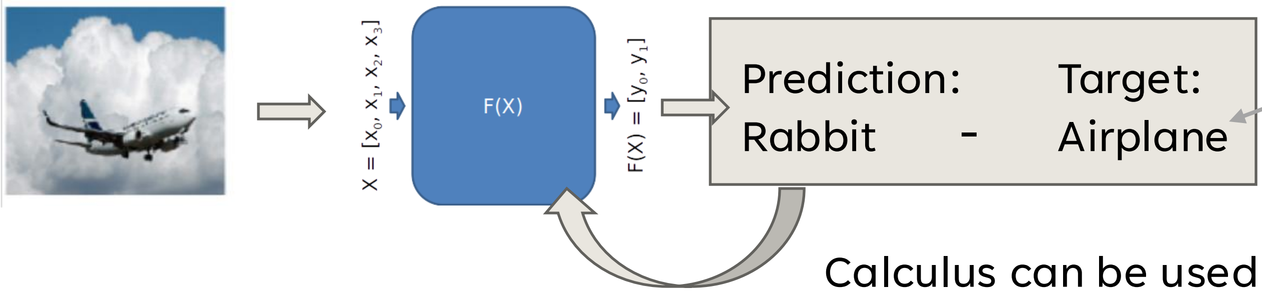
CSCI 4850/5850 – Neural Networks

Contrastive Learning

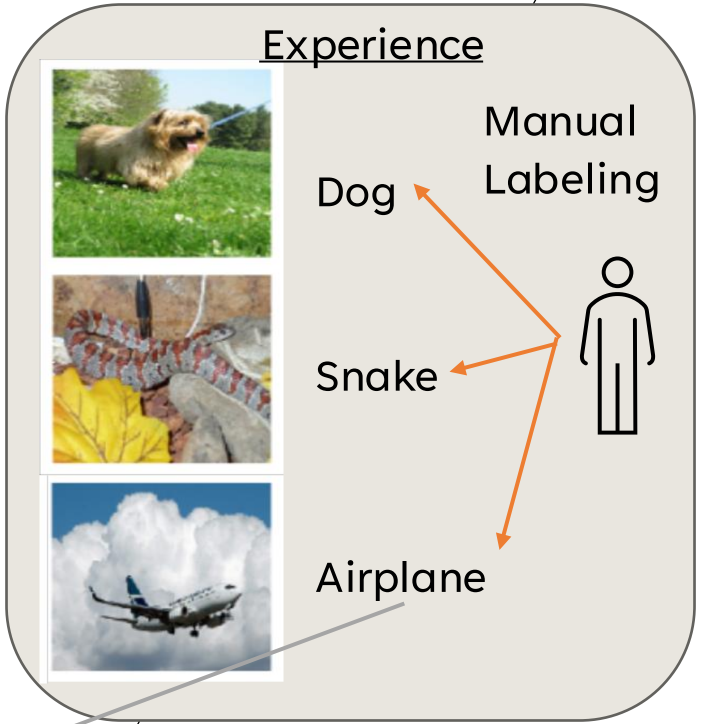
TRADITIONAL SUPERVISED LEARNING PROBLEMS

- Traditional approaches were the norm until about 2015
- The 2000-2015 deep learning models showed significant improvements because of data availability (storage, access, crowd-sourcing)

Typical structure was direct 1:1 correspondence between items.



<u>Task</u>: Image Recognition/Processing

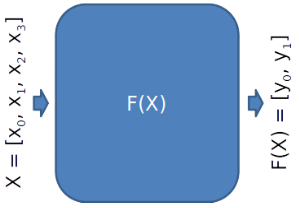


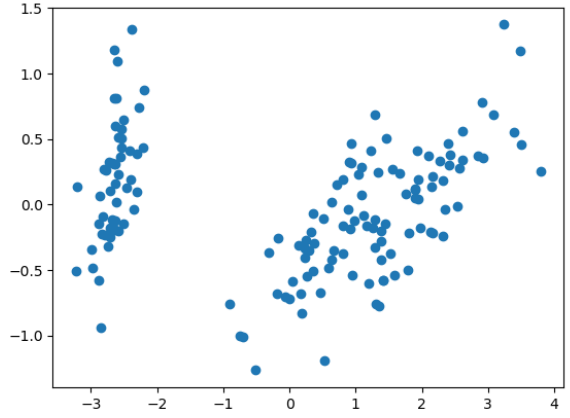
<u>Performance Measure</u>: Accuracy, F1-score (Gold Standard: Human Performance +)

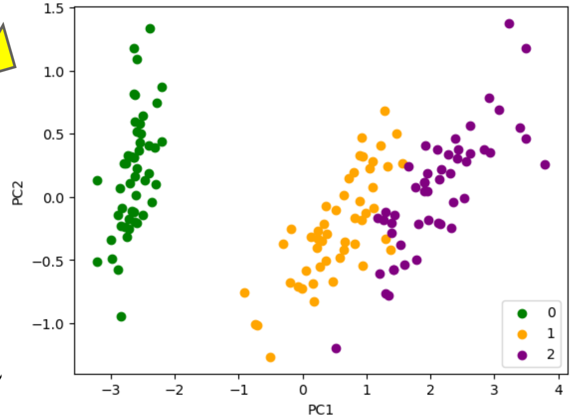
Calculus can be used to make a small change: Next time "Airplane" is more likely to be the prediction than before. [Training, Testing]

```
[6.9, 3.1, 5.4, 2.1, 2. ],
[7.7, 3. , 6.1, 2.3, 2. ],
[5.7, 3. , 4.2, 1.2, 1. ],
[5.1, 3.7, 1.5, 0.4, 0. ],
[5.6, 2.9, 3.6, 1.3, 1. ],
[6.2, 2.9, 4.3, 1.3, 1. ],
[5. , 3.2, 1.2, 0.2, 0. ],
[6.7, 3. , 5. , 1.7, 1. ],
```

Iris Data





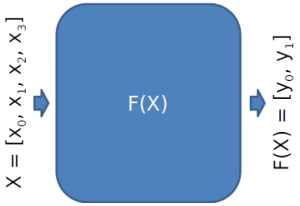


TRADITIONAL UNSUPERVISED LEARNING PROBLEMS

- Traditional approaches were the norm until about 2017
- Traditional methods selected some training procedure operating on the current data.
- From 2017-2022 deep learning models showed significant improvements due to *generative* training methods: don't just use the current data, learn to "make up" or "fill-in" data as part of the training process.
- Vaswani et al., 2017 - **Transformer** architecture
- Devlin et al., 2018 - BERT – Bidirectional Encoder Representations from Transfomers

Money in the ____

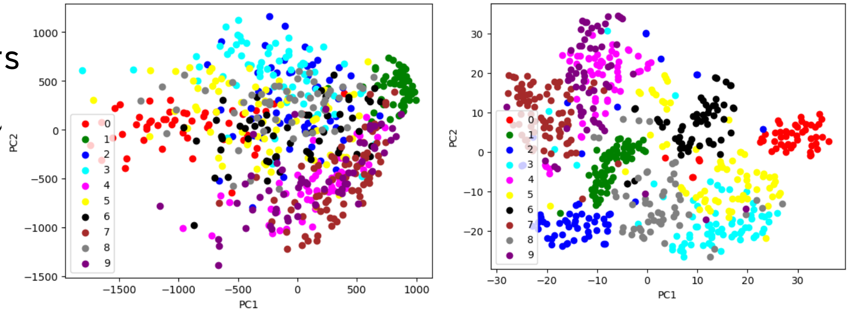
Short walk on the river ___



___ ___ ___ bank

_____ ____ __ ___ _____ bank

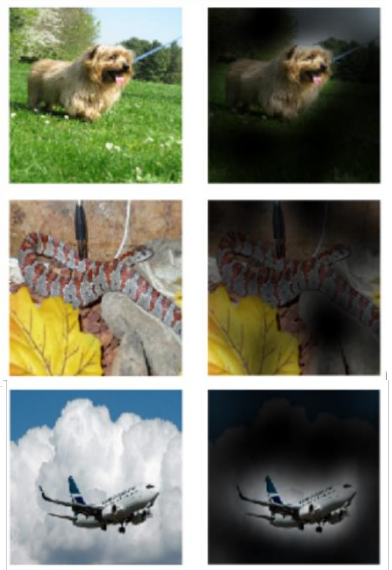
Better *contextualized meaning* from these models



MNIST Data

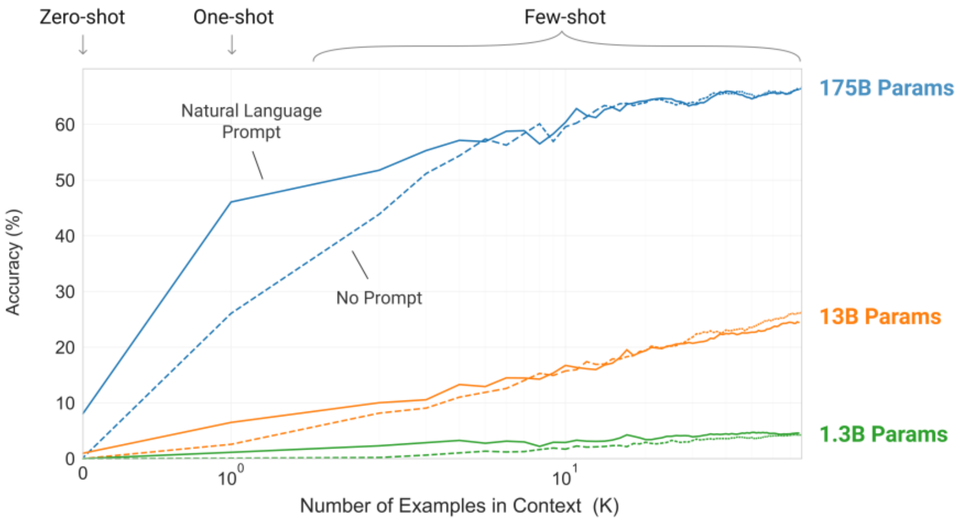
GPT (GENERATIVE PRETRAINED TRANSFORMER)

Transformer: Self-Attention



[Dosovitskiy et al., 2020](#)

- [Vaswani et al., 2017](#) - **Transformer** architecture
- [Radford et al., 2018](#) and [Brown et al., 2020](#)
- Simple *generative training* and *testing* procedure, perfectly suited for the *transformer* architecture.
- Very large model, very large data set

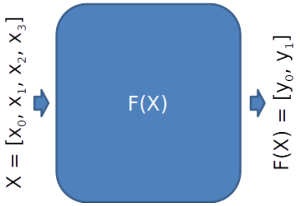


GPT-3 (Brown et al. 2020)

[To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:]
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

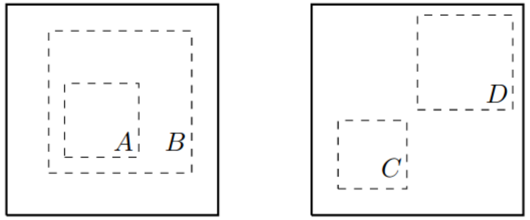
[A "yalubalu" is a type of vegetable that looks like a big pumpkin.
An example of a sentence that uses the word yalubalu is:]
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

1. The
2. Cat
3. Ran
4. Fast



1. Cat
2. Ran
3. Fast
4. <STOP>

The [P(duck), P(cat), P(fast), P(no), ...]
The cat [P(duck), P(cat), P(ran), ...]
The cat ran [P(fast), P(quickly), P(slowly, P(no) ...]



(a) Global and local views. (b) Adjacent views.

Right Stream

Left Stream

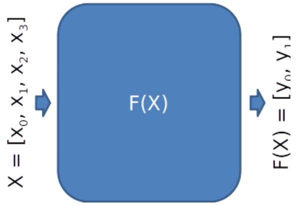
Targets

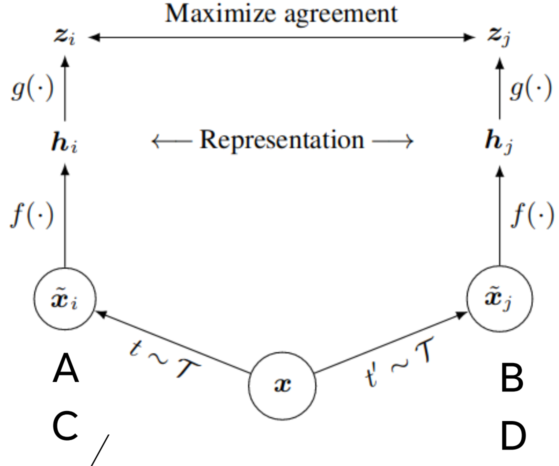
	B	D
A	1	0
C	0	1

\-

Predictions

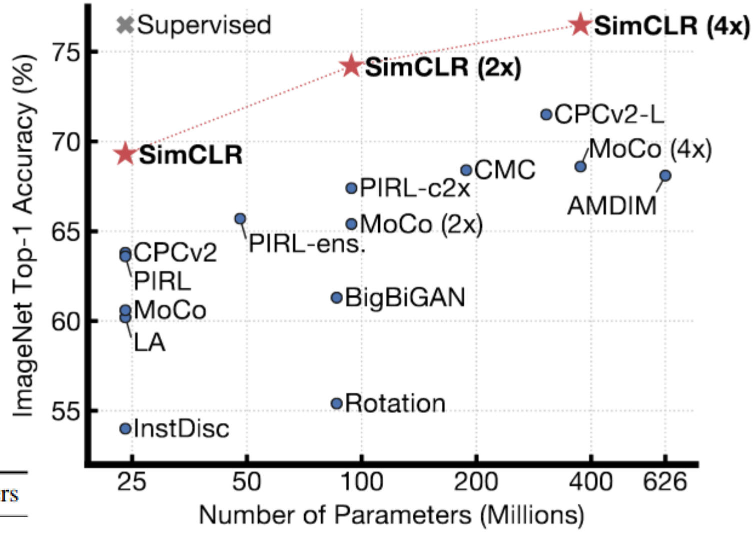
	B	D
A	1/2	1/2
C	1/4	3/4





CONTRASTIVE LEARNING

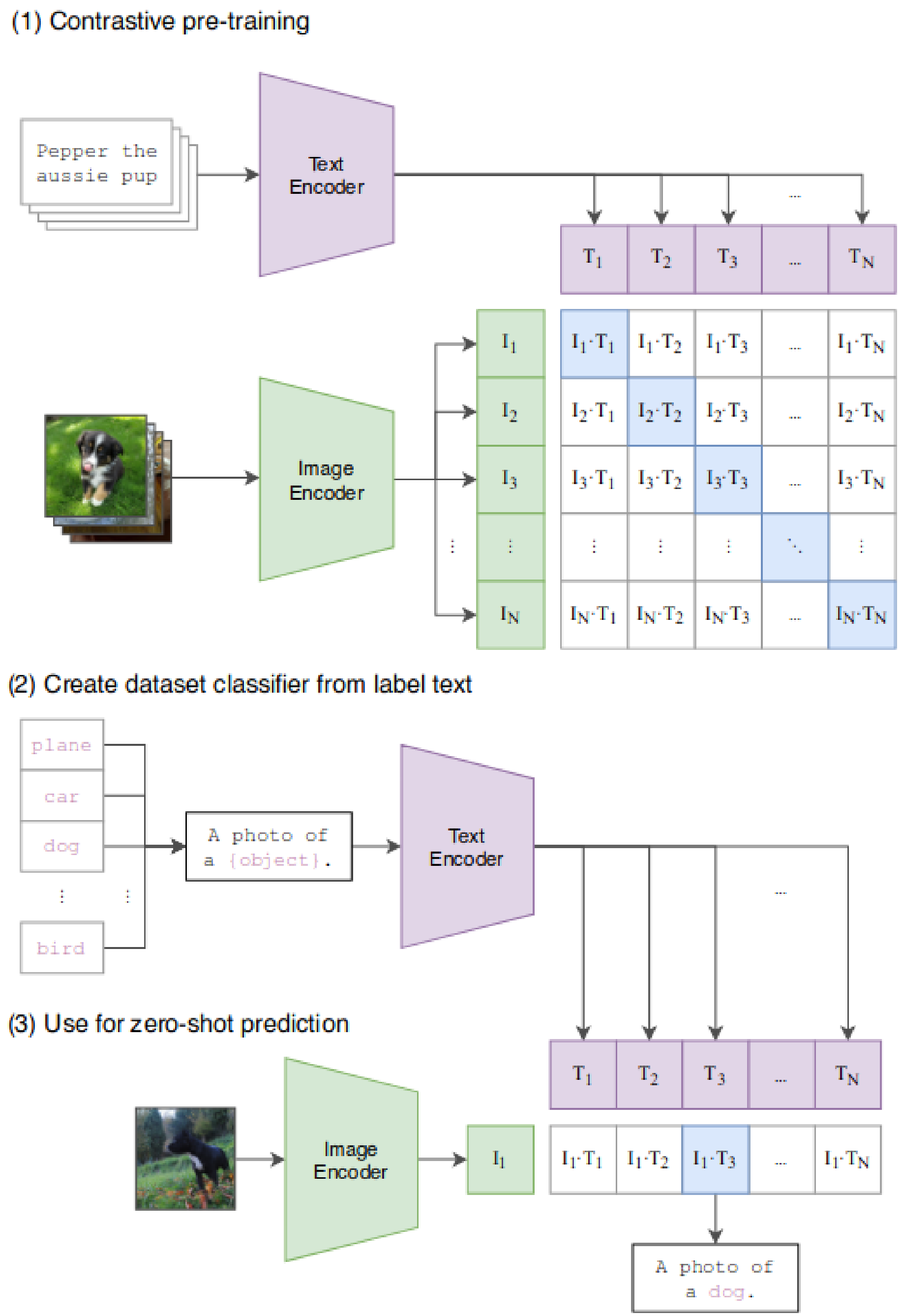
- Chen et al., 2020 "A Simple Framework for Contrastive Learning of Visual Representations"
- Technically, the process is *unsupervised* (more on that in a moment) because no targets are needed during the *training* process.
- A small sample of labeled images are needed during *testing* of the model (and deployment) making it *semi-supervised*
- Before 2020 we needed lots of *labeled* data: now we don't.



	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

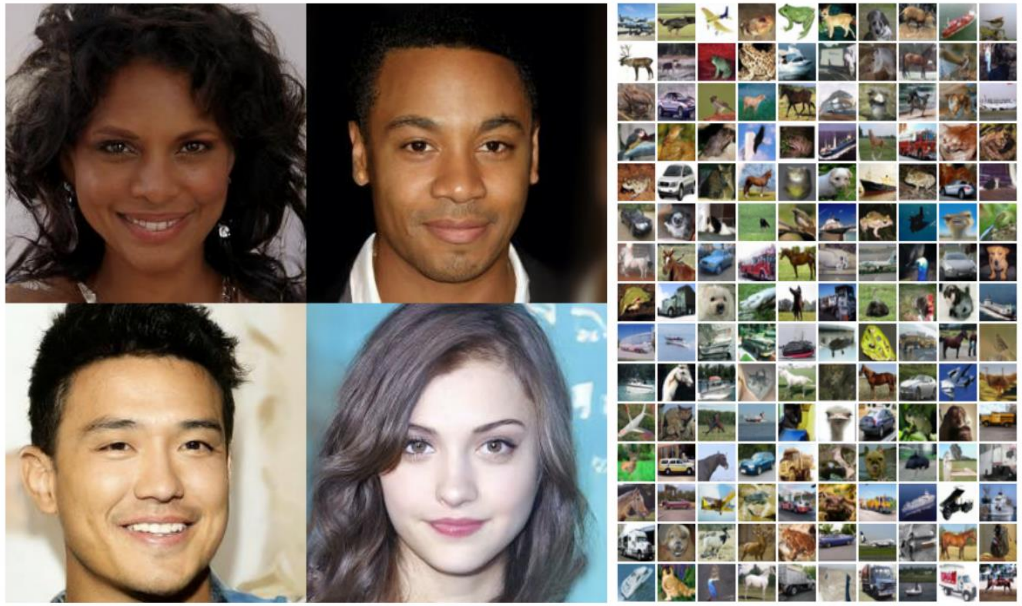
CONTRASTIVE LEARNING: ALIGNMENT ACROSS DOMAINS

- Radford et al., 2021
- Contrastive learning can be done to align the learned information from one domain to another related domain
- Image to text
- Text to image
- Doesn't have to just be these domains...
- Zero-shot prediction is possible, no additional training required for proper correct classification...

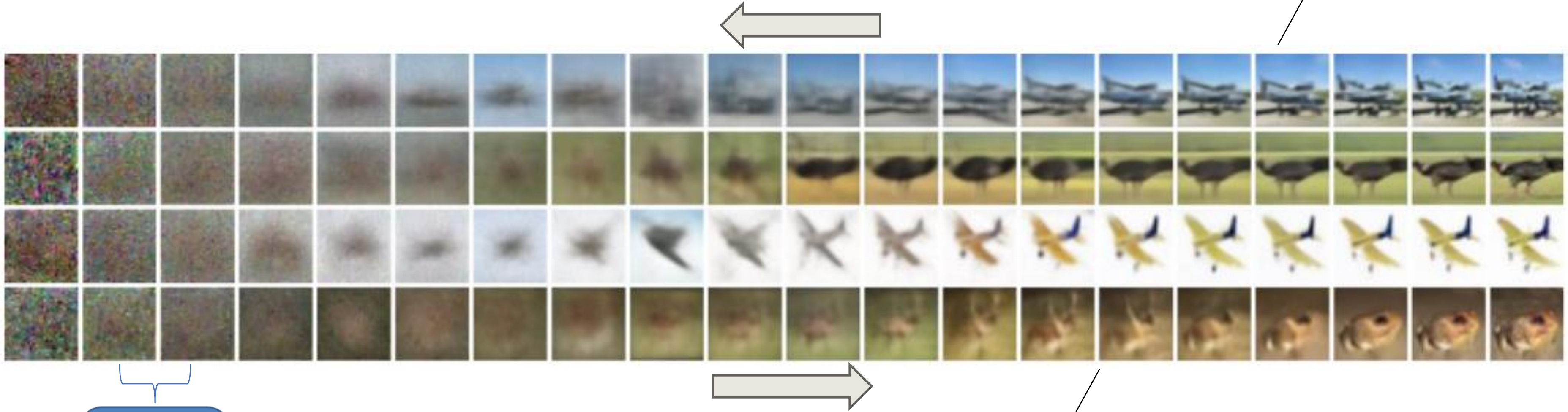


DIFFUSION MODELS

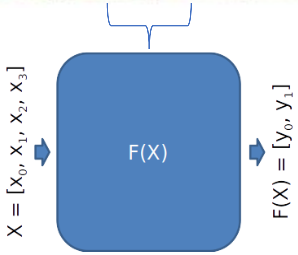
- Ho et al., 2020
- Denoising Diffusion Probabilistic Models (DDPM)
- An elegant solution to the *mode collaspe* issue with generative modeling tasks
- Sometimes called ***Stable Diffusion*** due to the correction of the mode collapse issue



Forward Process



Reverse Process

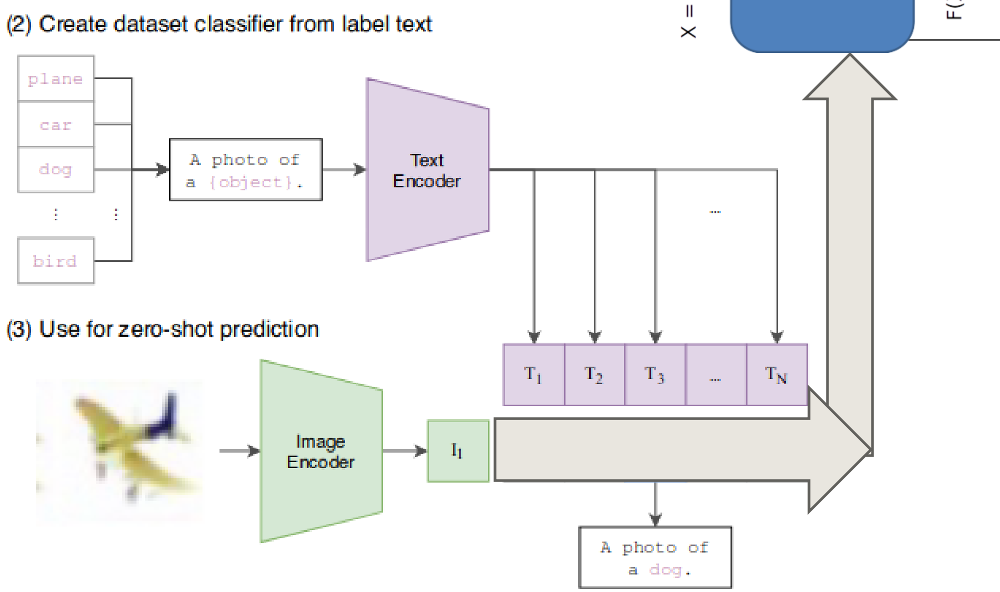


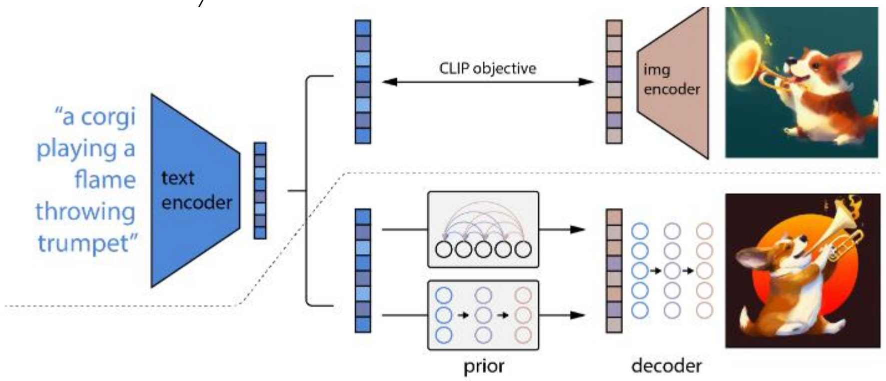
Separate Noise from the Signal (for each time step)

CONDITIONAL GENERATION WITH CONTRASTIVE EMBEDDINGS: DALL-E

- Ramesh et al., 2022
- Reverse diffusion process can be trained while including a *contrastive embedding*
- The text encoder can generate a *similar* contrastive embedding
- The diffusion model can use the text's contrastive embedding to extract a *similar* image
- Natural variation in the DDPM model and the large variation in training data allows for creative, generative modeling

Conditioned Diffusion Model






vibrant portrait painting of Salvador Dalí with a robotic half face


a shiba inu wearing a beret and black turtleneck


a close up of a handpalm with leaves growing from it


an espresso machine that makes coffee from human souls, artstation


panda mad scientist mixing sparkling chemicals, artstation


a corgi's head depicted as an explosion of a nebula