Introduction to Self-Supervised Learning

Roberto Souza

(Slides courtesy of Peyman Tahghighi)

W2025

Learning objectives

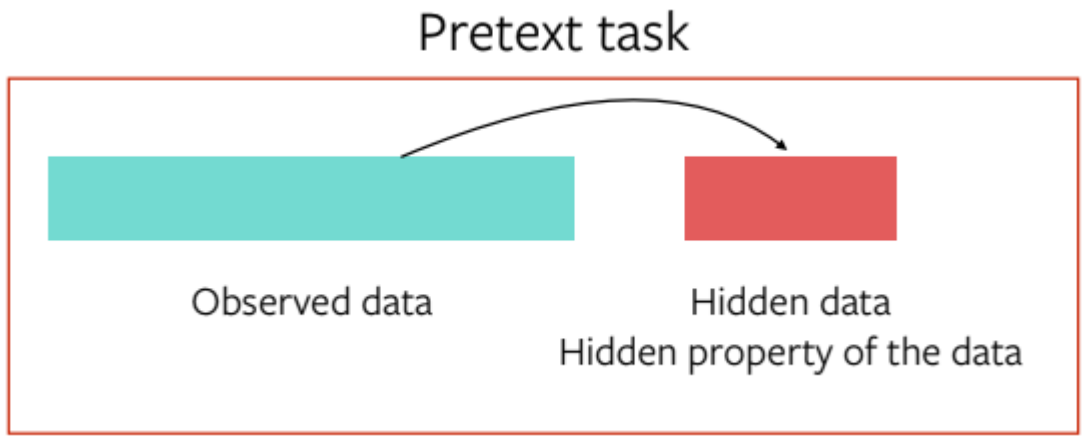
- What is self-supervised learning?
- How we can use it in NLP and Vision?
- How it can benefit us?

Self-supervised learning: The dark matter of intelligence

- Supervised learning is a bottleneck for building more intelligent generalist models that can do multiple tasks and acquire new skills without massive amounts of labelled data.
- As babies, we learn how the world works largely by observation.
- We learn new skills by short training, e.g. driving.
- A working hypothesis is that generalized knowledge about the world, or common sense, forms the bulk of biological intelligence in both humans and animals.
- Common sense helps people learn new skills without requiring massive amounts of teaching for every single task.

What is self-supervision?

- Obtain “labels” from the data itself and leverage the underlying structure in data.
- Predict unobserved or hidden parts.



Examples of data generation in NLP

GPT

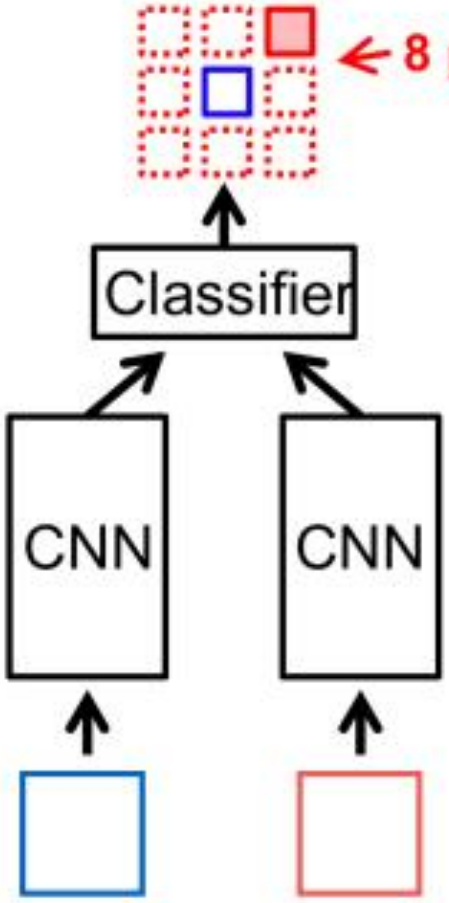
A	?							
A	quick	?						
A	quick	brown	?					
A	quick	brown	fox	?				
A	quick	brown	fox	jumped	?			
A	quick	brown	fox	jumped	over	?		
A	quick	brown	fox	jumped	over	the	?	
A	quick	brown	fox	jumped	over	the	lazy	?
A	quick	brown	fox	jumped	over	the	lazy	dog

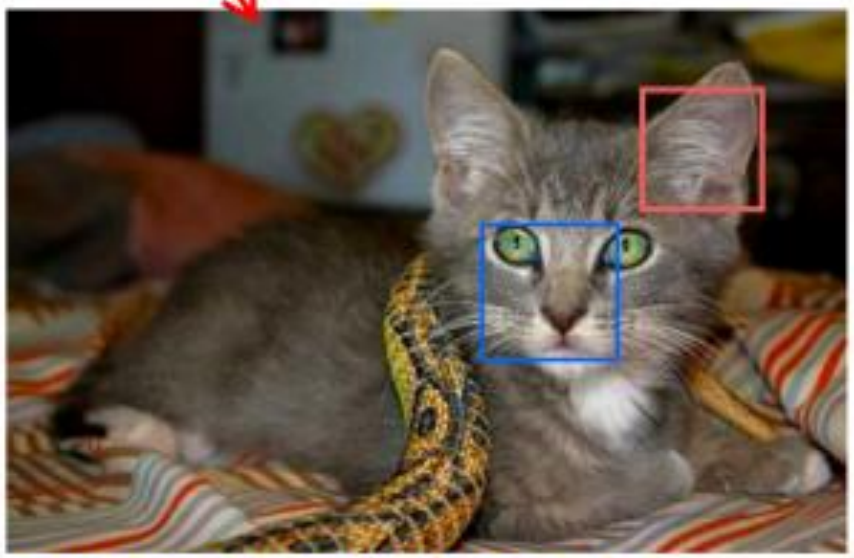
BERT

A	?	brown	fox	?	over	the	?	dog
A	quick	?	fox	jumped	over	?	lazy	?
A	quick	brown	?	jumped	over	the	?	dog
A	?	brown	fox	?	?	the	lazy	dog

Examples of data generation in vision

Relative position of patches





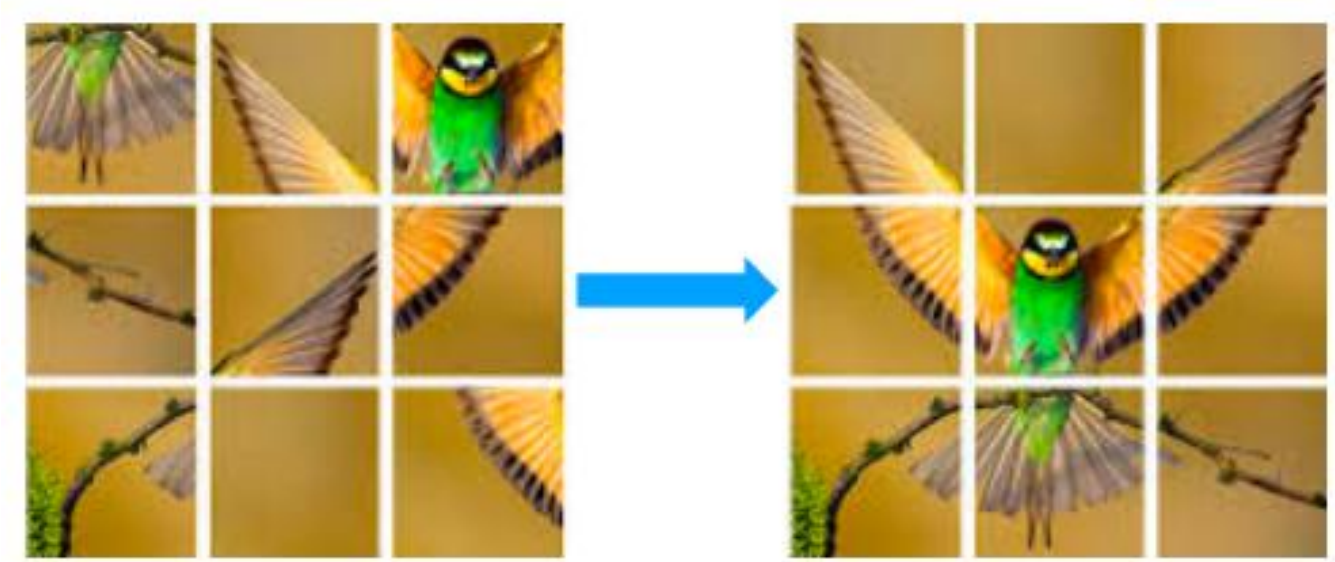
Randomly Sample Patch
Sample Second Patch

Input: Two patches
Output: 8-way classification

Doersch et al., 2015, Unsupervised visual representation learning by context prediction

Examples of data generation in vision

Jigsaw puzzle



Jigsaw puzzles
(Noorozi & Favaro, 2016)

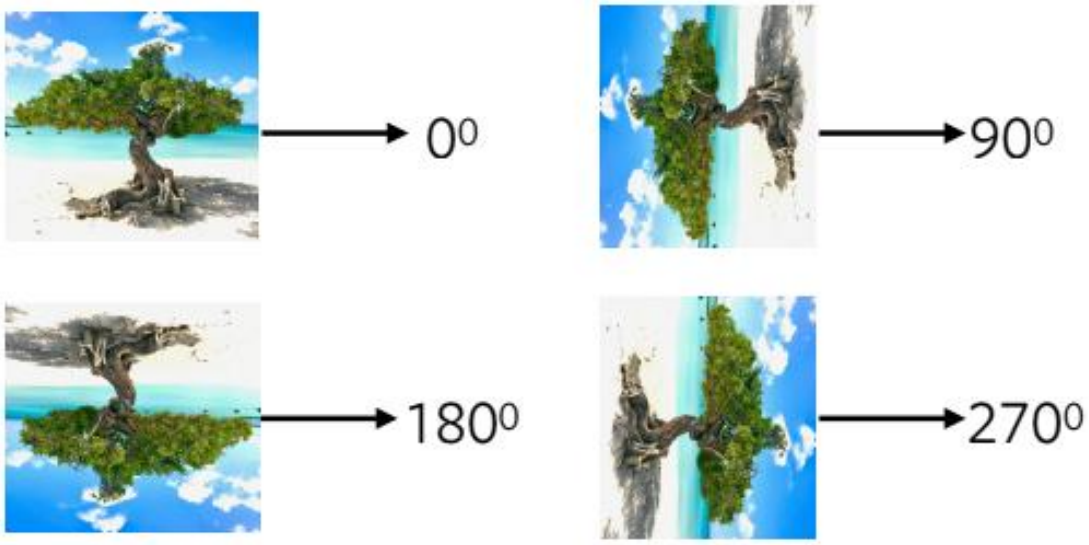
Input: nine patches
Permute using one of N permutations

Output: N-way classification

Set N << 9!

Examples of data generation in vision

Rotation prediction



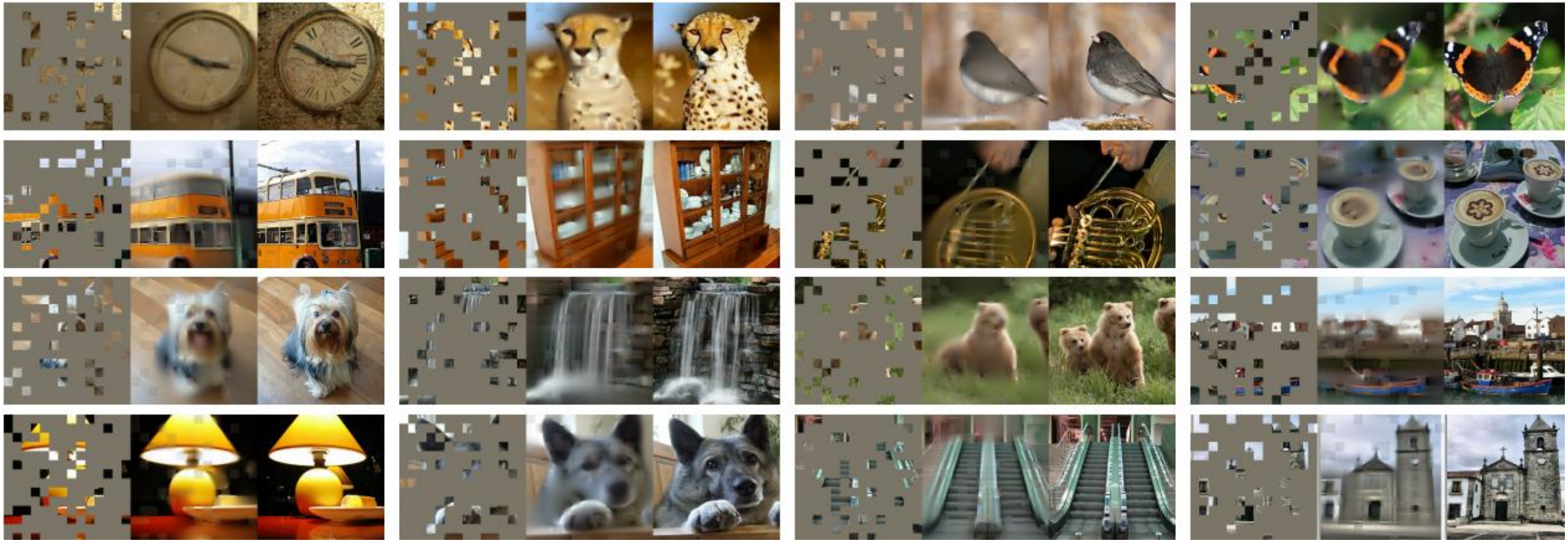
Input: image rotated by [0, 90, 180, 270]

Output: 4-way classification

Gidaris et al., 2018, Predicting Image Rotations

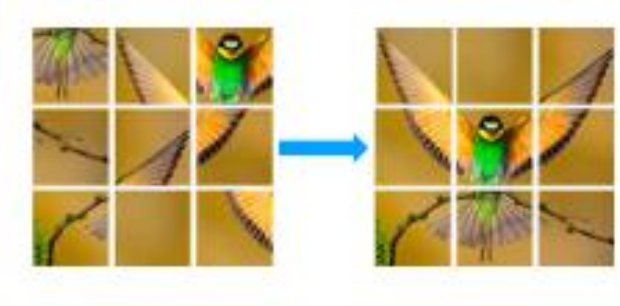
Examples of data generation in vision

Masking input



The hope of generalization

- We hope that the pre-text task and downstream task are aligned.

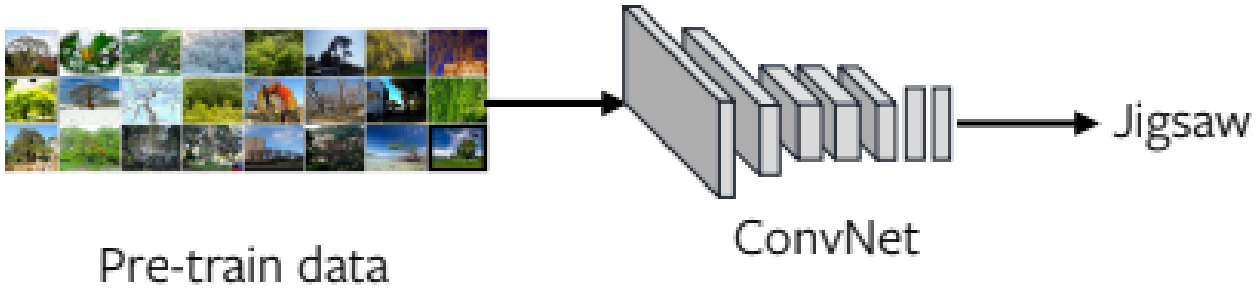


Pre-training

Self-supervised

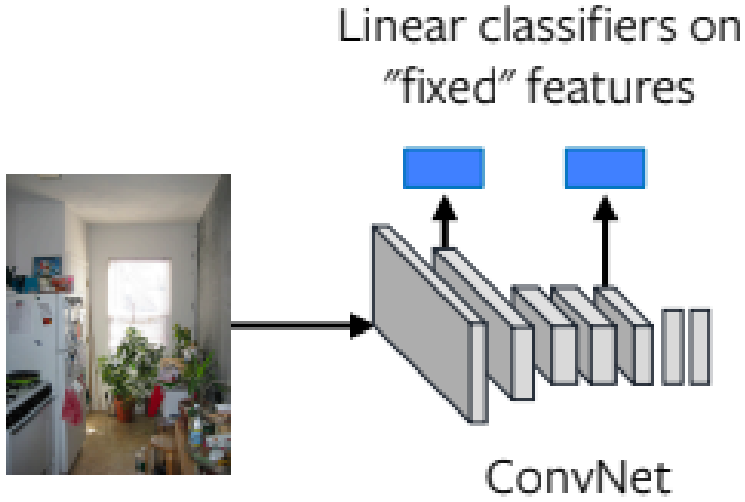
Transfer Tasks

The hope of generalization



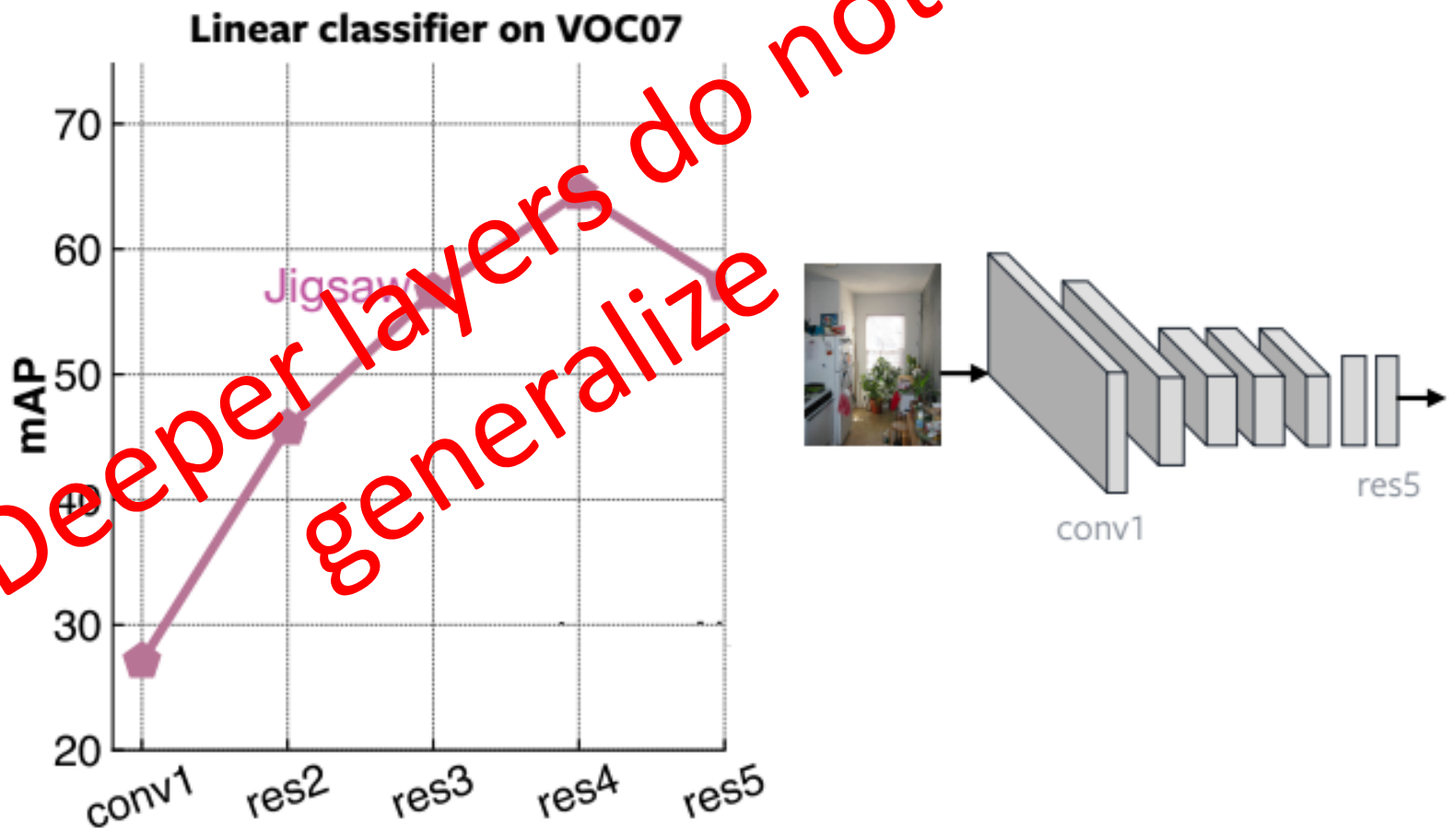
Pre-training

Weak or self-supervised



Transfer

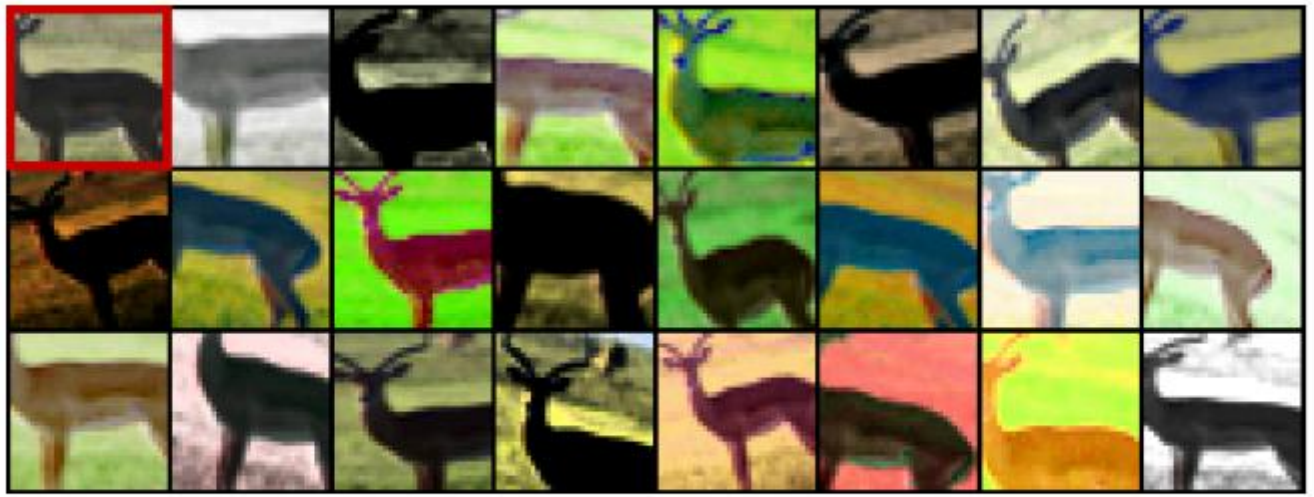
The hope of generalization



mAP = mean Average Precision
(Higher is better)

Pretrained features should

- Represent how images relate to each other.
- Be robust to “nuisance factors”
  - E.g., exact location, lighting, texture, color,…



Learn features such that:

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

Figure from Dosovitskiy et al., 2014

Self-supervision NLP vs Vision

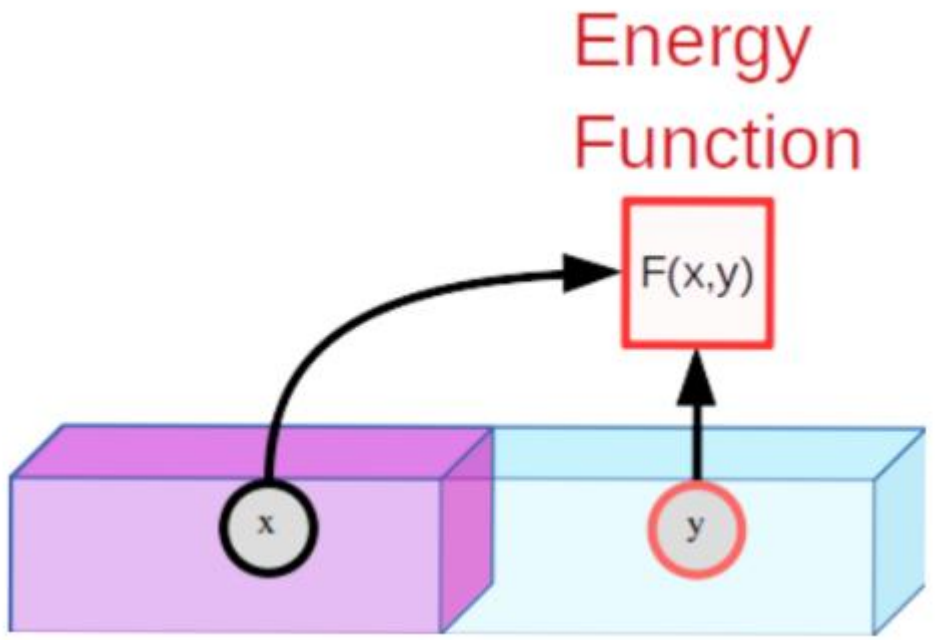
	NLP	Vision
Type	Discrete	Continuous
Uncertainty		

Number of possible outcomes can’t be enumerated in vision.

Energy-based model (EBM)

- An EBM is a trainable system that, given two inputs, x and y, tells us how incompatible they are with each other.
- Two steps of training:
  - Showing compatible x and y, produces low energy
  - For incompatible examples of y,

  give high energy



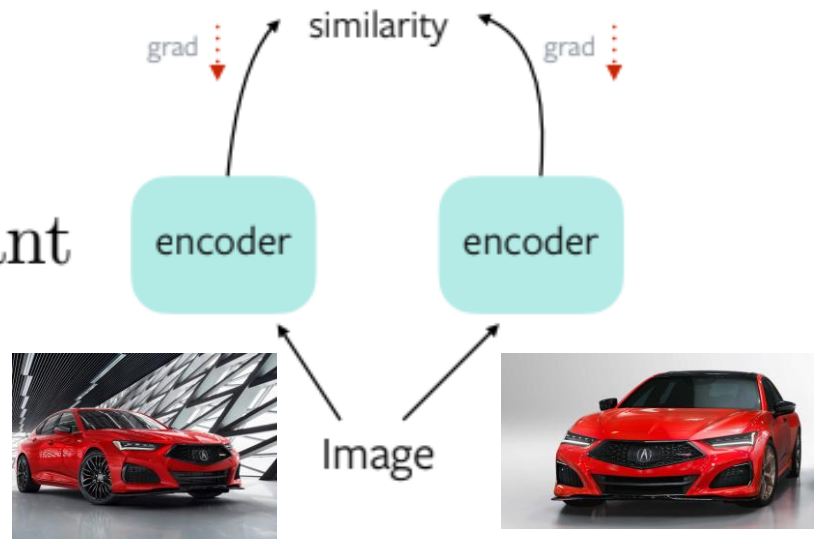
Join Embedding, Siamese networks

What could go wrong here?

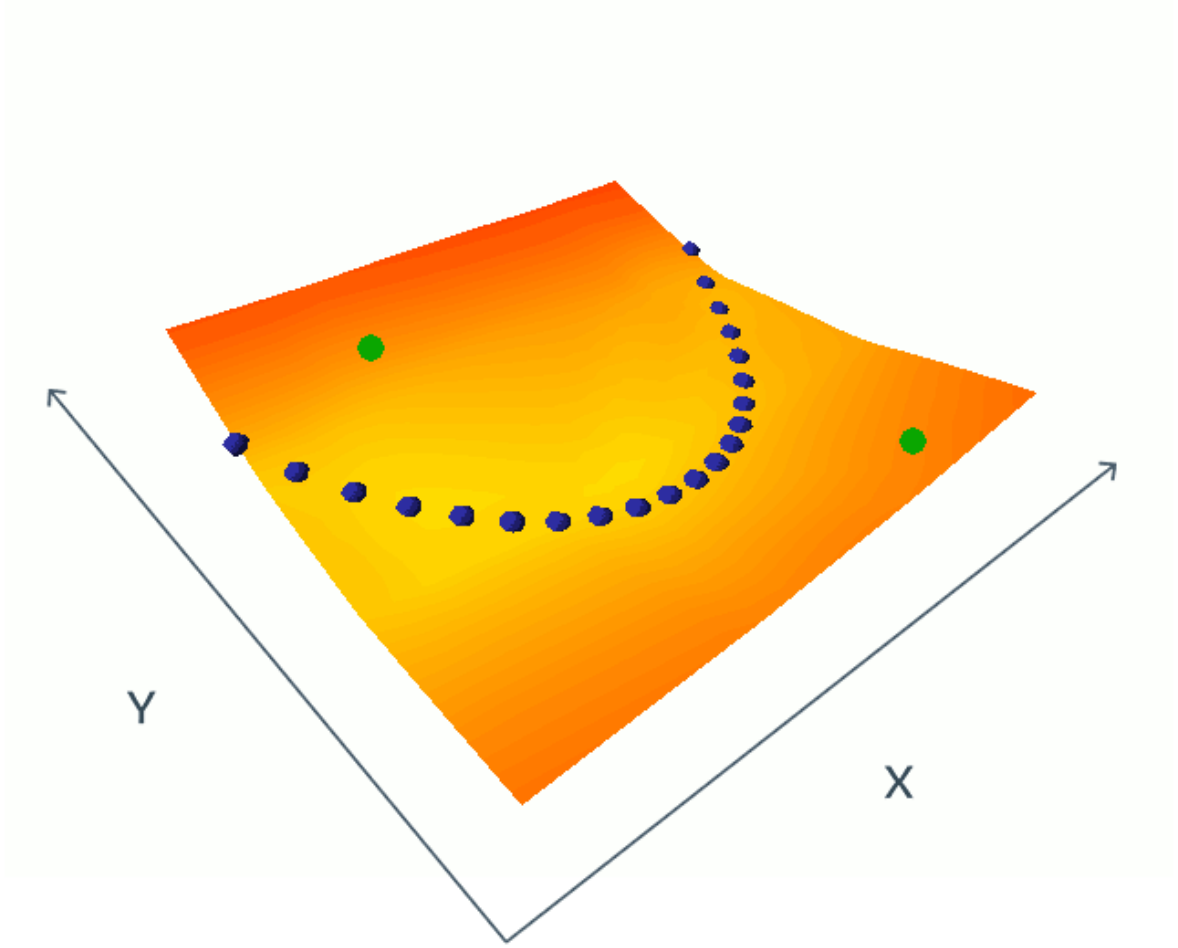
Network ignoring input and generating same output regardless.
Called **Collapse**

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

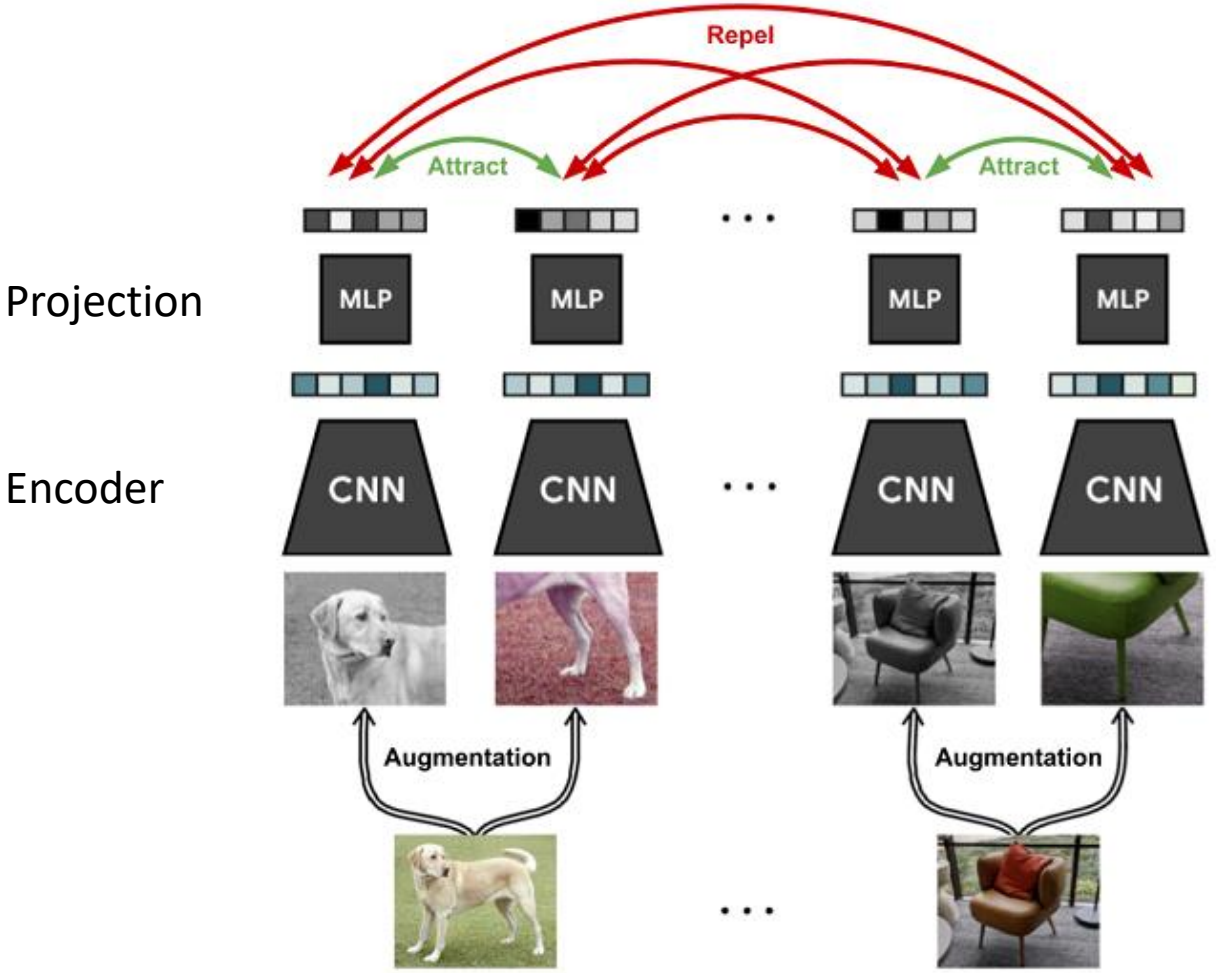
$$f_\theta(I) = \text{constant}$$



Contrastive energy-based SSL

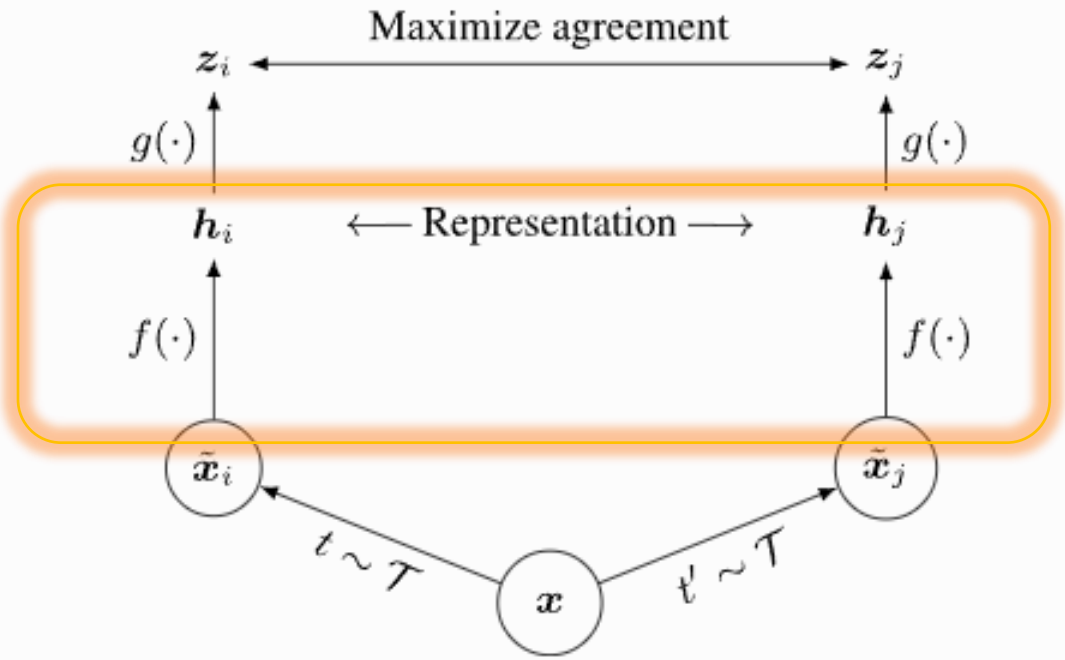


SimCLR: A Simple Framework for Contrastive Learning

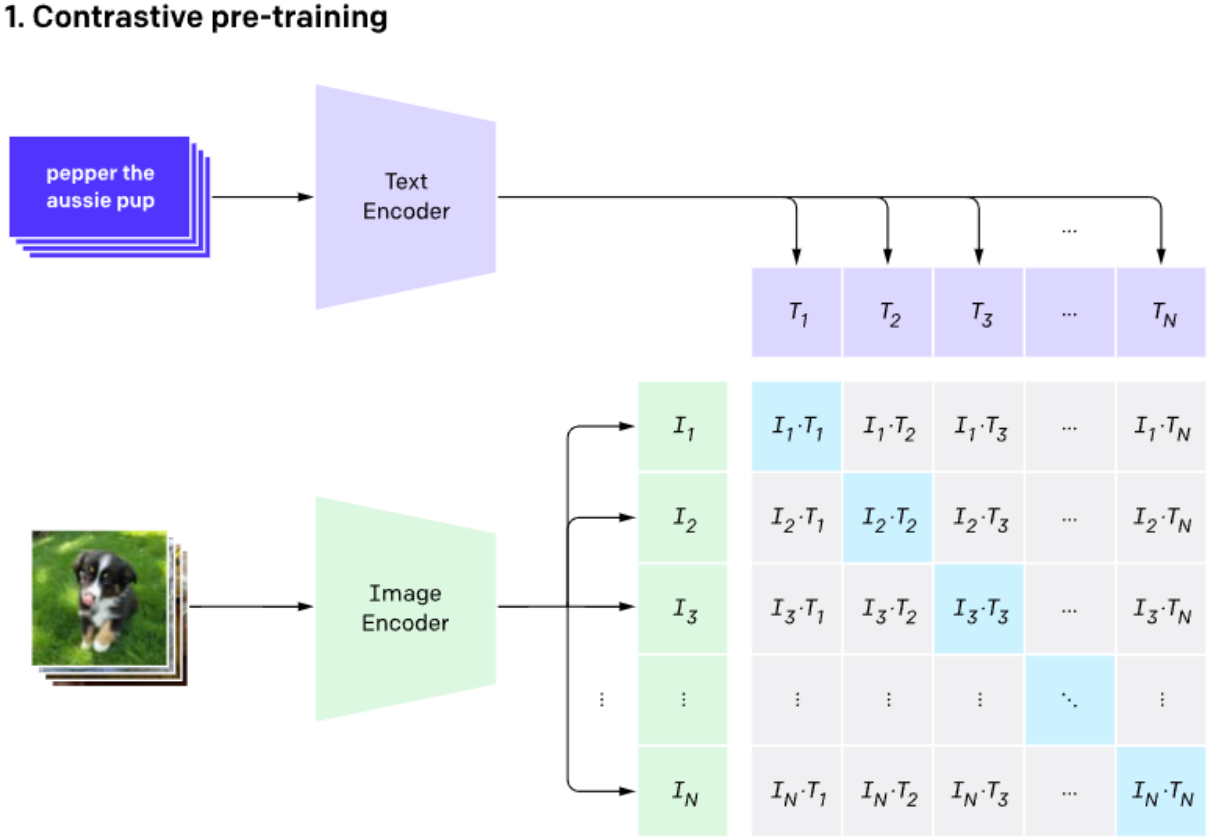


SimCLR: A Simple Framework for Contrastive Learning

Used a pre-trained weights



CLIP: Contrastive Language-Image Pretraining



Clip solves:

- Costly dataset: Can be trained using image-text pair found on internet.
- Adaptation: Easily adopted to other unseen datasets

CLIP: Contrastive Language-Image Pretraining

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

Summary

- Self-supervised learning means using the underlying structure of data to obtain data.
- Self-supervised learning is effective in both NLP and Vision
- Self-supervised learning is mostly used as pretraining phase

Useful resources

- https://www.youtube.com/watch?v=8L10w1KoOU8&t=486s&ab_channel=AlfredoCanziani

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