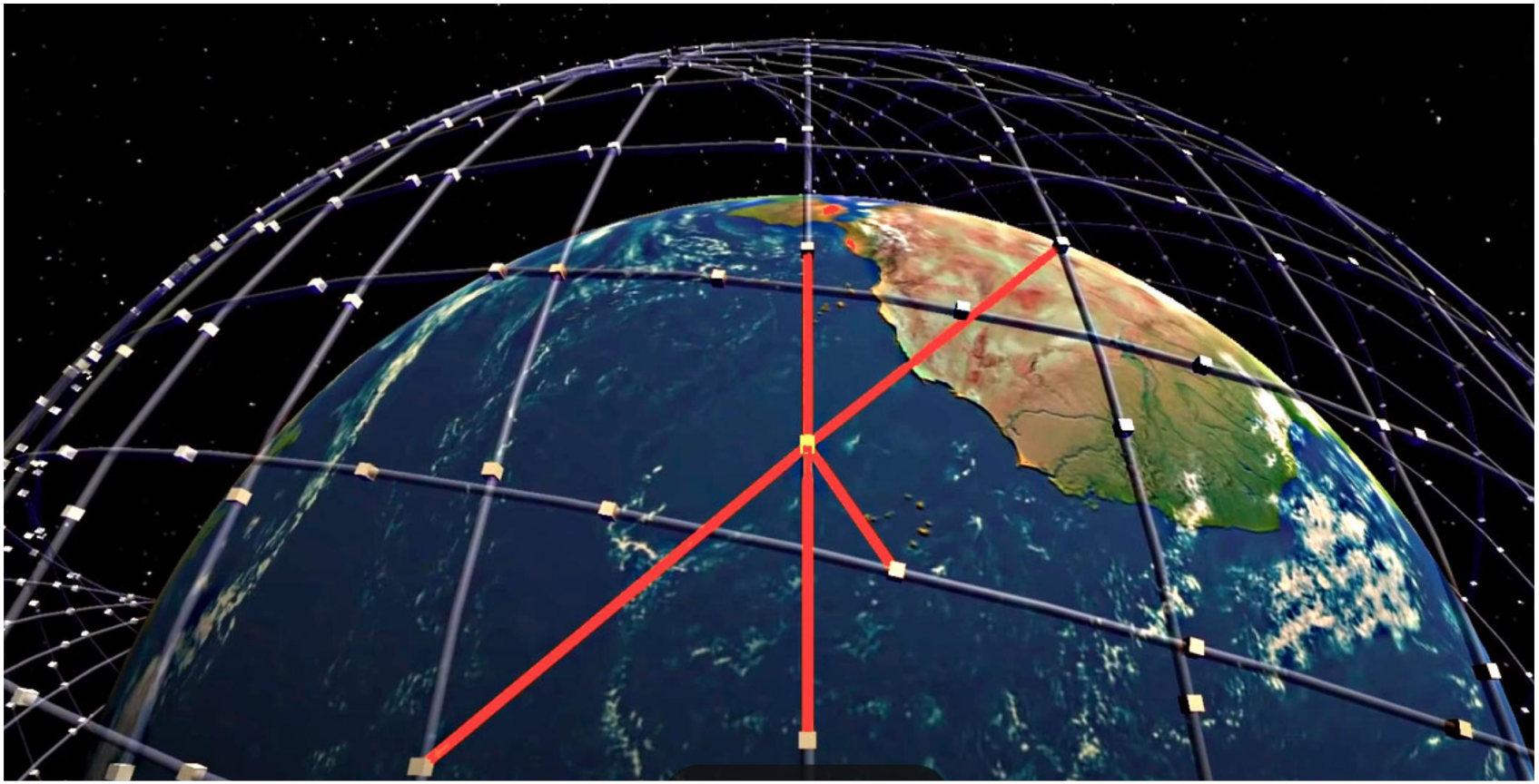


LEO Satellite Constellations

Well-known LEO satellite constellations currently in deployment:

- **Starlink** (SpaceX): 3,822 satellites launched; **12,000** planned; expandable to 42,000, elevation angle of 25 degrees
- **Kuiper** (Amazon):18 satellites, with **3,000** satellites planned
- **Lightspeed** (Telesat): 188 satellites, with **300** satellites planned; elevation angle of 10 degrees
- **Iridium** (Motorola): **66** satellites (final launch in 2019), providing global mobile and data services
- **OneWeb** (OneWeb and Airbus): **648** satellites, coming out from a bankruptcy filing in 2020, elevation angle of 25 degrees

Use of Inter-Satellite Laser Links to Form a Mesh Network in Space



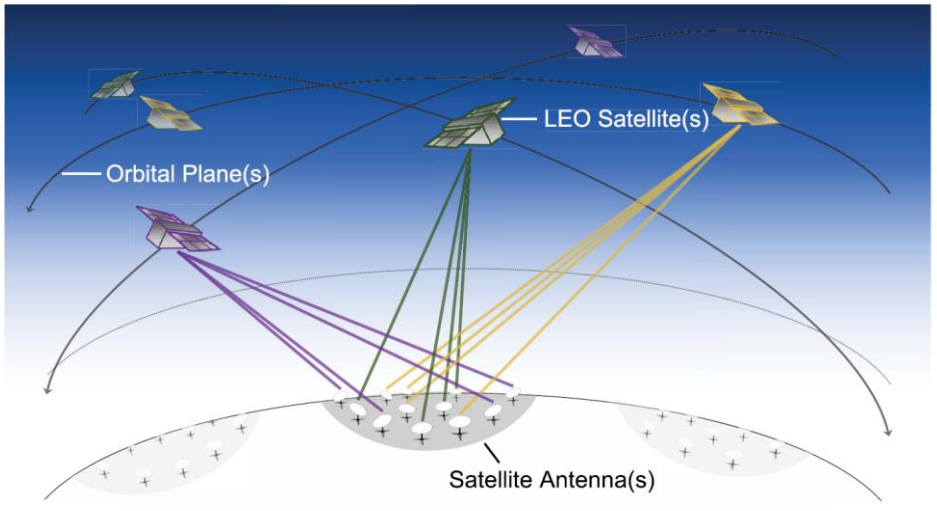
“SpaceX Is First With Inter-Satellite Laser Links in Low-Earth Orbit, but Others Will Follow” 2021

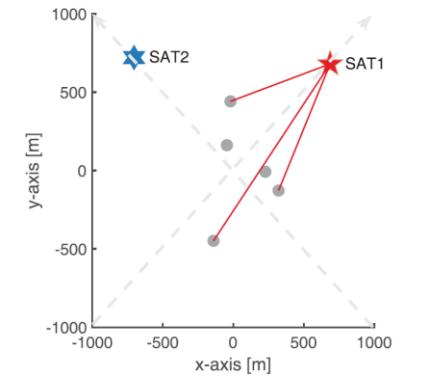
AI for LEO Satellite Constellations

- We can use AI in LEO satellite networking for:
  1. Forecasting wireless channel properties
  2. Spectrum sensing and frame classification
  3. Signal detection and decoding
  4. Learning **random access protocol** for ground user terminal (UT)'s association with satellites (a passing satellite is visible to a UT for only about 20 minutes)
- References
  - "Artificial Intelligence Techniques for Next-Generation Mega Satellite Networks," **2022**
  - "Learning Emergent Random Access Protocol for LEO Satellite Networks," **2023**

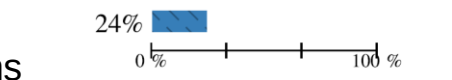
Dynamic Resource Management: User Terminal (UT) Association with LEO Satellite Base Stations

We can use AI's **reinforcement learning (RL)** to maximize satellite channel **utilization** among multiple agents (UTs) on the ground

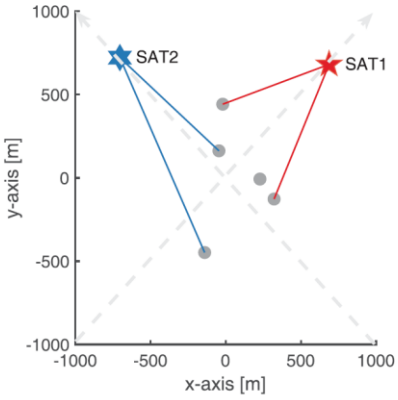




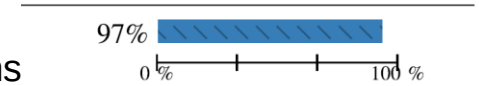
Bad associations



RL can make efficient UT-SAT associations automatically



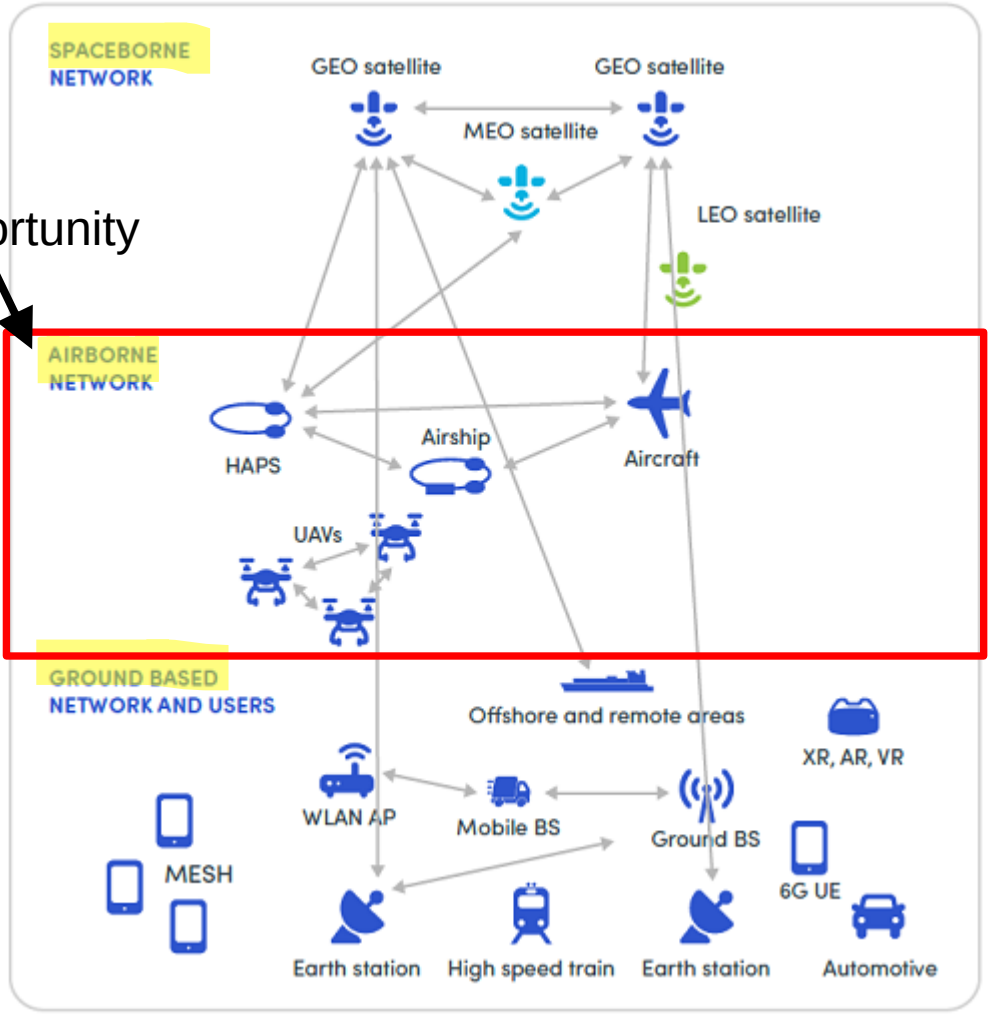
Good associations



Non-Terrestrial Network (NTN): Spaceborne, Airborne, and Ground Based

Important opportunity

Suited for high-resolution sensing and high-bandwidth communications



HAPS: helium-filled balloons (e.g., helium-filled balloons)

Non-Terrestrial Network Realities Call for New 5G Testing Approaches, 2021

台中精機

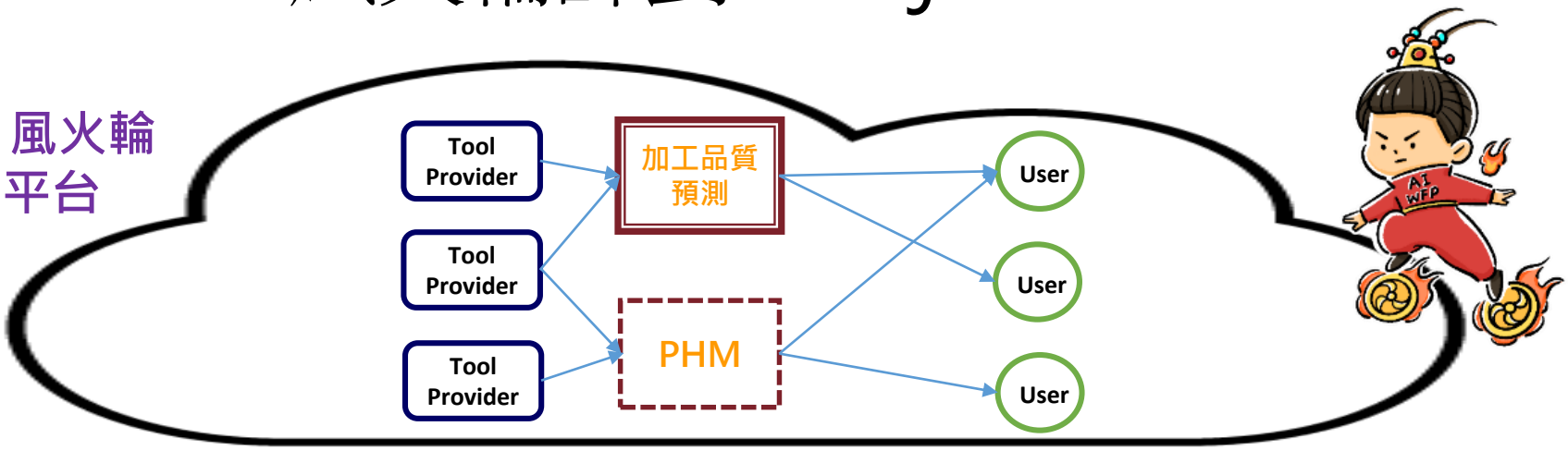
逢甲大學

台灣人工智慧學校

AI 風火輪

車床加工品質與刀具磨耗 AI 預診系統開發及應用

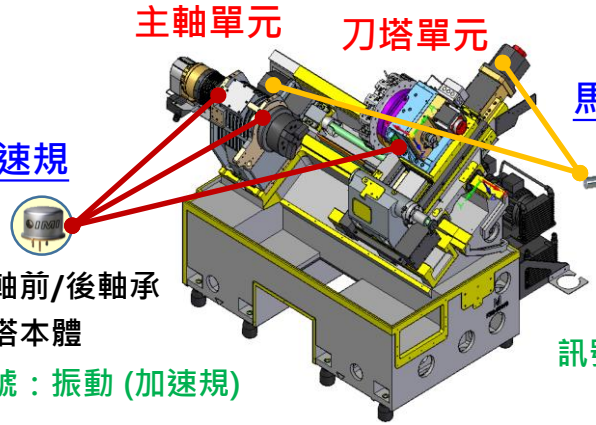
AI 風火輪計劃: Project Goal

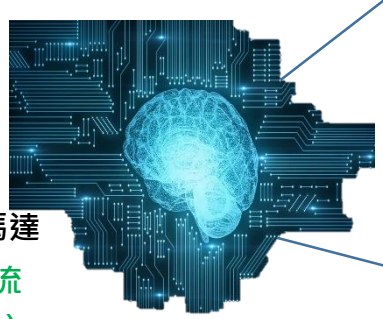


We have found that finetuning for specific tool providers and users can be **relatively simple** for this application; only some **bias-terms** of the model need to be modified

- 建構 AI 智慧製造開放平台 (簡稱 AI 風火輪)，以**開源模式**開發智能製造功能模組，共享平台基礎建設與功能模組
- 降低中小企業採用 AI 技術之進入門檻，協助台灣中小企業業者，解決製造現場在機台、製程良率、生產流程的痛點問題，如**加工品質特性預測、機台健康診斷 (PHM)** 等
- 透過此平台提供的 AI 功能模組，協助企業提升生產效率與品質良率，提高產品自身價值，為客戶提供更高附加價值之服務，**提升企業核心競爭力**

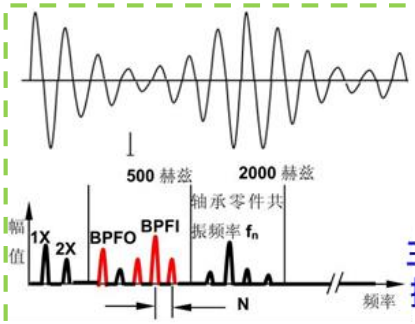
AI 風火輪計劃:
AI Predictions for Manufacturing

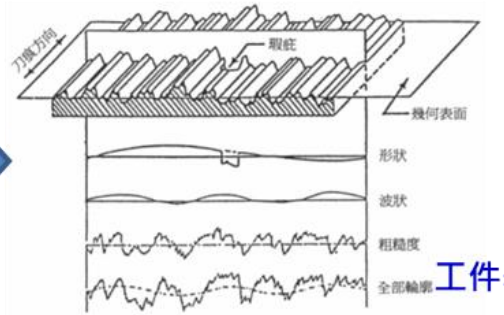




AI 模型演算需求能力：

1. 自主判斷當前工序 (訊號分類)
2. 依分類訊號各別評估加工品質精度 (不同工序) /刀具磨耗之預診器統





Pre-trained models to be adapted to 各種不同的加工產品:

台塑’s AI Efforts: 蒸餾塔 Example

台灣塑膠工業股份有限公司
FORMOSA PLASTICS CORPORATION

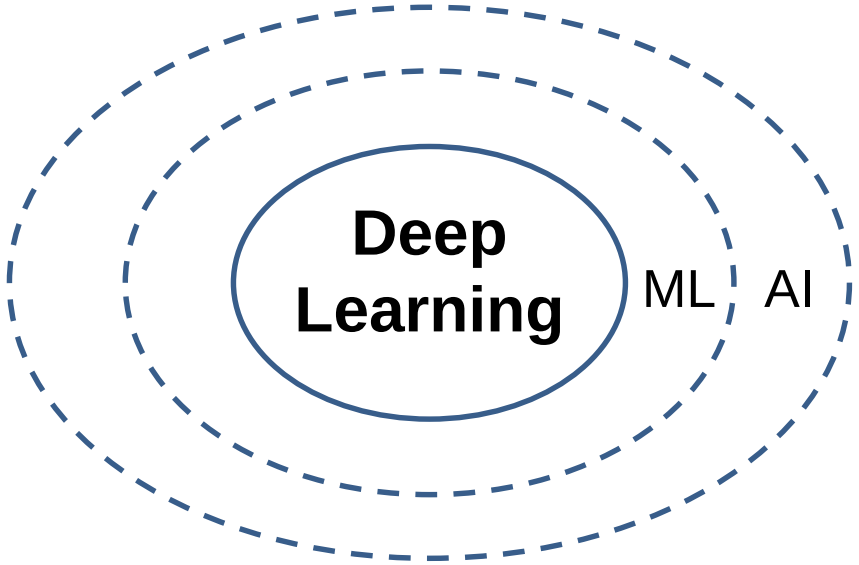
利用Aspen與製程數據開發
AI數位孿生模型

- 化工產業有大量的蒸餾塔用以分離不同的化合物，如台塑有**上百個蒸餾塔**，若能在維持產品純度的要求下節省蒸氣用量，**可以省下大量成本與碳排量**。
- 台塑數百個AI專案中，**蒸餾塔節能相關專案即占比50%以上**，且大多數蒸餾塔AI節能專案皆採用Aspen模擬，每案模擬器開發時程約1~3個月不等。**100案 * 2個月 = 200個月，相當耗時**。
- 模擬器開發完後，**產生一筆虛擬資料約30秒**，產生數千種條件組合需**數十小時**。
- 承上，過去專案已開發出模擬器，若能**整合既有的模擬器與製程資料建立蒸餾塔大型AI模型**，將知識保存在神經網路中，透過**AI模擬器與GPU加速則產生數千種組合只需數秒鐘**。

(Curtesy to 台塑)

An Intuitive Explanation of AI and Generative AI

Deep Learning Underlies Much of the Recent Excitement in Machine Learning (ML) and Artificial Intelligence (AI)



A Joke: “Only Machines Can Learn”

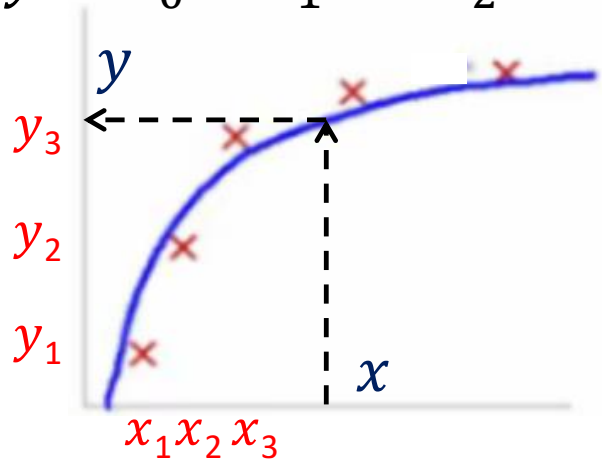
An Oversimplified Introduction to Supervised Deep Learning: Going Back to First Principles

Simply put, supervised deep learning is a glorified **curve-fitting** method:

- Given **training examples** of (x_i, y_i) pairs, where y_i is the target value (**label**) for a data point x_i, we learn a fitting curve (**model**) that can predict y for an **unseen** x

A high-school 3-parameter model:

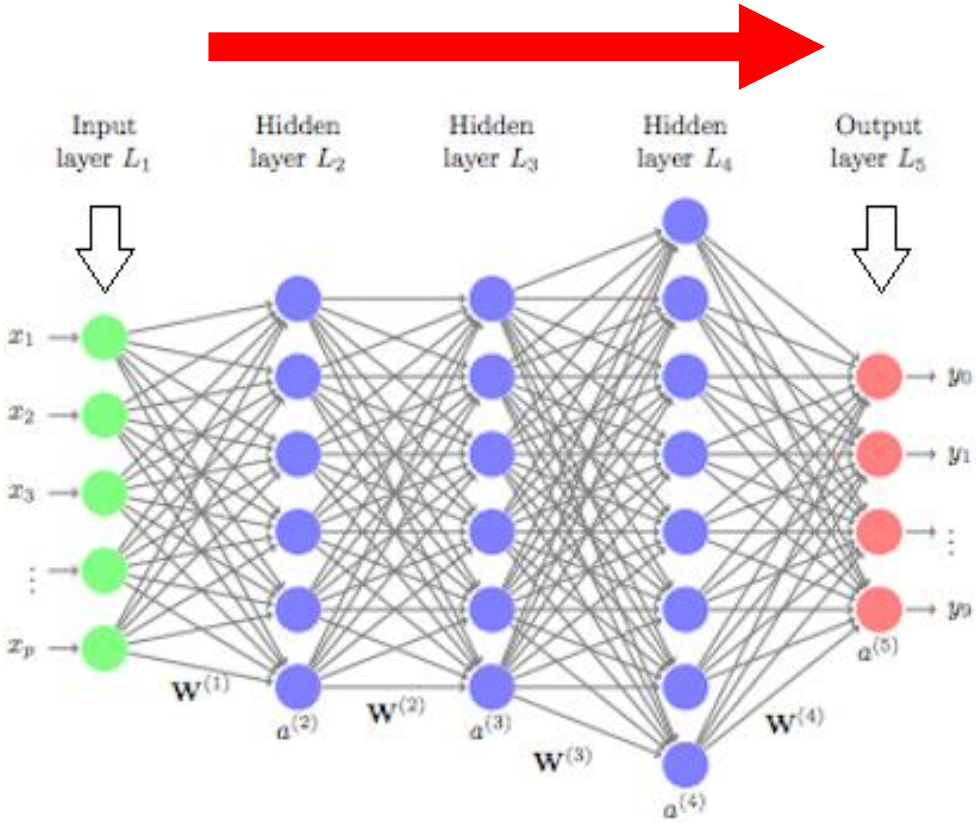
$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$



- For accurate prediction, we may use many model **parameters** θ_1, θ_2, ⋯ to fit training examples
- To avoid **overfitting**, we use sufficiently many training examples
- Deep learning automatically learns the values of a large # of parameters for a neural network via label fitting on a massive amount of training examples (**training dataset**)

Deep Neural Network

Forward Pass for **Prediction (Inference)**

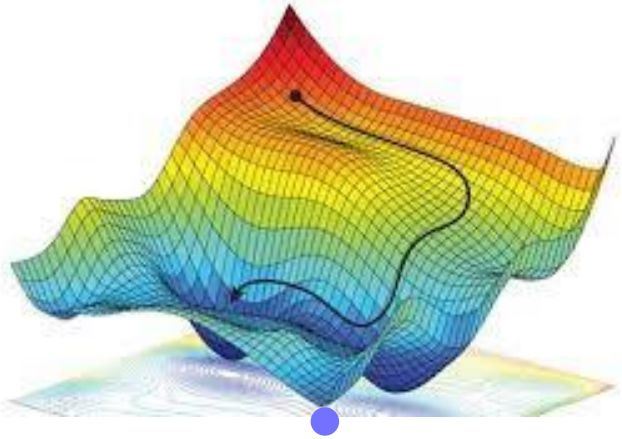


Backward Pass for **Learning (Training)**

- There are a number of layers between the input and output
- Every connection (line) is associated with a **parameter** (weight)
- A deep neural network uses many parameters to fit a training dataset by **minimizing the error (loss)** of model output in comparison with target labels

Advances in the Past Decade: Being Able to Use Massive Data and Computing

- We can acquire **massive amounts** of data generated by sensors, humans, and other means over computer networks
- We can use **parallel computing** to train large models (learn their parameters) based on this data by minimizing the error of the model output



Targeted point that minimizes the loss computed by Stochastic Gradient Descent (SGD)

In practice, we typically minimize over a high-dimensional parameter space, e.g., dimensionalities of millions or billions. (The 2-dimensional parameter space depicted above is just for illustration purposes)

Two Challenges

In the following slides, we discuss two challenges of supervised learning

1. **Data Biases**
   - To mitigate biases of the trained AI model, we need to **prepare data** by removing biases
2. **Data Labeling Costs**
   - **Human annotations** (e.g., a doctor's marking of a medical image to indicate areas of abnormality) can be expensive
   - To reduce labeling costs, we can pre-train a model without labels using **unsupervised** or **self-supervised** training. These **pre-trained** models compute embedding vectors of unlabeled training data that capture the semantics of the data. Then, we can **finetune** these models for downstream tasks such as classification using relatively small amounts of labeled data

A Data Bias Problem: A Well-known Example of Classifying a Sports Car

- A father decides to teach his young son what a sports car is, by using **examples**
- They stand on a motorway bridge and the father cries out “That’s a sports car!” each time when a sports car passes underneath
- After ten minutes, the father asks his son if he understands what a sports car is. The son says, “Sure, it’s easy.”
- An **old red VW Beetle** passes by, and the son shouts – “That’s a sports car!”. The father asks – “Why do you say that?”. “Because sports cars are red!”, replies the son

Sports Car

Old red VW Beetle

This example illustrates that AI models trained on biased data can lead to catastrophic failures in prediction

Models Pre-trained with Unlabeled Data to Compute Embedding Vectors

- **Large Language Models (LLMS)**, such as GPT and BERT, map input texts to embedding vectors that capture meanings of words, sentences, etc.

Texts → embeddings

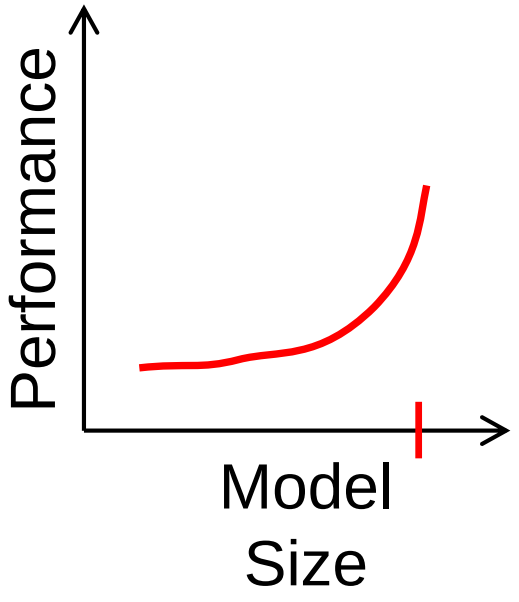
- **Multimodal models**, such as CLIP (Contrastive Language-Image Pretraining), map multimodal inputs (texts, images, audio, etc.) to embedding vectors that align multiple modalities. For example,

(Text, Image) pairs → text and image embeddings

We can identify **nearby content** by computing the **distance** between their embedding vectors

Pre-Trained Large Language Models (LLMs)

- We pre-train LLMs such as ChatGPT by learning to predict **next words** (more precisely, next tokens) from writings available on the Internet and other sources
- Several studies suggest that **pre-trained** LLMs can significantly improve their performance for various downstream tasks when the **model size** increases beyond a certain threshold
  - E.g., GPT-2/3 is large enough to generate fluent texts, e.g., **ChatGPT**
- Solomonoff Induction and **Kolmogorov Complexity in the 1960s** provide a theoretical foundation (in this theory, machine learning feature extraction is considered as compression)

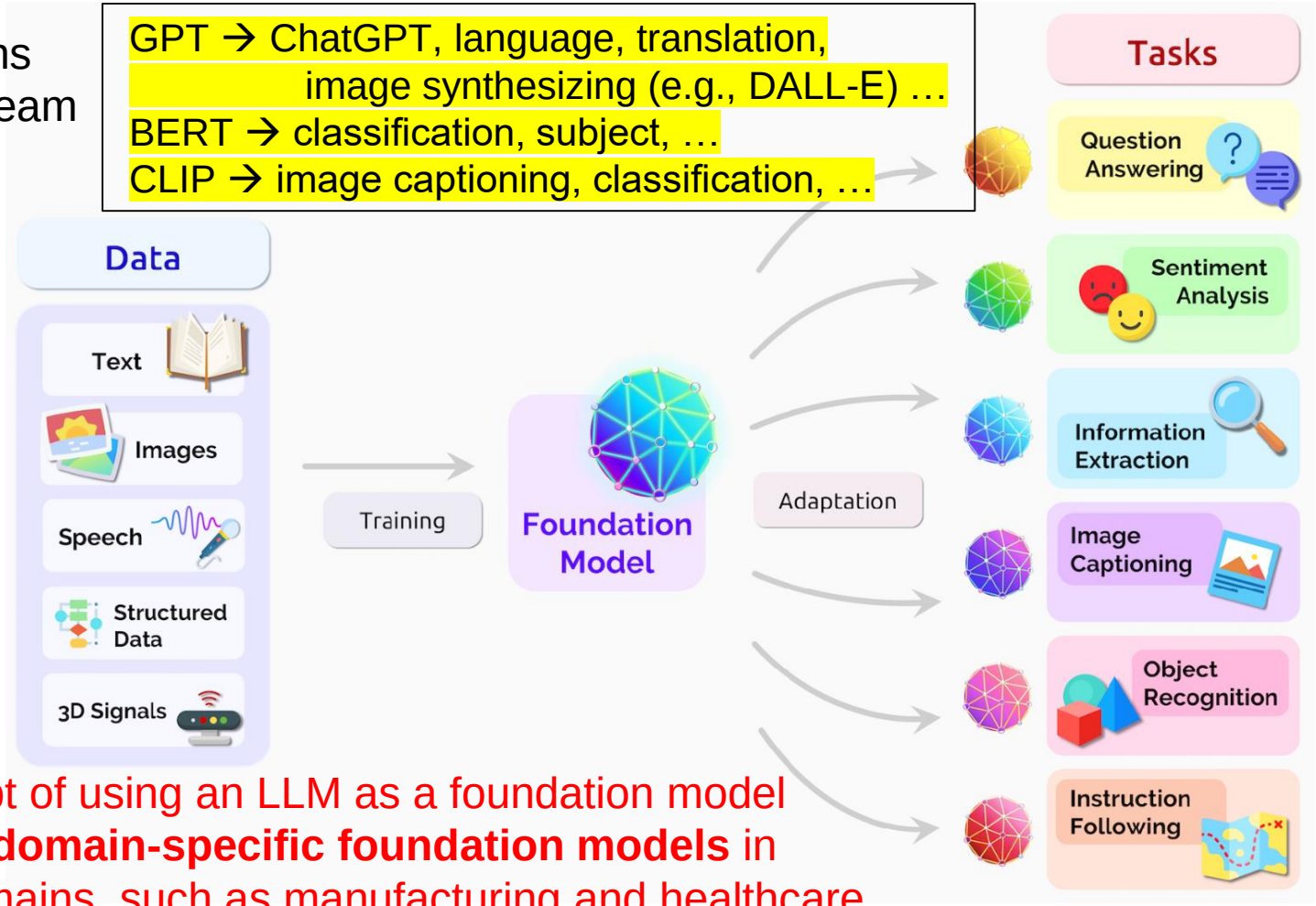


“Foundation Models”

These are pre-trained models that can be adapted to a variety of **downstream tasks**

Adaptions to downstream tasks:

GPT → ChatGPT, language, translation, image synthesizing (e.g., DALL-E) …
BERT → classification, subject, …
CLIP → image captioning, classification, …



The concept of using an LLM as a foundation model extends to **domain-specific foundation models** in various domains, such as manufacturing and healthcare

On the Opportunities and Risks of Foundation Models, 2021

ChatGPT: A Disruptive Technology

- By inputting **text prompts** to provide context, users get descriptive responses and engage in conversations with an LLM
- Compared to traditional virtual assistants (e.g., Amazon Echo), ChatGPT is **more accurate**, **more flexible**, **more natural** to the user, **easier** to use, and **simpler** to maintain
- Researchers and developers around the world are working hard to understand the **implications** of ChatGPT

ChatGPT Usage Example: Improving Writing

- User: Make the following text clearer and shorter
  - Yes 3/21 was a decided meeting date. However I'd like to ask if it is possible to move the meeting to Thursday the 23rd, otherwise we will stick to the 21st.
- AI:
  - Can we move the meeting to March 23rd instead of March 21st? If not possible, we'll keep the original date.

Whisper Usage Example: Improving Speech Input

- Conventional transcriber (poor in handling noise, accent, stuttering, etc.)
  - ME: 請告訴我**還是一次因**銀行失敗的因素及的產業創新公司的影響。
- LLM-powered **Whisper** transcriber
  - AI: 請告訴我**SVC**銀行失敗的因素, 及對產業和新創公司的影響。

Searching for Info Is Much Easier

- 最近我撰寫的專案報告花了我只平常時間的一半，同時也讓寫作過程變得更輕鬆且更愉悅
- 我收集上課的資料也比以前有效率多了
- E. g., using ChatGPT, I can get answers directly
  - ME: What is the system architecture of Starlink Low Earth Orbit satellite constellations?
  - ME: How does the attention mechanism in transformers for training LLMs work?
- In comparison, when using a conventional search engine, users need to sift through **a large pool of search results** and carefully assess which ones are relevant
  - When users perform this laborious and time-consuming task, search engine companies make money by getting them to read advertisements
  - It is ironic that they make money by not serving users well in the first place

ChatGPT and Students' Learning

Students can use ChatGPT for:

- Information retrieval and summarization
- Interactive exploration
- Content synthesis

If we do nothing, we can expect the following:

G1. High-performing students → even better

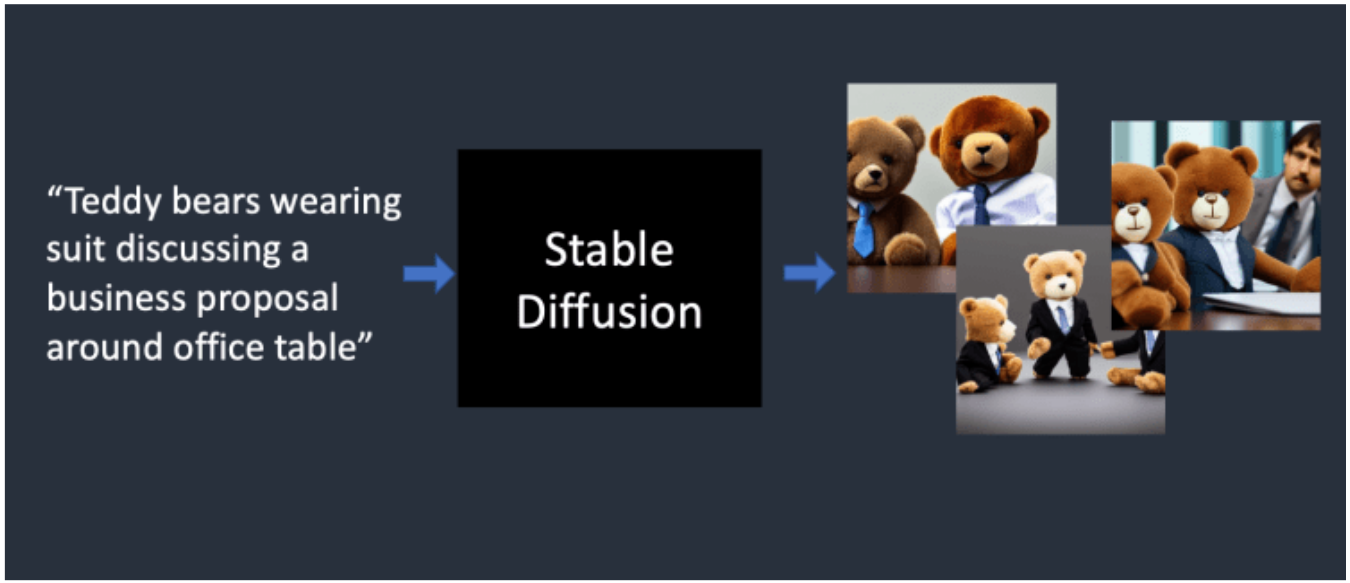
G2. Average-performing students → intellectually superficial

G3. Low-performing students → even less competitive

We would want to move students in G2 and G3 to G1

Generating Images with Diffusion Models

Stable Diffusion: Turning Text Prompts Into Images



" How does Stable Diffusion work?" 2023

Stable Diffusion: Forward and Reverse Processes

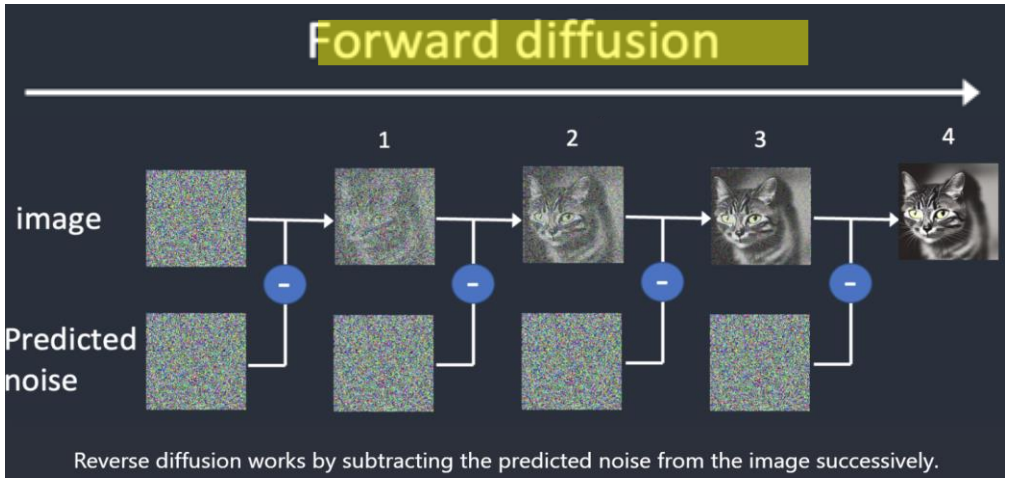
Adding noise in each step

Forward Process

Train a noise predictor (U-Net)
that predicts the total noise added to each step

Denoising in each step by removing predicted noise

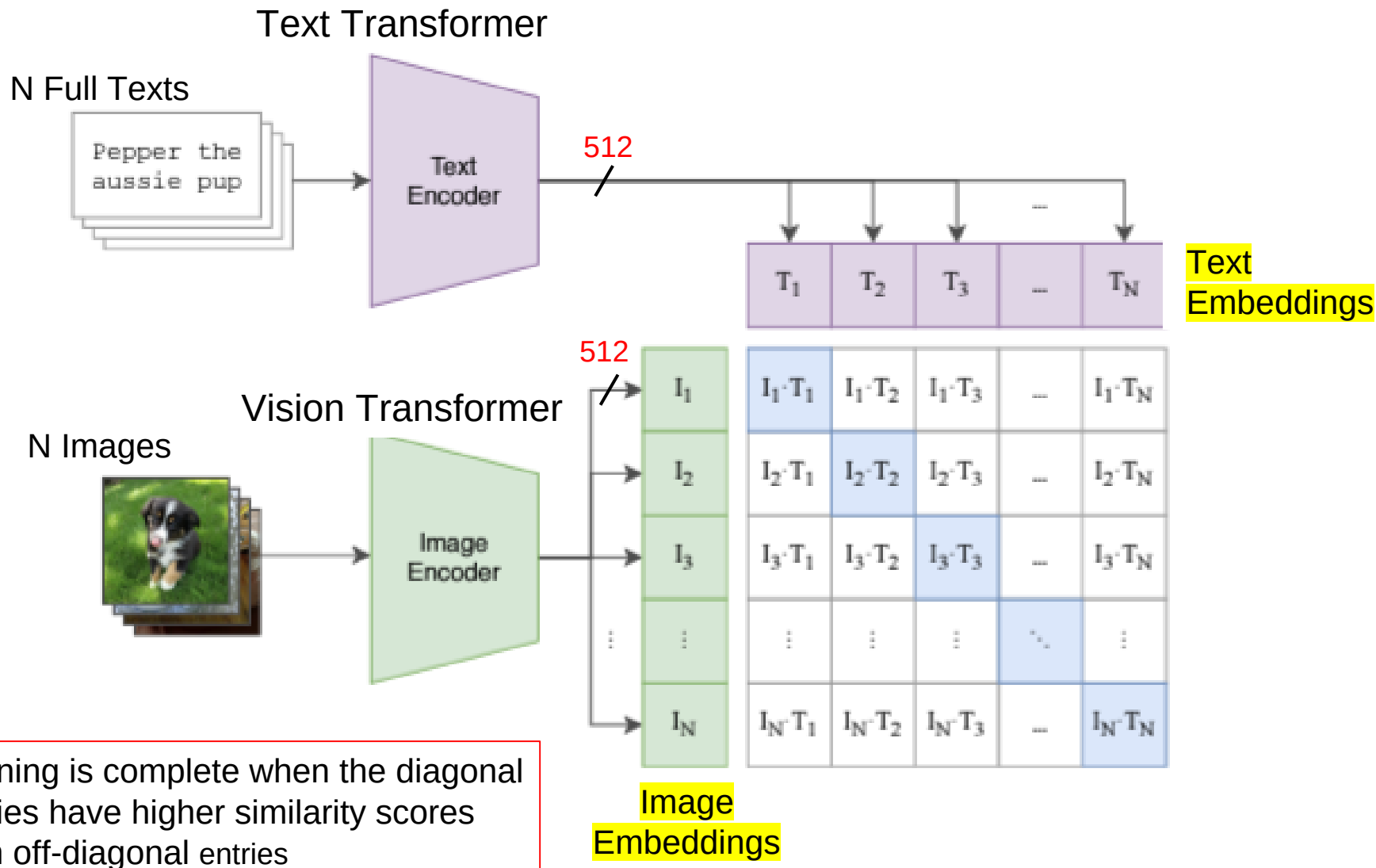
Reverse Process



Use a **text prompt** in the embedding space of a **visual language model** (e.g., CLIP) to **force** the generated image to be a CAT rather than any other image like a DOG

Remove predicted noise at each step

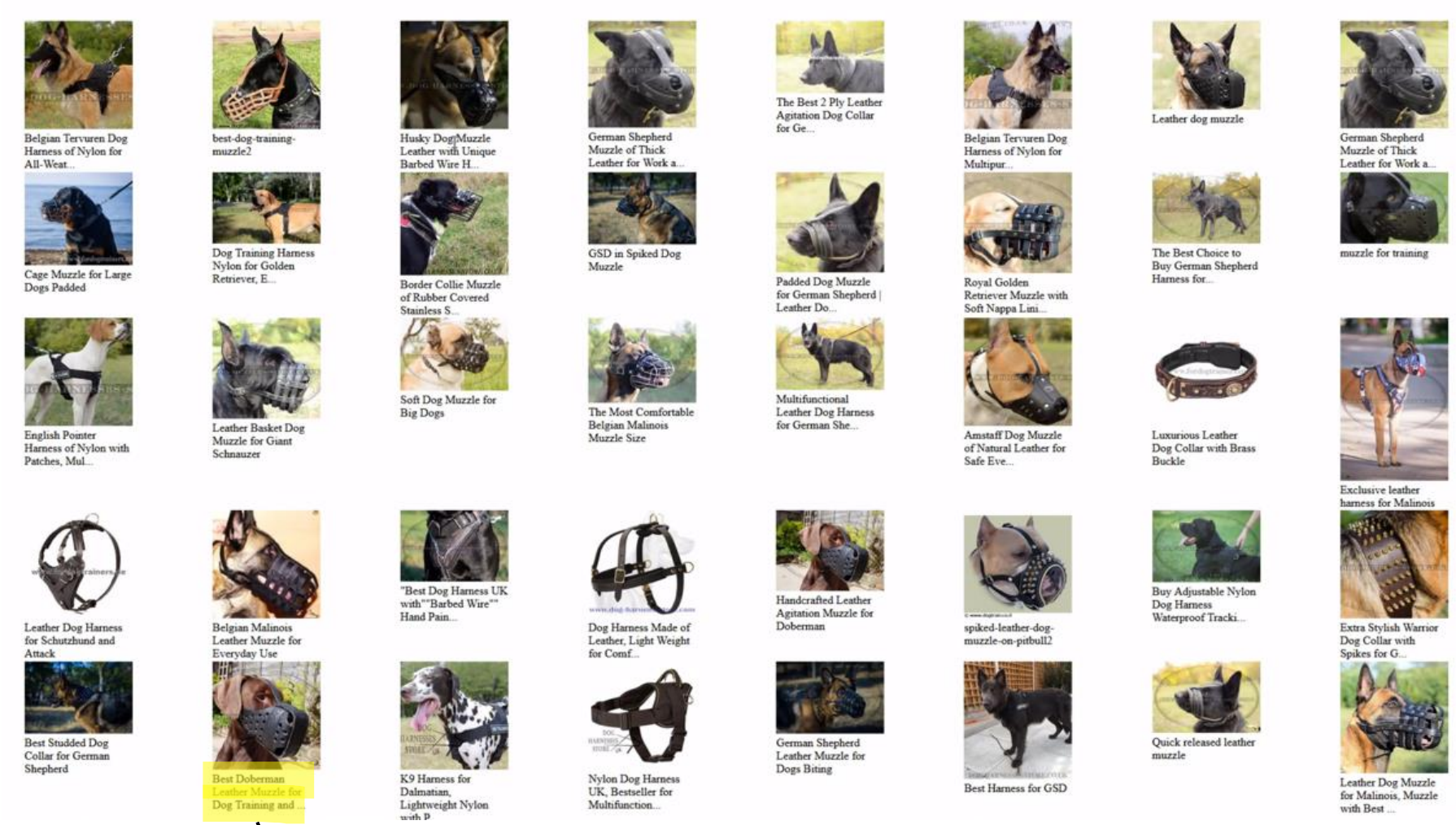
Contrastive Pre-training of CLIP: **Aligning** Text and Image Embeddings



Training is complete when the diagonal entries have higher similarity scores than off-diagonal entries

CLIP: Connecting text and images, 2021

E.g., Image-Text Pairs for Dogs in CLIP's Training Dataset



"Best Doberman Leather Muzzle for Dog Training and ..."

Synthetic Images Generating by Stable Diffusion

Text Prompt: “**Astronaut riding a horse**”

Unsuccessful case

Successful case (The horse can be made prettier via finetuning with pretty horses)

Application: Images Generated by Diffusion Models for 3D Human Reconstruction

Reconstructing highly accurate 3D human models with geometric details using only **8 RGB cameras**. Before, such results could only be achieved using nearly a **hundred cameras**

“DiffuStereo: High Quality Human Reconstruction via Diffusion-Based Stereo Using Sparse Cameras,” ECCV 2022

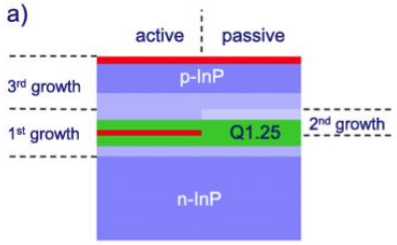
Towards Synthesizing Images in Training Models for Manufacturing

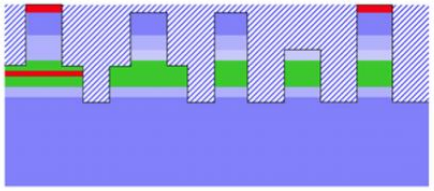
We can use optimized prompts to synthesize manufacturing-related images

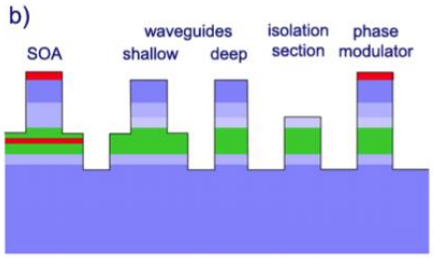
- Electric Water Pump Images

https://www.alamy.com/blue-electric-water-pump-3d-illustration-image244271827.html

- Wafer Cross-Section Structure Images







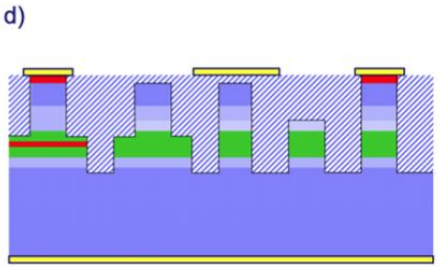


Figure 8. Cross-section of the wafer structure after the four process modules: (a) epitaxial growth, (b) waveguide etching, (c) passivation and planarization, (d) contacting and interconnect metallization.

An introduction to InP-based Generic Integration Technology, 2014

Responses Sensitive to Prompt Details

Text prompt: “**orange cat wearing bowtie**”

0.3674 

0.3595 

0.3574 

0.3536 

0.3468 

0.3434 

0.3431 

0.3418 

Just add an article

Text prompt:“**orange cat wearing a bowtie**”

0.3684 

0.3586 

0.3534 

0.3481 

0.3448 

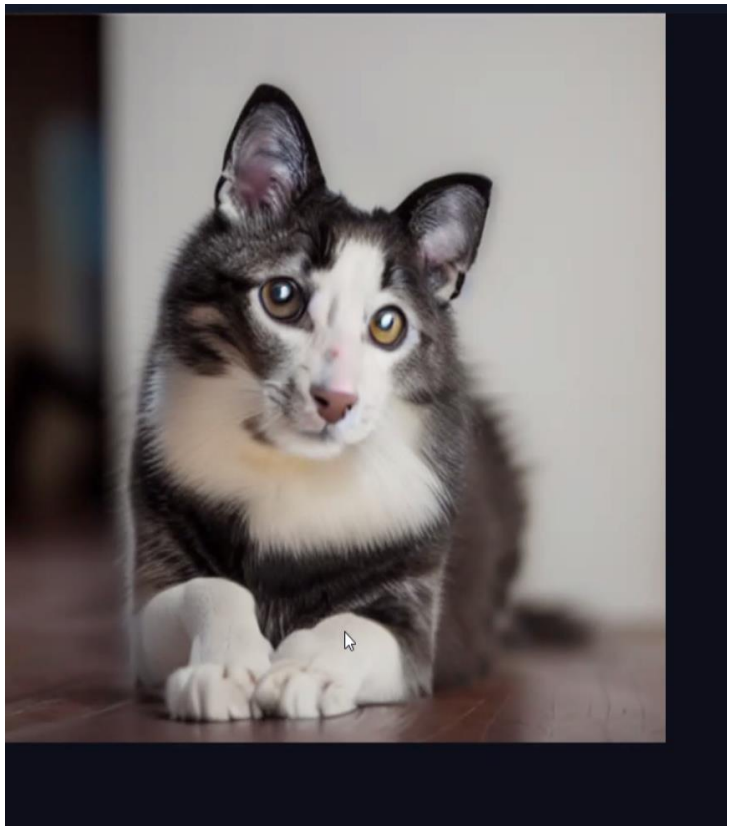
0.3444 

0.3441 

0.3429 

Crazy Images of Cat-Dog Hybrids

Prompt: cat-like dog, (feline 0.8), (canine 0.2)

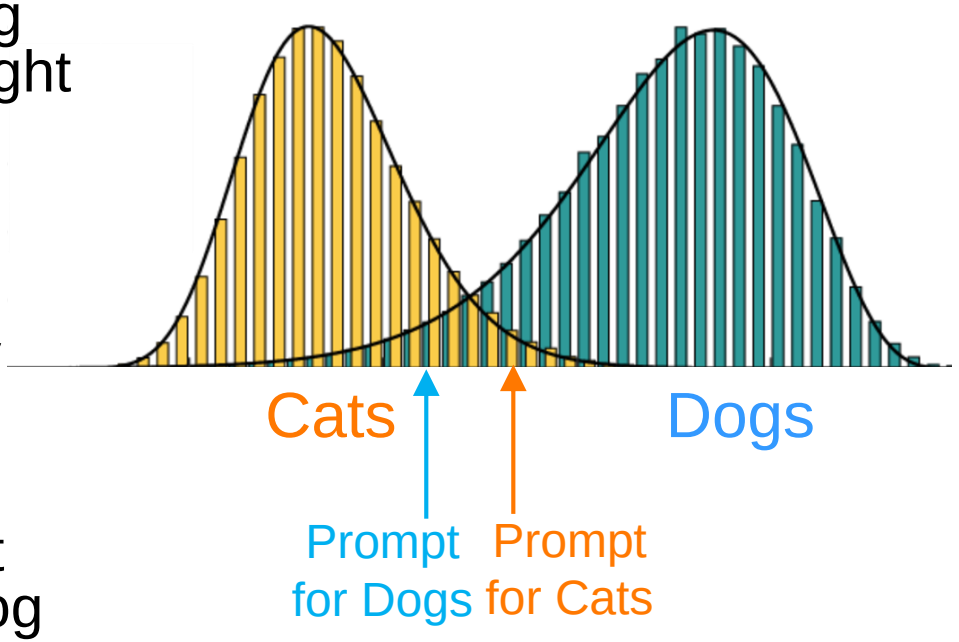


By increasing the weight on canine, images become more dog-like

Misaligned Prompts and an Illustration of How They Lead to Misclassification

- In this illustration, text prompts for cats and dogs are **not aligned** with their distributions
  - The dog prompt is to the left of the cat prompt, while dog images are mostly to the right of cat images
- Consequently, a classifier based on these misaligned prompts would **incorrectly** classify most cats as dogs
  - This is because in the embedding space most cat images are closer to the dog prompt, and similarly, most dog images are closer to the cat prompt

Distributions of embedding vectors (**1D visualization**)

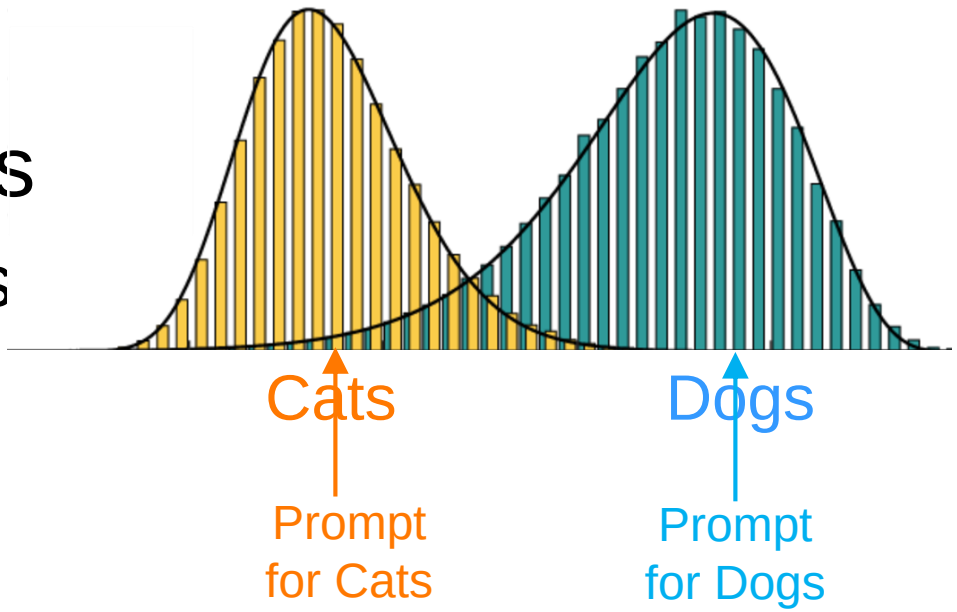


Misaligned prompts

Mitigating Misaligned Prompts with Cluster Medoids

- We would want to choose text prompts that correspond the **centers** (medoids!) of class distributions
  - A cluster **medoid** is a real data point most centrally located within a cluster

Distributions of embedding vectors (**1D visualization**)



Prompts are now aligned to class distributions

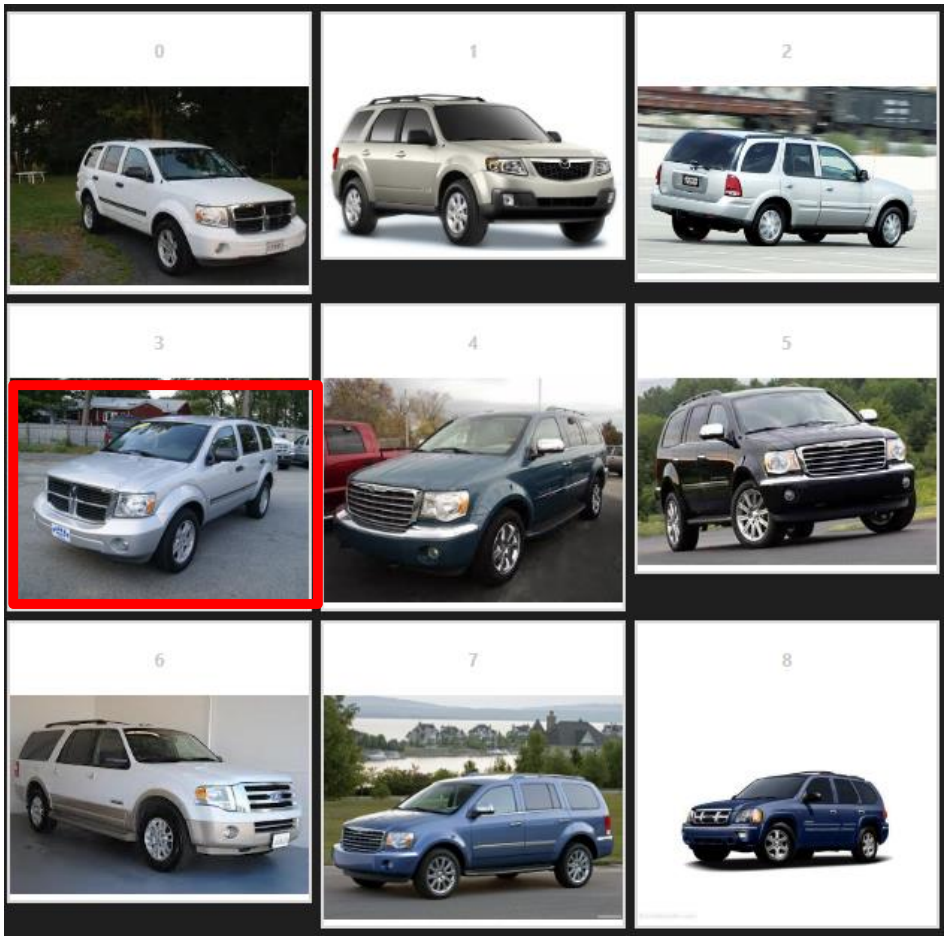
Empirical Results on Clustering Car Images

- Starting with an unlabeled set of images from a dataset (in our case, **Stanford Cars196** training set), we want to automatically generate categories with **unique category labels** given by hard prompts
  - This car image dataset has 196 different car models for a total of 16,185 car images
- In our experiments, we use K-medoids clustering algorithm for $\boldsymbol{K}$ **= 10** clusters to derive clusters 0, 1, …, 9
- We are interested in **automating** text prompt generation for defining classifiers based on the clustering of these car images in the CLIP **embedding space**
- We derive these text prompts (“medoid prompts”) using 10 images from each cluster
- Moreover, we **synthesize** additional images for these groupings

Example Images for a Cluster

Cluster medoid image

9 images in the cluster



Results Using Medoid Prompt

Medoid image

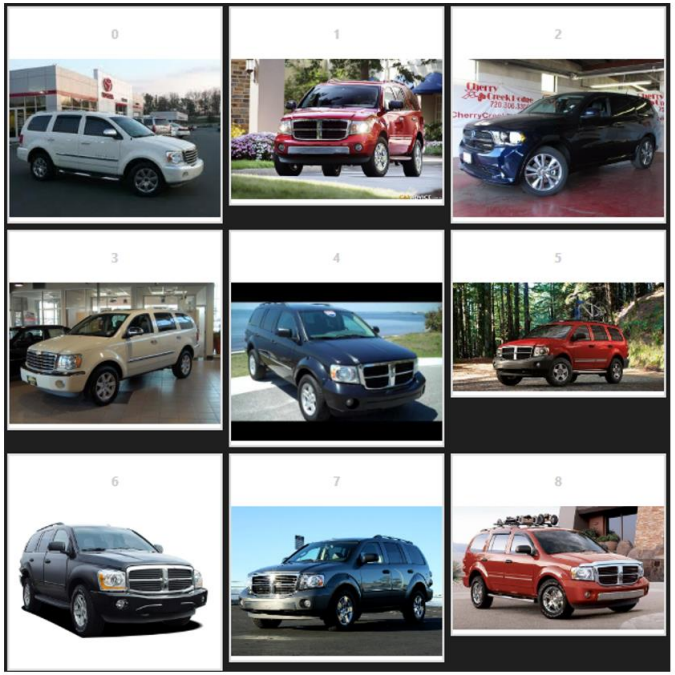
Medoid prompt:

comcast offers stolen dodge durango armada vans essentiheavenly nhljets ultimatefansday

(The prompt has tokens that are not human readable)

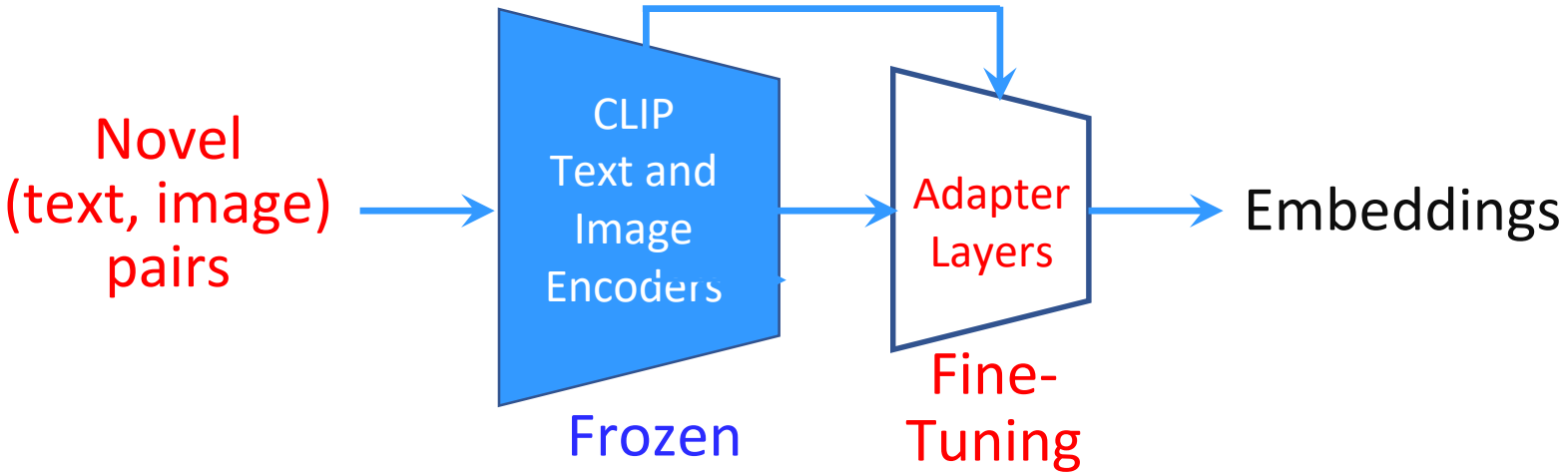
Images **synthesized** via Stable Diffusion using the medoid prompt

Closest images matching the medoid prompt from the Cars196 test set



Finetuning CLIP's Encoders

- When the input data (we assume it is an image here) is novel, existing medoids may not be very useful in deducing semantics
- In this case, we conduct research to **create new texts** to describe the novel data, resulting new (text, image) pairs for the novel datapoints



- We then **finetune** the CLIP encoders (using, e.g., an adapter layer shown in the diagram above) to incorporate novel inputs (images in this case) and their corresponding text descriptions

Performing the Tasks of Old Masters ("老師傅" or "黑手") via CLIP

Suppose that observed effects (inputs) in a factory are in the form of images (e.g., defect images). Then via CLIP an AI-assisted system can carry out tasks of old masters in recalling known solutions and continually adapting to new data:

Task 1: Is the input known or novel?

- Cluster images that have been seen in the embedding space of the current CLIP model
- Form medoid-text-prompt defined classes
- Determine whether a given image is in a known class or novel by checking if it is close to a given cluster medoid

Task 2: Classify the input or create a new class

- If the input is **known**, classify it and recall prior solution for this class
- If the input is **novel**, research the novel image and create a new class with a solution

Task 3: **Incorporate new classes** into the knowledge base

- Update the current CLIP model by finetuning a subset of the model with novel (text, image) pairs

Powerful Building Blocks Now Available for Composing Ambitious Systems

- Building blocks
  - Residual blocks
  - Attention blocks
  - Variational Autoencoder (VAE)
  - Pre-trained language models (e.g., GPT, BERT)
  - Pre-trained image-text models (e.g., CLIP)
  - …
- Composing application systems using these building blocks
  - Image classification using texts
  - Image-to-text and text-to-image retrieval
  - Stable Diffusion
  - Automatic generation of optimized prompts
  - …

Our task: Quicky **experiment** with AI systems composed of these building blocks to support various applications

Technology Challenges

Technology Challenges (1/2)

1. **Training skills**: Training large models (say, 200B parameters and beyond) in data centers is expensive in terms of equipment and electricity and requires substantial experience
2. **Cloud serving of massive inference demands**: E.g., handling real-time queries from millions of ChatGPT users (a cost of $0.02USD/chat has been reported!)
3. **Finetuning:** Finetune pre-trained models to incorporate local knowledge
4. **Edge training and inference**: Compressing and adapting models to edge devices constrained in memory, computing, and data
5. **Edge-cloud split**: Searching for networks and optimal edge-cloud splits for low-latency edge-cloud distributed computing using memory-constrained edges

Technology Challenges (2/2)

6. **Training and finetuning domain-specific foundation models**
7. **Finetuning for continual learning**: Finetuning of models for continuous incorporation of novel data
8. **User privacy**: User queries should not leave the organization
9. **Security**: Protection against adversarial input attacks
10. **Hallucination**: Mitigate hallucination when the model's training data may include both factual and fictional information
11. **Safety**: Screening training data and model output to avoid the possibility of sending improper responses to users

Conclusion

- The AI revolution is expected to be more **transformative** than the PC revolution
- Generative AI has **widespread** impacts
- Taiwan needs to seize this **once-in-a-50-years** opportunity to upgrade once again
- With sound government policies and effective execution, we can envision the world will look to Taiwan for **applying AI** to various industries and making powerful **chips for AI** computations