Self-supervised Learning (SSL)

Aapo Hyvärinen

for the NNDL course, 9 April 2025 (latter half)

- Motivation and diversity of unsupervised learning
- Self-supervised learning as a method of unsupervised learning
- Two main approaches in SSL:

1. Classification (contrastive learning)
2. Regression (autoencoders)

[Just a short introduction to SSL; they are central in the subsequent lectures]

Different kinds of machine learning

- Supervised learning: we have
  - "input" $\mathbf{x}$, e.g. images / medical measurements
  - "output" $\mathbf{y}$, e.g. content (cat or dog) / diagnosis

Different kinds of machine learning

- Supervised learning: we have
  - "input" $\mathbf{x}$, e.g. images / medical measurements
  - "output" $\mathbf{y}$, e.g. content (cat or dog) / diagnosis
- **Un**supervised learning: we have
  - only "input" $\mathbf{x}$

Different kinds of machine learning

- Supervised learning: we have
  - "input" $\mathbf{x}$, e.g. images / medical measurements
  - "output" $\mathbf{y}$, e.g. content (cat or dog) / diagnosis
- **Un**supervised learning: we have
  - only "input" $\mathbf{x}$
- Reinforcement learning: we have
  - An agent that *acts* in a world
  - based on input $\mathbf{x}$, takes actions a, obtaining reward r

Different kinds of machine learning

- Supervised learning: we have
  - "input" $\mathbf{x}$, e.g. images / medical measurements
  - "output" $\mathbf{y}$, e.g. content (cat or dog) / diagnosis
- **Un**supervised learning: we have
  - only "input" $\mathbf{x}$
- Reinforcement learning: we have
  - An agent that *acts* in a world
  - based on input $\mathbf{x}$, takes actions a, obtaining reward r

- *The rest of this course will be on unsupervised learning*

Importance unsupervised learning

- Early success stories in deep learning based on supervised learning
  - ImageNet competition (recognition of objects in images)
  - Google Neural Machine Translation in its original version
- Supervised learning needs category labels or targets
  - Is the object in the photograph a cat or a dog?
  - What is the French translation of the English sentence?
  - Is there a medical pathology in this patient or not?

Importance unsupervised learning

- Early success stories in deep learning based on supervised learning
  - ImageNet competition (recognition of objects in images)
  - Google Neural Machine Translation in its original version
- Supervised learning needs category labels or targets
  - Is the object in the photograph a cat or a dog?
  - What is the French translation of the English sentence?
  - Is there a medical pathology in this patient or not?
- Problem: labels may be
  - Expensive
    - Need work by human expert (translator, doctor)
  - Difficult /impossible to obtain
    - e.g. true medical condition may not be known

- Also: Generative AI is very different from such early deep learning: perhaps inherently unsupervised !

Unsupervised learning can have different goals

1) Useful features for supervised learning?
    - contrastive learning, autoencoders — many kinds of SSL

Unsupervised learning can have different goals

1) Useful features for supervised learning?
   - contrastive learning, autoencoders — many kinds of SSL
2) Reveal underlying structure in data, find latent quantities?
   - autoencoders (+ clustering, PCA)
   - independent component analysis, "disentanglement"

Unsupervised learning can have different goals

1) Useful features for supervised learning?
   - contrastive learning, autoencoders — many kinds of SSL
2) Reveal underlying structure in data, find latent quantities?
   - autoencoders (+ clustering, PCA)
   - independent component analysis, "disentanglement"
3) Accurate model of data distribution
   - Energy-based models (EBM)
   - (Variational Autoencoders)

Unsupervised learning can have different goals

1) Useful features for supervised learning?
   - contrastive learning, autoencoders — many kinds of SSL
2) Reveal underlying structure in data, find latent quantities?
   - autoencoders (+ clustering, PCA)
   - independent component analysis, “disentanglement”
3) Accurate model of data distribution
   - Energy-based models (EBM)
   - (Variational Autoencoders)
4) Sampling points from data distribution
   - Generative Adversarial Networks
   - EBM + Monte Carlo methods (e.g. diffusion)

Unsupervised learning can have different goals

1) Useful features for supervised learning?
   - contrastive learning, autoencoders — many kinds of SSL
2) Reveal underlying structure in data, find latent quantities?
   - autoencoders (+ clustering, PCA)
   - independent component analysis, "disentanglement"
3) Accurate model of data distribution
   - Energy-based models (EBM)
   - (Variational Autoencoders)
4) Sampling points from data distribution
   - Generative Adversarial Networks
   - EBM + Monte Carlo methods (e.g. diffusion)

- These goals are orthogonal, even contradictory!
  - Probably, no method can accomplish all
- Unsupervised learning needs a variety of methods
- All cases above (not in parentheses) treated in this course

What is self-supervised learning (SSL) then?

- Supervised learning:
  - We have "input" $\mathbf{x}$ and "output" $\mathbf{y}$
  - Goal: Find input-output mapping (regression)
- Unsupervised learning:
  - We have only "input" $\mathbf{x}$
  - Goal: Solve one of the many problems listed above

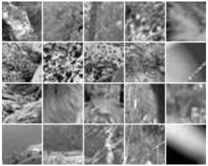
What is self-supervised learning (SSL) then?

- Supervised learning:
  - We have "input" $\mathbf{x}$ and "output" $\mathbf{y}$
  - Goal: Find input-output mapping (regression)
- Unsupervised learning:
  - We have only "input" $\mathbf{x}$
  - Goal: Solve one of the many problems listed above
- **Self**-supervised learning (SSL): we have
  - only "input" $\mathbf{x}$
  - *but we invent* $\mathbf{y}$ somehow and use supervised algorithms
    - "pretext task": supervised task not interesting in itself
  - (or: use $\mathbf{x}$ as output of regression, invent new input somehow)
- Goal of SSL: NN trained on pretext task learns something about $\mathbf{x}$, solving one of the unsupervised problems above
- Main categories
  - contrastive learning: uses classification (but terminology varies)
  - autoencoders: uses regression

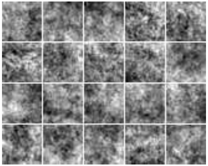
Contrastive learning 1: Noise-Contrastive Estimation

- Train a nonlinear classifier to discriminate observed data from some artificial noise
- For example, compare natural images with Gaussian noise

Real image windows



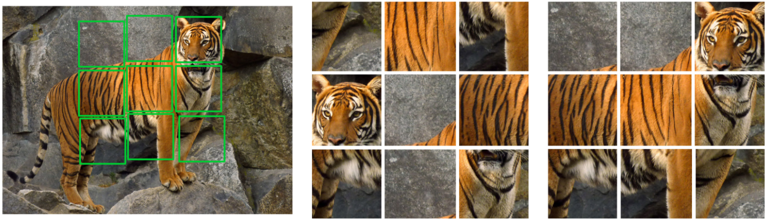
Gaussian noise



- To be successful in this pretext task:
  the classifier (NN) must "discover structure" in the data
- In particular, the NN should *learn features* from the data,
  a useful representation in the (last) hidden layer
- (maybe it learns something else too.... more on this later...)

Contrastive learning 2: Shuffling real data

- Take patches of real images, shuffle them
- Train NN to discriminate between real data and shuffled data
  Data points like in (b) are one class, points like (c) another class



(a) (b) (c)

(Norrozi and Favaro, 2016)

- The NN should learn useful features
- Easy to figure out a lot of similar tasks in computer vision
  - Corrupt real image data in some way

Autoencoders: SSL by regression

- Learn regression where
  - inputs are real data points, possibly corrupted
  - outputs are the *same* real data points
- A regression problem, as opposed to classification
- Basic case: predict $\mathbf{x}$ by $\mathbf{x}$! (Sounds absurd?)
- Somehow must make exact reconstruction impossible: NN should not learn identity mapping $\mathbf{g}(\mathbf{x}) = \mathbf{x}$
- Case 1: Restrict the architecture of the NN
  - Create a "bottleneck": a hidden layer with few units
- Case 2: Penalize hidden layer
  - Force hidden units to be mostly zero ("sparse")
- Case 3: Corrupt the input
  - Case 3a: Add Gaussian noise to the input
  - Case 3b: Remove some variables (pixels) from the input
  - Case 3c: Remove color from input images (make gray-scale)
    - Learn to colorize images
- (Examples above will be treated in the next lecture, except 3c)

Classic method: Principal component analysis (PCA)

- Consider a linear function instead of neural network
- Dimension reduction as: $\mathbf{s} = \mathbf{W}\mathbf{x}$, with $\mathbf{W} \in \mathbb{R}^{m \times n}$, $m = \dim(\mathbf{s}) < \dim(\mathbf{x}) = n$.
- Reconstruct as: $\hat{\mathbf{x}} = \mathbf{A}\mathbf{s}$, with $\mathbf{A} \in \mathbb{R}^{n \times m}$
- Least-squares criterion above simplies to

$$E\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = E\|\mathbf{x} - \mathbf{A}\mathbf{W}\mathbf{x}\|^2 \tag{1}$$

- One definition of linear PCA, can be interpreted as SSL
- Equivalent definition: Find direction that maximizes variance

$$\max_{\|\mathbf{w}\|=1} E(\mathbf{w}^T \mathbf{x})^2 \tag{2}$$

  and repeat, constraining new $\mathbf{w}$ to be orthogonal to those already found
- Example where same goal can be achieved by SSL or other unsupervised learning

SSL by prediction of future: autoregressive models

- The most classical time-series task is prediction
- Say, predict $\mathbf{x}(t)$ from $\mathbf{x}(t-1), \mathbf{x}(t-2), \ldots$; t is time index
- You may want to predict the weather, the stock market, ecological variables, etc. etc.
- Widely used in natural language processing (with variations)
  - Train a NN to predict the next word
  - .... as just seen in Arto's lecture
- Useful in itself for generating new data, especially text
  - ... and of course is predicting future weather etc.
- Could be called SSL (?)
  - Justification: the predicting NN learns a useful representation in the hidden layer(s)

Typical application of SSL in (image) classification

- Goal: train a NN to classify photographs, say, cats vs. dogs
- Problem: We do not have a lot of labelled data
  - i.e. not many photographs where we know it is a cat or a dog
- But: We have a lot of unlabelled data
  - e.g. easy to download a lot of photographs from the internet
- Solution:
  1. Train NN by SSL to learn features using big unlabelled dataset
  2. Use learned features as input to a simple classifier (even linear), trained with small labelled dataset
- Supervised problem in SSL is called *pretext task*
  - E.g. discriminate between noise and real data
  - Pretext task is pointless in itself; not our actual goal
- Actually discriminating cats vs. dogs is called *downstream task*
- The final measure of performance of this kind of SSL: classification accuracy in downstream task
- (Foundation model $\approx$ big pretext task already trained by somebody else, and publicly (?) distributed)

Typical application of SSL in image generation

- A state-of-the-art framework for generating images
1. Learn model of data distribution by SSL (an autoencoder)
2. Generate data by a sampling method (e.g. MCMC)
- Some examples with dynamics:

Iterative image generation, final results in right-most column.

From (Yang and Ermon, 2019)

Connection between SSL and unsupervised learning?

- Different viewpoints exist in the literature:

1. Hype viewpoint: SSL is something completely new and does things that nobody has been able to do earlier
2. My viewpoint: SSL is one *technique* for *un*supervised learning
   - SSL uses supervised learning algorithms to achieve unsupervised learning
3. Another viewpoint:
   SSL is something between supervised and unsupervised;
   it is not unsupervised since the algorithms are supervised

- Difference between #2 and #3 partly a question of whether we consider *goal* of learning or *algorithm* used
- A lot of confusion in the literature due to hype in #1
- Remember: Not all unsupervised learning is SSL
  - E.g.: maximum likelihood estimation of generative model
  - E.g.: PCA example above
  - SSL often computationally efficient, but statistically inferior

Conclusion

- Unsupervised learning has many utilities, especially:
  - Feature extraction: helping "downstream" supervised learning
  - Generative AI: model data distribution for generating new data
- Self-supervised learning is a framework for performing unsupervised learning
  - goal is unsupervised learning; but algorithms are supervised
- In this course, we will mainly do unsupervised learning by SSL
  - simple algorithmically: no new algorithms needed (?)
  - very fashionable ;)
  - but other kinds of unsupervised learning do exist
- Many of, the following lectures show how SSL performs probabilistic modelling
  - Modelling data distribution ("energy-based modelling")
  - Learning generative models ("generative adversarial networks")
  - ... but sometimes purely heuristic ("basic autoencoders")