Self-Supervised Learning

COMPSCI 760
2024 Semester 1

Olivier Graffeuille

ogra439@aucklanduni.ac.nz

Outline

1. Motivation
2. Background
3. Methods for SSL
    1. What is a generative model?
    2. Generative models for SSL
    3. Contrastive methods for SSL

Outline

1. ***Motivation***
2. Background
3. Methods for SSL
    1. What is a generative model?
    2. Generative models for SSL
    3. Contrastive methods for SSL

Issues with Supervised Learning

- Supervised learning models have had great success in modelling various tasks in recent years
- However, these systems require *massive* amounts of *carefully labelled* data
  - Practically, it is impossible to label everything in the world
  - Some things don't have sufficient data (e.g. low-resource languages)
  - Supervised learning models tend to have narrow intelligence

Human vs Machine Learning

- It's clear that humans are able to learn things with far fewer examples
  - A self-driving car must drive off a cliff in a simulation thousands of times before learning not to
  - Children recognising animals
- We seem to have a lot of "background" knowledge about the world: "Common sense"?
- This ability to learn quickly comes from millions of years of "pre-training" via evolution

Self-Supervised Learning

- Allows models to learn from unlabelled data
  - Substantially more unlabelled data
  - Important to learn subtle knowledge and less common concepts
- Self-Supervised Learning is a Machine Learning "Paradigm"
  - Other topics recently covered including data streams, domain generalisation etc. can be though of as "settings" of ML

Motivation

"Self-Supervised Learning is one of the most promising ways to build such background knowledge and approximate a form of 'common sense' in AI systems"

- Yann LeCun

https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Outline

1. Motivation
2. ***Background***
3. Methods for SSL
    1. What is a generative model?
    2. Generative models for SSL
    3. Contrastive methods for SSL

Self-Supervised Learning (SSL)

- Learning by observation?
- We still want to make a prediction!
    - But without a human-defined label

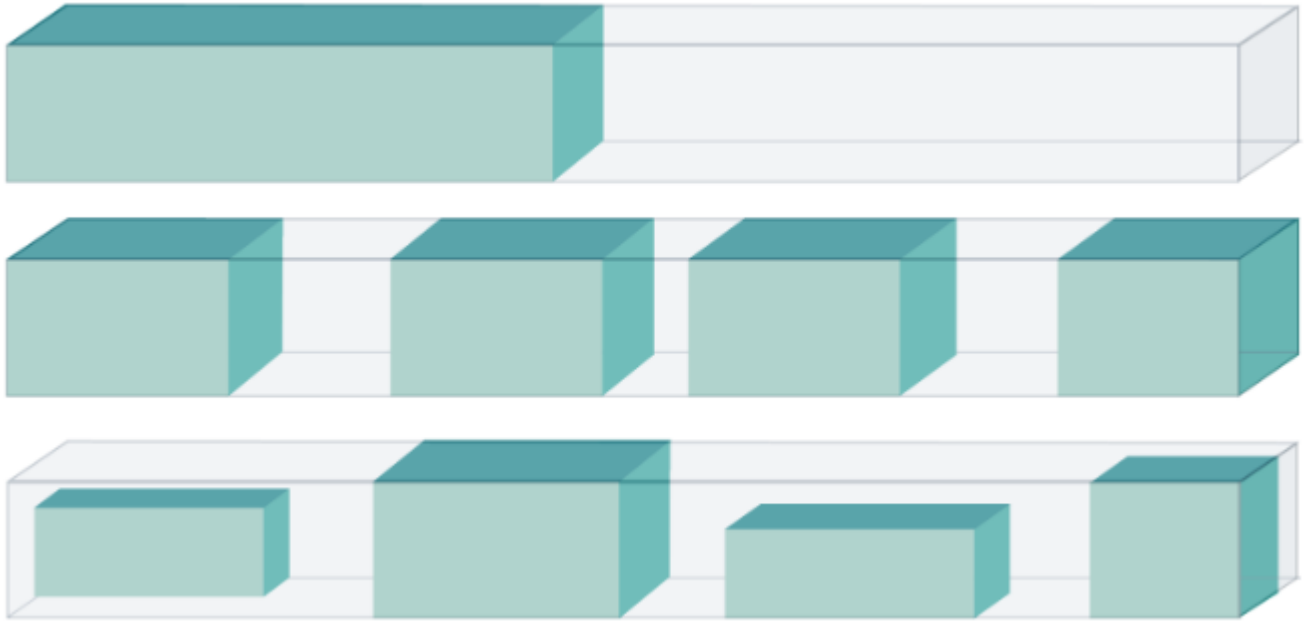
Self-Supervised Learning (SSL)

Learns to predict a part of the data from another part of the data.
Two steps:

1. Obtain "labels" from the data itself "semi-automatically"
    - Could be incomplete, transformed, distorted, or corrupted etc.
2. Predict something about the data from other parts of the data
    - E.g. the machine learns to 'recover' all or some parts of its original input
    - "Repairing" the data

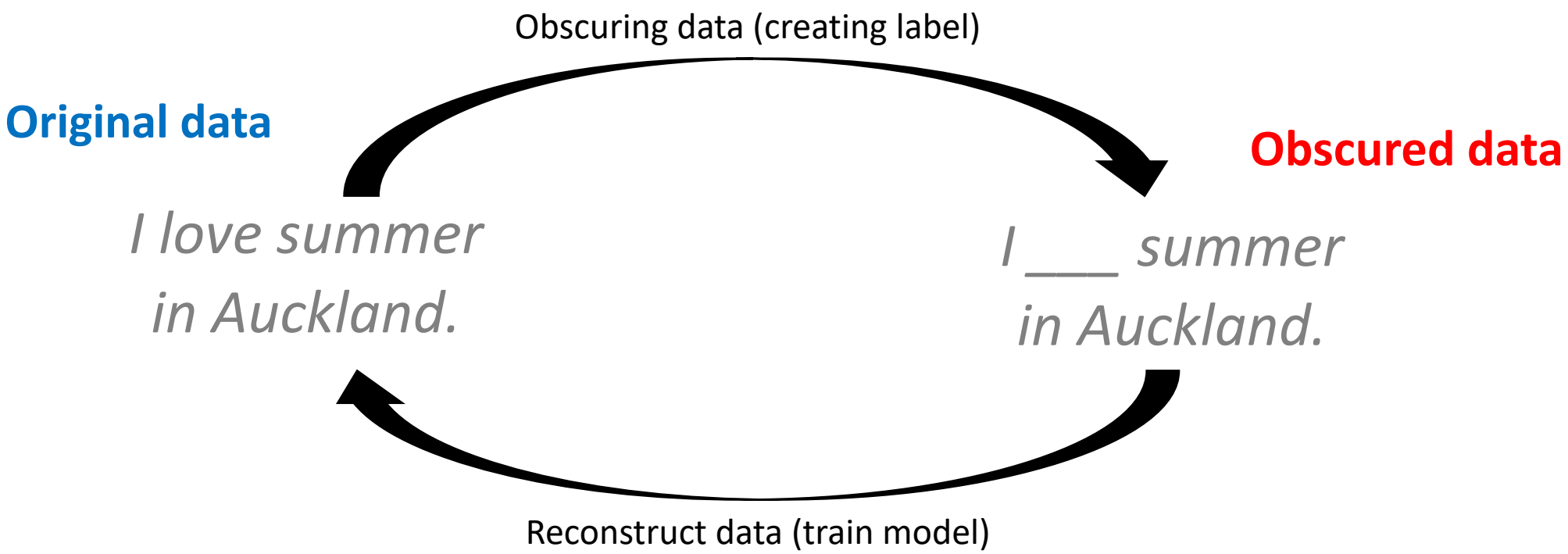
Obtaining Labels

- One recipe: hide part of the label



https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

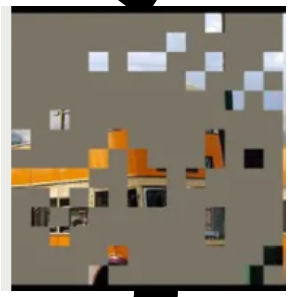
Obtaining Labels: Examples



Obtaining Labels: Examples

Obscuring data (creating label)

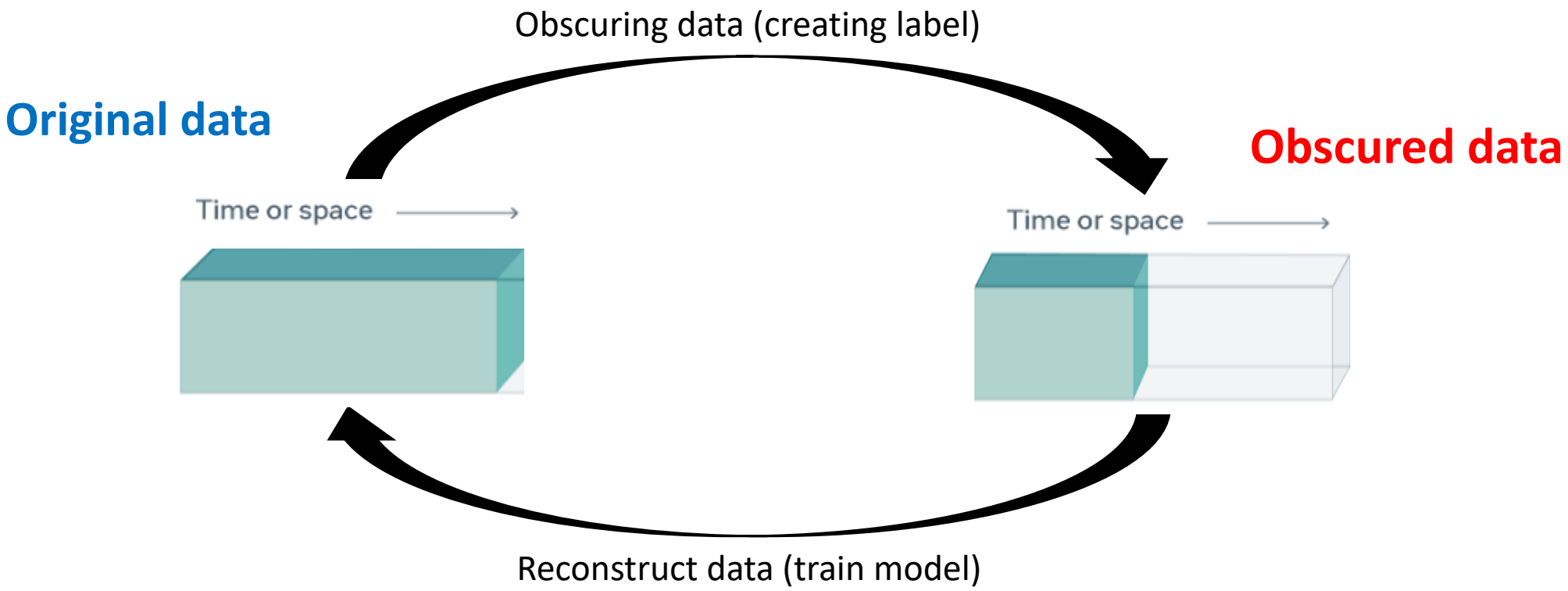
Original data



Obscured data

Reconstruct data (train model)

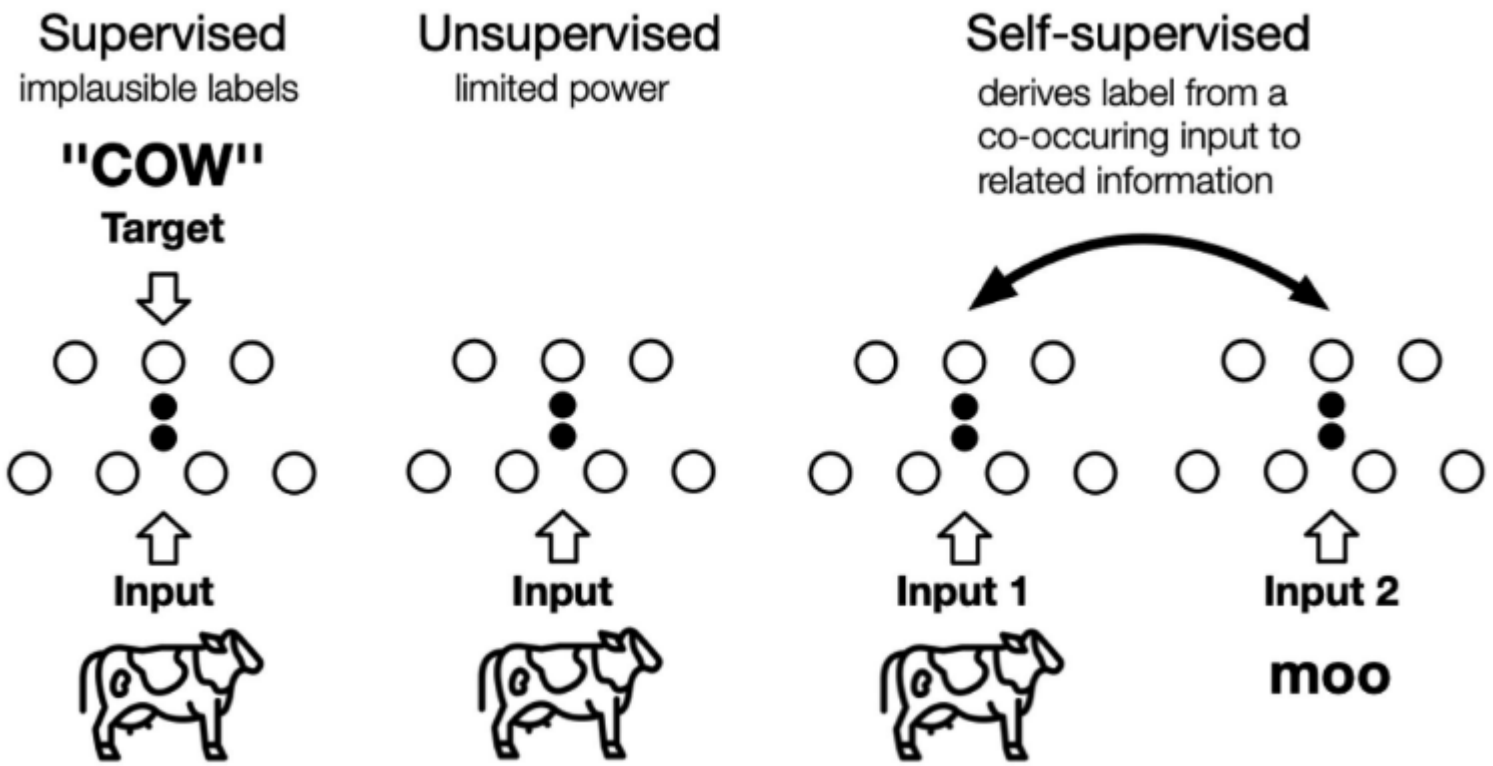
Obtaining Labels: Examples



Unsupervised Learning?

- If we are learning from unlabelled data, is this not just unsupervised learning?
    - E.g. clustering, dimensionality reduction, anomaly detection
- "Self-Supervised Learning" is used to imply that we are still learning from supervision
    - SSL is not unsupervised at all
    - We are still making predictions and learning from many labels
        - Arguably, we are making richer predictions

Unsupervised Learning?



Obtaining Labels

- One of the challenges of supervised learning is human bias in labels
- There is still a human bias in SSL labels, since humans are defining the labelling process. Do we want the model to:
  - Fill in gaps?
  - De-noise?
  - Predict the future?

https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Why is this useful?

- This is not just useful to repair data
- It is useful for learning **meaningful representations** from huge quantities of data

High-level representation of a $1 bill

Low-level representation of a $1 bill

Why is this useful?

- This is not just useful to repair data
- It is useful for learning **meaningful representations** from huge quantities of data
  - If we can predict the future of a sentence, we have probably learnt the representations required to label that sentence

"Despite its flaws, I have come to ___ this product." Self-prediction

⋮

"Despite its flaws, I have come to love this product." -> Positive sentiment Classification

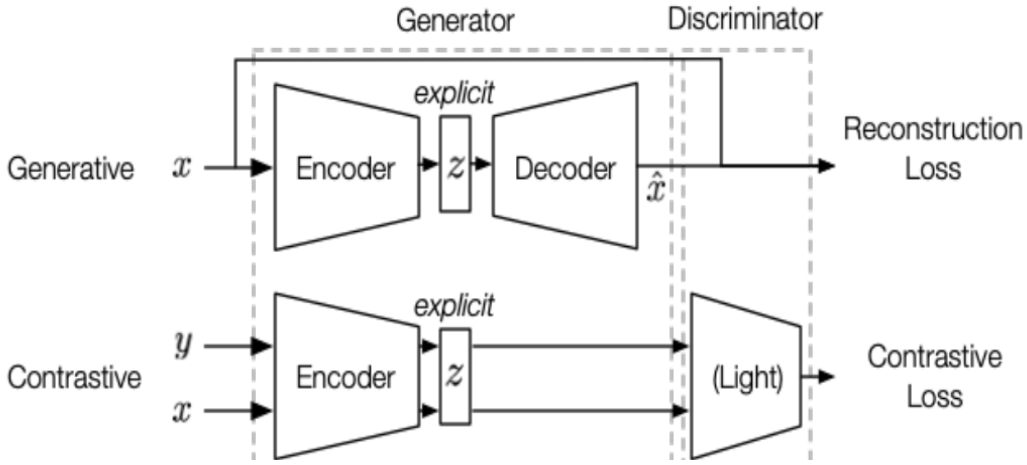
  - SSL can have more feedback signals from labels than supervised

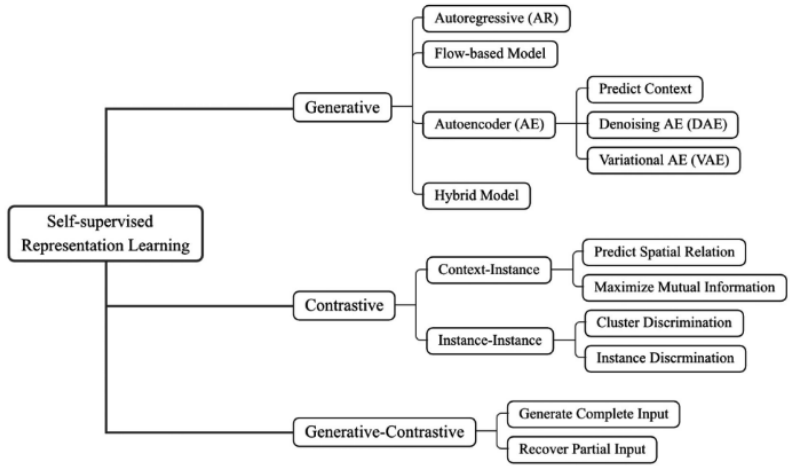
Outline

1. Motivation
2. Background
3. ***Methods for SSL***
    1. What is a generative model?
    2. Generative models for SSL
    3. Contrastive methods for SSL

Self-Supervised Learning Approaches

- Generative Methods
  - Reconstruct the original data
  - Typically encoder/decoder architecture
- Constrastive Methods
  - Compare embeddings of input data
  - Typically encoder architecture
- Generative-contrastive methods





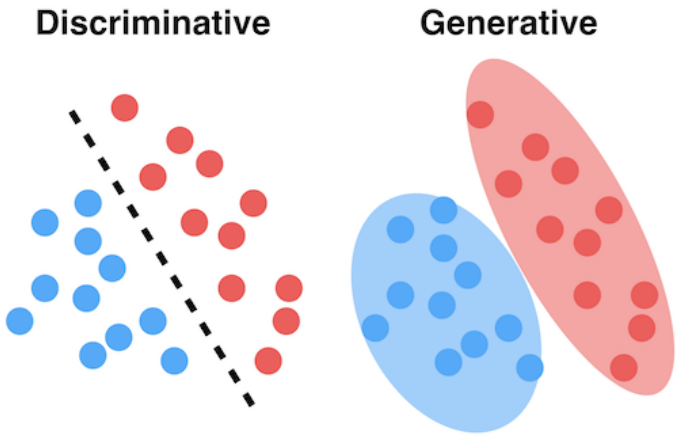
Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1), 857-876.

Outline

1. Motivation
2. Background
3. Methods for SSL
    1. ***What is a generative model?***
    2. Generative models for SSL
    3. Contrastive methods for SSL

Generative vs. Discriminative Methods

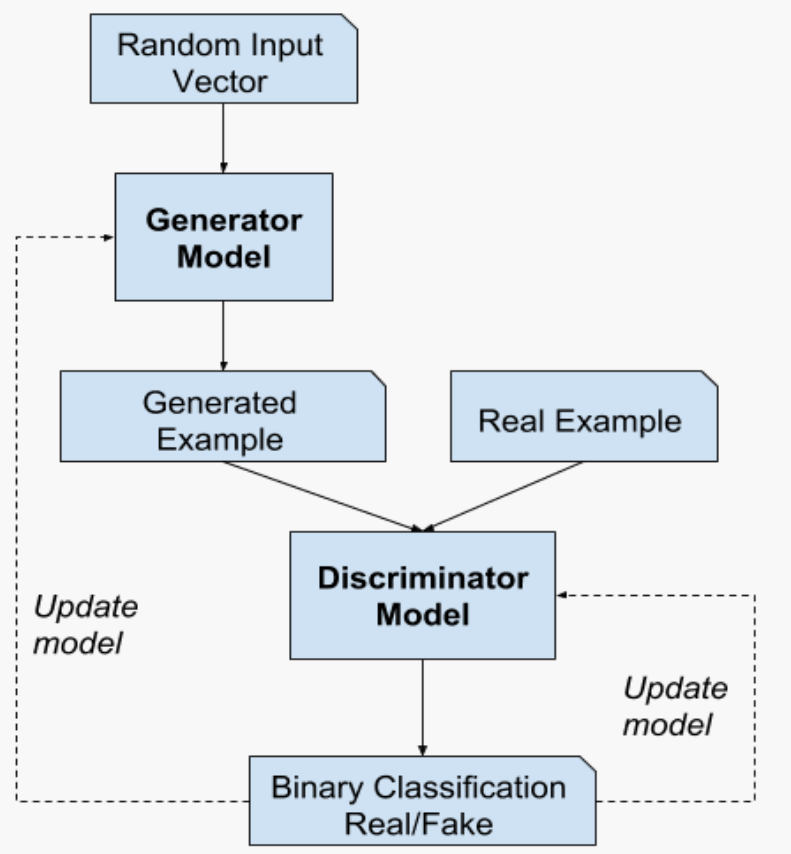
- You will be used to discriminative models:
  learn conditional probability P(y|X)
- Generative models achieve a different goal:
  learn joint probability P(X,y)
- In other words: learn the data probability density directly



Generative Adversarial Networks (GANs)

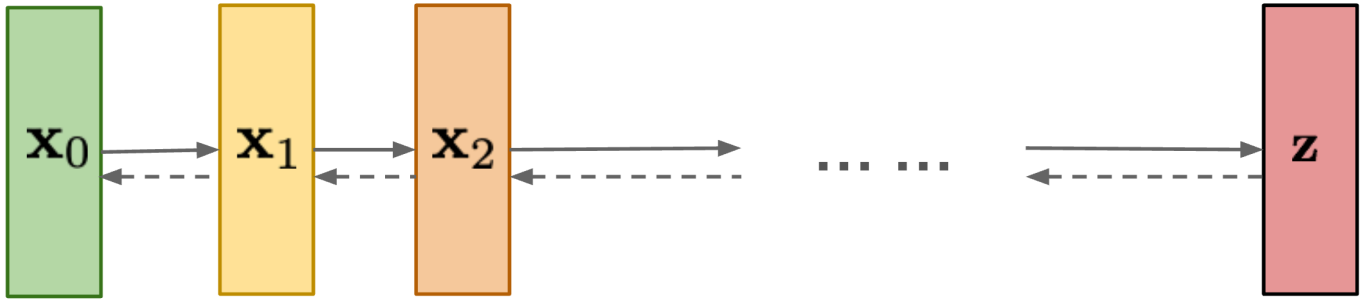
- Generator
  - Generate fake samples, trying to fool the discriminator
- Discriminator
  - Discriminating real and fake samples

- Disadvantages:
  - Difficult to train, unstable
  - Mode collapse



Diffusion Models

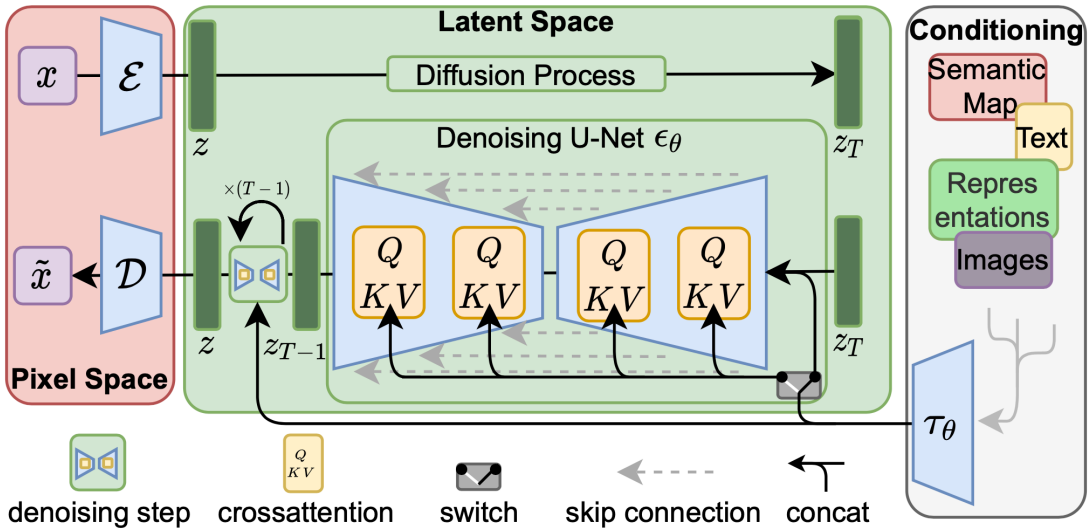
1. Add some noise to data, train model to remove noise
2. Repeat many times
3. Input noise
4. ???



Diffusion models:
Gradually add Gaussian noise and then reverse

Diffusion Models

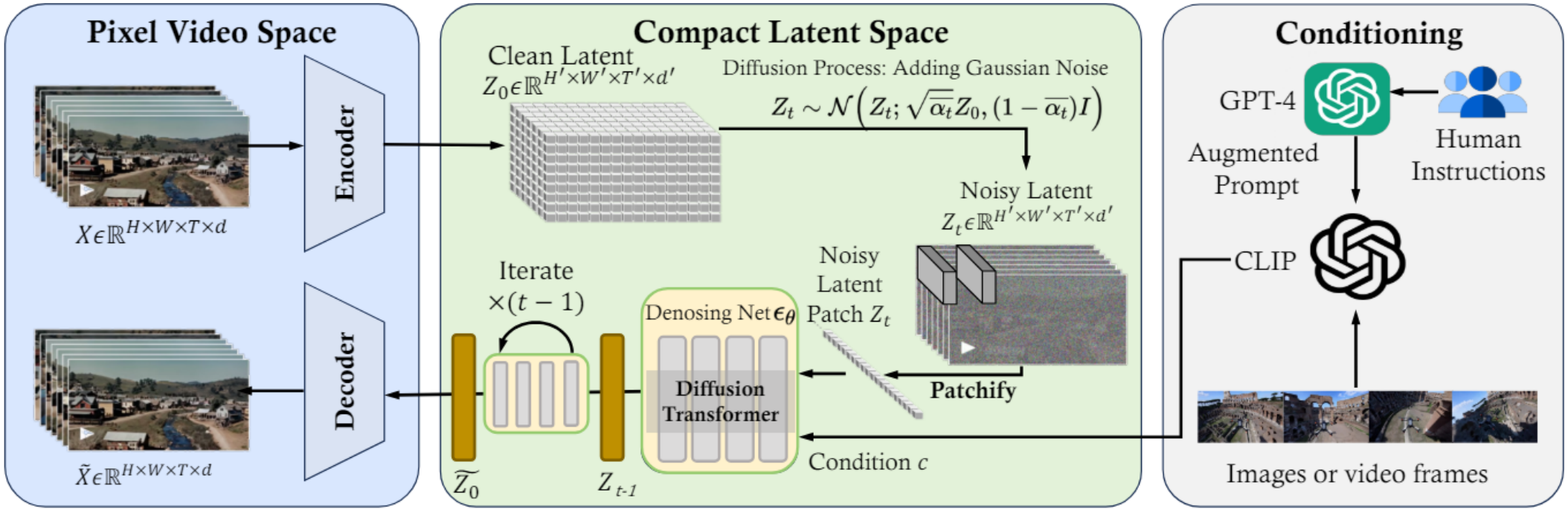
- Overall architecture involves more steps
- Conditioning: allows us to "direct" the generating process



Diffusion Models

https://stability.ai/blog/stable-diffusion-public-release

Video Diffusion Models



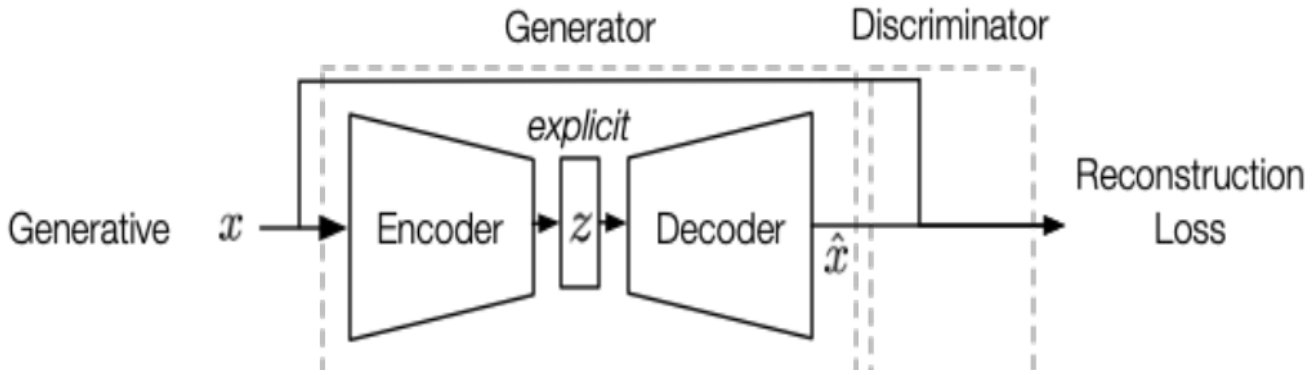
Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., ... & Sun, L. (2024). Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv preprint arXiv:2402.17177*.

Outline

1. Motivation
2. Background
3. Methods for SSL
    1. What is a generative model?
    2. ***Generative models for SSL***
    3. Contrastive methods for SSL

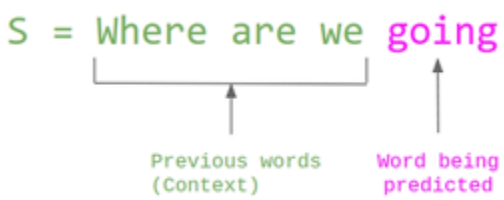
Generative Methods

- Reconstructs a part of the input data
- More exactly:
  - Train an encoder to encode input x into an explicit vector, and a decoder to reconstruct x from z



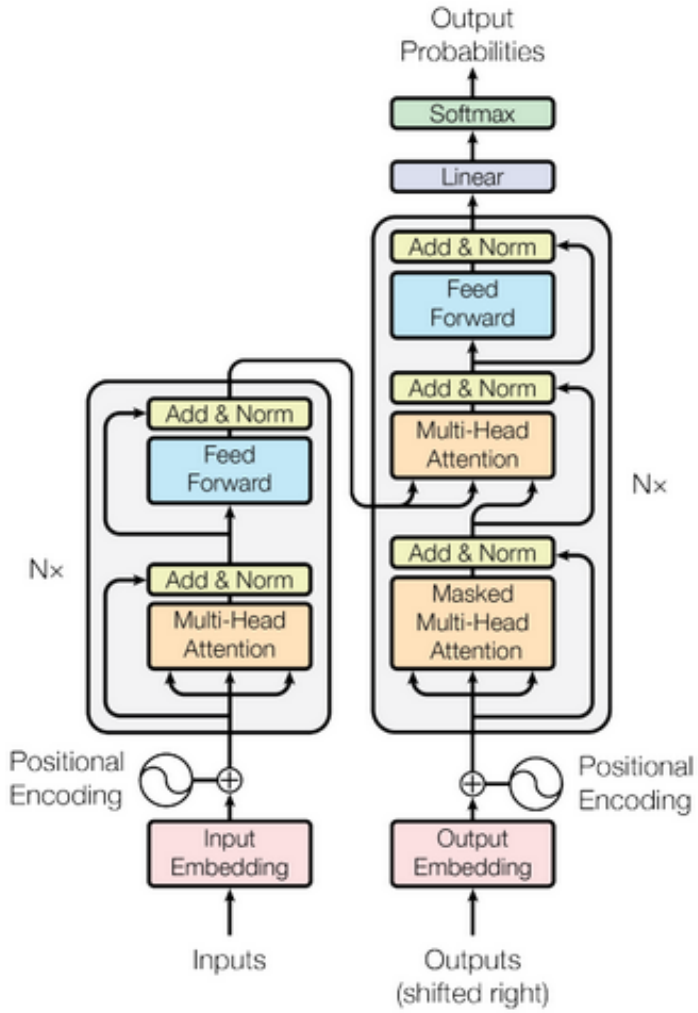
Autoregressive models

- A feed-forward network that predicts future values based on past values
- RNN, LSTM and transformers (and timeseries)



Transformers, GPTs

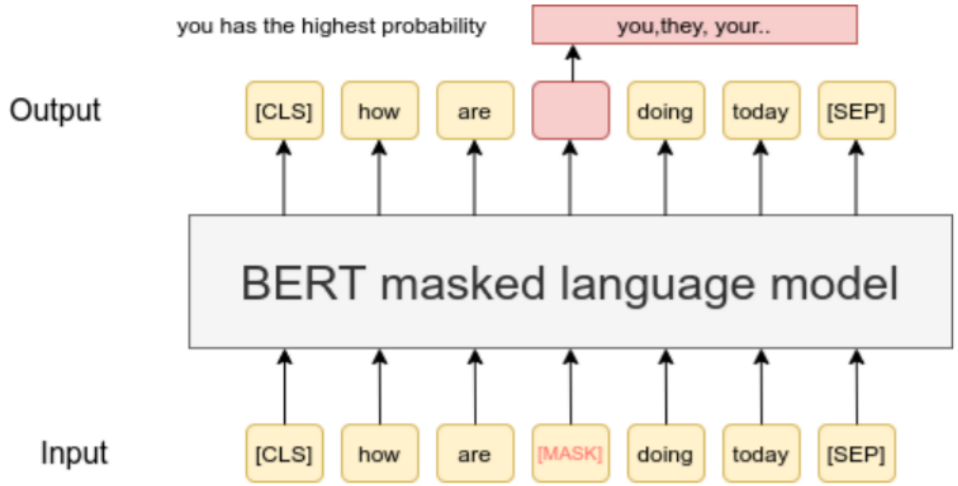
- Most famous examples of autoregressive models
  - Trained on a large portion of the internet
  - Trained to predict the next word
- Fine-tuned with Reinforcement Learning with Human Feedback (RLHF)
  - To behaves more like a useful chatbot, rather than just completing text



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Masked Language Models (MLM)
E.g. BERT

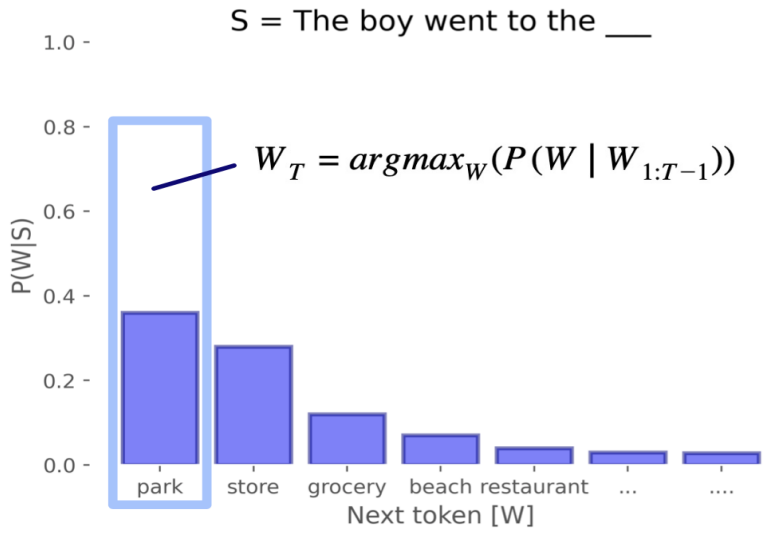
- Mask a part of the input with special [MASK] token
- Get model to determine missing token



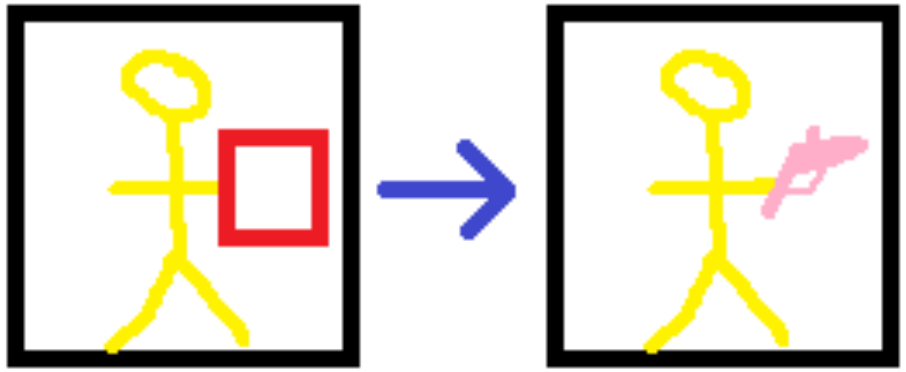
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

SSL for Language vs Vision Tasks

- Let's think of how we generate text versus images:



Language Generation



Vision Generation

https://blog.allenai.org/a-guide-to-language-model-sampling-in-allennlp-3b1239274bc3?gi=ead91fb1c828

SSL for Language vs Vision Tasks

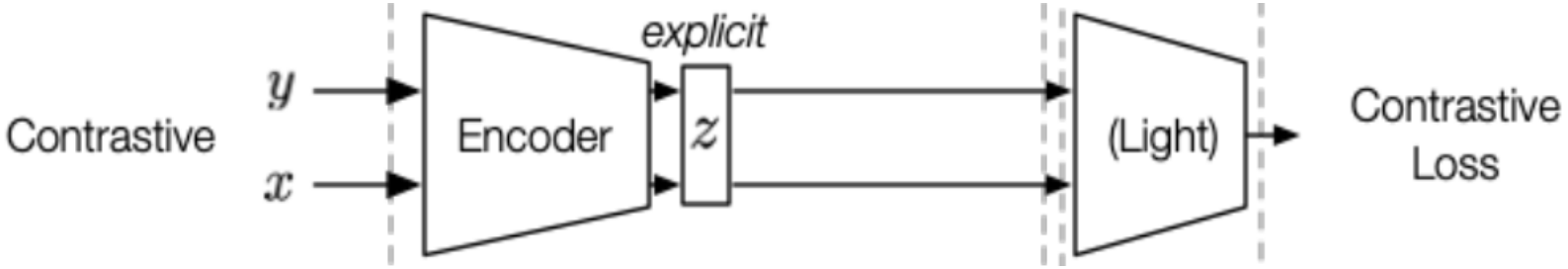
- Language tasks are much better able to model uncertainty
  - We can give a distribution over every possible word (maybe 50,000 possibilities?)
  - We cannot generate a distribution over every possible image ($2^{512x512x3}$ possibilities)

Outline

1. Motivation
2. Background
3. Methods for SSL
    1. What is a generative model?
    2. Generative models for SSL
    3. ***Contrastive methods for SSL***

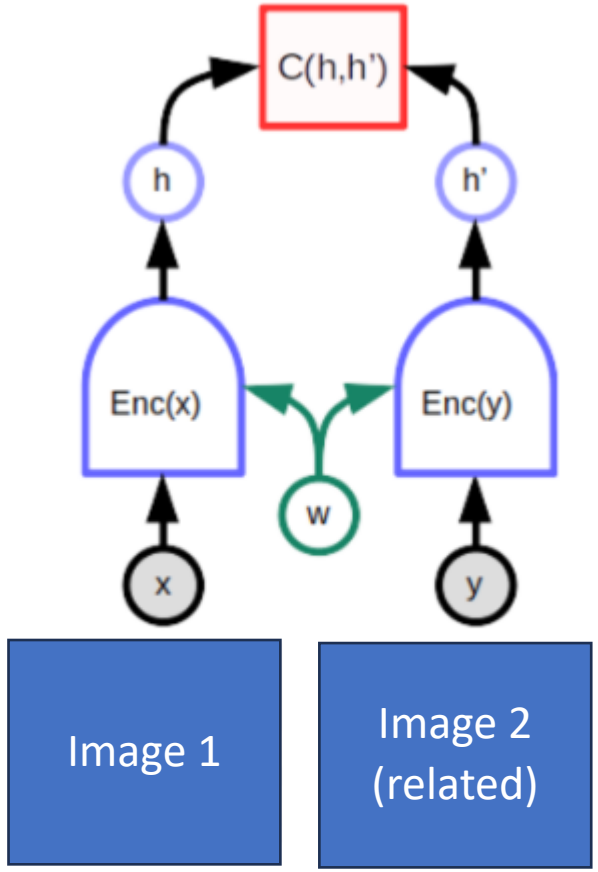
Contrastive Methods

- Train an encoder to encode input x into an explicit vector z to measure similarity
  - No longer generating hidden part of the input
  - Instead, we predict a hidden property (latent variables) about the input
- The goal is still to learn meaningful hidden representations



Origins: Siamese Networks

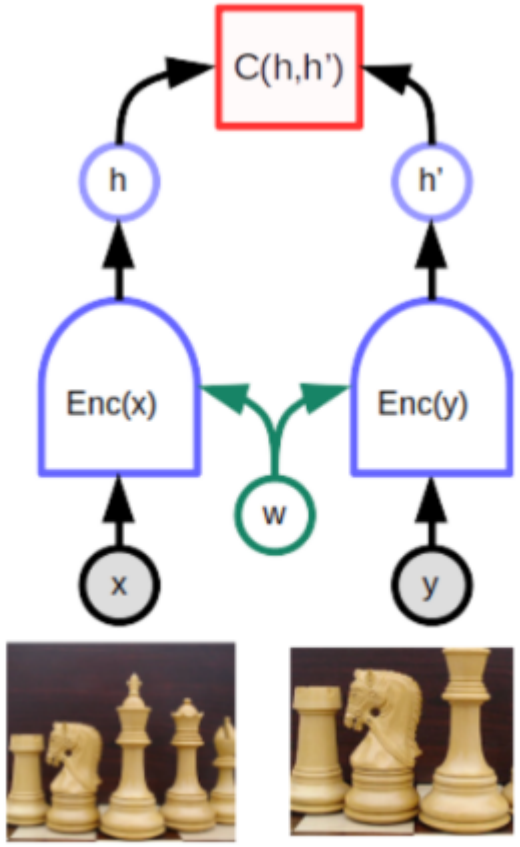
- Pass two *related* inputs through two identical neural networks
- Compare the generated latent embeddings generated by each network



https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

How do choose related inputs?

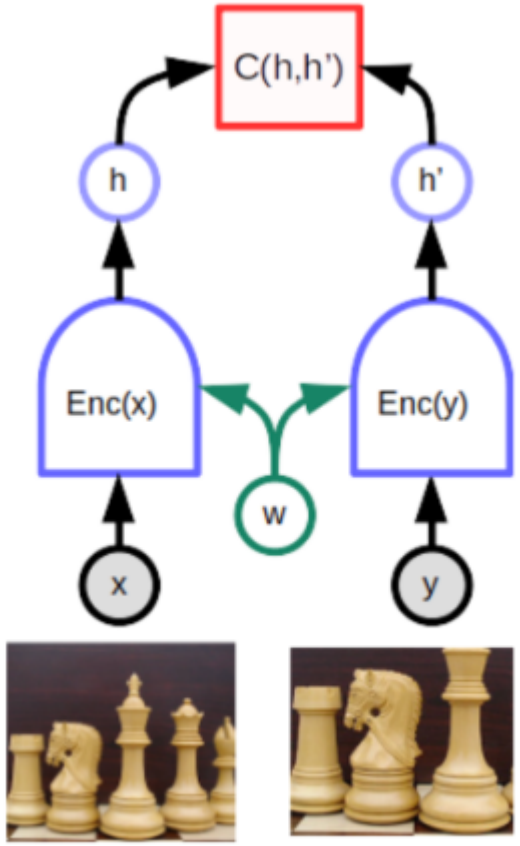
- Remember, we have no labels
- We can find two related inputs by transforming an image into two inputs
  - (most common) two crops of same image
  - Changing colour, noise, etc…
  - Transformation is important
- We have a problem…



https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

The Problem: Mode Collapse

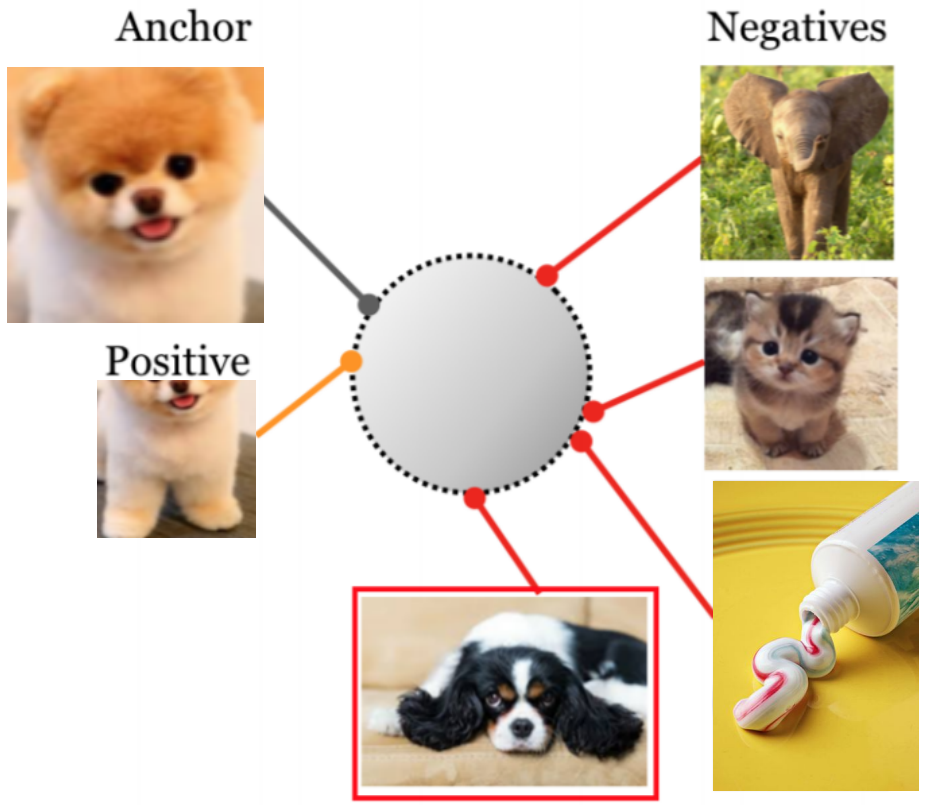
- Nothing is stopping the network from generating a trivial embedding
  - Ignoring the input
  - Output the same embedding to get a good loss score
- This is called mode collapse
  - Think back to GANs…



https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

The Solution: Contrastive Learning

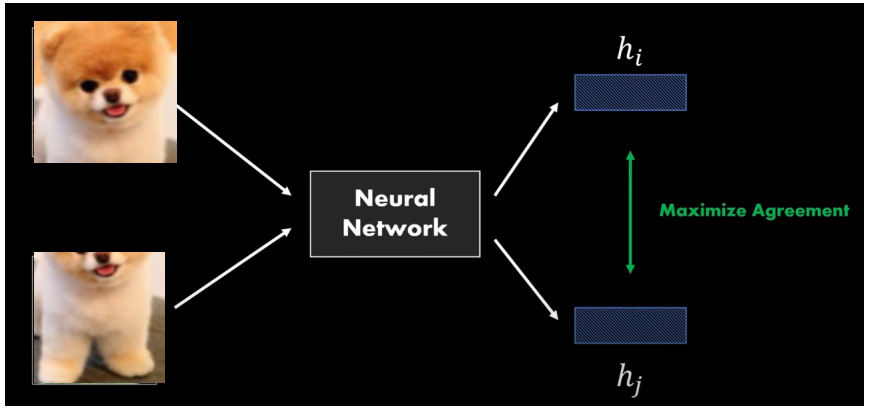
- Take two *unrelated* inputs and make sure they generate different embeddings
- How to find unrelated inputs?
  - Remember: still no labels
  - We just transform any other image

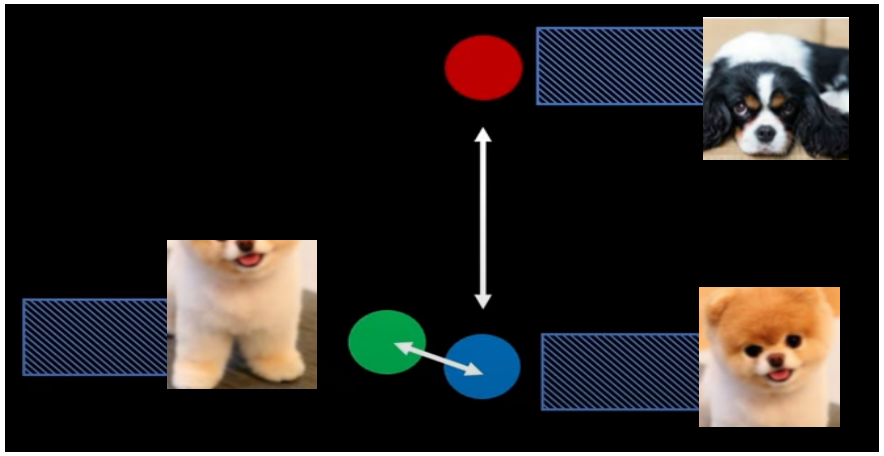


The Solution: Contrastive Learning

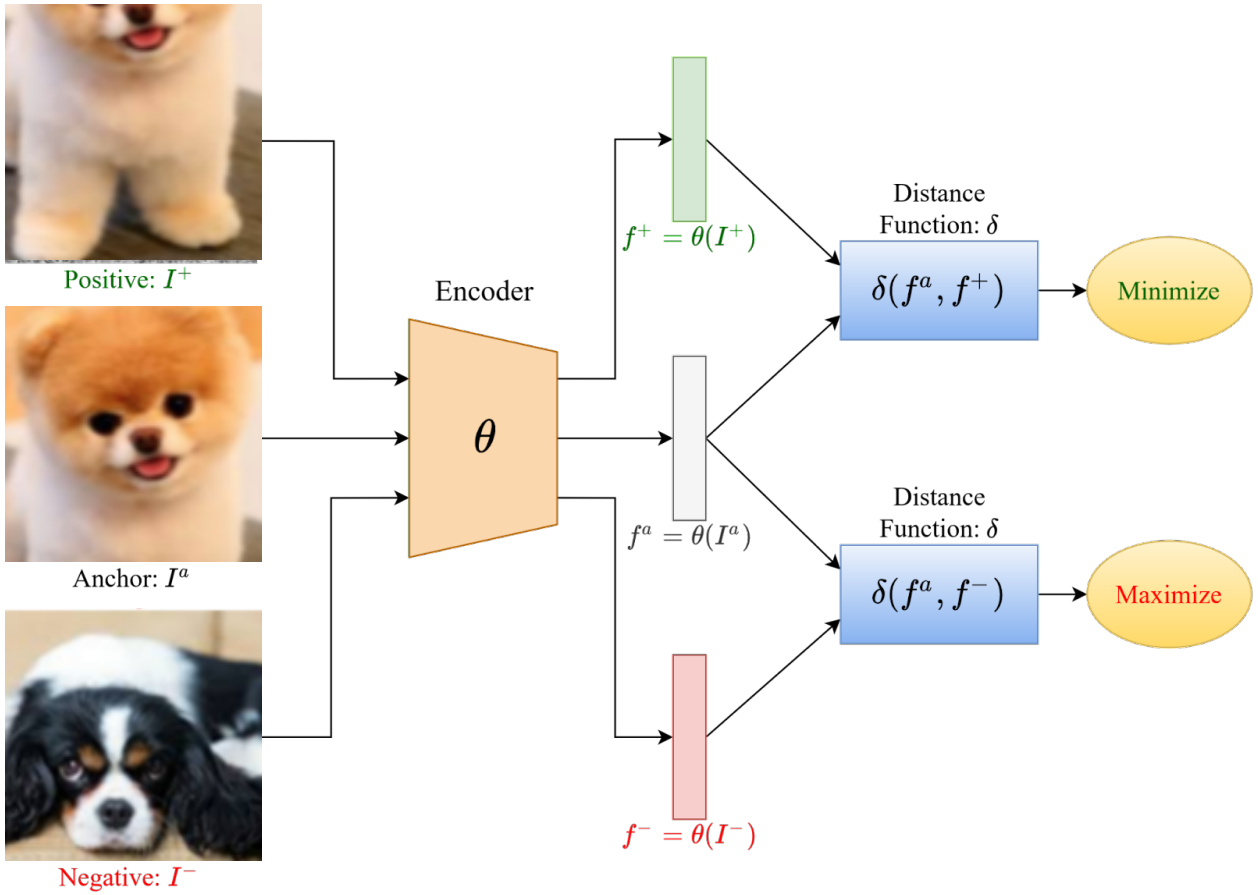
Simple implementation:

1. Get anchor data point
    - Choose positive sample
    - Choose negative sample
2. Calculate Euclidean distance between output space of input pairs
3. Minimise distance for related pairs
4. Maximise distance for unrelated pairs



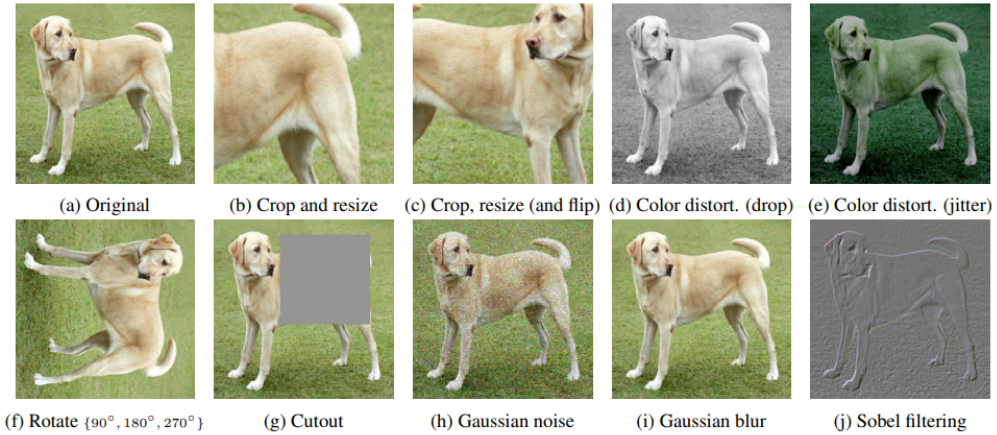


The Solution: Contrastive Learning



SimCLR
(Simple Contrastive Learning of visual Representations)

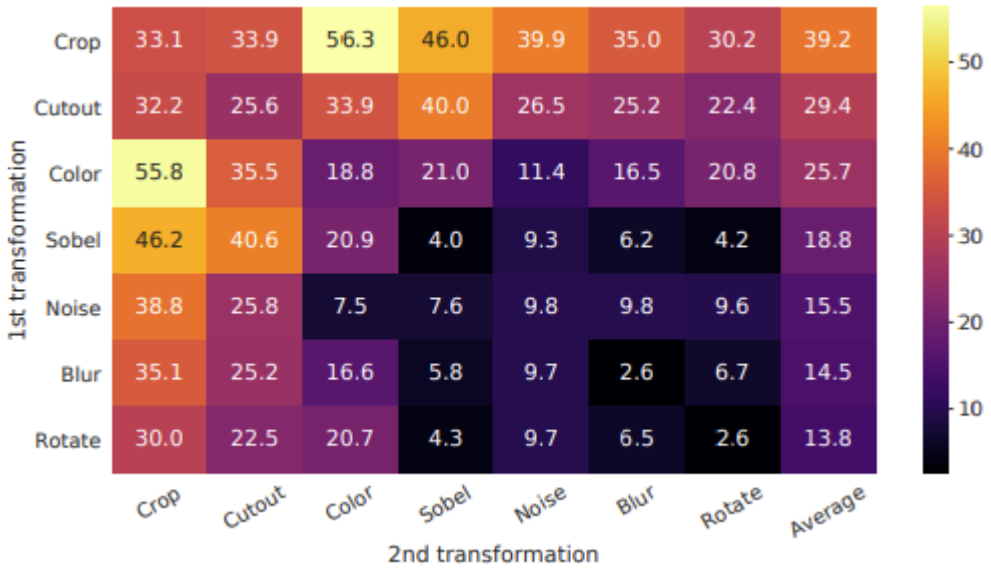
- For positive and negative examples, generate many transformations of each image
  - Focus on transformations: not just cropping
  - Focus on hard positive samples



Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

SimCLR
(Simple Contrastive Learning of visual Representations)

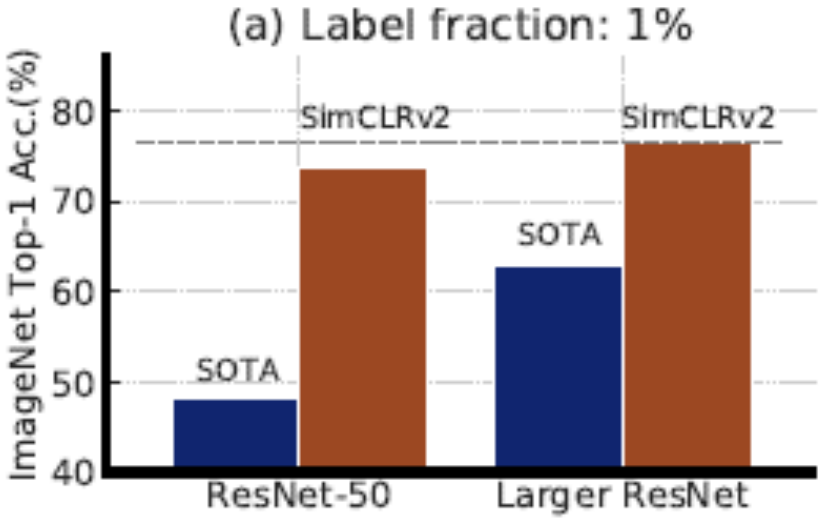
- What transformations to use?



Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

SimCLRv2

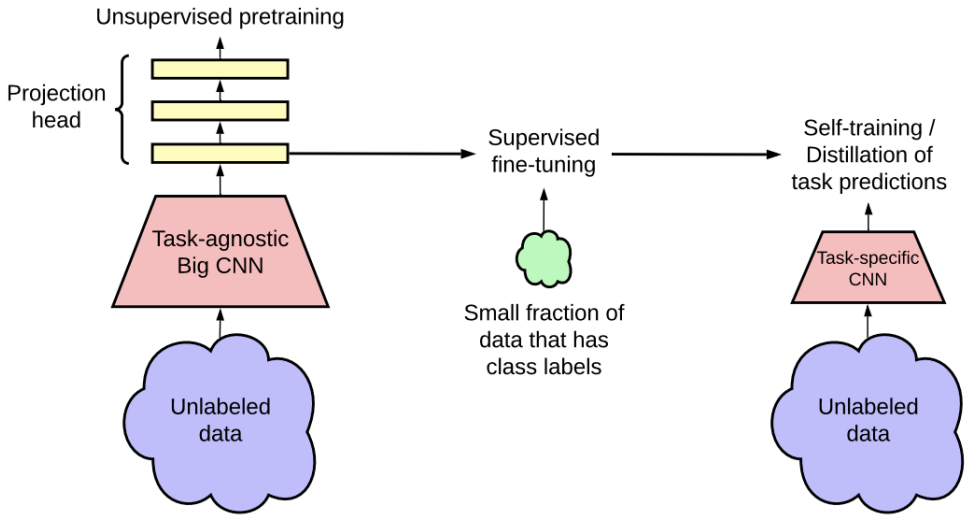
- Develop SimCLR further by applying it to supervised downstream tasks
- Works very well



Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *Advances in neural information processing systems* 33 (2020): 22243-22255.

SimCLRv2

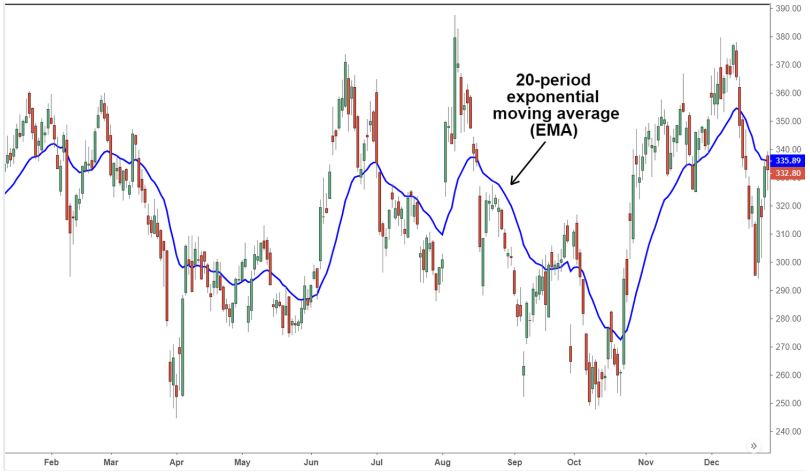
- Phase 1
  - Initial self-supervised pre-training (same as SimCLR with projection head)
- Phase 2
  - A supervised fine-tuning for specific tasks
- Phase 3
  - Unsupervised self-distillation further improves performance



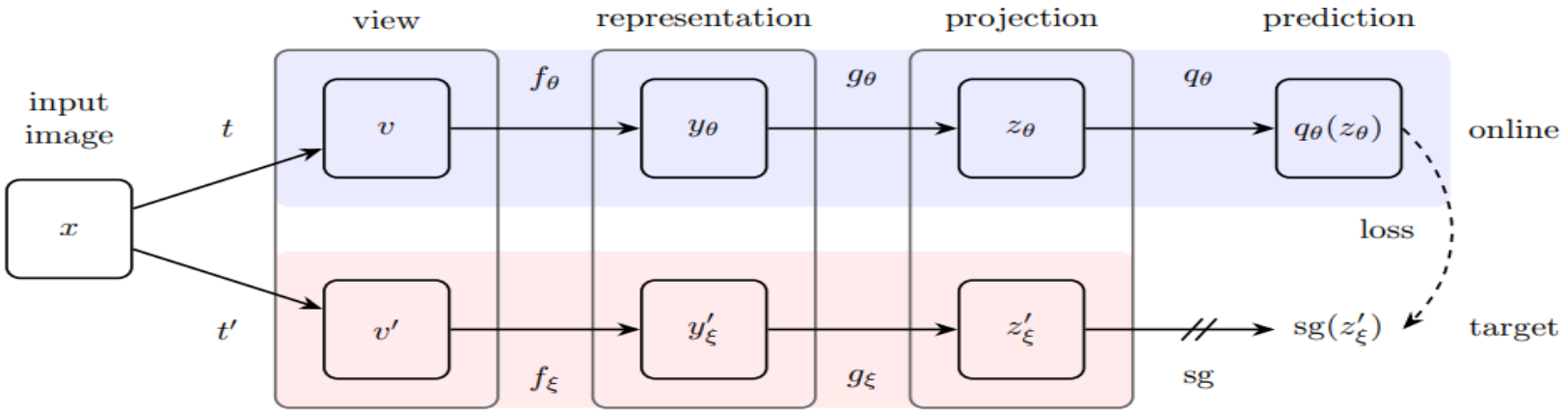
Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *Advances in neural information processing systems* 33 (2020): 22243-22255.

Contrastive Methods

BYOL
(Bootstrap Your Own Learning)



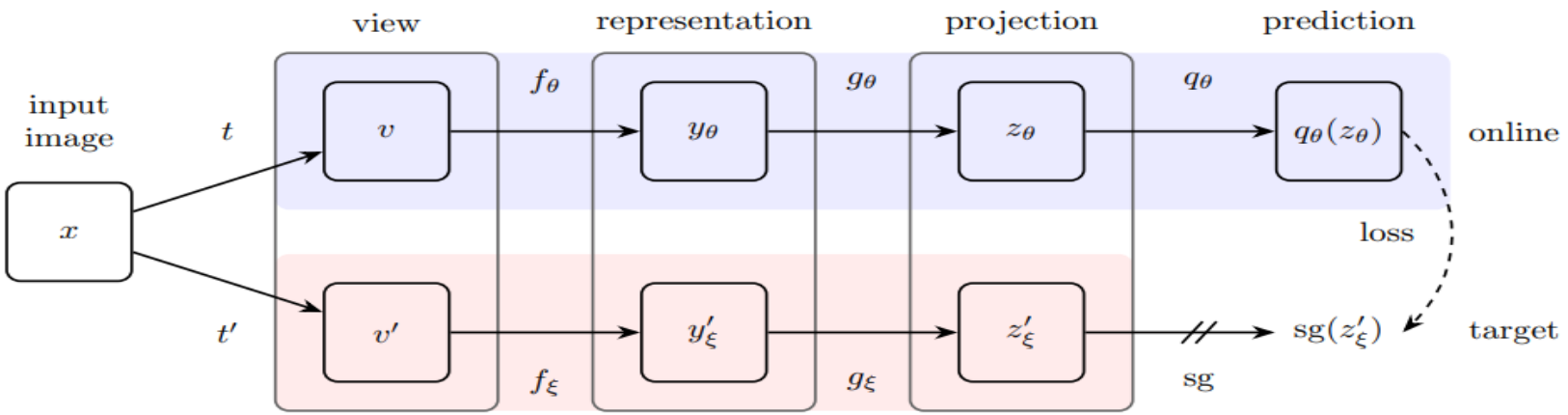
- Uses two models
  - Online model is a typical neural network, trained continuously
  - Target model is more stable: exponential moving average



Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271-21284.

BYOL
(Bootstrap Your Own Learning)

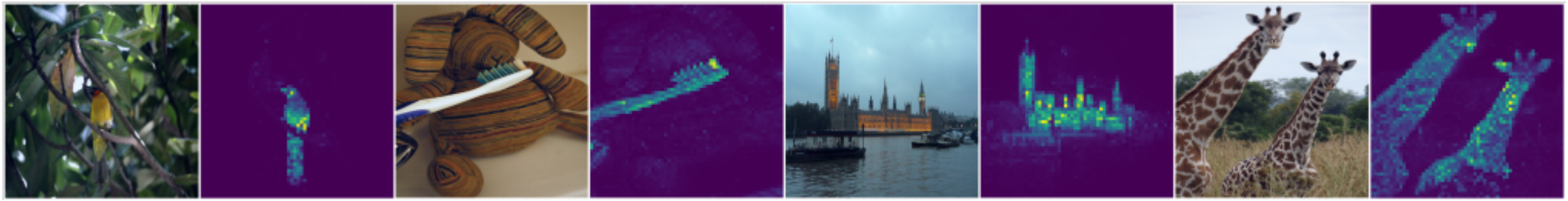
- No negative examples to stop collapse
- Collapse seems to be stopped by stability from target network



Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271-21284.

Emerging Properties in Self-Supervised Vision Transformers

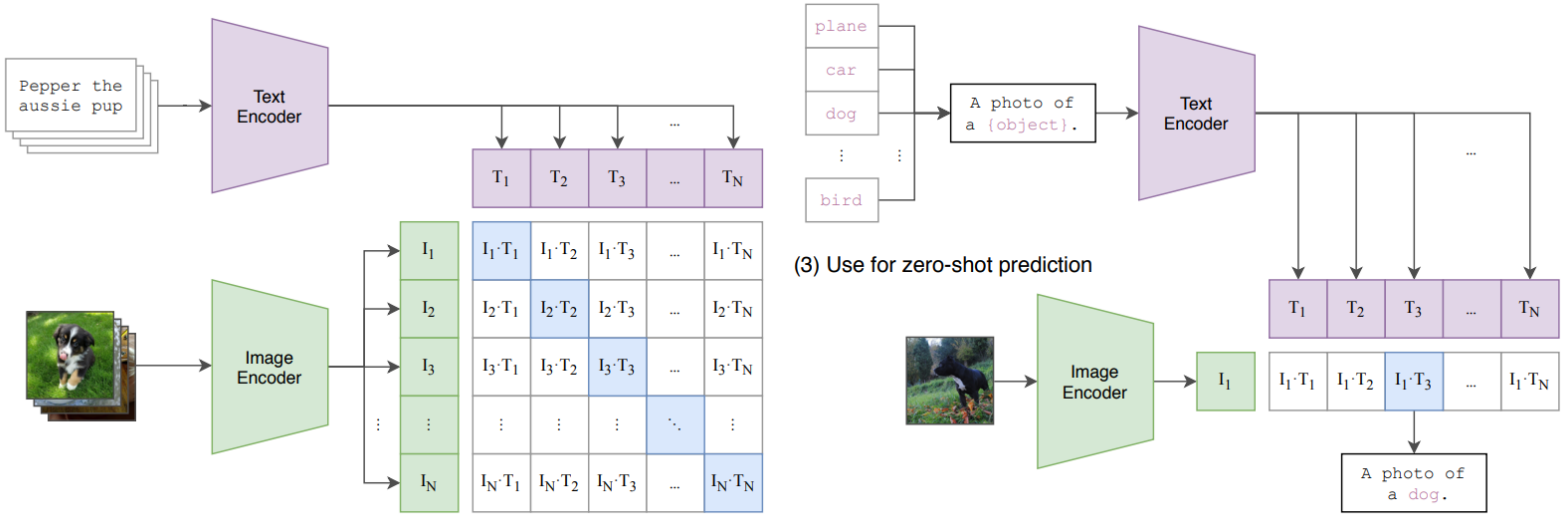
- Training Vision Transformers with Self-Supervised Learning
  - Attention tokens learns informative class features
  - Applying kNN to learnt representation achieves 78.3% top-1 accuracy on ImageNet, without finetuning



Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650-9660).

Multi-Modal Representations

- CLIP aligns text representations with visual representations
- A different form of supervision using image captions



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Some thoughts

- SSL has been behind all the recent AI hype
- Huge quantities of unlabelled data have allowed for more "intelligent" systems
- Existing systems are (presumably) learning from most of human knowledge
    - How will this scale in the future?

Announcements

- We'll announce which research paper readings are examinable next week (recap session)
- SET evaluations
  - Set evaluations are out for a couple of weeks
  - Please give us useful feedback (including what you liked!)