# Self-supervised Learning

## Representation Learning

Vladimir Zaigrajew

2025-03-12

Vladimir Zaigrajew - vladimir.zaigrajew.dokt@pw.edu.pl

Tymoteusz Kwieciński - tymoteuszkwiecinski@gmail.com

You can find us in Room 316, MINI, $PW$

Remember every information you can find on our Github Repo:

Figure 1: QR code to course Github Repo

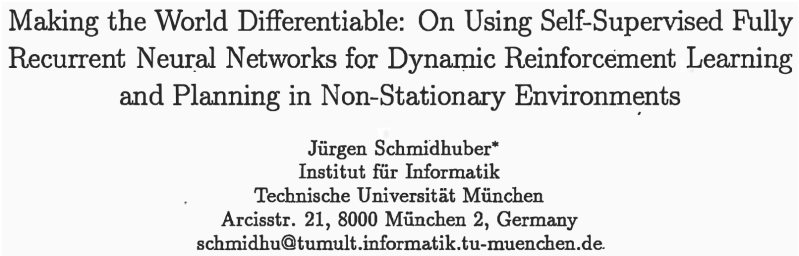
Figure 2: QR code to our Github Repo

# Recap from last lecture

- Representation Learning is a subfield of machine learning that instead of learning mapping $f : X \to Y$ learns mapping $f : X \to Z$ where $Z$ is a latent space.
- We can learn representation in a semi-supervised/self-supervised way removing the need for labels.
- Models learn from native data representation (e.g. images, text, audio) but this is not always the most efficient way.
- We have traditional methods of representation learning (PCA, t-SNE, UMAP) and deep learning methods.
- We can use representation learning for clustering, classification, regression, and generative tasks.
- Each deep learning model has its own representation learning method, explicitly or implicitly.

**The term self-supervised learning was coined at 1990!**

> ... the model should not only predict the reinforcement units but also the other input units

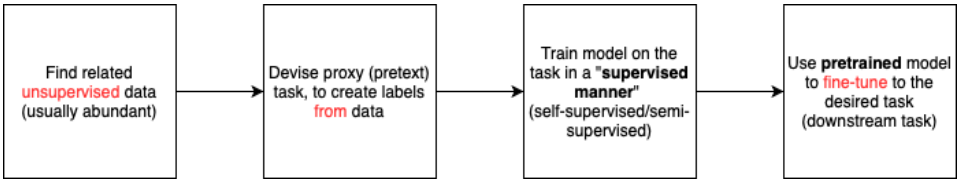


Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments

Jürgen Schmidhuber*
Institut für Informatik
Technische Universität München
Arcisstr. 21, 8000 München 2, Germany
schmidhu@tumult.informatik.tu-muenchen.de

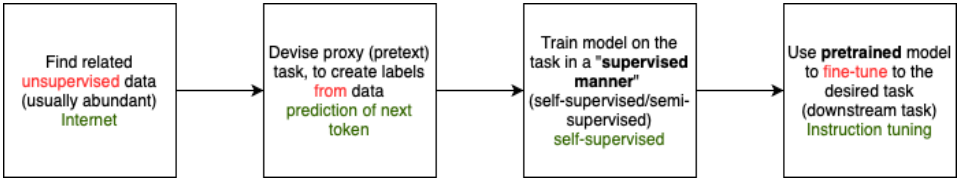
1

[1] Schmidhuber, J. (1990). Making the world differentiable: on using self supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments (Vol. 126). Inst. für Informatik.



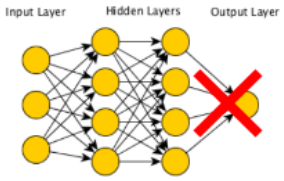

**Pretext task** - a supervised task that is used to learn representation of the data.
**Fine-tuning** - a supervised task that is used to adapt the model to a specific task.
**Downstream task** - a specific task that is used to evaluate the performance of the model.
**Transfer learning** - applying knowledge gained from one task to improve performance on a different, but related task.

# SSL overview - GPT3 Example

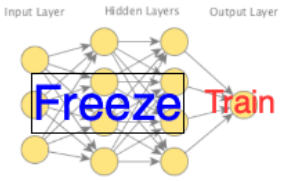



**Pretext task** - training the model to predict the next word in a sentence.
**Fine-tuning** - prepare data in instruction-following format and train the model to follow instructions.
**Downstream task** - assessing the model's performance on specific applications (e.g. summarization, translation, etc.).

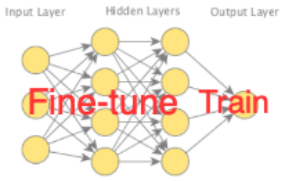
**Downstream tasks** are specific applications where models are deployed, utilizing the learned representations to perform targeted predictions or generate outputs based on new input data.




Pretrained Model




Freeze pretrained model
Train classification head
Accuracy **90%**
Train time **1h**



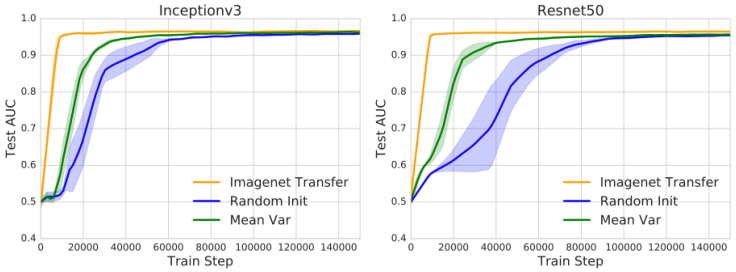

Fine-tune pretrained model
Train classification head
Accuracy **95%**
Train time **3h**

# Transfer learning: Using pretrained weights to different task

What is more efficient?

Teach a 40 year old mathematician to play chess or teach a young child to play chess?

Transfer learning can improve speed and test accuracy (**marginal!**)




[2]

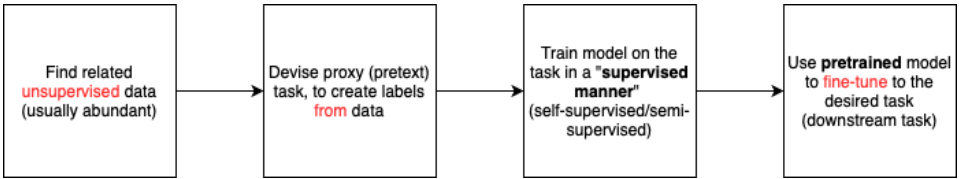
[2]Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.

- **Linear Probing (Evaluation)** - train a linear classifier on top of the frozen representation. Trained on MNIST dataset, train a linear classifier on top of the frozen representation to classify digits.
  - Top-1 accuracy: accuracy of the model on the test set.
  - Top-5 accuracy: accuracy of the model on the test set when the correct label is in the top 5 predictions.
- **KNN evaluation** - use the learned representation to perform k-nearest neighbors classification. Having the learned representation, we can use it to perform k-nearest neighbors classification on the test set.
- **Transfer learning** - fine-tune the model on a specific task and assess its effectiveness on unseen data.

# Evaluation - Generative models

- **Transfer learning** - fine-tune the model on a specific task and assess its effectiveness on unseen data.
- **Zero-shot evaluation** - evaluate the model on a specific task without any fine-tuning.
- **Few-shot evaluation** - evaluate the model on a specific task with a small amount of fine-tuning.

# SSL overview - one more time




**Pretext task** - a supervised task that is used to learn representation of the data.
**Downstream task** - a specific task that is used to evaluate the performance of the model.
**Transfer learning** - applying knowledge gained from one task to improve performance on a different, but related task.

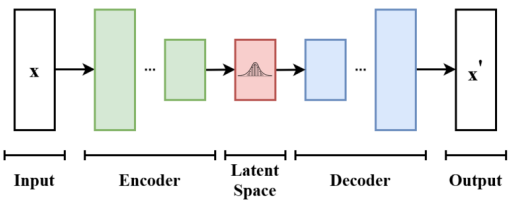
Pretext task?

Downstream task?

From: AAAI-87 Proceedings. Copyright ©1987, AAAI (www.aaai.org). All rights reserved.

**Modular Learning in Neural Networks**

Dana H. Ballard
Department of Computer Science
University of Rochester, Rochester, New York 14627




3

[3]"Variational autoencoder." Wikipedia, Wikimedia Foundation, en.wikipedia.org/wiki/Variational autoencoder.

# Autoencoders (1987)

Pretext task? Reconstruct Input

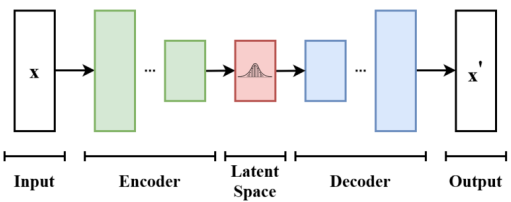
Downstream task? Use bottleneck representation

- Variational Autoencoders (VAE)
- Denoising Autoencoders (DAE)
- Contrastive Autoencoders (CAE)
- Adversarial Autoencoders (AAE)
- **Masked Autoencoders (MAE)**
- **Sparse Autoencoders (SAE)**

From: AAAI-87 Proceedings. Copyright ©1987, AAAI (www.aaai.org). All rights reserved.

**Modular Learning in Neural Networks**

Dana H. Ballard
Department of Computer Science
University of Rochester, Rochester, New York 14627




4

[4]"Variational autoencoder." Wikipedia, Wikimedia Foundation, en.wikipedia.org/wiki/Variational autoencoder.

Pretext task?

Downstream task?


Figure 3: Fully automatic colorization results on ImageNet/ctest10k.

Larsson, Gustav, Michael Maire, and Gregory Shakhnarovich. "Learning representations for automatic colorization." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016.

# Colorization (2016)

Pretext task? Predicting from grayscale input colorization per pixel

Downstream task? Trained VGG-16 without classification head


Figure 4: Fully automatic colorization results on ImageNet/ctest10k.

Larsson, Gustav, Michael Maire, and Gregory Shakhnarovich. "Learning representations for automatic colorization." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016.

# Jigsaw puzzles: Placing image patches in the right place (2016)

Pretext task?

Downstream task?

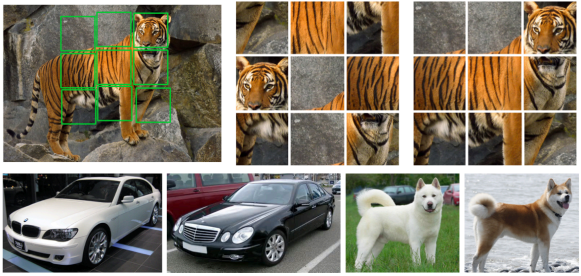
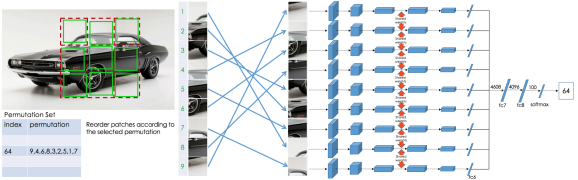


Fig. 2: Most of the shape of these 2 pairs of images is the same (two separate instances within the same categories). However, some low-level statistics are different (color and texture). The Jigsaw puzzle solver learns to ignore such statistics when they do not help the localization of parts.





Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Cham: Springer International Publishing, 2016.

# Jigsaw puzzles: Placing image patches in the right place (2016)

Pretext task? Predicting the position of image patches

Downstream task? "we use the CFN weights to initialize all the conv layers of a standard AlexNet network"

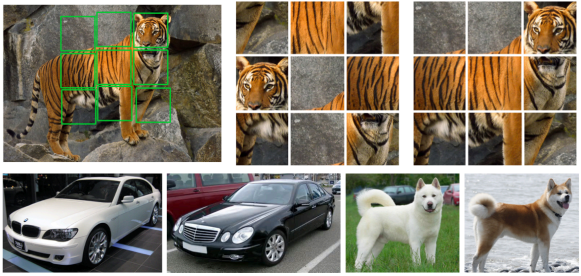
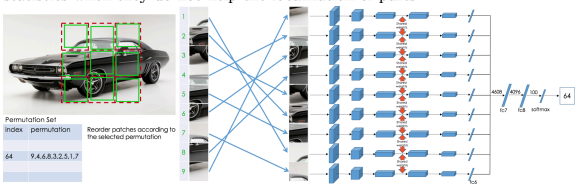


Fig. 2: Most of the shape of these 2 pairs of images is the same (two separate instances within the same categories). However, some low-level statistics are different (color and texture). The Jigsaw puzzle solver learns to ignore such statistics when they do not help the localization of parts.

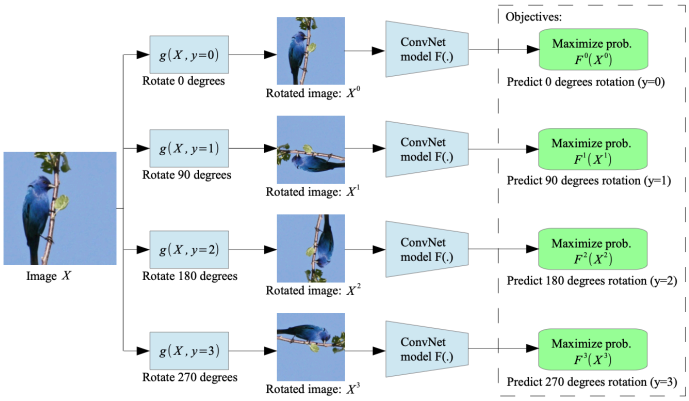



Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." European conference on computer vision. Cham: Springer International Publishing, 2016.

# Rotation Prediction (2018)

Pretext task? Predicting the rotation of an image: 0, 90, 180, 270 degrees

Downstream task?
"ConvNet model that is trained on the self-supervised task of rotation recognition RotNet model"
we learn classifiers on top of the feature maps generated by each conv. block of each RotNet model




Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." arXiv preprint arXiv:1803.07728 (2018).

# Is SSL realy so good in transfer learning?

The ranking is **not** consistent across different methods, **nor** across architectures.

The answer is: **it depends!**

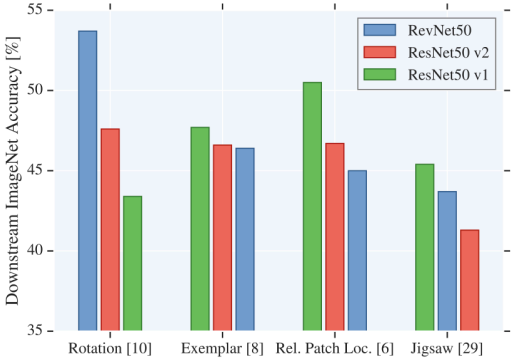



Figure 1. Quality of visual representations learned by various self-supervised learning techniques significantly depends on the convolutional neural network architecture that was used for solving the self-supervised learning task. In our paper we provide a large scale in-depth study in support of this observation and discuss its implications for evaluation of self-supervised models.

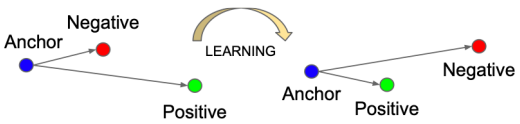
---

Kolesnikov, Alexander, Xiaohua Zhai, and Lucas Beyer. "Revisiting self-supervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

# Contrastive learning (2015/2020)

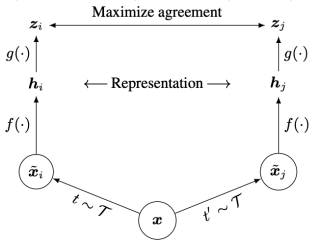
Learn representations by **maximizing agreement** between *differently augmented* views of the same image via a contrastive loss in the latent space $z$

Triplet loss (2015)




Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Normalized Temperature-scaled Cross Entropy (NT-Xent) Loss (2020)

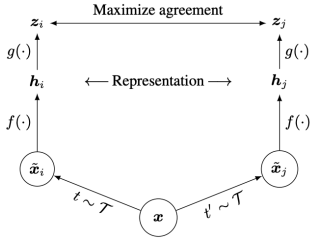



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PmLR, 2020.

# Contrastive learning (2020)

Learn representations by **maximizing agreement** between *differently augmented* views of the same image via a contrastive loss in the latent space $z$

Normalized Temperature-scaled Cross Entropy (NT-Xent) Loss (2020)




Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PmLR, 2020.

Version v2 in the same year

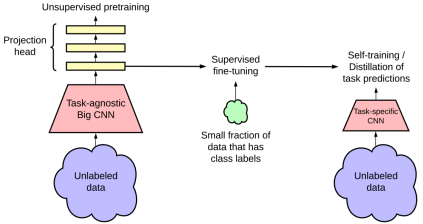



Figure 3: The proposed semi-supervised learning framework leverages unlabeled data in two ways: (1) task-agnostic use in unsupervised pretraining, and (2) task-specific use in self-training / distillation.

## **Bigger models** and **distillation**

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

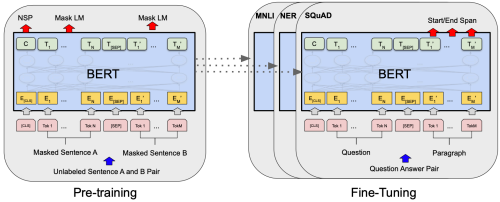
Table 3: ImageNet accuracy of models trained under semi-supervised settings. For our methods, we report results with distillation after fine-tuning. For our smaller models, we use self-distilled ResNet-152 (3×+SK) as the teacher.

| Method | Architecture | Top-1 Label fraction 1% | Top-1 Label fraction 10% | Top-5 Label fraction 1% | Top-5 Label fraction 10% |
|---|---|---|---|---|---|
| Supervised baseline [30] | ResNet-50 | 25.4 | 56.4 | 48.4 | 80.4 |
| *Methods using unlabeled data in a task-specific way:* | | | | | |
| Pseudo-label [11, 30] | ResNet-50 | - | - | 51.6 | 82.4 |
| VAT+Entropy Min. [37, 38, 30] | ResNet-50 | - | - | 47.0 | 83.4 |
| Mean teacher [39] | ResNeXt-152 | - | - | - | 90.9 |
| UDA (w. RandAug) [14] | ResNet-50 | - | 68.8 | - | 88.5 |
| FixMatch (w. RandAug) [15] | ResNet-50 | - | 71.5 | - | 89.1 |
| S4L (Rot+VAT+Entropy Min.) [30] | ResNet-50 (4×) | - | 73.2 | - | 91.2 |
| MPL (w. RandAug) [2] | ResNet-50 | - | 73.8 | - | - |
| CowMix [40] | ResNet-152 | - | 73.9 | - | 91.2 |
| *Methods using unlabeled data in a task-agnostic way:* | | | | | |
| InstDisc [17] | ResNet-50 | - | - | 39.2 | 77.4 |
| BigBiGAN [41] | RevNet-50 (4×) | - | - | 55.2 | 78.8 |
| PIRL [42] | ResNet-50 | - | - | 57.2 | 83.8 |
| CPC v2 [19] | ResNet-161(∗) | 52.7 | 73.1 | 77.9 | 91.2 |
| SimCLR [1] | ResNet-50 | 48.3 | 65.6 | 75.5 | 87.8 |
| SimCLR [1] | ResNet-50 (2×) | 58.5 | 71.7 | 83.0 | 91.2 |
| SimCLR [1] | ResNet-50 (4×) | 63.0 | 74.4 | 85.8 | 92.6 |
| BYOL [43] (concurrent work) | ResNet-50 | 53.2 | 68.8 | 78.4 | 89.0 |
| BYOL [43] (concurrent work) | ResNet-200 (2×) | 71.2 | 77.7 | 89.5 | 93.7 |
| *Methods using unlabeled data in both ways:* | | | | | |
| SimCLRv2 distilled (ours) | ResNet-50 | 73.9 | 77.5 | 91.5 | 93.4 |
| SimCLRv2 distilled (ours) | ResNet-50 (2×+SK) | 75.9 | 80.2 | 93.0 | 95.0 |
| SimCLRv2 self-distilled (ours) | ResNet-152 (3×+SK) | **76.6** | **80.9** | **93.4** | **95.5** |

# Input Masked Models for text (2019) and images (2021)

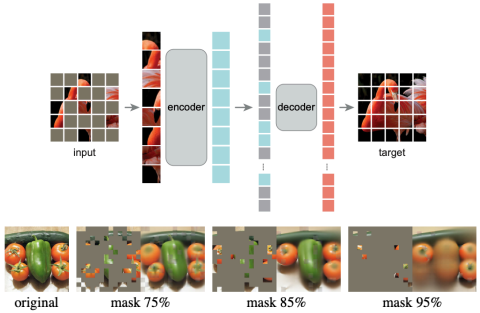
Learn representations by **predicting masked parts of the input** or **prediction of the next part of the input**. Transformer architecture!

Text (Bert or GPT)




Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." 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). 2019.

Image (MAE)



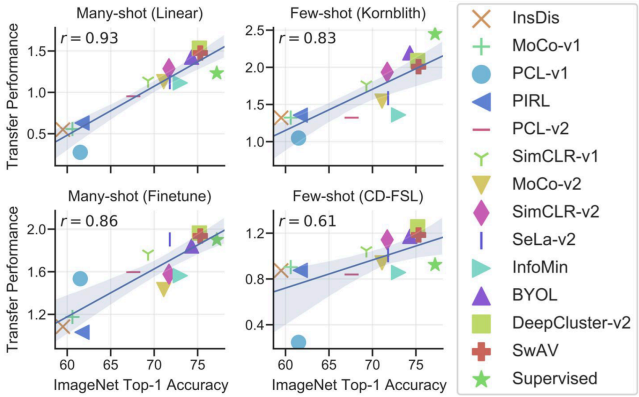

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

# What about now?

The best SSL models outperform supervision

No single self-supervised method dominates overall

SSL induces better classifier calibration




Finally, self-supervised learning is not only **efficient** but also **effective**!

[4]Ericsson, Linus, Henry Gouk, and Timothy M. Hospedales. "How well do self-supervised models transfer?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

# Distillation - student and teacher relation (2021)

Teacher and student share the same architecture

"centering prevents one dimension to dominate but encourages collapse to the uniform distribution, while the sharpening has the opposite effect."

The stop gradient operation is used to **only** update the student model. The teacher model is an **exponential moving average** of the student model.

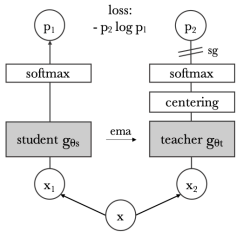



Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views ($x_1$, $x_2$) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a $K$ dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

# Distillation - student and teacher relation (2021)

Teacher and student share the same architecture

"centering prevents one dimension to dominate but encourages collapse to the uniform distribution, while the sharpening has the opposite effect."

The stop gradient operation is used to **only** update the student model. The teacher model is an **exponential moving average** (EMA) of the student model.

**Algorithm 1** DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

# Current the state of the art model - DINOv2 (2023)

Do we reached the end of the road?

| Method | Arch. | Data | Text sup. | kNN | linear | | |
|---|---|---|---|---|---|---|---|
| | | | | val | val | ReaL | V2 |
| **Weakly supervised** | | | | | | | |
| CLIP | ViT-L/14 | WIT-400M | ✓ | 79.8 | 84.3 | 88.1 | 75.3 |
| CLIP | ViT-L/$14_{336}$ | WIT-400M | ✓ | 80.5 | 85.3 | 88.8 | 75.8 |
| SWAG | ViT-H/14 | IG3.6B | ✓ | 82.6 | 85.7 | 88.7 | 77.6 |
| OpenCLIP | ViT-H/14 | LAION-2B | ✓ | 81.7 | 84.4 | 88.4 | 75.5 |
| OpenCLIP | ViT-G/14 | LAION-2B | ✓ | 83.2 | 86.2 | 89.4 | 77.2 |
| EVA-CLIP | ViT-g/14 | custom* | ✓ | **83.5** | 86.4 | 89.3 | 77.4 |
| **Self-supervised** | | | | | | | |
| MAE | ViT-H/14 | INet-1k | × | 49.4 | 76.6 | 83.3 | 64.8 |
| DINO | ViT-S/8 | INet-1k | × | 78.6 | 79.2 | 85.5 | 68.2 |
| SEERv2 | RG10B | IG2B | × | – | 79.8 | – | – |
| MSN | ViT-L/7 | INet-1k | × | 79.2 | 80.7 | 86.0 | 69.7 |
| EsViT | Swin-B/W=14 | INet-1k | × | 79.4 | 81.3 | 87.0 | 70.4 |
| Mugs | ViT-L/16 | INet-1k | × | 80.2 | 82.1 | 86.9 | 70.8 |
| iBOT | ViT-L/16 | INet-22k | × | 72.9 | 82.3 | 87.5 | 72.4 |
| DINOv2 | ViT-S/14 | LVD-142M | × | 79.0 | 81.1 | 86.6 | 70.9 |
| | ViT-B/14 | LVD-142M | × | 82.1 | 84.5 | 88.3 | 75.1 |
| | ViT-L/14 | LVD-142M | × | **83.5** | 86.3 | 89.5 | 78.0 |
| | ViT-g/14 | LVD-142M | × | **83.5** | **86.5** | **89.6** | **78.4** |

Table 4: **Linear evaluation on ImageNet-1k of frozen pretrained features.** We report Top-1 accuracy on the validation set for publicly available models trained on public or private data, and with or without text supervision (text sup.). For reference, we also report the kNN performance on the validation set. We

[4]Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

# Closing Thoughts

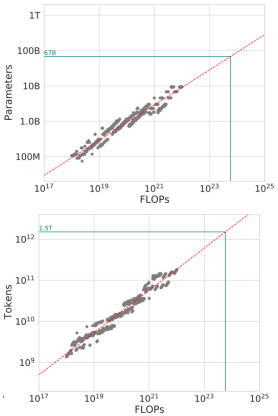
Representations are downstream *task-dependent*, *architecture-dependent*

SSL can result in more transferable features than supervised transfer learning

No single self-supervised method dominates overall

Almost all current best performing models used models that were pretrained with SSL

Only big tech companies can afford to train such models from scratch




Hoffmann, Jordan, et al. "Training compute-optimal large language models." arXiv preprint arXiv:2203.15556 (2022).

For more, read this lecture from the Lab of HHU Dusseldorf (clickable link).