learning with few data

Marcus Liwicki, Machine Learning
Luleå University of Technology

bit.ly/2023-nldl-tutorial

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

LULEÅ UNIVERSITY OF TECHNOLOGY

are you working on your
PhD
or finished recently ?

did you ever

did you ever

feel insignificant

did you ever

feel insignificant

doubt your skills

did you ever

feel insignificant

doubt your skills

or

feel challenged ?

unchallenged

You are not alone !

Marcus Liwicki, Machine Learning
Luleå University of Technology

ELLIS member, WASP member
IEEE senior member, IAPR award winner, …

bit.ly/2023-nldl-tutorial

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

agenda

motivation
prior
approaches
end to end learning
- transfer learning
- clustering

representation learning
- auto-encoding
- contrastive learning

comparative summary
remarks on contrastive learning

and some spices in-between:
what I have learned during my life as presenter

agenda

motivation
prior
approaches
end to end learning
- transfer learning
- clustering

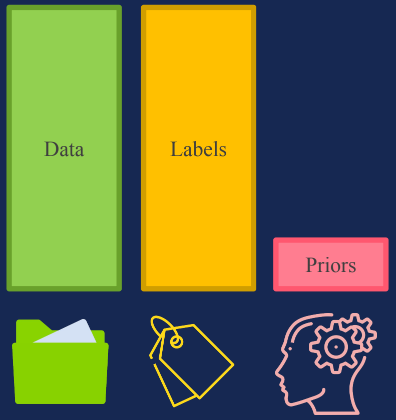
representation learning
- auto-encoding
- contrastive learning

comparative summary
remarks on contrastive learning

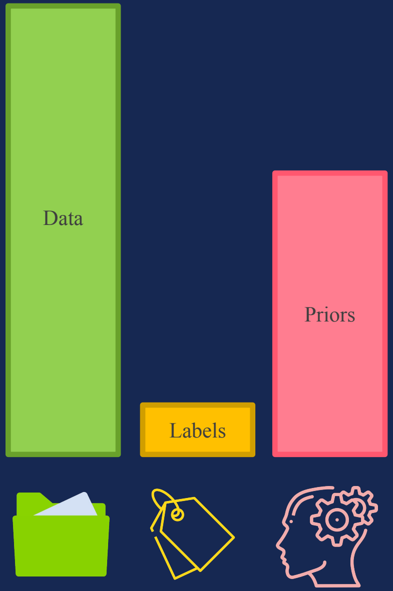
and some spices in-between:
what I have learned during my life as presenter

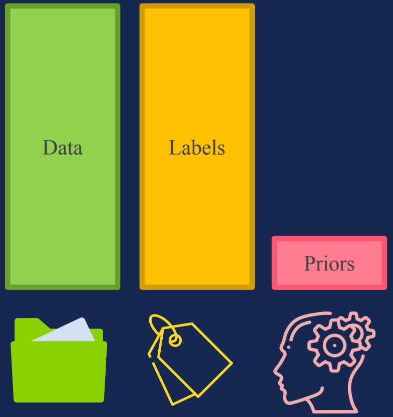
machine learning needs data

machine learning (ideal)

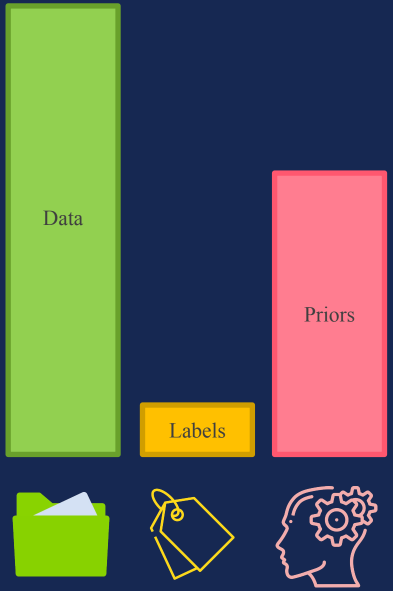


reality





minimize human supervision

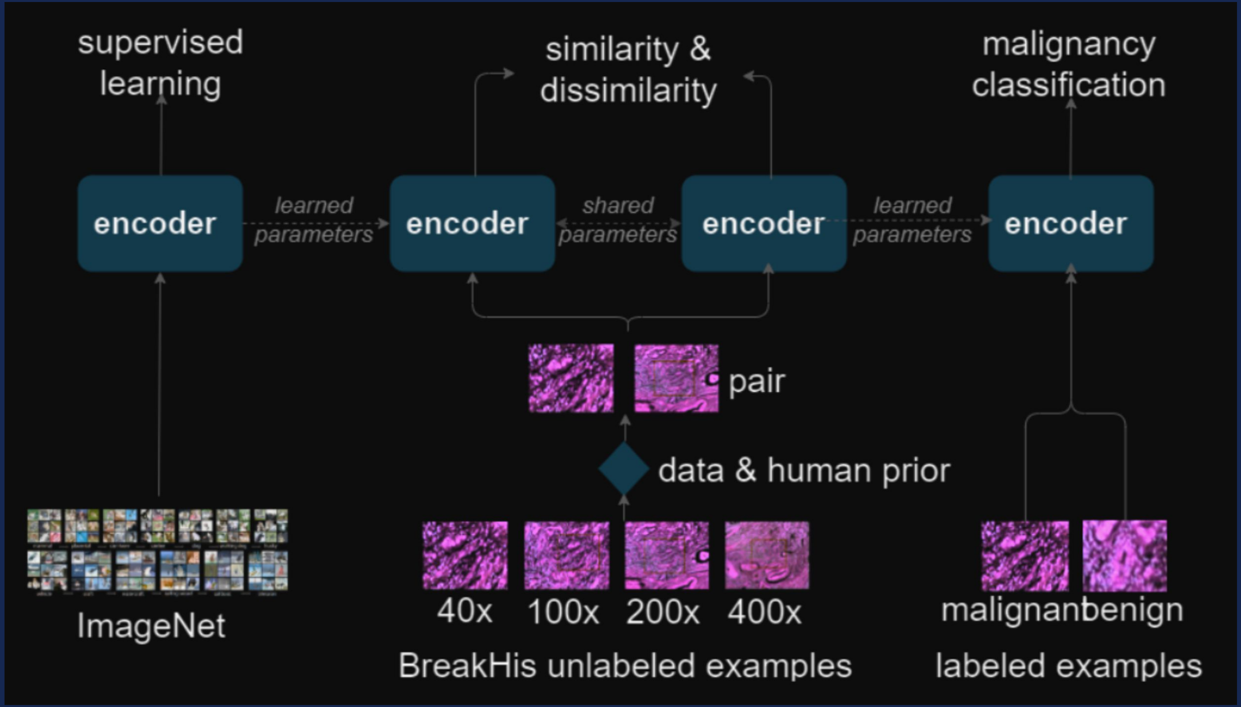


how?

1. adding more unlabeled data or synthetic data
2. incorporating more prior (knowledge)

there are so many priors hidden in structure

there are so many priors hidden in structure



including priors

92.15% (SotA 88.2%)

Better than Google

prior

experience (from earlier experiments)

proven architectures, meta parameters, …

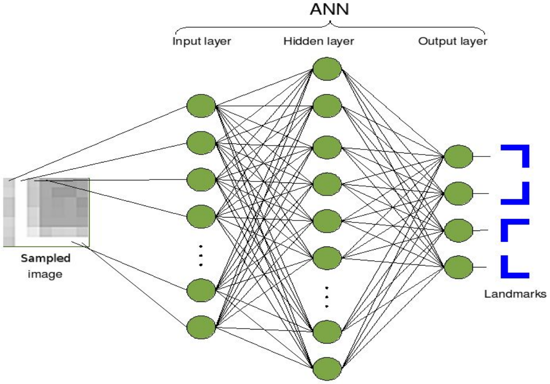
knowledge (human reasoning)

correlating the given input details and identifying discriminative features

data (intrinsic or human induced)

sequential correlation, local correlation

filenames folder structures, taxonomies



x001-t14.xml
x001-t15.xml

time to learn something about presentations ;)

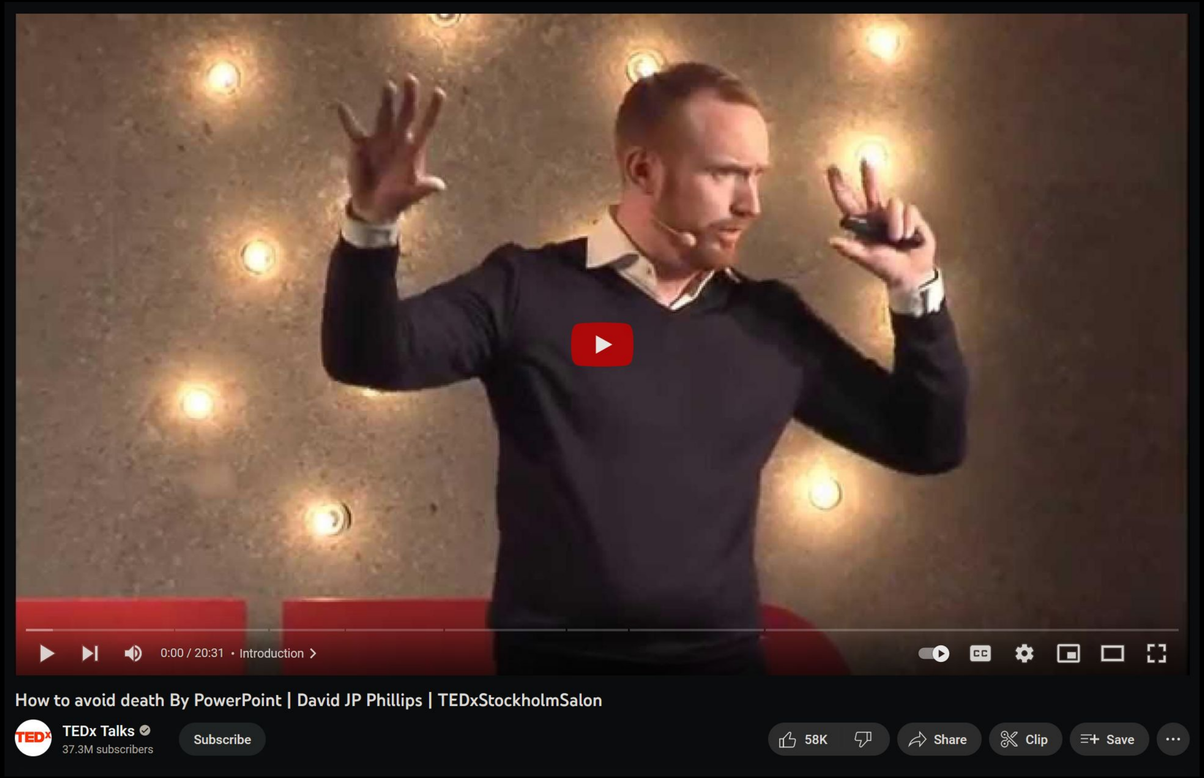
should we use dark background ?

or white ?

ok, enough of the torture

but why did so many of you torture each other?

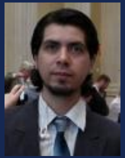
Contrast is important



How to avoid death By PowerPoint | David JP Phillips | TEDxStockholmSalon

equity in the machine learning group

Marcus


Pedro

Konstantina

Gustav

Fotini

Christian

Kanjar

Vibha

Fredrik

5 GENDER EQUALITY

Notice
something?
Almost 4**0%**
woman

Priyamvada

György

Saleha

Rajkumar

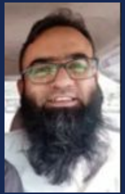
Oluwatosin

Homam

Mattias

Nosheen

Sana


Ali

András

Richa

Karl

Carl

Prakash

Lama

Elisa

machine learning for the welfare of society

thanks to previous and current PhDs

Michele Alberti

Vinay Pondenkandath

Gustav G. Pihlgren

Prakash Ch. Chhipa

overview of approaches

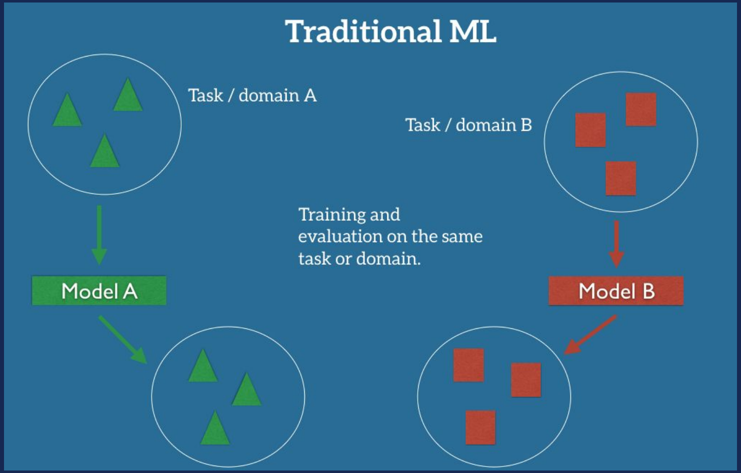
end to end learning

- transfer learning *(A Survey on Deep Transfer Learning - 2018)*
    - Utilizing pretrained models and finetuning on application specific data
    - Required less data to fine tune than training it from scratch
- clustering – *(Deep Clustering for Unsupervised Learning of Visual Features - 2018)*
    - Labelled data not required

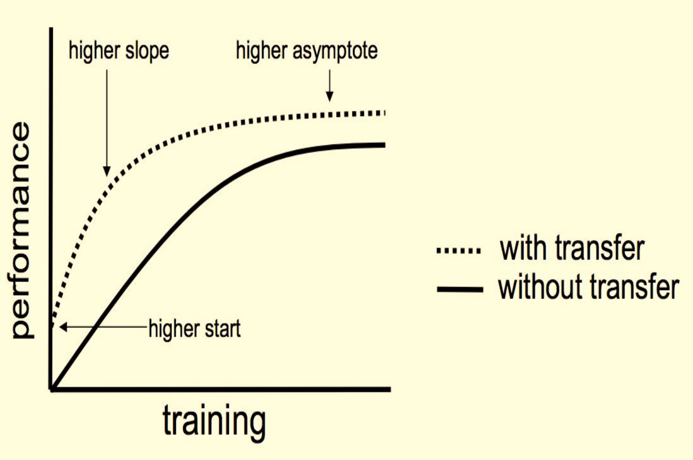
representation learning

- auto-encoding – *(Variational Autoencoder for Deep Learning of Images, Labels and Captions, 2016)*
    - Questionable if this is a good way to go – *(A Pitfall of Unsupervised Pre-Training, 2017)*
- contrastive learning *(SimCLR - July 2020, SwAV – October 2020)*
    - Pretraining mechanism which utilizes application specific unlabeled data
    - Also compute intensive but possibility to scale down

transfer learning



Source: https://ruder.io/transfer-learning/

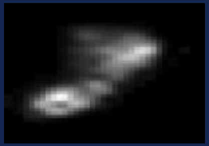


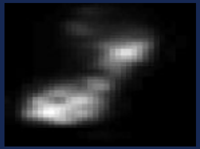
Source: https://machinelearningmastery.com/transfer-learning-for-deep-learning/

remarks

- successful but only initial layers with low-level features are common & useful across applications
- no possibility for unlabeled data

ImageNet pretraining works outside of natural images

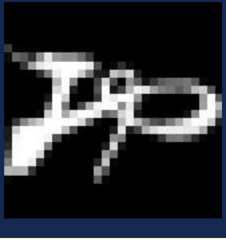


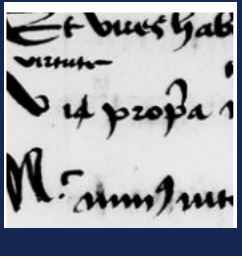


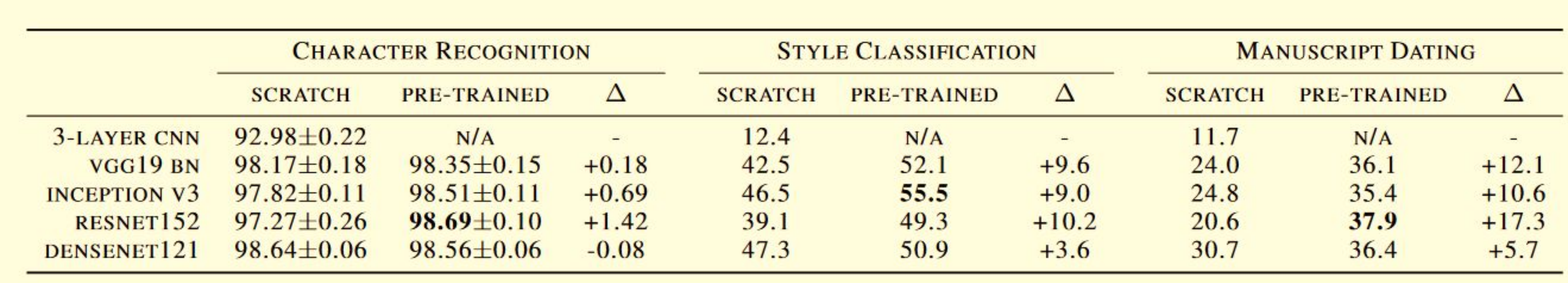
footsteps for person identification
(88 % for 13 persons, previous SotA 77 %)

MS Singh, V Pondenkandath, B Zhou, P Lukowicz, M Liwicki
Transforming sensor data to the image domain for deep learning—An application to footstep detection, IJCNN 2017

ImageNet pre-training works often well







	CHARACTER RECOGNITION			STYLE CLASSIFICATION			MANUSCRIPT DATING		
	SCRATCH	PRE-TRAINED	Δ	SCRATCH	PRE-TRAINED	Δ	SCRATCH	PRE-TRAINED	Δ
3-LAYER CNN	92.98±0.22	N/A	-	12.4	N/A	-	11.7	N/A	-
VGG19 BN	98.17±0.18	98.35±0.15	+0.18	42.5	52.1	+9.6	24.0	36.1	+12.1
INCEPTION V3	97.82±0.11	98.51±0.11	+0.69	46.5	**55.5**	+9.0	24.8	35.4	+10.6
RESNET152	97.27±0.26	**98.69**±0.10	+1.42	39.1	49.3	+10.2	20.6	**37.9**	+17.3
DENSENET121	98.64±0.06	98.56±0.06	-0.08	47.3	50.9	+3.6	30.7	36.4	+5.7

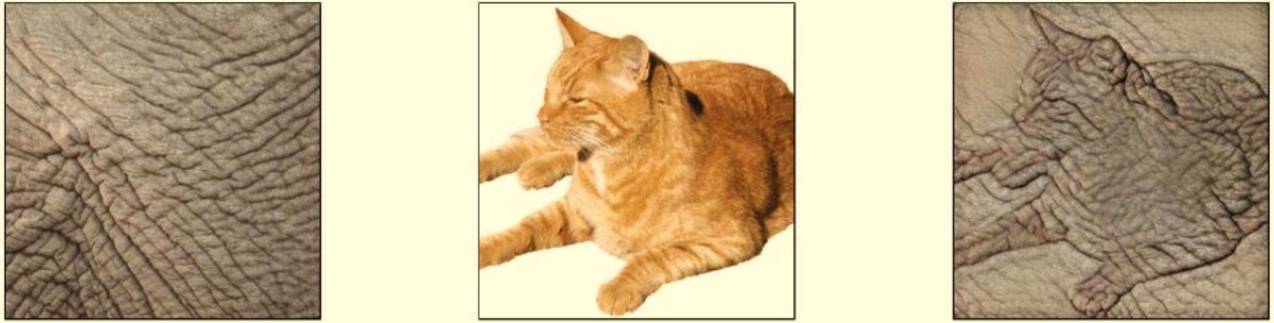
Linda Studer, Michele Alberti, Vinaychandran Pondenkandath, Pinar Goktepe, Thomas Kolonko, Andreas Fischer, Marcus Liwicki, Rolf Ingold:
A Comprehensive Study of ImageNet Pre-Training for Historical Document Image Analysis, ICDAR, 2019

shortcomings – ImageNet transfer learning

ImageNet-trained CNNs are biased towards texture

- Strongly biased towards recognizing textures rather than shapes

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018, September). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations.



(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

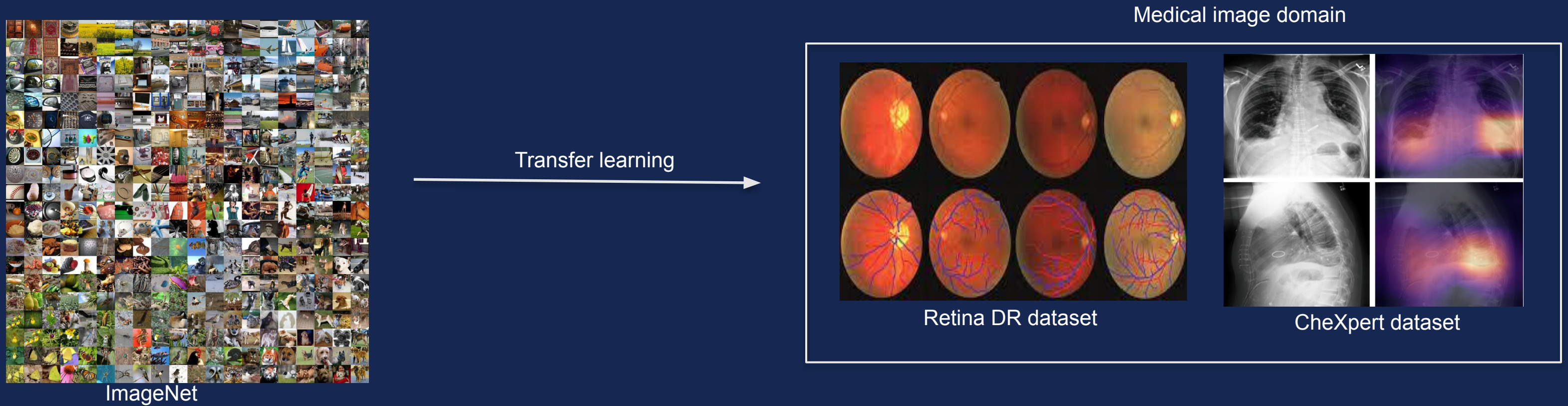
(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

ImageNet transfer learning in medical images

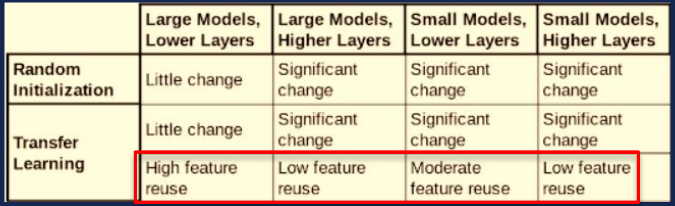


ImageNet transfer learning does not significantly affect performance on medical imaging tasks

- Ref: *Transfusion: Understanding Transfer Learning for Medical Imaging*

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

- Task specific learning - only initial layers with low-level features are useful



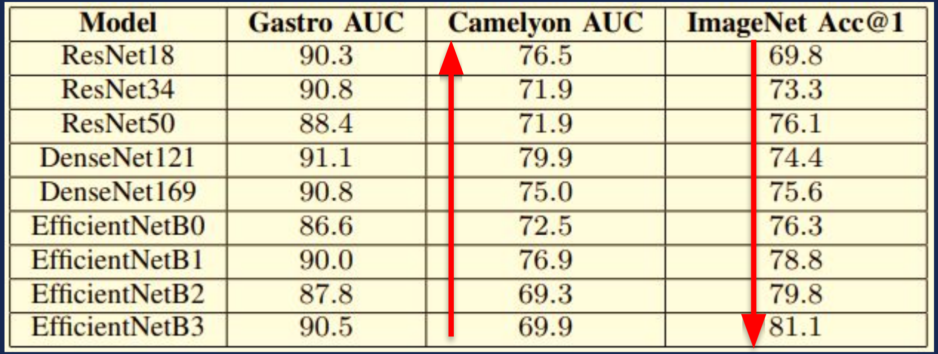
	Large Models, Lower Layers	Large Models, Higher Layers	Small Models, Lower Layers	Small Models, Higher Layers
Random Initialization	Little change	Significant change	Significant change	Significant change
Transfer Learning	Little change	Significant change	Significant change	Significant change
	High feature reuse	Low feature reuse	Moderate feature reuse	Low feature reuse

Adapted from https://ai.googleblog.com/2019/12/understanding-transfer-learning-for.html

ImageNet transfer learning in histopathology

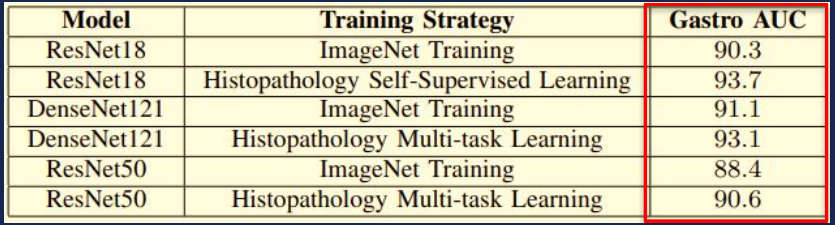
Sharmay, Y., Ehsany, L., Syed, S., & Brown, D. E. (2021, July). HistoTransfer: Understanding Transfer Learning for Histopathology. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-4). IEEE.

Gastrointestinal, breast cancer



Model	Gastro AUC	Camelyon AUC	ImageNet Acc@1
ResNet18	90.3	76.5	69.8
ResNet34	90.8	71.9	73.3
ResNet50	88.4	71.9	76.1
DenseNet121	91.1	79.9	74.4
DenseNet169	90.8	75.0	75.6
EfficientNetB0	86.6	72.5	76.3
EfficientNetB1	90.0	76.9	78.8
EfficientNetB2	87.8	69.3	79.8
EfficientNetB3	90.5	69.9	81.1

ImageNet vs. SSL



Model	Training Strategy	Gastro AUC
ResNet18	ImageNet Training	90.3
ResNet18	Histopathology Self-Supervised Learning	93.7
DenseNet121	ImageNet Training	91.1
DenseNet121	Histopathology Multi-task Learning	93.1
ResNet50	ImageNet Training	88.4
ResNet50	Histopathology Multi-task Learning	90.6

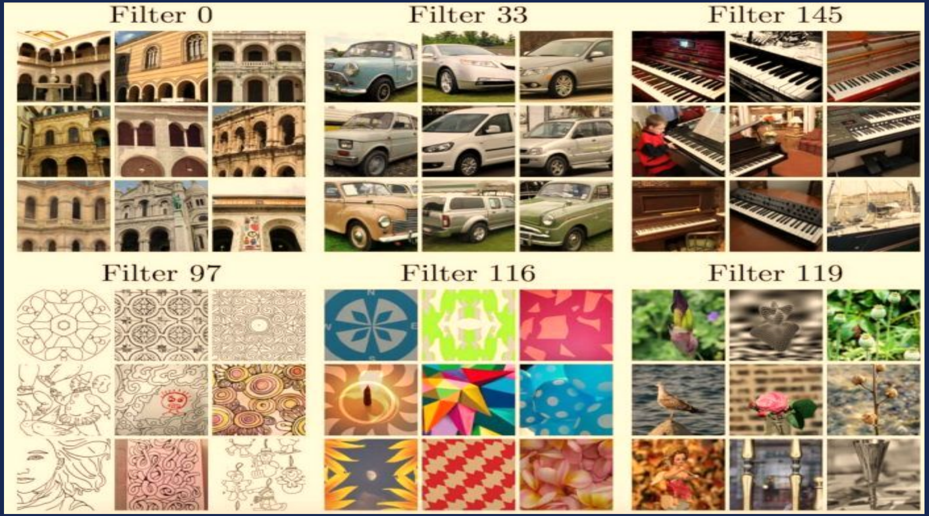
Why ImageNet supervised transfer learning is sub-optimal?

Possibly, ImageNet trained model is overfitted for natural scenes

Optimized for dataset specific characteristics

clustering

group features with k-means and update the weights to optimize for these assignments



Source: https://neurohive.io/en/state-of-the-art/deep-clustering-approach/

remarks

- Compute intensive when applied on images
- Non robust feature representation when feature extracted with pretrained models

agenda

motivation

prior

approaches

end to end learning

- transfer learning
- clustering

representation learning

- auto-encoding – and alternatives
- contrastive learning

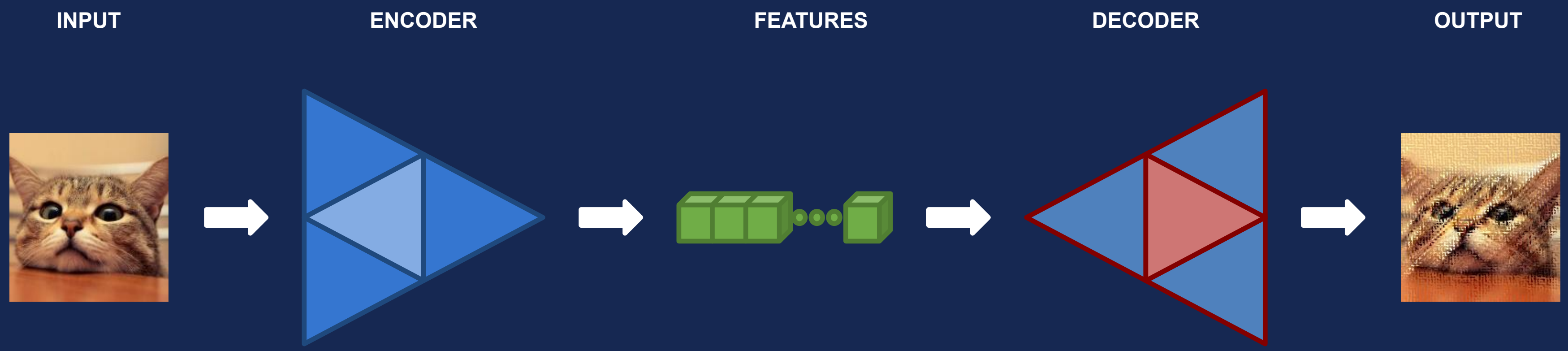
comparative summary

remarks on contrastive learning

and some spices in-between:

what I have learned during my life as presenter

Auto-Encoding – pre-training



INPUT

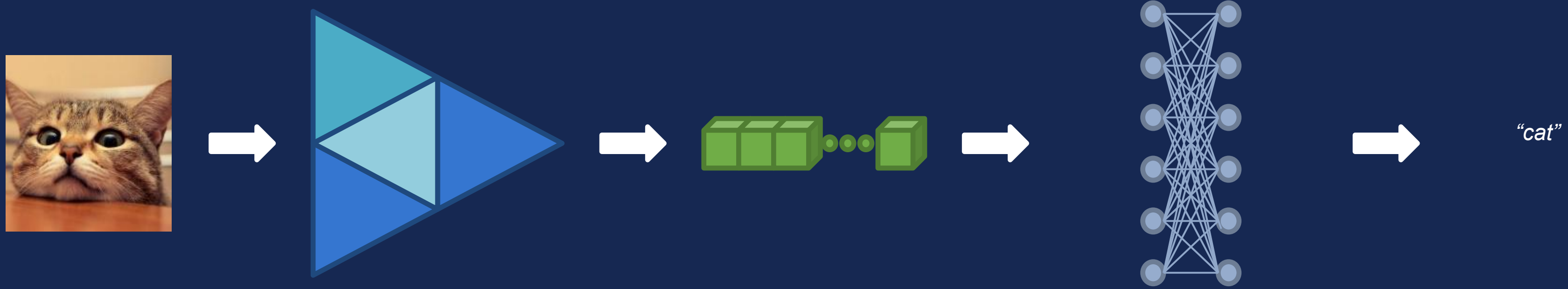
ENCODER

FEATURES

DECODER

OUTPUT

Auto-Encoding – classification



INPUT

ENCODER

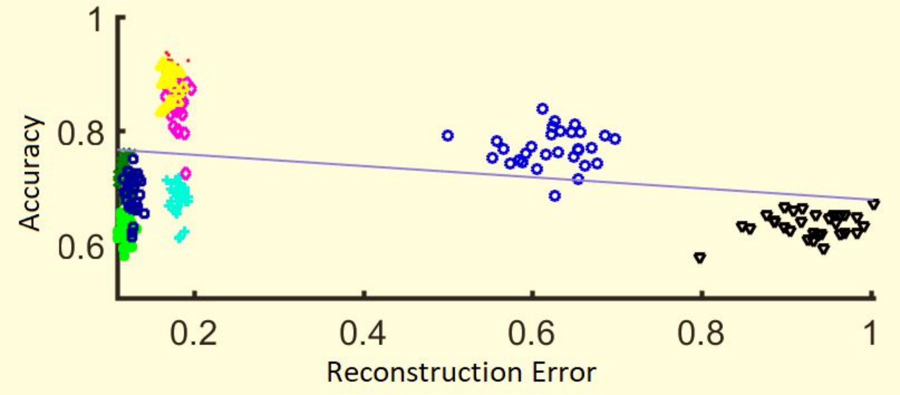
FEATURES

CLASSIFIER

OUTPUT

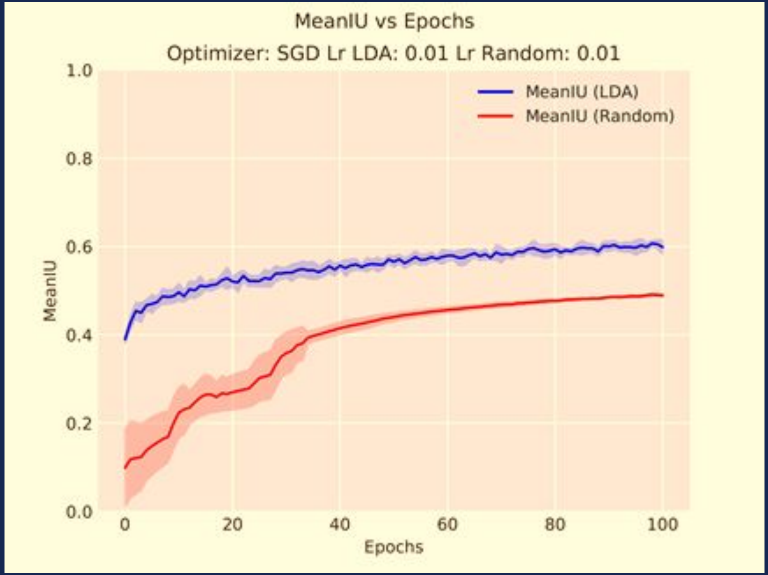
“cat”

a pitfall of unsupervised pre-training, 2017



a good auto-encoder (low reconstruction error) does not necessarily lead to better accuracy

alternative: use PCA or LDA for initialization



Will they converge ?

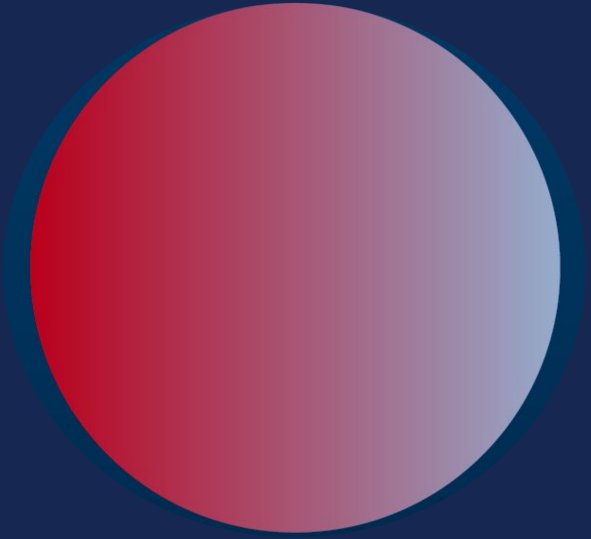
No !

Better local minima ?

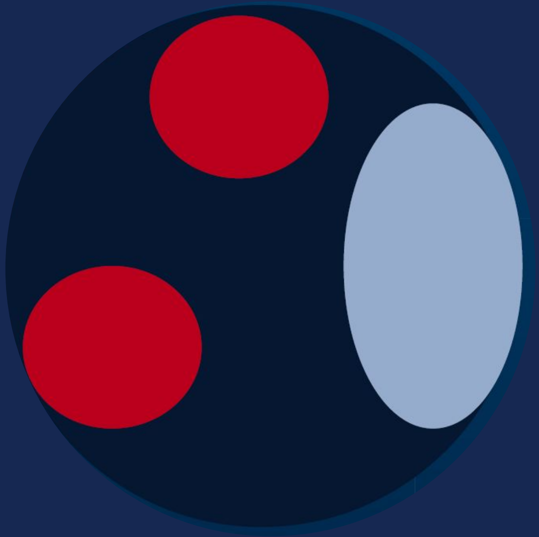
Michele Alberti, Mathias Seuret, Vinaychandran Pondenkandath, Rolf Ingold, Marcus Liwicki
Historical Document Image Segmentation with LDA-Initialized Deep Neural Networks. ICDAR 2017

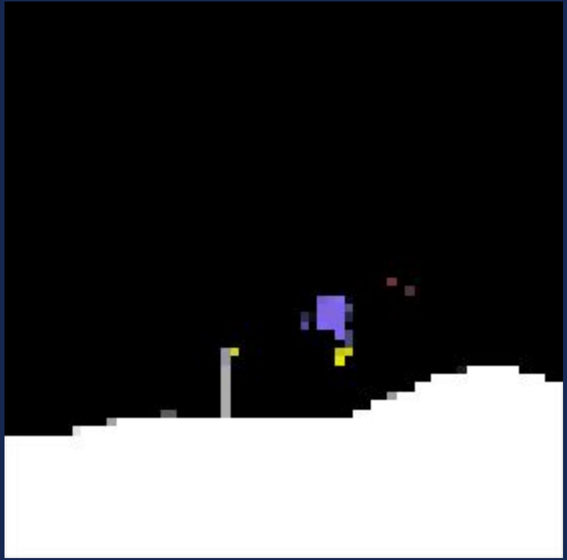
auto-encoding limitation

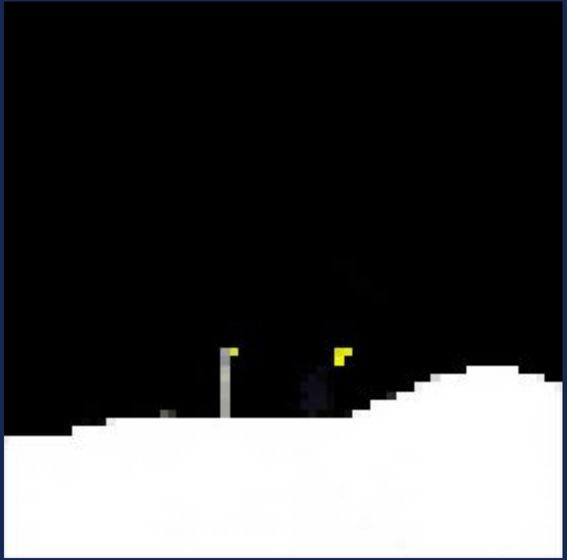
what we want



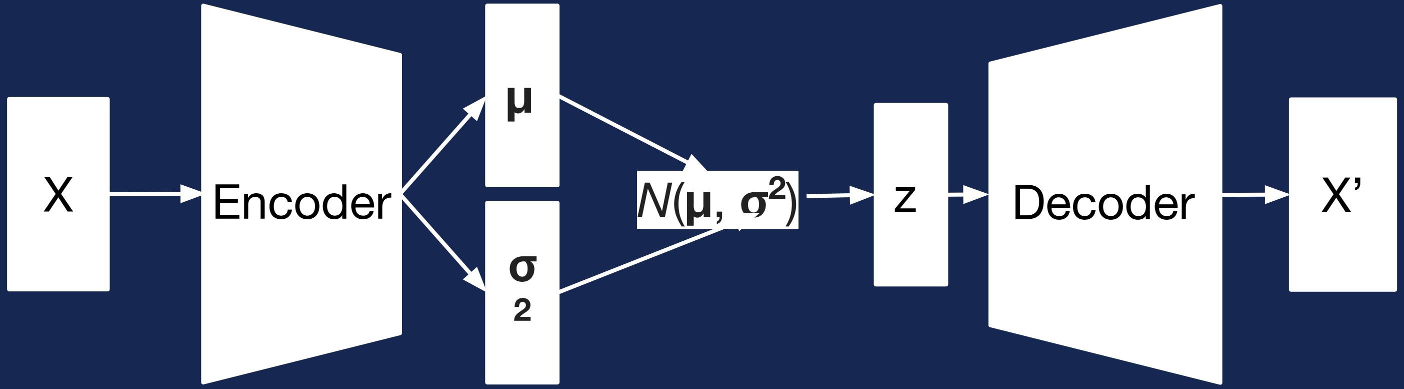
what we might get





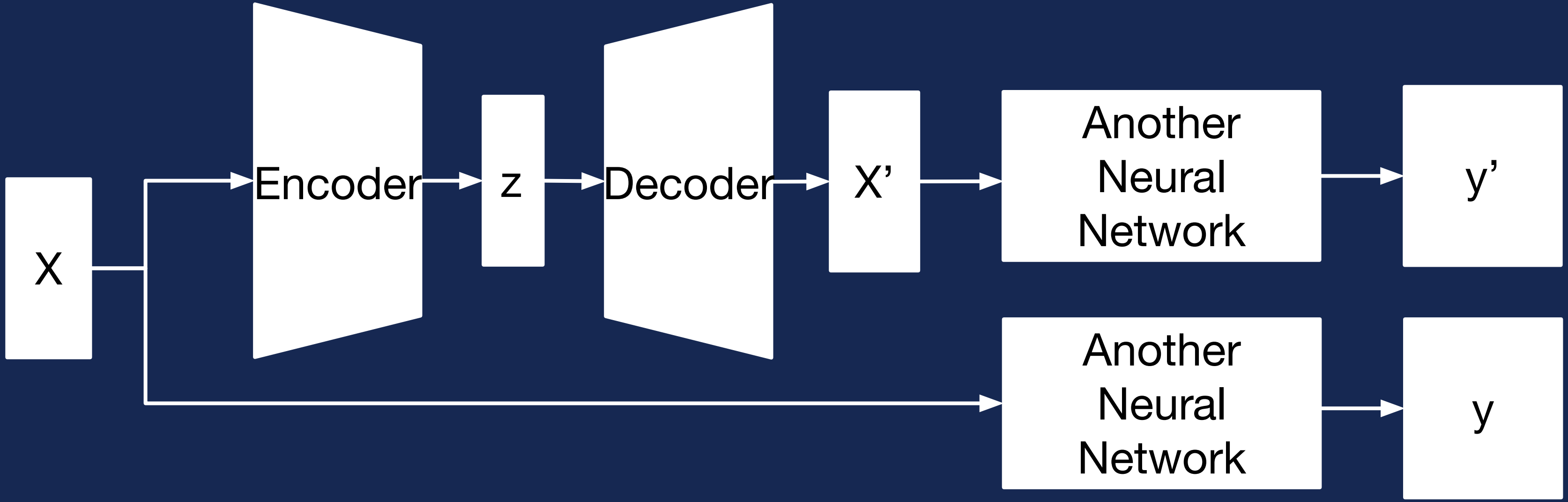


variational auto-encoders

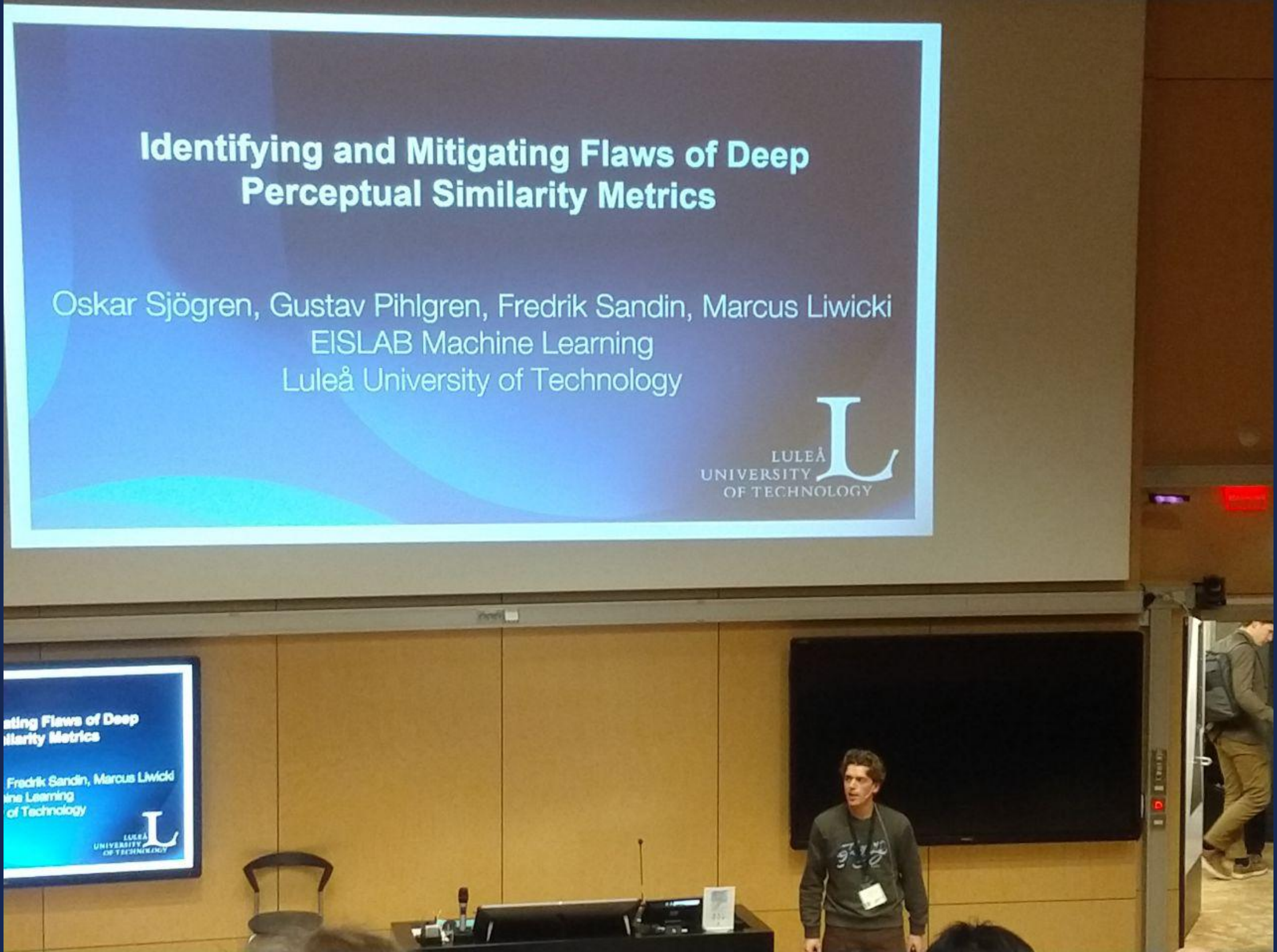


Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes.“ 2013

perceptual loss



Thorough investigation :
Improving image autoencoder embeddings with perceptual loss, 2020
And Oskar Sjögren (yesterday)



Identifying and Mitigating Flaws of Deep Perceptual Similarity Metrics

Oskar Sjögren, Gustav Pihlgren, Fredrik Sandin, Marcus Liwicki
EISLAB Machine Learning
Luleå University of Technology

try it out …

bit.ly/2023-nldl-tutorial

https://github.com/guspih/Perceptual-Autoencoders

https://github.com/guspih/Perceptual-Encoding

https://github.com/guspih/deep_perceptual_similarity_analysis

Contrastive Learning (CL)

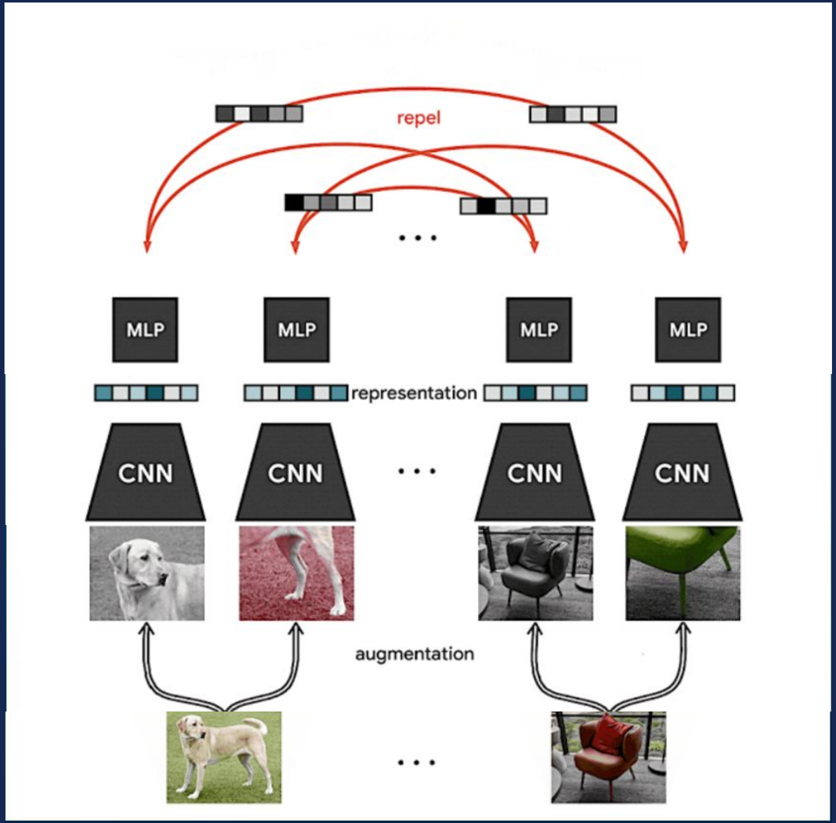
Self-Supervised Method:

Allows model to learn generic representations on unlabeled data

Method:

Learn similarity between augmented representation from same image

Learn dissimilarity otherwise



Source: https://ai.googleblog.com/2020/04/advancing-self-supervised-and-semi.html

(not so) recent work in Contrastive Learning

Simple Framework for Contrastive Learning (SimCLR)

A Simple Framework for Contrastive Learning of Visual Representations (SimCLR v1), ICML - 2020

Big Self-Supervised Models are Strong Semi-Supervised Learners (SimCLR v2), NeurIPS – 2020

Momentum Contrast Learning (MOCO)

Momentum Contrast for Unsupervised Visual Representation Learning (MOCO v1), CVPR - Mar 2020

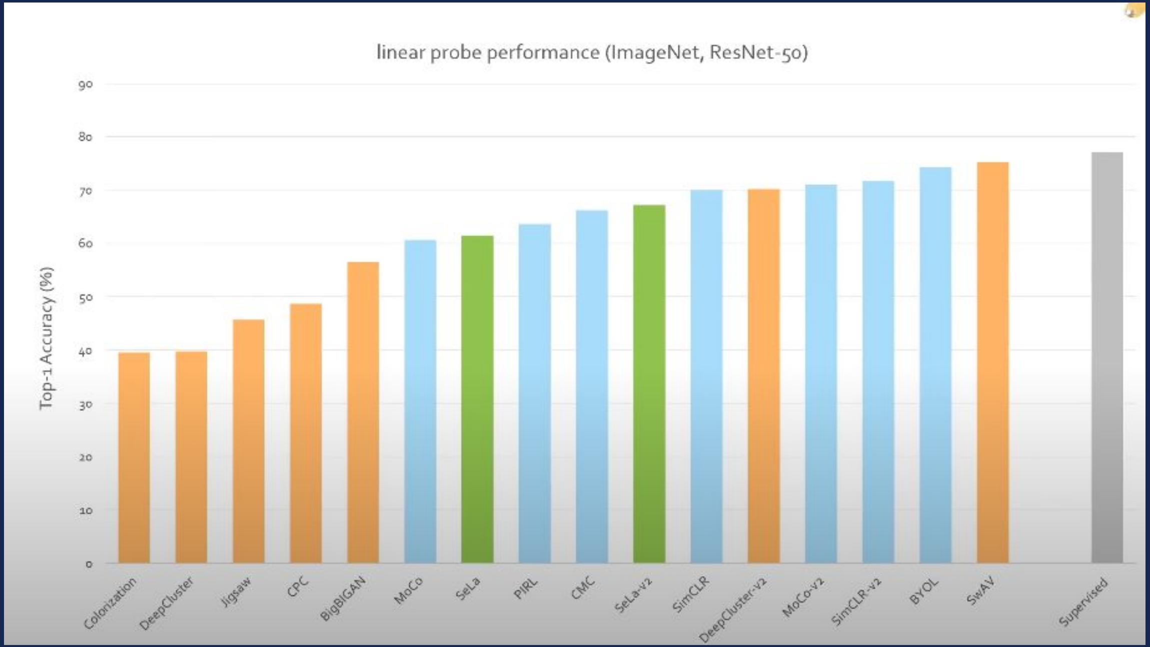
Improved Baselines with Momentum Contrastive Learning (MOCO v2), ?? Arxiv Oct- 2020

Bootstrap Your Own Latent A New Approach to Self-Supervised Learning, NeurIPS - 2020

Contrastive Learning with Clustering

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments (SwAE), Arxiv 2020

Comparative Summary on SOTA



Source (IARAI): https://www.youtube.com/watch?v=Bn66HnBxXFM

🟦 Contrastive Learning

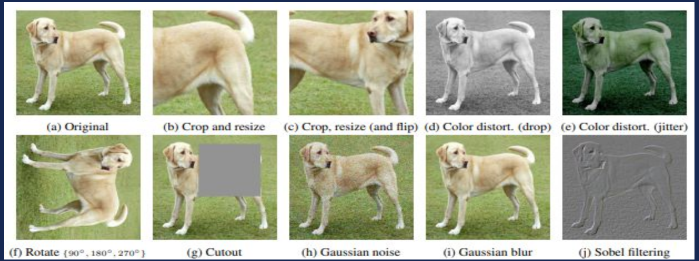
🟧 Clustering + Self-supervised

🟩 Self-Labelling

- Remarks
  - Priors (augmentation mechanism) is more important than learning method
  - Obtains performance approx. equal to supervised methods with 10% labelled data

it’s easy on natural images

distorted views (augmented views) of input visual



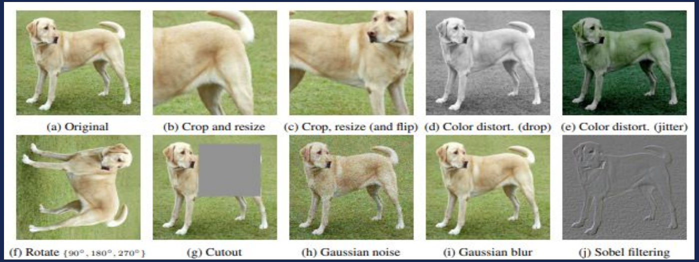
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Human prior for visual	Relevant Augmentation
Size	Resize
Shape	Crop, Flip
Foreground-Background	Blur, Noise, Color schemes, filtering
Angle	Flip, Rotation
Color spectrum	Contrast, saturation

challenge in adapting SOTA self-supervised methods in another specialized domain (Not so natural visual concepts)

- Joint-embedding based self-supervised methods has following core components:
  - Distorted views (augmented views) of input visual ~ Helps in learning generalized representation about visual concepts to network
  - Objective function - similarity metrics selection in loss function



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Human prior for visual	Relevant Augmentation
Size	Resize
Shape	Crop, Flip
Foreground-Background	Blur, Noise, Color schemes, filtering
Angle	Flip, Rotation
Color spectrum	Contrast, saturation

- Enabling comprehensive distorted views for natural visual concepts is easy with human prior using obvious knowledge of visual world
- Thus, state-of-the-art methods in self-supervised learning are mainly optimized for natural visual

- What about the other vision domain beyond natural visual concepts i.e., medical images, remote sensing imagery, non-obvious visual concepts? – *It makes existing state-of-the-art methods sub-optimal due to insufficiency of human prior for distorted view* – next slide

But does not work in other domains

Distorted views (augmented views) of input visual

(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

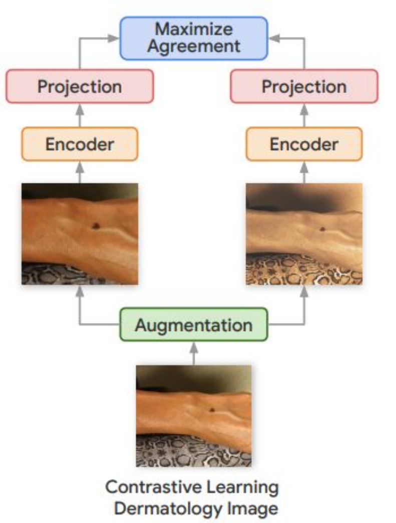
(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Human prior for visual	Relevant Augmentation
Size	Resize
Shape	Crop, Flip
Foreground-Background	Blur, Noise, Color schemes, filtering
Angle	Flip, Rotation
Color spectrum	Contrast, saturation

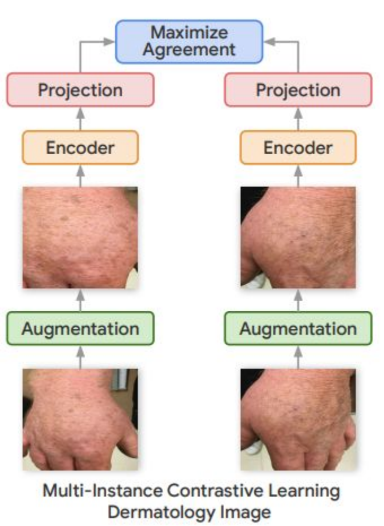
medical images, remote sensing imagery, non-obvious visual concepts

insufficiency of human prior for distorted view

Use two views of same patient



Contrastive Learning
Dermatology Image



Multi-Instance Contrastive Learning
Dermatology Image

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3478-3488).

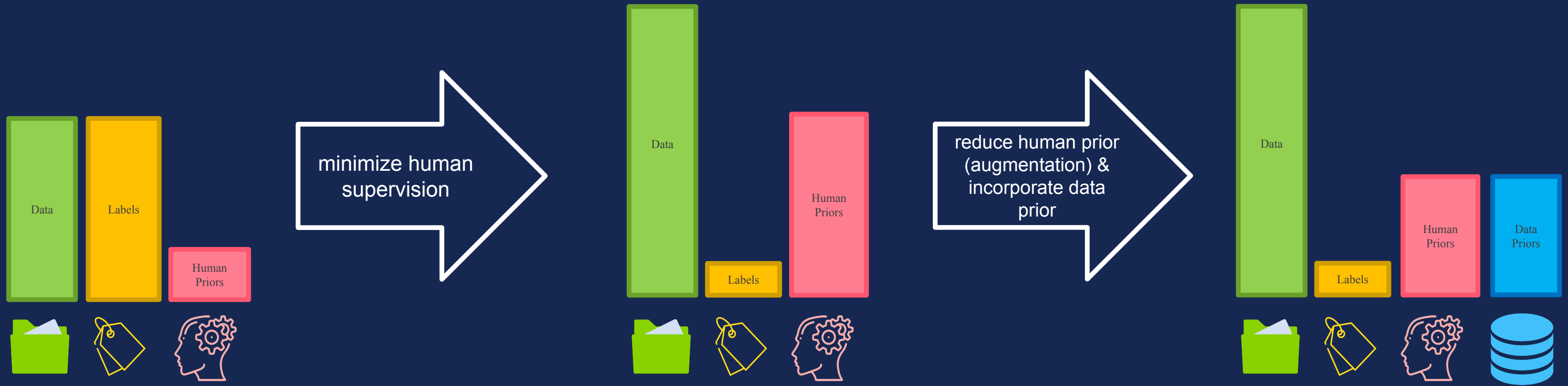
But wait … did we use labels ?

Our Approach – Shifting focus from human prior to data prior

Supervised approach

Self-supervised approach (on natural visual concepts)

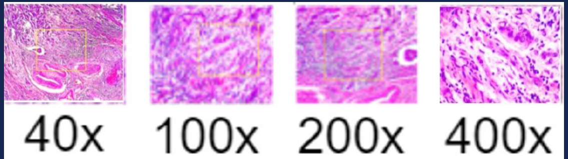
Adapting self-supervised approach on specialized domain



minimize human supervision

reduce human prior (augmentation) & incorporate data prior

let us use the data prior



data (prior) magnification levels (in BreakHis data) are utilized to generate both views for SSL input

the only human prior used in magnification sampling

Achieves state-of-the-art results with only 20% labels on classification

Chhipa, P. C., Upadhyay, R., Pihlgren, G. G., Saini, R., Uchida, S., & Liwicki, M. (2022). Magnification Prior: A Self-Supervised Method for Learning Representations on Breast Cancer Histopathological Images. arXiv preprint arXiv:2203.07707.

ideas for data prior

temporal proximity

spatial proximity

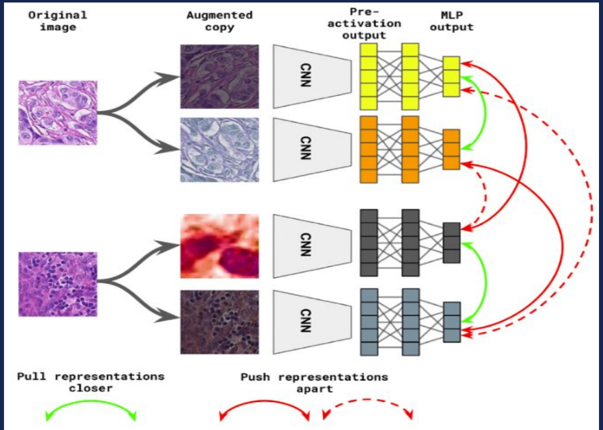
sequential co-occurrence (BERT)

different modalities

more ?

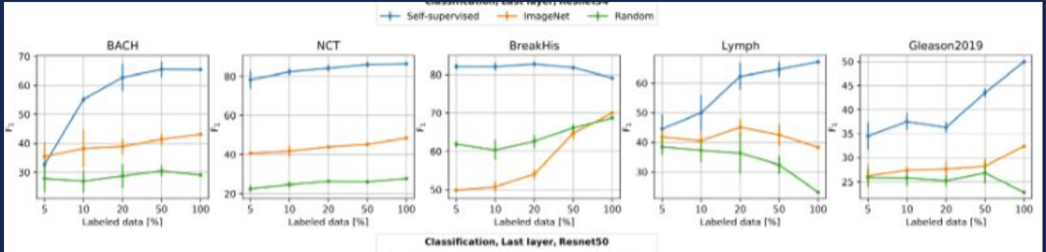
Adapting SSL on histopathology images

- Contrastive learning on collectively 57 datasets



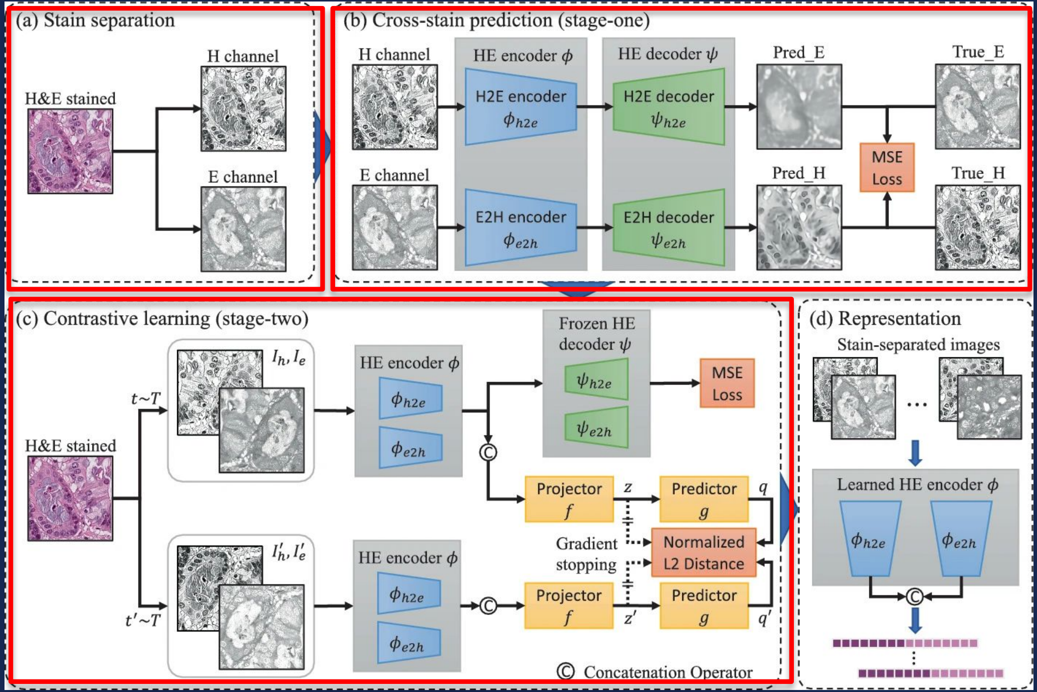
- Adapts SSL to histopathology domain by combining massive and diverse datasets
- Outperform over ImageNet (supervised) transfer learning with significant margins on multiple target datasets
  - Multiple downstream tasks
  - BACH, NCT, BreakHis, Lymph, many more

Ciga, O., Xu, T., & Martel, A. L. (2022). Self supervised contrastive learning for digital histopathology. Machine Learning with Applications, 7, 100198.

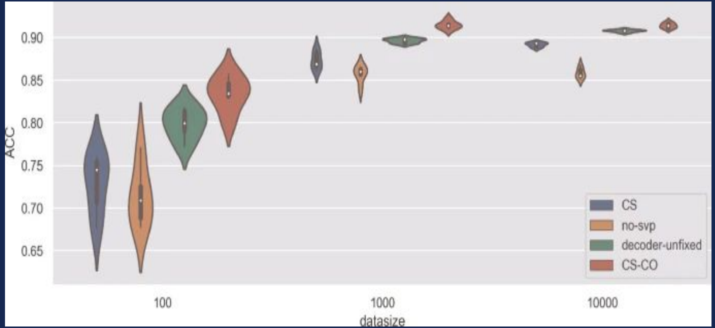


SSL on histopathology using H&E staining (domain specific details)

Framework of the proposed CS-CO method – multi-stage pretraining



- Dataset – human colorectal cancer dataset (NCT-CRC-HE-100K)



- Combining cross-staining prediction with contrastive learning works well

Yang, P., Hong, Z., Yin, X., Zhu, C., & Jiang, R. (2021, September). Self-supervised visual representation learning for histopathological images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 47-57). Springer, Cham.

Curious, what more we can learn about presentation techniques ?

Btw., should we use slide numbers ?

typical issues, I observe at scientific conferences :

typical issues, I observe at scientific conferences :

unconfident posture

typical issues, I observe at scientific conferences :

unconfident posture

filler sounds

typical issues, I observe at scientific conferences :

unconfident posture

filler sounds

angle and interaction

agenda

motivation
prior
approaches
end to end learning
- transfer learning
- clustering

representation learning
- auto-encoding
- contrastive learning

comparative summary
remarks on contrastive learning

And some spices in-between:
What I have learned during my life as presenter

summary

end to end learning

- transfer learning
- clustering

representation learning

- auto-encoding
- PCA, LDA
- perceptual loss
- contrastive learning

meta learning (not covered today)

remarks on contrastive learning

Method	Contrastive Learning Key Factor	Contribution	Limitation
SimCLR V1.0	K1: Similarity learning for positive pairs K2: Dissimilarity learning for negative pairs	Established benchmark performance on unsupervised contrastive learning	1. 'Large batch size' due to positive + negative pair 2. 'Mass gradient computation & backprop issue' due to all (+ve & -ve) pairs
SimCLR V2.0	K1 + K2 on Task agnostic Big n/w which used in distillation for task specific small n/w	+ Added enablement of semi-supervised learning through distillation	Same as SimCLR V1.0 + usage of bigger networks
MOCO V1.0	K1 + K2 over momentum encoder where CL as dynamic dictionary lookup	Revealed unsupervised contrastive learning with smaller batch size and lessor backpropagation of gradients	1. 'Mass gradient computation & backprop issue' due to all (+ve & -ve) *pairs (same as SimCLR because as q-encoder backpropagates)* 2. Overhead of dynamic dictionary queue
MOCO V2.0	MOCO V1.0 + 2-layer MLP projection head	Stronger baseline, outperformed on SimCLR and MOCO v1.0.	1. 'Mass gradient computation & backprop issue' due to all (+ve & -ve) pairs *same as SimCLR because q-encoder and k-encoder both backpropagates* 2. Overhead of dynamic dictionary queue
BYOL	K1+ momentum encoding + two separate networks (online and target)	Achieves self supervised CL without negative pair. Establishes benchmarks in semi-supervised approach. Robust for smaller batch size.	1. Complex pipeline with large number of pruning. Makes it challenging for concept utilization.
SwAE	K1 + Swapped" prediction mechanism + cluster assignment	Achieves self supervised CL without negative pair. Claims state of art in unsupervised image clustering.	1. Relatively complex loss computation due to swapped prediction 2. Additional online cluster assignment swapping
DINO	Distillation transformers	Self attention without supervision Moderate computation power	1. More research required 2. Authors are not self-critical
Barlow Twins	Redundancy reduction	minimize covariance across embedding dimension Maximize invariance across sample	

Remarks on Contrastive Learning

CL is leading the self-supervision & potential push for semi-supervised

CL in current state is compute intensive

batch size is huge

SimCLR, performance increase, when batch size of 2048
Reason: large number of negative pairs
□requires array of GPUs and sophisticated parallel processing

knowledge distillation (BYOL 2020, SimSiam 2020) do not use negative pairs
□batch size 512
However, embedding output size in range of 4096

For non natural images, smaller batch size is already good (128)
Reason: not RGB images, but simpler

Remarks on Contrastive Learning

CL is leading the self-supervision & potential push for semi-supervised

CL in current state is compute intensive (batch size, negative pairs, & gradients) which makes it challenging for direct (as-it-is) applications. Needs (Research Potential) to be tailored for custom and small-scale application requirement.
Contrastive methods are sensitive to the choice of image/data augmentation.

Leveraging utilization of application specific but unlabeled data.

CL can be benchmarking framework (Different methods for different applications) for semi-supervised and even supervised task.

Thanks to my colleagues

There is so much more, I could share

https://irdta.eu/deeplearn/2023su/

bit.ly/2023-nldl-tutorial

LULEÅ
UNIVERSITY
OF TECHNOLOGY