Self-Supervised Learning: Training Targets and Loss Functions

Tutorial at NLDL 2023, Tromsø, Norway

Prof. Zheng-Hua Tan, zt@es.aau.dk
Aalborg University, Denmark
Pioneer Centre for AI, Denmark

Chat with you.com

AALBORG UNIVERSITY
DENMARK

- Q: What is self-supervised learning

- A: Self-supervised learning is a machine learning process where the model trains itself to learn one part of the input from another part of the input [1][2]. It is also known as unsupervised representation learning, and is a subset of unsupervised learning. Self-supervised learning obtains supervisory signals from the data itself [3], often leveraging the underlying structure in the data [3]. It is used to create representations of data that can be used for downstream tasks, such as classification or object detection. The general technique of self-supervised learning is to create a task from the input data [3], such as predicting the rotation angle of an image, and then to use the model to solve the task. The model is then able to learn useful features from the data using the task as a form of supervision.

Chat with you.com – cont'd

AALBORG UNIVERSITY
DENMARK

- Q: What is the relationship between northern lights and deep learning

- A: The relationship between the northern lights and deep learning is that deep learning can be used to analyze and understand the physical mechanisms behind the northern lights. By studying the physical properties of the aurora, researchers can use deep learning to gain insights into the dynamics of the northern lights, and develop models that can better predict the behavior and characteristics of the aurora. In addition, deep learning can also be used to identify and analyze auroral features that can help us better understand the aurora and its behavior.

Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Learning paradigms
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

The talk discusses selective topics in self-supervised learning, by no means aiming to be comprehensive. Given the vast literature of SSL, many excellent works will inevitably be outside of the coverage of this talk.

Learning with or without supervision

kaola

kangaroo

"A picture is worth a thousand words."

$\{\boldsymbol{x}_n, y_n\}$ vs
$\{\boldsymbol{x}_n\}$
in terms of amount of supervision
& learnt representations

Learning with or without supervision – cont'd

AALBORG UNIVERSITY
DENMARK

Corpus:

The zoo is nice.
The kaola looks very cute. — positive review

The zoo seems too small.
The location is too far away. — negative review

N-gram language model:

$P(\text{the}) = 0.2;$
$P(\text{zoo} \mid \text{the}) = 0.5;$
$P(\text{is} \mid \text{the, zoo}) = 0.5;$
$P(\text{nice} \mid \text{the, zoo, is}) = 1;$

- Data itself contains abundant, implicit supervisory signals
- Difficult to obtain sufficient labeled data in many domains

Supervised learning pipeline

AALBORG UNIVERSITY
DENMARK

Statistical supervised learning pipeline (looping as in MLOps)

1. Collect data
2. Provide labels or training targets, manually
3. Define a loss function
4. Optimize a selected model
5. Apply the model for inference

The targets are fixed:

- continuous numbers
- discrete categories

- L1 or L2 for estimating numbers (regression)
- Cross-entropy or margin-based for categorizing (classification)

Self-supervised learning (SSL) pipeline

AALBORG UNIVERSITY
DENMARK

Statistical *SSL* pipeline (looping as in MLOps)

1. Collect SSL data for a *pretext* task
2. Generate training targets, automatically (*self-found*) → The targets may not be fixed
3. Define a loss function
4. Pre-train a selected model
5. Collect data for a downstream task
6. Provide training targets, manually
7. [Define a loss function]
8. [Fine-tune the model] → Omitted for zero-shot learning
9. Apply the model for inference

From supervised learning to SSL

AALBORG UNIVERSITY
DENMARK

- Supervised learning
  - Tasks
    - Regression
    - Classification
  - Training targets:
    - Continues numbers
    - Discrete categories
  - Loss function examples
    - L1 and L2
    - Likelihood
    - Cross-entropy

- SSL
  - Pretext tasks
    - Regression (AR), reconstruction
    - Classification, contrast
  - Training targets (self-found)
    - Continues numbers
    - Discrete categories, data instances (in batches)
  - Loss function examples
    - L1 and L2
    - Likelihood
    - Cross-entropy
    - Contrastive loss

From supervised learning to SSL – cont'd

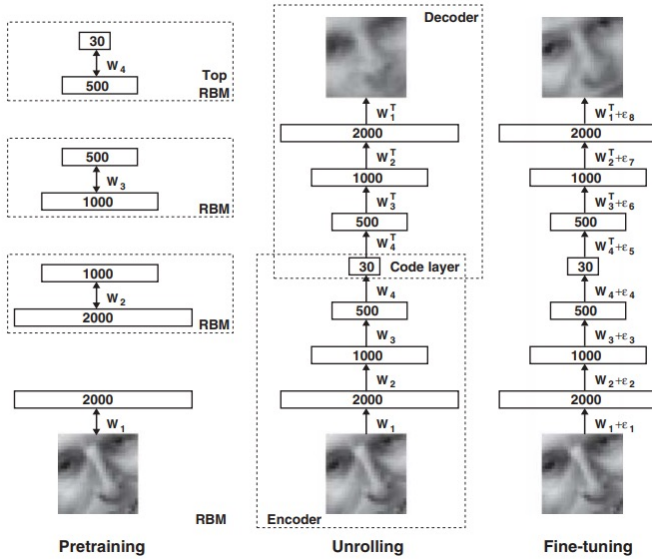
AALBORG UNIVERSITY
DENMARK

- Supervised learning's challenges
  - data efficiency, generalization error
  - much labeled data is needed

- Data, unlabeled for a specific task in hand, is massively available
  - Unsupervised pre-training and supervised fine-tuning
  - Self-supervised learning

Pre-training and fine-tuning

- Pretraining restricted Boltzmann machines for creating a deep autoencoder, then fine-tuned [Hinton & Salakhutdinov, 2006]

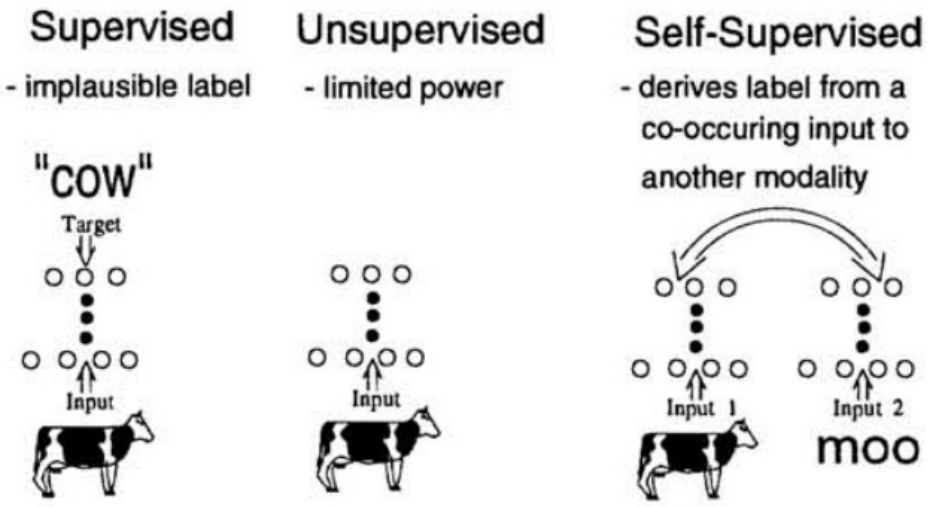


- Not end-to-end training with explicit supervisory signals

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.

Self-supervised learning

- Three learning paradigms [de Sa, 1994]



- # SSL publications = x^a with $a \approx 2$ [Liu, 2021]

de Sa, V. R. (1994). Learning classification with unlabeled data. Advances in neural information processing systems, 112-112.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering.

Training paradigms

AALBORG UNIVERSITY
DENMARK

Methods	Given	Purpose	Data
Supervised	$\{\boldsymbol{x}_n, y_n\}_N$	Find $\tilde{y}$ for $\boldsymbol{x}$	Much labeled data
Semi-supervised	$\{\boldsymbol{x}_n, y_n\}_N$ & $\{\boldsymbol{x}_m\}_M$ $M>>N$	Find $\tilde{y}$ for $\boldsymbol{x}$	Mixture of labeled and unlabeled data
Un-supervised	$\{\boldsymbol{x}_n\}_N$	Find the structure of $\{\boldsymbol{x}_n\}_N$	Unlimited, unlabeled data
Self-supervised	$\{\boldsymbol{x}_n\}_N$	Find latent representation $\boldsymbol{z}_n$ for $\boldsymbol{x}_n$ in a supervised learning fashion with $\{\boldsymbol{y^{ssl}}_n\}$ derived from $\{\boldsymbol{x}_n\}$	Unlimited, unlabeled data

Self-supervised learning: A machine learning paradigm that learns a general representation from unlabeled data and via auto-labeling for downstream tasks with or without fine-tuning

Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Supervised, unsupervised and self-supervised learning
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

SSL for text – a huge success

- Language modeling: distribution estimation
- Text corpus: $\{\boldsymbol{x}_n\}_{\mathrm{N}}$
  where $\boldsymbol{x}_n$ is a sequence of tokens $(s_1, \ldots, s_M)$
- The joint probability is factorized as

$$P(\boldsymbol{x}_n) = \prod_{m=1}^{M} P(s_m | s_1, \ldots, s_{m-1}; \boldsymbol{\theta})$$

- The objective is to maximize the log-likelihood:

$$L(\boldsymbol{x}_n) = \sum_{m=1}^{M} log P(s_m | s_{m-k}, \ldots, s_{m-1}; \boldsymbol{\theta})$$

  where k the size of the context window.

Corpus (each sentence is $\boldsymbol{x}_n$):

The kaola looks very cute.
The zoo is nice.

The zoo seems too small.
The location is too far away.

N-gram language model:
P(the) = 0.2;
P(zoo | the) = 0.5;
P(is | the, zoo) = 0.5;
P(nice | the, zoo, is) = 1;

The joint probability:
P(the, zoo, is, nice)

Training targets and loss functions

AALBORG UNIVERSITY
DENMARK

- The training target is the next (discrete) token s_m
- For softmax activation function, the loss function becomes the standard cross-entropy.

Generative Pre-trained Transformer (GPT-1)

AALBORG UNIVERSITY
DENMARK

- Unsupervised pre-training via maximizing log-likelihood:

$$L_{SSL}(\boldsymbol{x}_n) = \sum_{m=1}^{M} logP(s_m|s_{m-k}, \dots, s_{m-1}; \boldsymbol{\theta}_{SSL})$$

- Supervised fine-tuning: log-linear classifier
  - One linear output layer is added on the pre-trained transformer model to predict y_n:

$$P(y_n|\boldsymbol{x}_n) = softmax(h_n^l W_y)$$

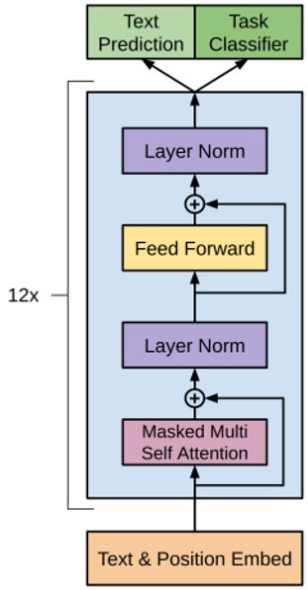
    where h_n^l is the final transformer block's activation.
  - The objective is to maximize

$$L_{SL}(\boldsymbol{x}_n) = \sum_{(\boldsymbol{x}_n, yn)} logP(y_n|\boldsymbol{x}_n; \boldsymbol{\theta}_{SL})$$

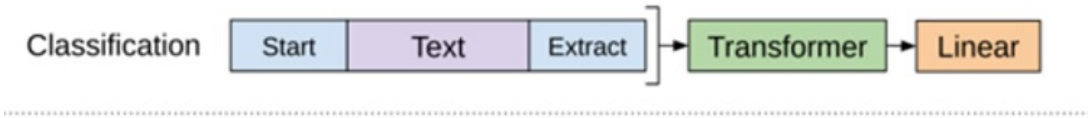
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

GPT-1 – cont'd

- Transformer architecture and training objectives (L_{SSL}, L_{SL})



- Downstream tasks
  - Classification
  - Similarity
  - Entailment
  - Multiple choice



Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Large-scale pre-trained models

AALBORG UNIVERSITY
DENMARK

- GPT-1 (117m parameters)
- GPT-2 (1.5b) models $\mathrm{P}(\mathrm{output} \mid \mathrm{input}, \mathrm{task})$ [Radford, 2019]
- GPT-3 (175b) [Brown, 2020], (generated news with close to 50% acc for human to detect)
- BERT: Bidirectional Encoder Representations from Transformers ($\mathrm{BERT}_{\mathrm{Large}}$: 340m parameters) [Devlin, 2018] (use a masked language model pre-training objective)
- Wu Dao (悟道) (1.75 trillion parameters) [Ding 2021] (a multimodal AI model)
- ChatGPT [OpenAI, 2022]

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., ... & Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, *34*.

OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2022. URL: https://openai. com/blog/chatgpt/ (visited on 08/01/2023).

Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Supervised, unsupervised and self-supervised learning
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

Predictive coding

AALBORG UNIVERSITY
DENMARK

- In a similar way as text, speech is a sequential signal
  - continuous magnitudes than discrete words, though

- The concept of using past sequence of magnitudes to estimate the next sequence of magnitudes is in the core of predictive coding [Elias, 1955]

- This inspires several predictive coding-based SSL methods
  - autoregressive predictive coding (APC)
  - contrastive predictive coding (CPC)

Elias, P. (1955). Predictive coding--I. IRE transactions on information theory, 1(1), 16-24.

Contrastive predictive coding (CPC)

AALBORG UNIVERSITY
DENMARK

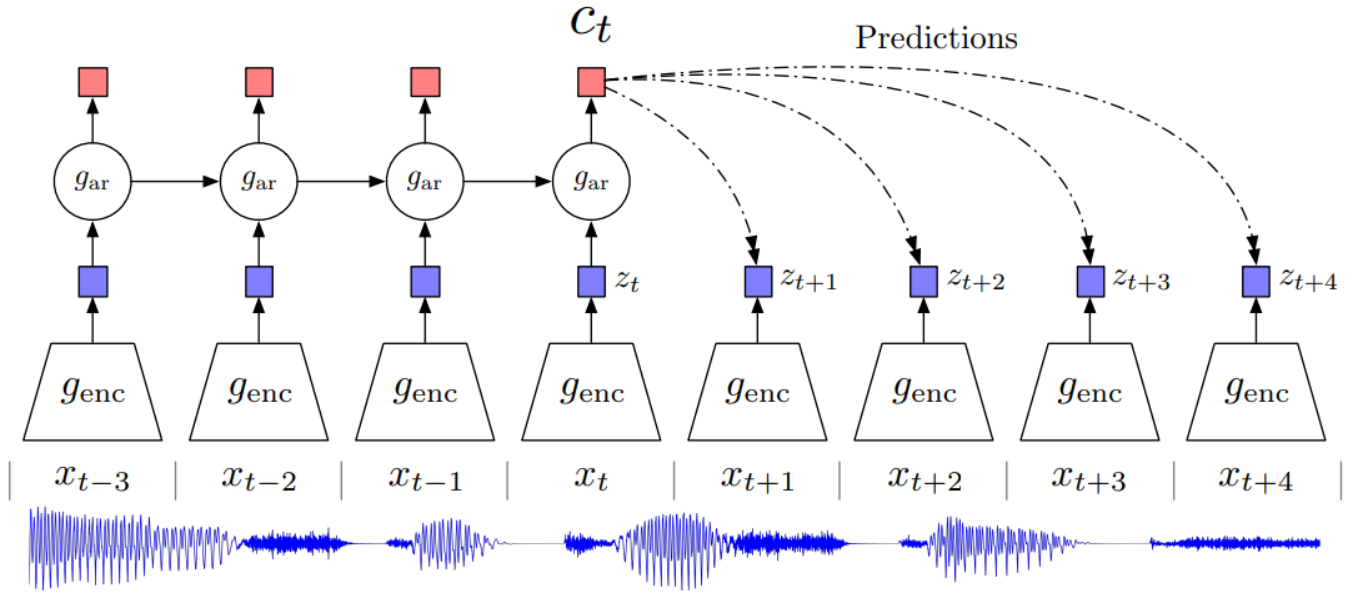
- CPC predicts a future frame $\boldsymbol{x}_{n+m}$ based on history $\boldsymbol{H} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, using autoregressive models.
  - CPC predicts the future frame in latent space, not in data space
  - Not using a regression loss to directly predict $\boldsymbol{x}_{n+m}$, CPC uses a contrastive loss to learn representations most discriminative between $\boldsymbol{x}_{n+m}$ and a set of randomly sampled frames $\{\boldsymbol{x}_i\}$.

- CPC encodes the target $\boldsymbol{x}$ (future) and context $\boldsymbol{c}$ (present) to maximize the mutual information of the original signal $\boldsymbol{x}$ and the context $\boldsymbol{c}$: $I(\boldsymbol{x};\ \boldsymbol{c})$

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Contrastive predictive coding - cont'd

- Architecture of CPC model (applicable to audio, text and image)



Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

InfoNCE loss – widely used

AALBORG UNIVERSITY
DENMARK

- In CPC, both the encoder and AR model are trained jointly to optimize a loss (called InfoNCE), which is based on NCE
  - NCE: noise-contrastive estimation (classifying data vs noise samples) [Gutmann, 2010]
- InfoNCE loss [Oord, 2018]: given *N* random samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, containing one positive sample from $p(\boldsymbol{x}_{t+k}|\boldsymbol{c}_t)$, and N-1 negative samples from the "proposal" distribution $p(\boldsymbol{x}_{t+k})$, it optimizes

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

where $$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

- The loss is the categorical cross-entropy of classifying the positive sample correctly.

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. The 13th international conference on AI and statistics.

Pre-training with autoregressive predictive coding (APC)

AALBORG UNIVERSITY
DENMARK

- Retain as much information about the original signals as possible

- At each time step, the encoder produces a prediction $\tilde{\boldsymbol{y}}_n$, a future frame.
- L1 loss

$$\sum_{n=1}^{N-m} |\boldsymbol{y}_n - \tilde{\boldsymbol{y}}_n|, \boldsymbol{y}_n = \boldsymbol{x}_{n\ +\ m}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ is the training-target sequence, and $\tilde{\boldsymbol{y}} = (\tilde{y_1}, \tilde{y_2}, \ldots, , \tilde{y_N})$ is the predicted sequence.

- It outperforms several strong baselines including CPC, in the reported experiments.

Chung, Y. A., & Glass, J. (2020, May). Generative pre-training for speech with autoregressive predictive coding. In ICASSP 2020, IEEE.

Factorized hierachical VAE with APC loss

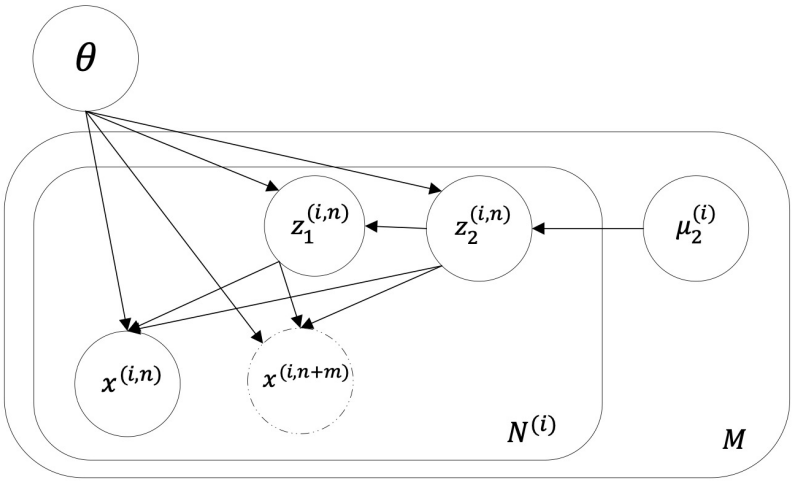


Fig. 2. Generative model of FHVAE-APC. The new variable introduced in FHVAE-APC is highlighted using the dashed circle.

Xie, Y., Arildsen, T., & Tan, Z. H. (2021, October). Disentangled speech representation learning based on factorized hierarchical variational autoencoder with self-supervised objective. MLSP 2021. IEEE.

A probabilistic view of APC losses

AALBORG UNIVERSITY
DENMARK

- The l_2 loss: $L = \frac{1}{2}\sum_{t=1}^{T-k}\|x_{t+k} - Wh_t\|_2^2$
- The l_2 norm can be seen as an isotropic Gaussian. Let's define [Yang]

$$p(x_{t+k}|h_t) \propto \exp\left(-\frac{1}{2}(x_{t+k} - Wh_t)^T I^{-1}(x_{t+k} - Wh_t)\right).$$

Then, the l_2 loss function can be rewritten as [Yang]

$$L = -\sum_{t=1}^{T-k} logp(x_{t+k}|h_t) = -\sum_{t=1}^{T-k} logp(x_{t+k}|x_1, \dots, x_t)$$

which has a likelihood interpretation.

- The l_1 loss is equivalent to maximizing the likelihood under the Laplacian (double exponential) distribution.
- Laplacian distribution is a good approximation for clean speech [Petsatodis].

Yang, G. P., Yeh, S. L., Chung, Y. A., Glass, J., & Tang, H. (2022). Autoregressive Predictive Coding: A Comprehensive Study. IEEE Journal of Selected Topics in Signal Processing.
Petsatodis, T., Boukis, C., Talantzis, F., Tan, Z. H., & Prasad, R. (2011). Convex combination of multiple statistical models with application to VAD. IEEE Transactions on Audio, Speech, and Language Processing, 19(8), 2314-2327.

A probabilistic view of APC losses - cont'd

AALBORG UNIVERSITY
DENMARK

- A von Mises-Fisher distribution can be considered too

$$p(x_{t+k}|h_t) \propto \exp\left(x_{t+k}^T W h_t\right)$$

where x_{t+k} no longer needs to be an "output" of a neural network.

- More generally, we can use a neural network g to process x_{t+k}, and then we have

$$p(x_{t+k}|h_t) \propto \exp\left(g(x_{t+k}^T) W h_t\right)$$

which reminds us of contrastive predictive coding (CPC) that optimizes

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

where

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

Yang, G. P., Yeh, S. L., Chung, Y. A., Glass, J., & Tang, H. (2022). Autoregressive Predictive Coding: A Comprehensive Study. IEEE Journal of Selected Topics in Signal Processing.

Contrastive loss functions

AALBORG UNIVERSITY
DENMARK

Contrastive learning aims at minimizing the distance between positive samples and maximizing distance between negative samples, SSL or supervised, (which LDA also aims to do)

- Contrastive loss [Hadsell, 2006]
- Triplet loss [Weinberger, 2009]
- N-pairs [Sohn, 2016]
  - It applies the softmax function to each positive pair relative to all other pairs. The N-pairs loss is also known as InfoNCE [Musgrave, 2020].

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. IEEE CVPR.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of machine learning research, 10(2).

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems, 29.

Musgrave, K., Belongie, S., & Lim, S. N. (2020, August). A metric learning reality check. In European Conference on Computer Vision (pp. 681-699). Springer, Cham.

($C_{N,2}$ + 1)-pair loss function for noise-robust KWS

- Noise-robust keyword spotting based on res15 architecture

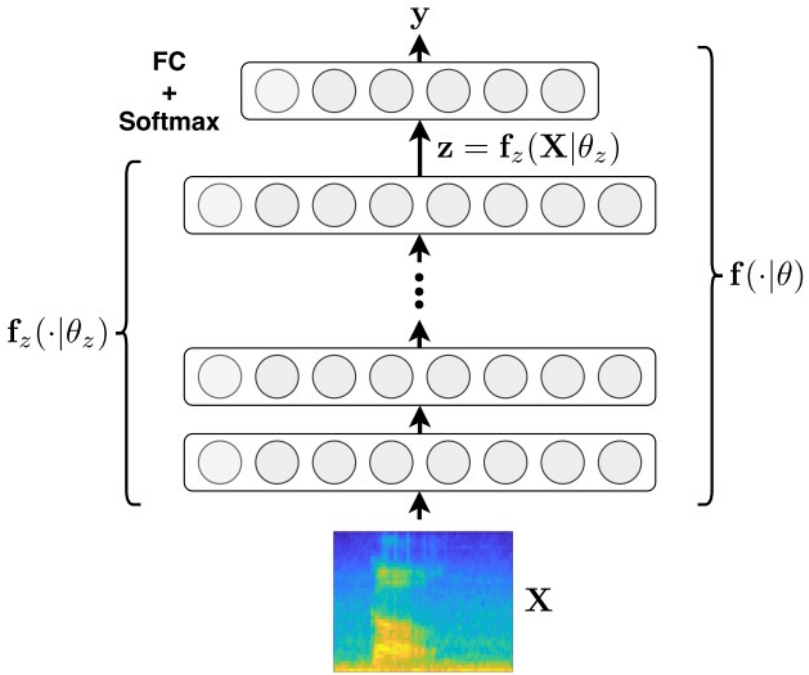


Fig. 1. General state-of-the-art KWS approach. "FC + Softmax" stands for fully-connected layer with softmax activation. See the text for further details.

López-Espejo, I., Tan, Z. H., & Jensen, J. (2021). A Novel Loss Function and Training Strategy for Noise-Robust Keyword Spotting. IEEE/ACM Transactions on ASLP.

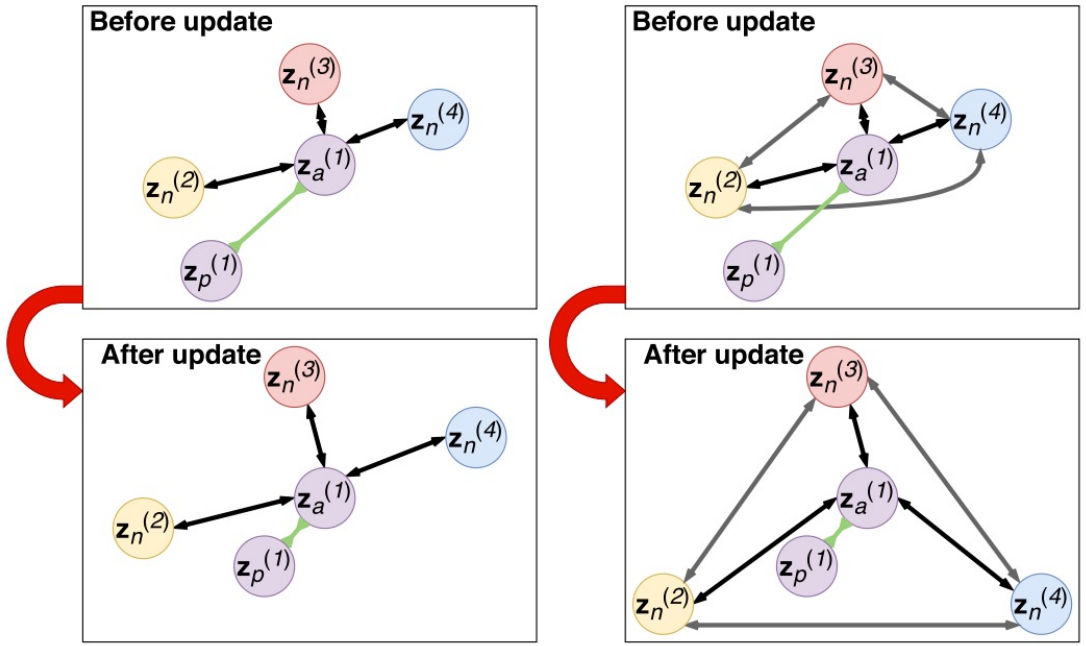
($C_{N,2}$ + 1)-pair loss function for noise-robust KWS - cont'd

- In N-pair loss, N−1 negative examples are pushed away from the anchor example.
- However, distance between these negative examples is not under control.
- To make negative examples distant from each other, a ($C_{N,2}$ + 1)-pair loss, also maximizes the distance among the N−1 negative examples (knowing, or assuming, they are negative towards each other) as

$$\mathcal{L}'_{(C_{N,2}+1)\text{-pair}} = \log\left(1+\exp\left\{\mathcal{D}\left(\mathbf{z}_a^{(i)},\mathbf{z}_p^{(i)}\right) -\lambda\sum_{\substack{j=1\\ j\neq i}}^{N}\left[\mathcal{D}\left(\mathbf{z}_a^{(i)},\mathbf{z}_n^{(j)}\right)+\sum_{\substack{k>j\\ k\neq i}}^{N}\mathcal{D}\left(\mathbf{z}_n^{(j)},\mathbf{z}_n^{(k)}\right)\right]\right\}\right);$$

where $\boldsymbol{z}_a^{(i)}$ is anchor embedding belonging to class i, $\boldsymbol{z}_p^{(i)}$ and $\boldsymbol{z}_n^{(j)}$ are positive and negative embeddings

($C_{N,2}$ + 1)-pair loss function – cont'd



(d) N-pair loss (e) $(C_{N,2}+1)$-pair loss

where $\boldsymbol{z}_a^{(i)}$ is anchor embedding belonging to class i, $\boldsymbol{z}_p^{(i)}$ and $\boldsymbol{z}_n^{(j)}$ are positive and negative embeddings

Noise-robust keyword spotting accuracy

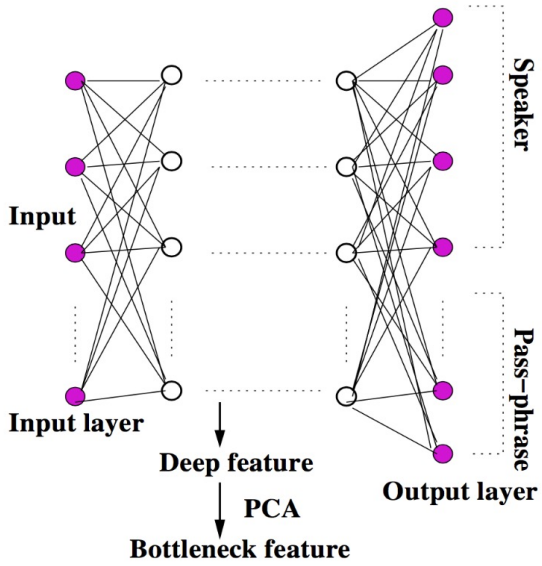
AALBORG UNIVERSITY
DENMARK

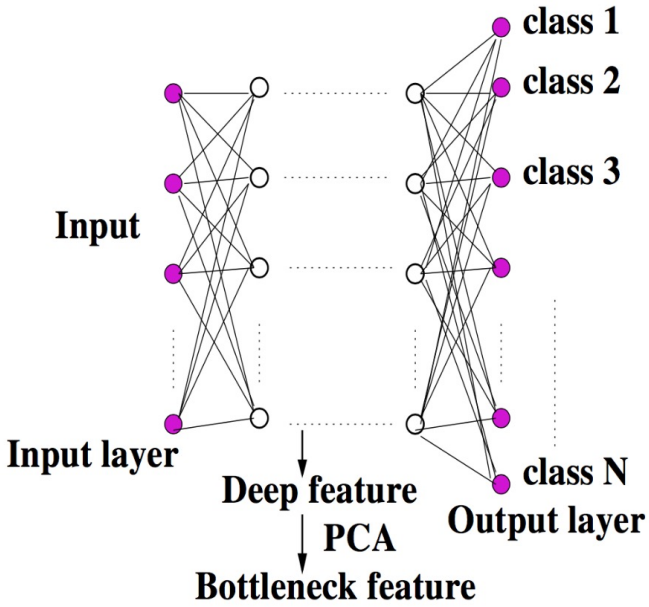
- Average accuracy with 95% confidence intervals

Unseen noises	Baseline	81.22 ± 0.69
	Contrastive loss	81.32 ± 0.86
	Triplet loss	81.71 ± 0.52
	Quadruplet loss	82.15 ± 0.27
	N-pair loss	82.70 ± 0.52
	$(C_{N,2} + 1)$-pair loss	83.53 ± 0.30

Time contrastive learning (TCL) for speech

- TCL based bottleneck features for speaker verification
- Speech is a quasi-stationary signal

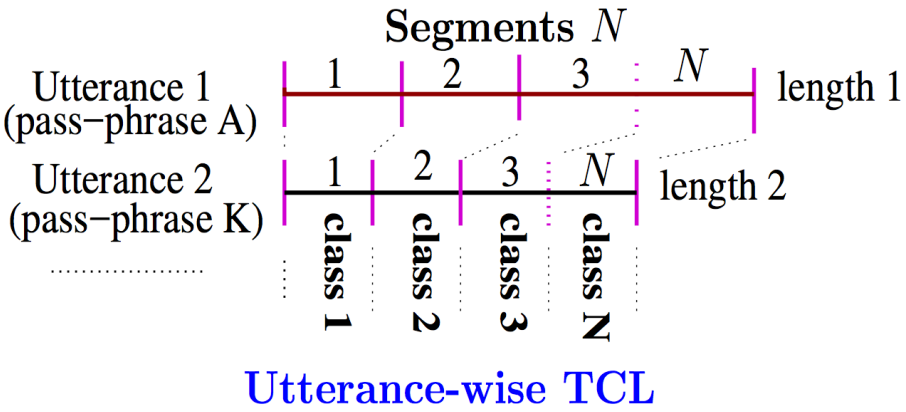




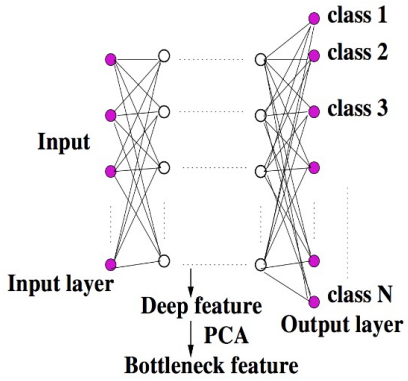
Sarkar, A. K., Tan, Z. H., Tang, H., Shon, S., & Glass, J. (2019). Time-contrastive learning based deep bottleneck features for text-dependent speaker verification. IEEE/ACM TASLP, 27(8), 1267-1279.

Time contrastive learning – training

- Targets: segmentation and **segment-based clustering** (and relabeling)



Utterance-wise TCL



- Loss: Cross-entropy
- Nonlinear ICA interpretation [Hyvarinen, 2016]
- Key elements: discrimination in latent space, positive and negative samples, CE loss, clustering for obtaining training targets, MFCCs (some similar concepts in HuBERT, MelHuBERT)

Hyvarinen, A., & Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. NeurIPS.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM TASLP*.

Lin, T. Q., Lee, H. Y., & Tang, H. (2022). MelHuBERT: A simplified HuBERT on Mel spectrogram. arXiv preprint arXiv:2211.09944.

Time contrastive learning – results

AALBORG UNIVERSITY
DENMARK

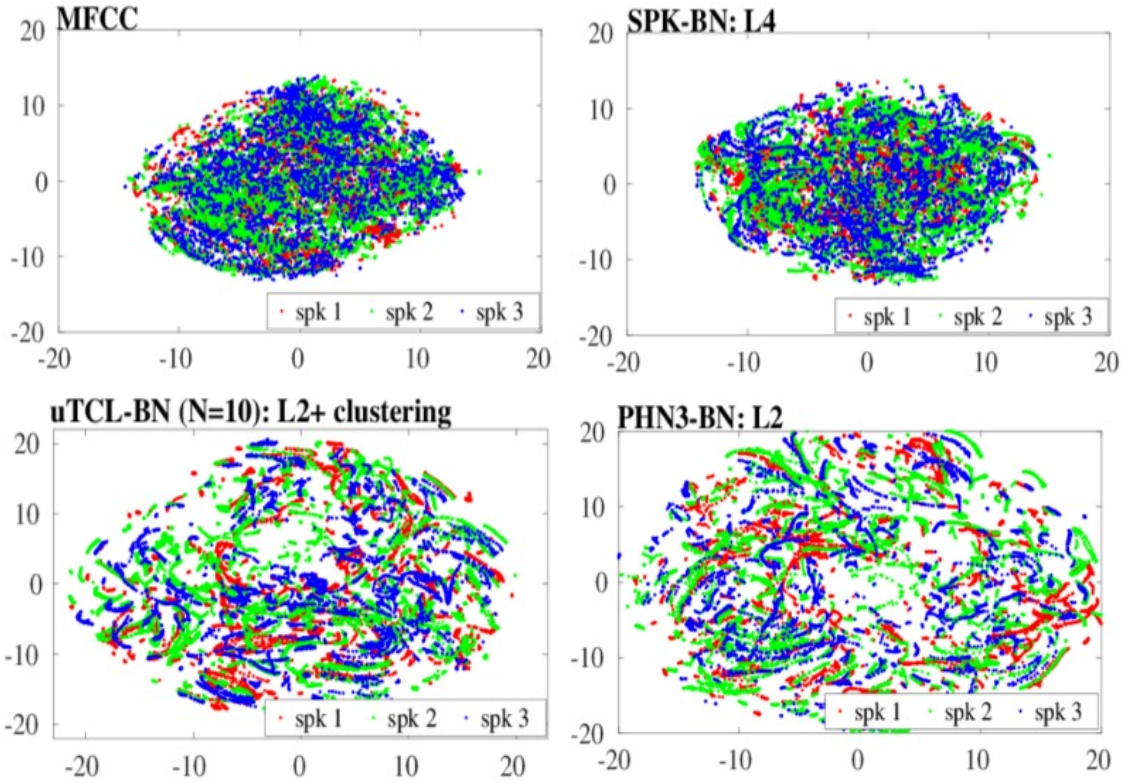
- Equal error rate (EER) performance for text-dependent speaker verification

Feature	DNN Lyr.	# of classes	Average (EER /minDCF)
MFCC			3.19/1.35
SPK-BN	L2	300	3.13/1.16
	L4		**2.91/1.13**
PHN-BN3	L2	39	**1.81/0.76**
(ASR force-alignment)	L4		2.39/1.11
+clustering	L2		1.82/**0.69**
	L4		2.89/1.26
uTCL-BN	L2	10	1.89/0.76
	L4		16.61/8.59
+ clustering	L2		**1.79/0.65**
	L4		5.31/2.96

Our recent work shows

- using Gaussian error linear unit (GELU) instead of Sigmoid activation function, further reduces EER by 20% relatively
- Promising results for noise robust VAD

Scatter plots of frame features using T-SNE toolkits

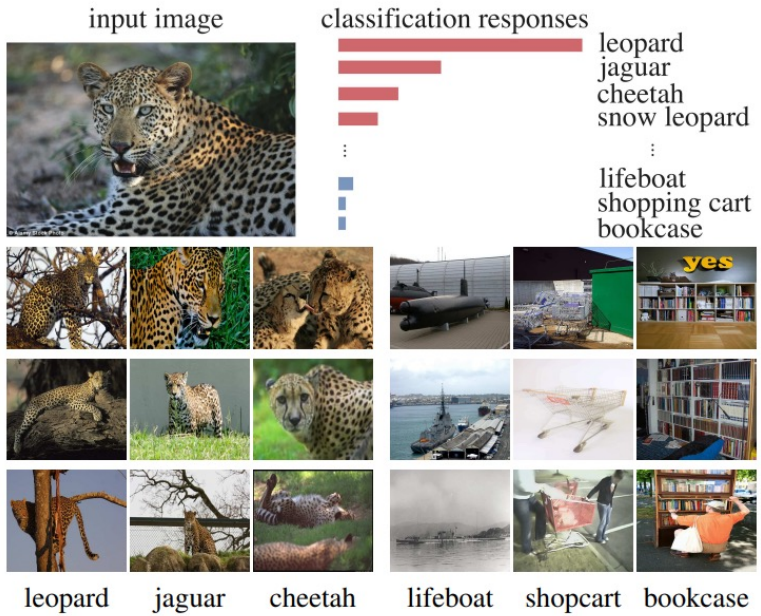


Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Supervised, unsupervised and self-supervised learning
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

From class-wise supervision to instance-wise



Apparent similarity is learned not from semantic annotations, but from the visual data themselves [Wu, 2018]

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE CVPR.

From class-wise supervision to instance-wise – cont'd

AALBORG UNIVERSITY
DENMARK

- Treat each image instance as a distinct class of its own
- Challenge
  - In ImageNet has1.2 million instead of 1,000 classes
  - Simply extending softmax is infeasible
- Instead, approximate the full softmax distribution with noise-contrastive estimation (NCE) [Gutmann & Hyvärinen, 2010], to maximize distinction between instances via a novel non-parametric softmax formulation

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)}$$

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log P(i|f_{\boldsymbol{\theta}}(x_i))$$

where the memory bank $V = \{\mathbf{v}_j\}$, initialized as unit vectors

InfoNCE in CPC

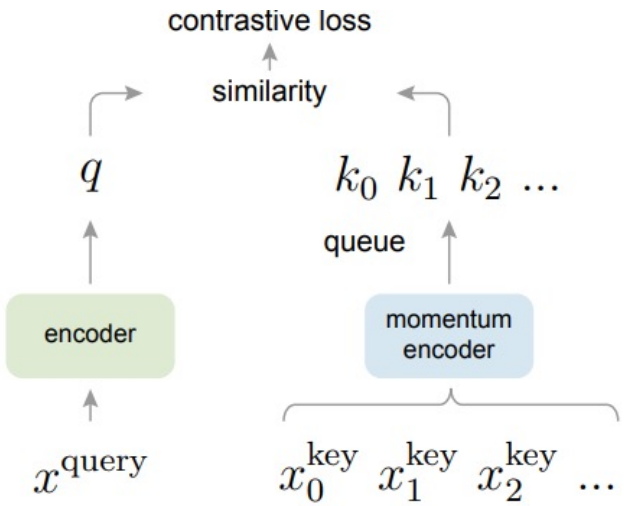
$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

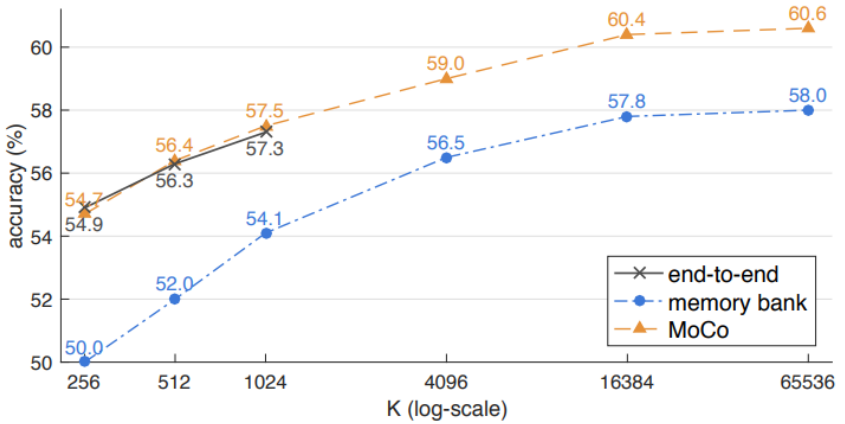
$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE CVPR.

Momentum contrast (MoCo)

- Use queue instead of memory bank to find training targets
- Build a dynamic dictionary with a queue and a moving-average enc.
- InfoNCE loss

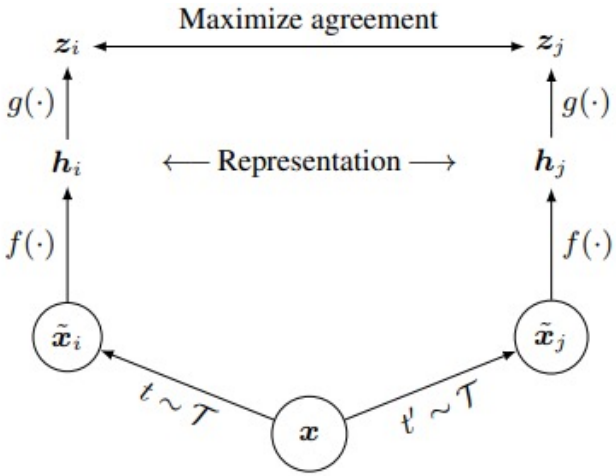


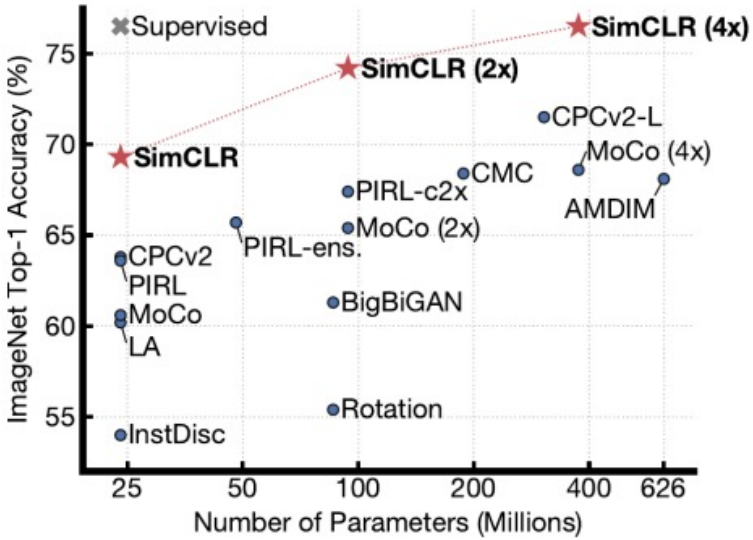


He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. IEEE CVPR.

A simple framework for contrastive learning (SimCLR)

- Introduce a learnable nonlinear transformation between the representation and the contrastive loss
  + data augmentations, larger batch sizes and more training steps
- NT-Xent (the normalized temperature-scaled cross entropy loss), also known as InfoNCE or N-pairs loss

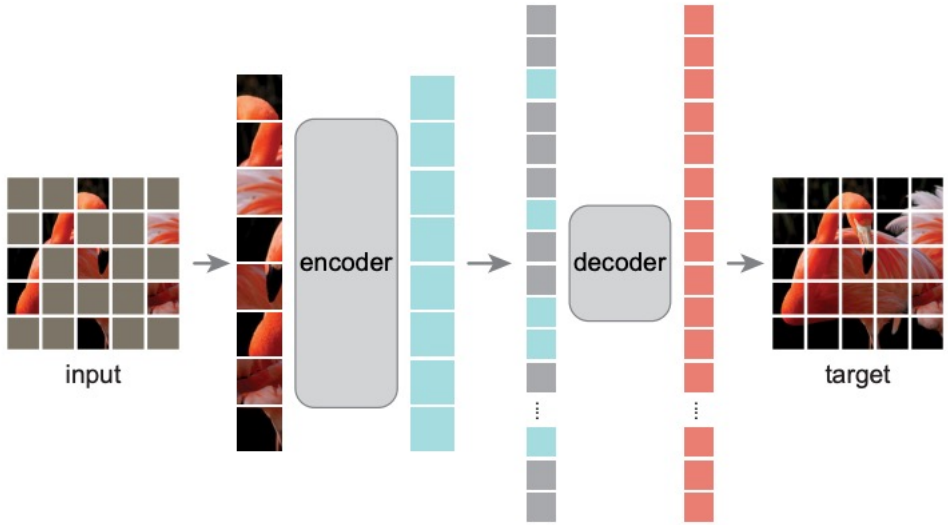




Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. ICML.

Masked autoencoder

- Mask random patches of the input image and reconstruct the missing pixels
  - An asymmetric encoder-decoder architecture, with an encoder operating only on the visible patches, & a lightweight decoder reconstructing the original image from the latent representation and mask tokens.
  - Masking a high proportion of the input image, e.g., 75%



He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. CVPR.

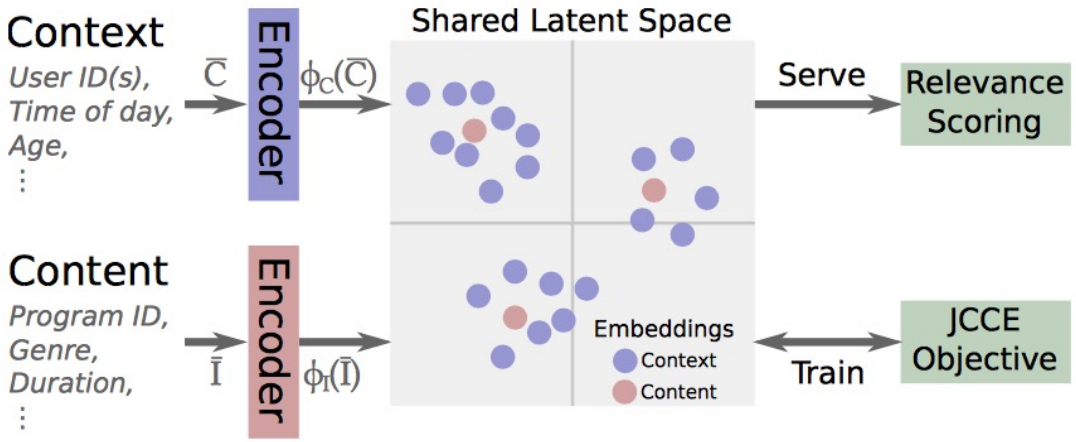
Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Supervised, unsupervised and self-supervised learning
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

Relaxed N-pairs loss for recommendations

- Representation learning for context-aware recommendations
- Learns joint embeddings of context and content



Kristoffersen, M. S., Shepstone, S. E., & Tan, Z. H. (2020). Context-aware recommendations for televisions using deep embeddings with relaxed N-Pairs loss objective. arXiv preprint arXiv:2002.01554.

Kristoffersen, M. S., Wieland, J. L., Shepstone, S. E., Tan, Z. H., & Vinayagamoorthy, V. (2019). Deep Joint Embeddings of Context and Content for Recommendation. In Context-Aware Recommender Systems Workshop: In conjunction with RecSys 2019.

Relaxed N-pairs loss for recommendations – cont'd

AALBORG UNIVERSITY
DENMARK

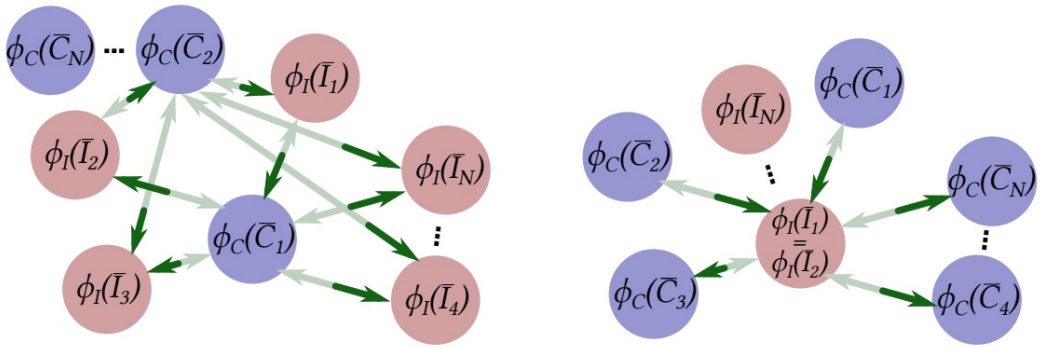
- N-pairs loss

$$\mathcal{L}_{\mathrm{NP}}(a,p) = \frac{1}{N}\sum_{i}^{N} -\log\left(P_i\right) = \frac{1}{N}\sum_{i}^{N} -\log\left(\frac{e^{a_i^{\mathsf{T}} p_i}}{\sum_{j}^{N} e^{a_i^{\mathsf{T}} p_j}}\right)$$

where $a = \{a_i\}_{i=1}^{N}$ is a set of N anchor embeddings and $p = \{p_i\}_{i=1}^{N}$ is a set of N corresponding positive embeddings. The N-pairs mini-batch construction makes it simple to compute CE loss as each anchor has exactly one positive example.

- To trade-off the number of negatives, relaxed N-pairs loss is proposed to include multiple positive examples in the training targets

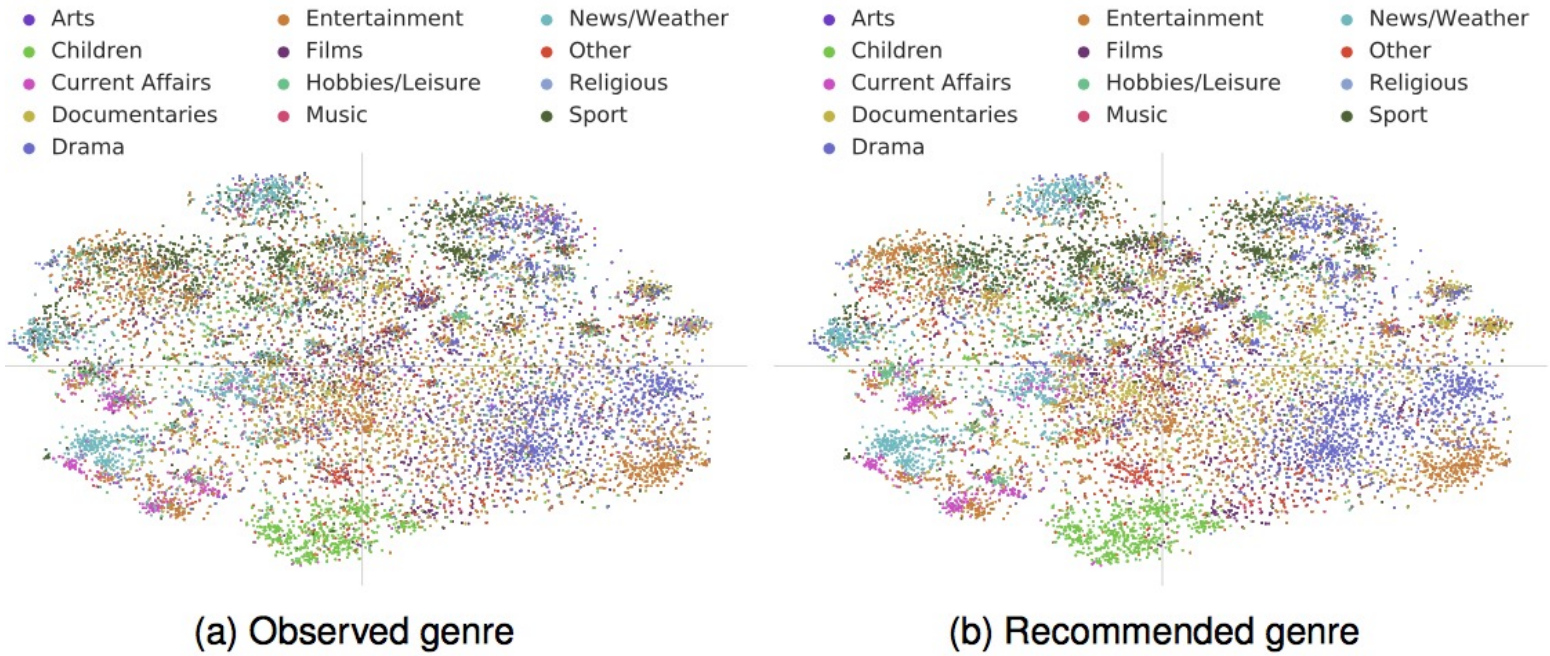
$$P_i' = \frac{\sum_{j \in X_i} e^{{a_i}^{\mathsf{T}} p_j}}{\sum_{k}^{N} e^{{a_i}^{\mathsf{T}} p_k}}.$$



(a) N-Pairs (b) Relaxed N-Pairs

Relaxed N-pairs loss for recommendations – cont'd

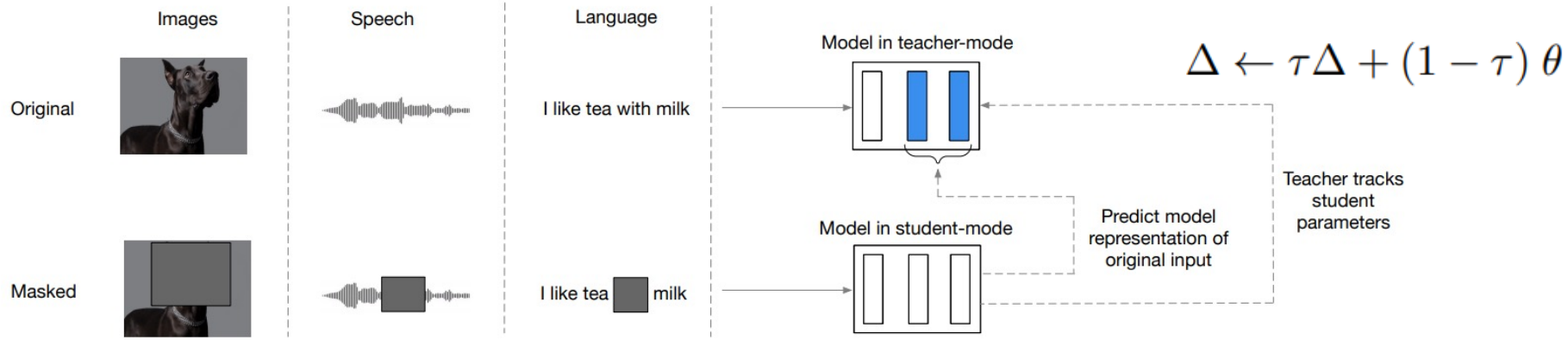
- Good match between observed and recommended
- Clusters shown as a by-product (via learning metric)



(a) Observed genre (b) Recommended genre

Modality independent SSL methods

- Contrastive predictive coding (CPC) is applicable to audio, text and images
- Data2vec is a general SSL framework for these domains too [Baevski, 2022]



- Given contextualized training targets, data2vec uses a smooth L1 loss to regress these targets

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - f_t(x)| \le \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

Baevski, A., Hsu, W. N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555.

Data2vec for keyword spotting

AALBORG UNIVERSITY
DENMARK

Table 4: Summary of results for the three KWT models. SC denotes Data2Vec pretraining using Speech Commands pretraining set, and LS denotes pretraining using Librispeech 100-hour clean training set. Full indicates model trained on the full original Speech Commands V2 training set without pretraining.

Model	Test accuracy			
	Baseline	Data2Vec		Full
		SC	LS	
KWT-1	0.8622	0.9294	0.9436	**0.9638**
KWT-2	0.8575	**0.9507**	0.9447	0.9498
KWT-3	0.8398	**0.9529**	0.9458	0.9079

HS Bovbjerg, ZH Tan (2022). Improving Label-Deficient Keyword Spotting Using Self-Supervised Pretraining

Agenda

AALBORG UNIVERSITY
DENMARK

- Introduction
  - Learning paradigms
  - Training targets and loss functions
- Domains
  - Text
  - Speech and audio
  - Images
  - Multimodal
- Discussions

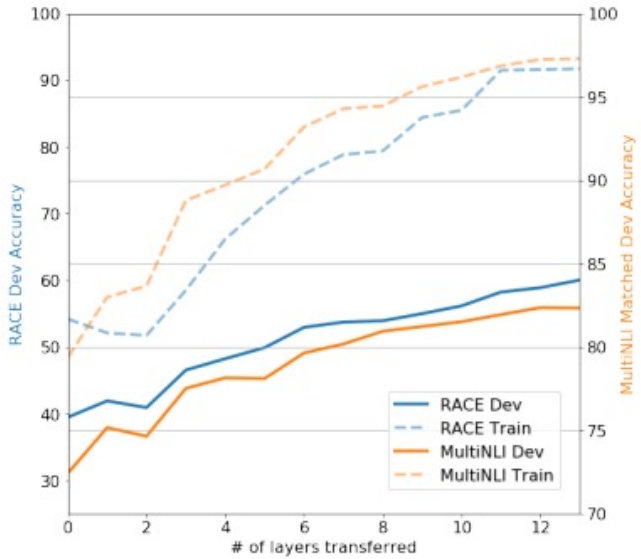
Training targets, loss functions and bias

AALBORG UNIVERSITY
DENMARK

- Self-supervised learning helps with data efficiency and generalization
- The chosen training targets and loss functions induce biases and impact on the performance of generalization
- Learnt representations are target/dependent and layer-dependent

Impact of number of layers transferred (GPT-1)

- Each layer in the pre-trained model contains useful functionality for solving target tasks

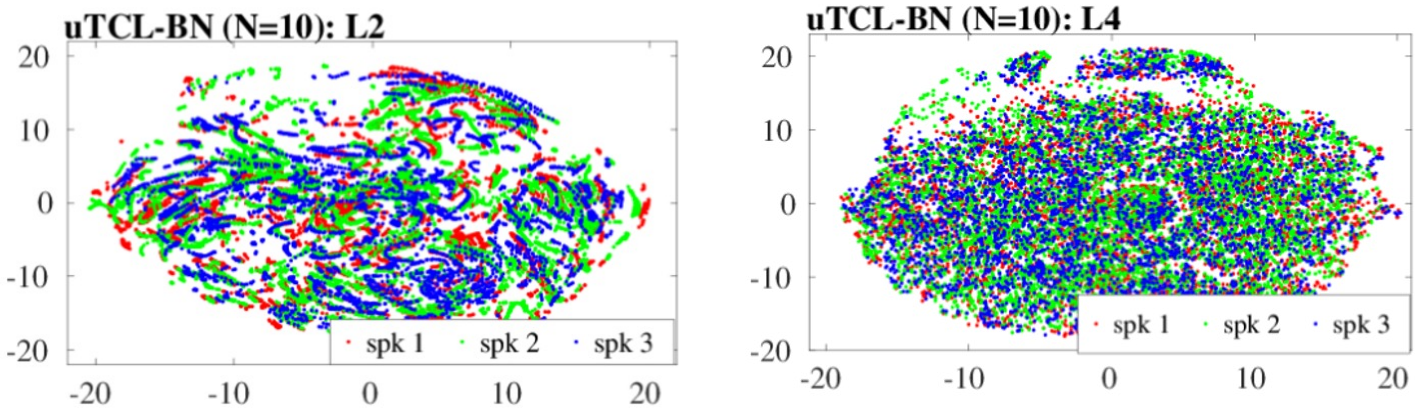


- General training targets and loss functions, e.g., autoregressive predictive (APC) ones, generate more less-target-specific representations

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Impact of number of layers transferred (time-contrastive learning for speech)

- Scatter plots of frame-level features using T-SNE

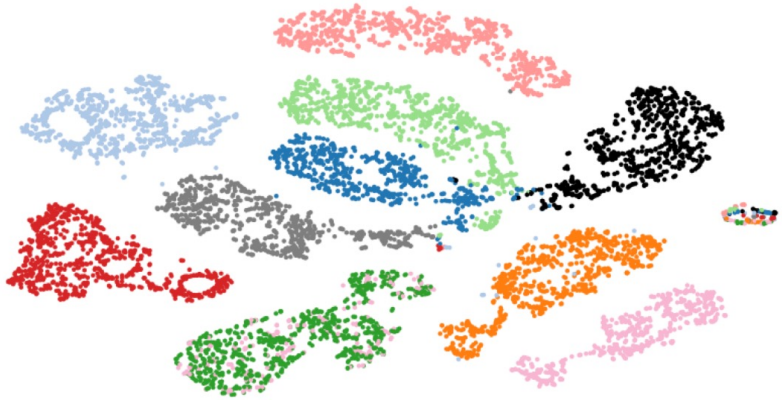


Sarkar, A. K., Tan, Z. H., Tang, H., Shon, S., & Glass, J. (2019). Time-contrastive learning based deep bottleneck features for text-dependent speaker verification. IEEE/ACM TASLP, 27(8), 1267-1279.

Representation learned by CPC

AALBORG UNIVERSITY
DENMARK

- t-SNE visualization of speech representations for a subset of 10 speakers (out of 251). Every colour represents a different speaker.



- Depending on the training targets where they are picked up
- The window size (maximum context size for the GRU) has a big impact on the performance, and longer segments would give better results. Here is slightly larger than 1s speech.
- Speaker versus speech recognition

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Semi-supervised learning

AALBORG UNIVERSITY
DENMARK

- Learning from both unlabeled and labeled samples
- Approaches [van Engelen, 2020].
  - Unsupervised pre-processing
    - Pre-training, e.g. restricted Boltzmann machines (RBMs)
    - Feature extraction with generative models, e.g. denoising autoencode, VAEs
    - Cluster-then-label
  - Intrinsically semi-supervised methods: extending supervised methods to include unlabelled samples in the objective function, assuming smoothness or low-density.
    - Maximum-margin methods, e.g. semi-supervised SVMs
    - Adding consistency regularization losses computed on unlabeled data, measuring discrepancy between predictions made on perturbed unlabeled data points.
    - Pseudo-labeling, predicting approximate classes on unlabeled data from a model trained only on labeled data

van Engelen and Hoos. A survey on semi-supervised learning. Machine Learning 2020.

Self-supervised semi-supervised learning

AALBORG UNIVERSITY
DENMARK

- Semi-supervised learning learns from both unlabeled & labeled samples, typically assumed to be sampled from the same or similar distributions [Zhao, 2019].
- Self-supervised learning defines **pretext** tasks, using **only unlabeled data,** to learn **representations** for downstream tasks.

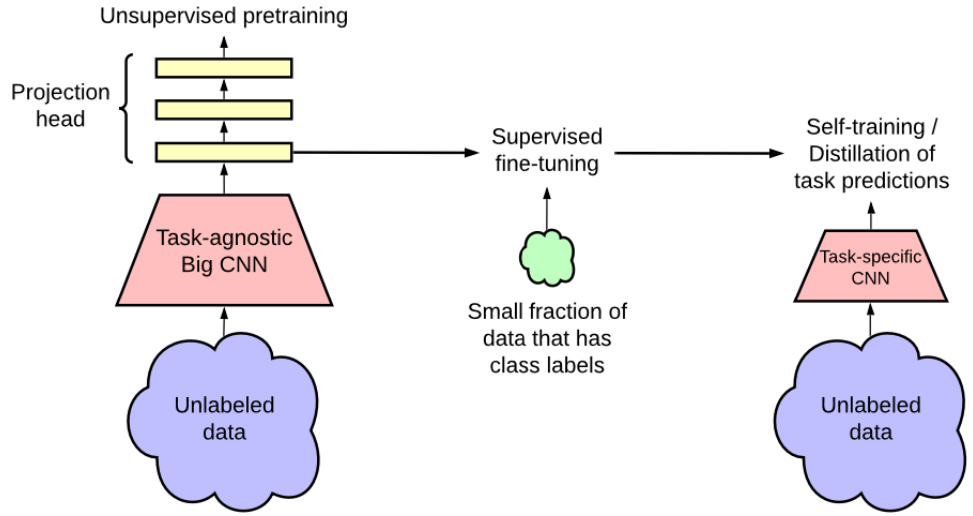
- S4L includes self-supervised loss in the supervised learning objective function

$$\min_{\theta} \; \mathcal{L}_l(D_l, \theta) + w\mathcal{L}_u(D_u, \theta)$$

Zhai et al. S4L: Self-Supervised Semi-Supervised Learning. ICCV 2019.

Big self-supervised models are strong semi-supervised learners

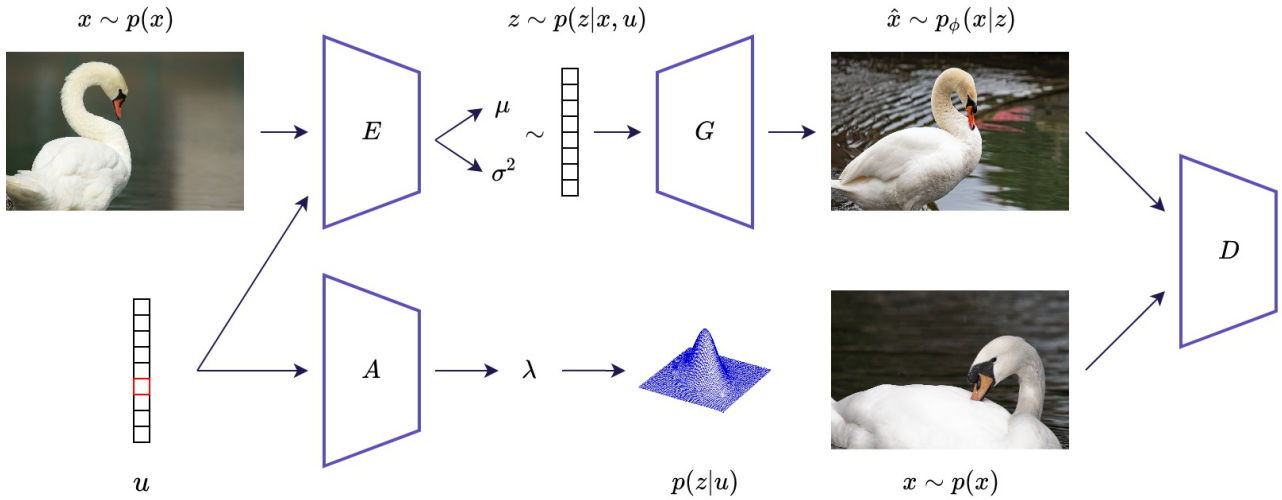
- Three steps
  - Unsupervised pretraining of a big ResNet model using SimCLRv2
  - Supervised fine-tuning on a few labeled examples
  - Distillation with unlabeled examples for refining and transferring the task-specific knowledge



Chen, et al. Big Self-Supervised Models are Strong Semi-Supervised Learners. NeurIPS 2020.

Identifiable VAE-GAN

- An identifiable Variational Autoencoder (VAE) based Generative Adversarial Network (GAN) model [Dideriksen, 2022]
- Recover true latent variables for meaningful representations



$$\mathcal{L}_{iVAE-GAN} = \mathcal{L}_{prior} + \mathcal{L}_{GAN}$$

$$\mathcal{L}_{prior} = -KL(q_\phi(z|x,u)||p_\theta(z|u))$$

$$\mathcal{L}_{GAN} = V(D,G)$$
$$= \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]$$
$$+ \mathbb{E}_{z\sim p_z(z|u)}[\log(1 - D(G(z)))]$$

- Inspired by [Khemakhem, 2020]

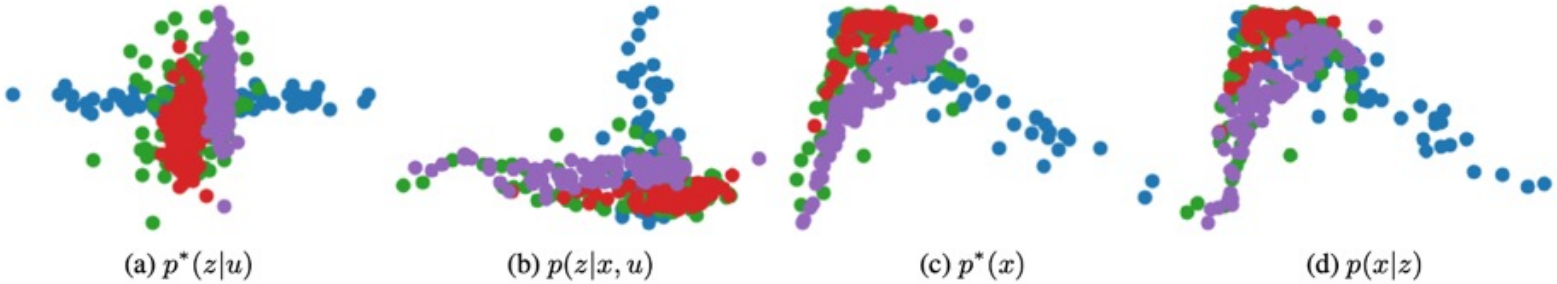
BU Dideriksen, K Derosche, ZH Tan, iVAE-GAN: Identifiable VAE-GAN Models for Latent Representation Learning, IEEE Access, 2022

Khemakhem, I., Monti, R., Kingma, D., & Hyvarinen, A. (2020). Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. Advances in Neural Information Processing Systems.

Identifiable VAE-GAN – simulation results

AALBORG UNIVERSITY
DENMARK

2-Dimensional data and latent spaces



original generating latent variables

latent variables recovered by iVAE-GAN

input data

data generated by iVAE-GAN

BU Dideriksen, K Derosche, ZH Tan, iVAE-GAN: Identifiable VAE-GAN Models for Latent Representation Learning, IEEE Access, 2022

AALBORG UNIVERSITY
DENMARK

Thank you for your attention.

Thank my co-authors as in cited papers.

Agenda

- Introduction: learning paradigms, training targets and loss functions
- Domains: text, speech and audio, images, multimodal
- Discussions