Self-supervised learning

Contrastive | non-contrastive

Maciej Żelaszczyk

April 27, 2022

PhD Student in Computer Science
Division of Artificial Intelligence and Computational Methods
Faculty of Mathematics and Information Science

m.zelaszczyk@mini.pw.edu.pl

Warsaw University of Technology

State of deep learning

- Enormous success.
- Mostly relies on CNNs (vision) and Transformers (NLP).
- Relatively large models (e.g. 1.75 trillion parameters).
- Computationally expensive.
- Architectures geared toward dataset or task.
- Supervised learning.

Supervised vs. unsupervised

Supervised:

- We have explicit labels and use them to guide training.
- Requires huge datasets.
- Extensive training.
- Annotating is costly.
- Limit to how much data we can obtain.
- Does not scale.
- Ignores physical world.
- RL makes this ridiculous.
- Driving a car off a cliff.

Supervised vs. unsupervised

How do children learn?

- A lot of evolutionary knowledge.
- Vision, hearing, touch etc. in place.
- Extensive observation.
- Build a model of the world.
- Model vs. physical world.
- Surprise, curiosity guide learning.
- Continuous refinement of model.
- Limited reinforcement learning.
- All initial learning is unsupervised.

Supervised vs. unsupervised

Unsupervised:

- In practice, very little labelled data available.
- Need to create model of world, confront it with reality.
- Update model when it does not agree with reality.
- Exploit physical structure of world to obtain links.
- Learn from little external reward.
- Learn from very few labelled examples.

Importance of unsupervised learning

What if importance of various kinds of learning is like a cake?

- Pure reinforcement learning = cherry.
- Supervised learning = icing.
- Unsupervised/self-supervised/predictive learning = génoise.

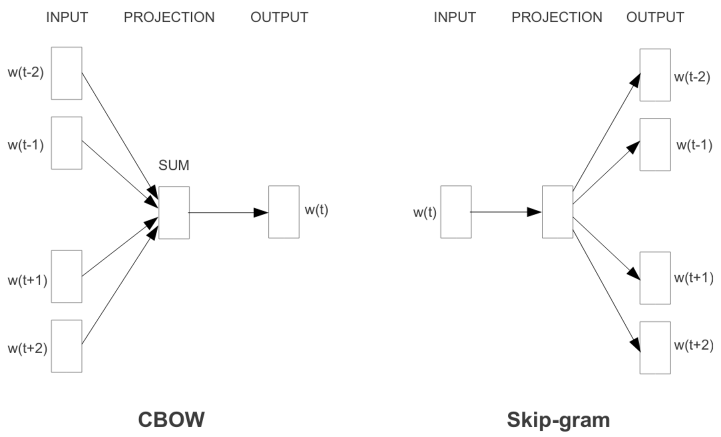
Source: LeCun, Y., *The Next Step Towards Artificial Intelligence*

Self-supervised learning

Self-supervised is the new unsupervised:

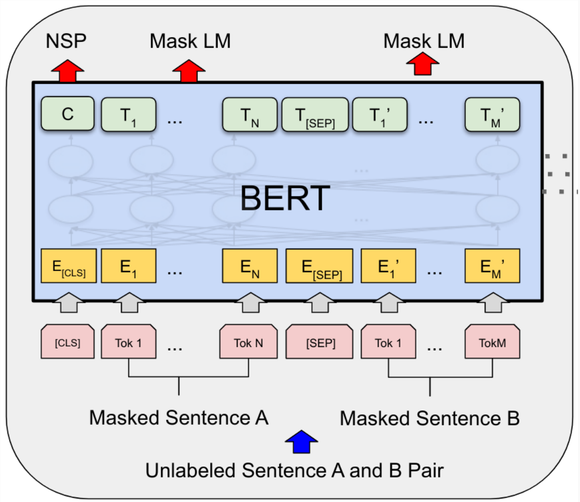
- Supervised: data and labels.
- Unsupervised: data without labels.
- Self-supervised: use data as labels.
- In reality: use data transformations to obtain labels.
- Predict characteristics based on transformed data and the obtained labels.
- Quite different from standard labels.
- Pre-train to use on downstream tasks.

NLP as success story for SSL



Source: [Mikolov et al., 2013]

NLP as success story for SSL



Source: [Devlin et al., 2019]

NLP as success story for SSL

Why NLP?

- Sentences naturally represented as sequences.
- There is significant structure to the data.
- Possible to approximately identify the vocabulary.
- Predicting a masked word from the context can be cast as a classification problem.
- Manageable dimensionality.
- We can use quite similar techniques as for supervised learning.
- Softmax, loss function, etc.
- Supervised training with SSL pre-training beats vanilla supervised training.
- Useful for downstream tasks.

SSL for vision

Why is it more difficult to use predictive self-supervised learning for vision?

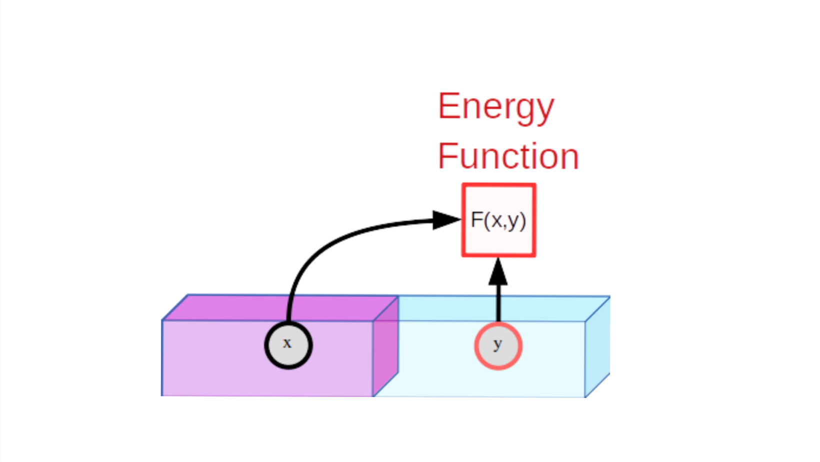
- For images, it is considerably harder to divide them into meaningful parts than for sentences.
- There is no immediate analogue of a vocabulary.
- Predicting a masked part of an image from the context is not easily cast as a classification problem.
- In particular: dimensionality blows up for the predictive problem.
- We cannot use techniques from supervised learning out of the box.
- Predictive problem not completely out of the question, only significantly harder.

Energy-based models

A potentially unifying view of self-supervised learning methods.

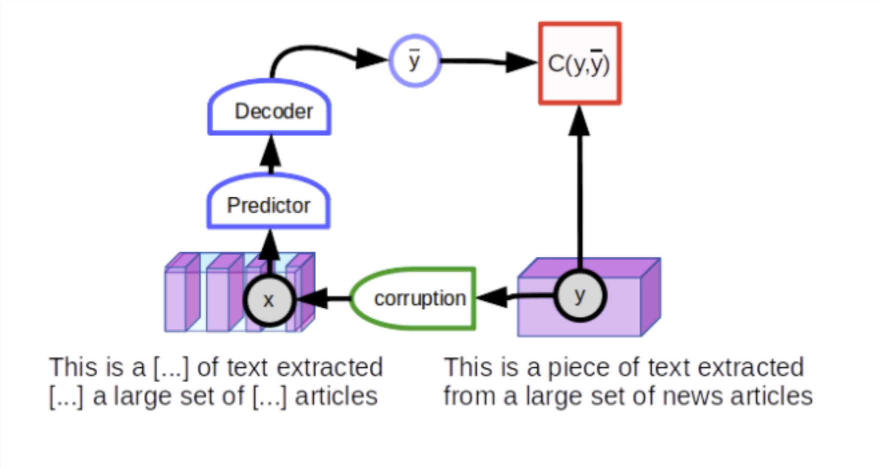
- Trainable system.
- Outputs whether two inputs $\mathbf{x}$ and $\mathbf{y}$ are compatible.
- Scalar assessment of agreement between inputs - *energy*.
- $F(\mathbf{x}, \mathbf{y})$ - *energy function*.
- High energy function values for incompatible inputs, low energy values for compatible inputs.

Energy-based models



Source: LeCun, Y. and Misra, I., *Self-supervised learning: The dark matter of intelligence*

SSL in NLP as EBM



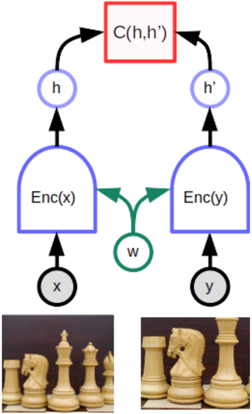
Source: LeCun, Y. and Misra, I., *Self-supervised learning: The dark matter of intelligence*

Joint embedding architecture

A more concrete application for vision.

- Two identical or close to identical encoders.
- One processes $\mathbf{x}$, the other $\mathbf{y}$.
- Each produces an embedding, $\mathbf{h_x}$ and $\mathbf{h_y}$, respectively.
- $C(\mathbf{h_x}, \mathbf{h_y})$ - distance between embeddings - energy function.
- Relatively easy to train the system to output low energy for transformed version of the same image, different views of the same object, etc - *positive samples*.
- We need *negative samples* as well to avoid *collapse*.
- *Collapse* - system ignores input and outputs the same assessment.
- Approaches to avoid collapse: contrastive and non-contrastive.

Joint embedding architecture



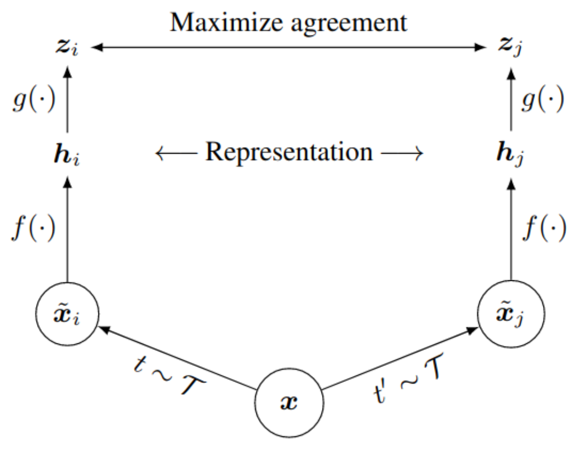
Source: LeCun, Y. and Misra, I., *Self-supervised learning: The dark matter of intelligence*

Contrastive SSL

Circumvent the collapse problem.

- Apart from positive samples, we specifically construct negative samples.
- Ensure the energy is low for positive samples.
- Ensure the energy is high for negative samples.
- There are less-than-obvious difficulties.
- One challenge: What if the difference between positive samles and negative samples is to stark?
- Training might quickly allow the system to distinguish between positive and negative samples without additional benefits.
- We need difficult negative samples.
- Costly to construct.

SimCLR



[Chen et al., 2020]

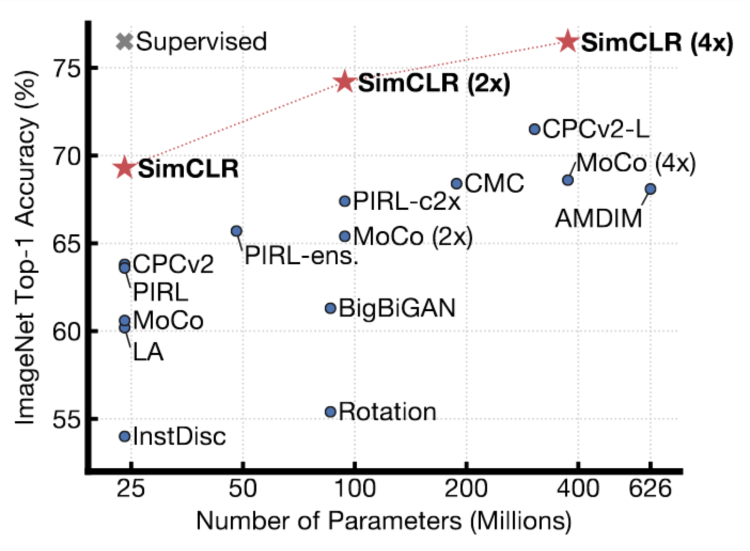
SimCLR

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

[Chen et al., 2020]

SimCLR



[Chen et al., 2020]

SimCLR

Architecture	Label fraction 1%		10%		100%	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
ResNet-50	49.4	76.6	66.1	88.1	76.0	93.1
ResNet-50 ($2\times$)	59.4	83.7	71.8	91.2	79.1	94.8
ResNet-50 ($4\times$)	64.1	86.6	74.8	92.8	80.4	95.4

Table B.2. Classification accuracy obtained by fine-tuning the SimCLR (which is pretrained with broader data augmentations) on 1%, 10% and full of ImageNet. As a reference, our ResNet-50 ($4\times$) trained from scratch on 100% labels achieves 78.4% top-1 / 94.2% top-5.

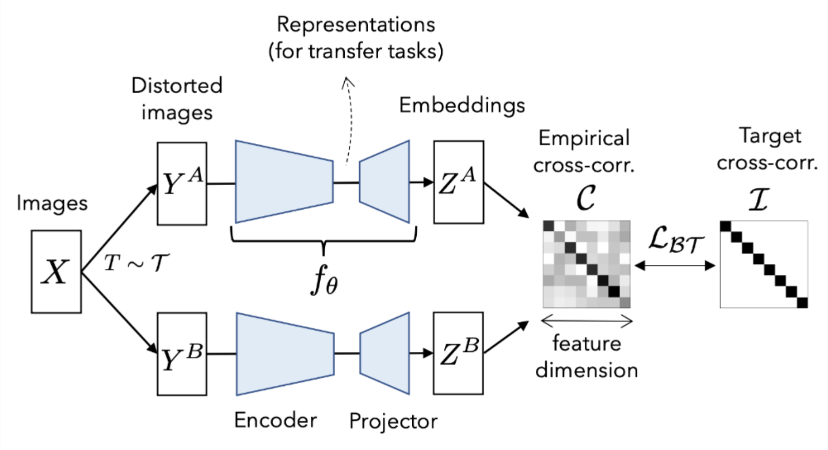
[Chen et al., 2020]

Non-contrastive SSL

More diverse in approaches than contrastive methods.

- Does not rely on explicit negative samples.
- Might allow for computationally more efficient learning.
- One approach is to use regularization to constrain the parameters of the models and the representation space.
- Relatively little research done on non-contrastive methods but this is changing.

Barlow Twins



[Zbontar et al., 2021]

Barlow Twins

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} {\mathcal{C}_{ij}}^2}_{\text{redundancy reduction term}}$$

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

[Zbontar et al., 2021]

Barlow Twins

Method	Top-1	Top-5
Supervised	76.5	
MoCo	60.6	
PIRL	63.6	-
SimCLR	69.3	89.0
MoCo v2	71.1	90.1
SimSiam	71.3	-
SwAV (w/o multi-crop)	71.8	-
BYOL	<u>74.3</u>	91.6
SwAV	<u>75.3</u>	-
Barlow Twins (ours)	<u>73.2</u>	91.0

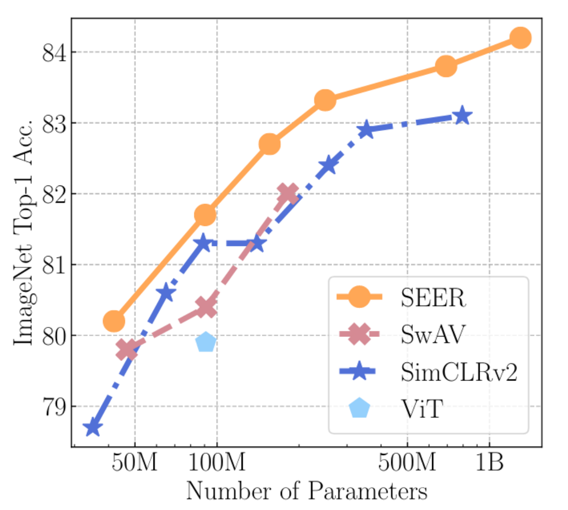
[Zbontar et al., 2021]

Barlow Twins

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised	25.4	56.4	48.4	80.4
PIRL	-	-	57.2	83.8
SimCLR	48.3	65.6	75.5	87.8
BYOL	53.2	68.8	78.4	89.0
SwAV	53.9	**70.2**	78.5	**89.9**
Barlow Twins (ours)	**55.0**	69.7	**79.2**	89.3

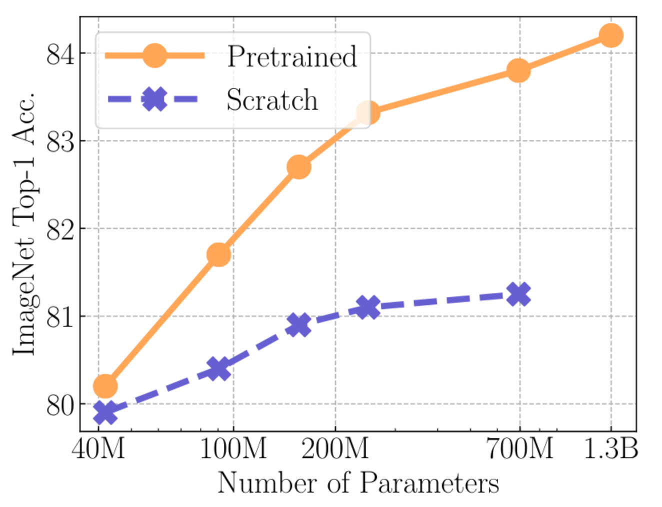
[Zbontar et al., 2021]

SEER



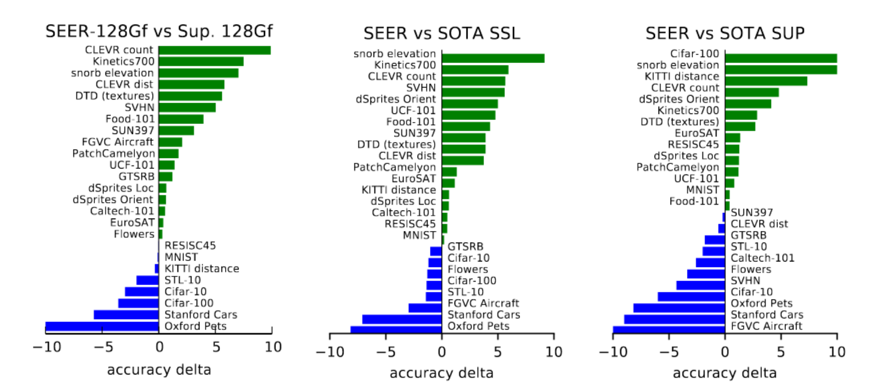
[Goyal et al., 2021]

SEER



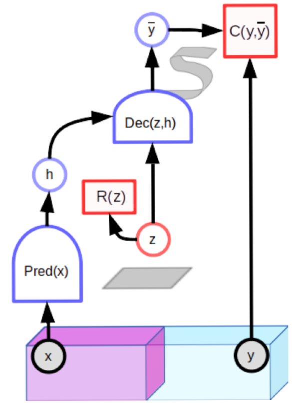
[Goyal et al., 2021]

SEER-10B



[Goyal et al., 2022]

Latent-variable predictive models



Source: LeCun, Y. and Misra, I., *Self-supervised learning: The dark matter of intelligence*

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). **A simple framework for contrastive learning of visual representations.** In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). **BERT: Pre-training of deep bidirectional transformers for language understanding.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., and Bojanowski, P. (2021).
Self-supervised pretraining of visual features in the wild.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P. (2022).
Vision models are more robust and fair when pretrained on uncurated images without supervision.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021).
Barlow twins: Self-supervised learning via redundancy reduction.
In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.