NTNU

Norwegian University of Science and Technology

Green Learning: Reusing Generative Models for Different Tasks.

IMF PhD Seminar – 19th April 2023

Martin Ludvigsen

Department of Mathematical Sciences, NTNU.

Figure: Green learning?

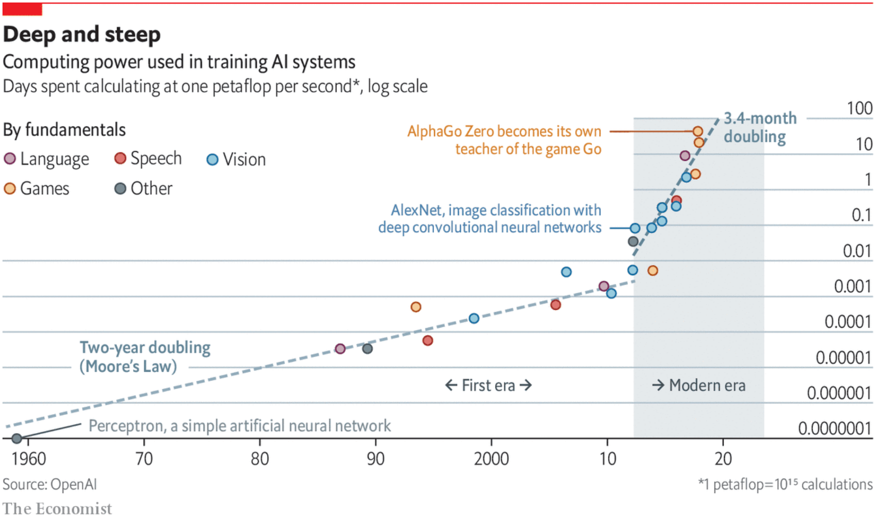


Figure: Energy costs of training AI models is increasing! Source: `https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/`

What is a generative model?

Mapping that generate signals/data from a low-dimensional space

$$g : \mathcal{H} \to \mathcal{U}$$

- $\mathcal{H}$ latent space of dimension d
- $\mathcal{U}$ signal space of dimension m
- Usually $d \ll m$
- Goal is to have $g(h) = u$ for some $h \in \mathcal{H}$, where $u \in \mathcal{U}$.
- Statisticians: Can do everything with probability distributions
- Numericists: Parallels to Galerkin method and interpolation

NTNU

Norwegian University of Science and Technology

Why?

Generative models allow us to learn effective representations of data.

Some plug-and-play applications:

- Feature extraction
- Dimensionality reduction
- Clustering
- Classification
- Data imputation
- Compression
- Anomaly detection

Inverse problems

- Denoising
- Deblurring/defiltering
- Compressed sensing
- Source separation
- ... And more

NTNU

Norwegian University of Science and Technology

Knowledge-based
Human experts in domain-specific fields create models and methods.

Generative
Data-driven methods where the goal is to generate data, then apply this to problems.

Discriminative
Data-driven methods where the goal is solving specific problems.

Non-Negative Matrix Factorization (NMF)

Parameterize generative function as non-negative linear combination

$$g(h) = Wh, \quad W \in \mathbb{R}_+^{m \times d}, h \in \mathbb{R}_+^d.$$

The columns of W spans out a convex cone $C(W)$, which hopefully represents likely signals $\mathcal{U}$. W is also called a dictionary or a basis.

Training: Store data column-wise in U

$$\min_{W \geq 0} \|U - WH(U, W)\|_F^2 + \mu_W |W|_1$$
$$H(U, W) = \arg\min_{H \geq 0} \|U - WH\|_F^2 + \mu_H |H|_1,$$

where $|.|_1$ is entry-wise 1-norm and μ_W, μ_H are sparsity parameters.

Main Tools

NTNU

Norwegian University of Science and Technology

Define projection-like operator onto cone spanned by columns of W:

$$P_{C(W)}(u) = Wh^*(u), \quad h^*(u) = \arg\min_{h \geq 0} \|u - Wh\|_2^2 + \mu_H \|h\|_1,$$

where $\mu_H > 0$ is a sparsity parameter[1].

Sparse non-negative least square problem, easy to solve!

Distance/Likelihood[2]:

$$D_{C(W)}(u) = \|u - P_{C(W)}(u)\|_2$$

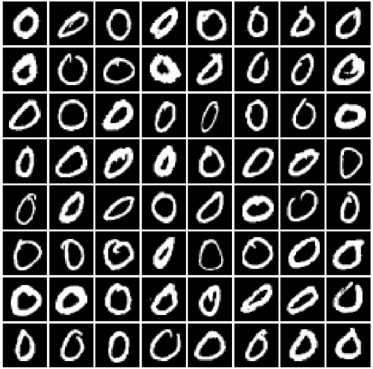
[1] With sparsity parameter this *technically* is not a projection...

[2] Obtain distribution $\propto e^{-D_{C(W)}(u)^2}$. Need sparsity parameter for this to be proper!

NTNU

Norwegian University of Science and Technology

MNIST



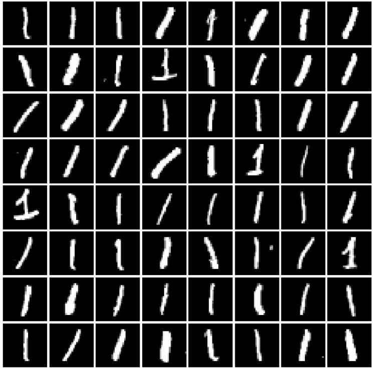


Figure: Example MNIST pictures. Each image contains $m = 28 \times 28 = 784$ pixels.

NMF bases

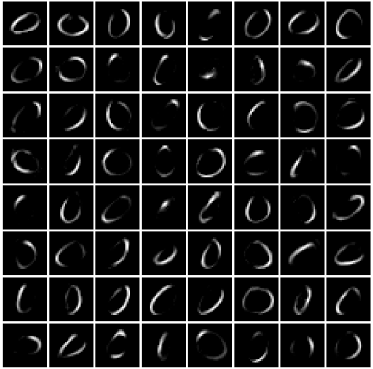
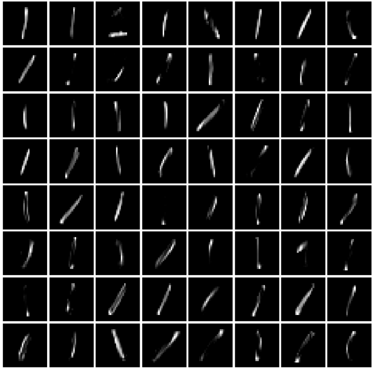
 

Figure: NMF bases (columns of W) trained with $N = 5000$ datapoints and $d = 64$. All results in this presentation are created using these bases. Training takes ~ 5 seconds on a modern laptop.

Projection example

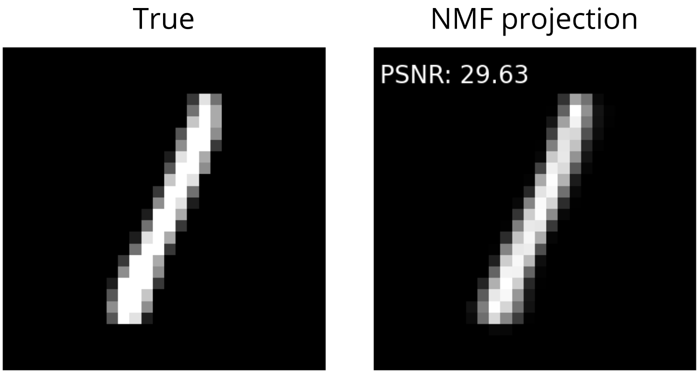


Figure: Projections are usually blurrier, and lose out on fine details. 784 pixels $\rightarrow$ 64 latent variables. Projection is optimization in latent space instead of signal space

Classification Training

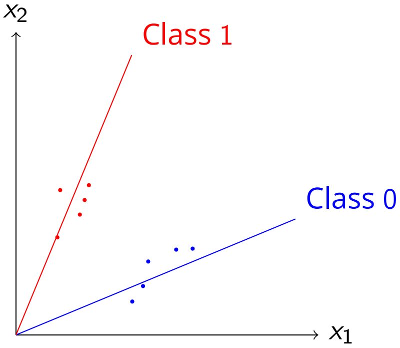


Figure: Train a generative function (a cone), for each class of signal.

Classification Testing

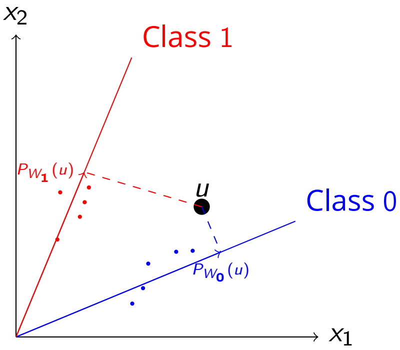


Figure: Project a new datapoint onto the cones $\rightarrow$ classify to closest cone.

Classification results

- Use pre-trained dictionaries!
- 2000 test data points of zero and one digits $\rightarrow$ **99.8**% accuracy!
- Using more digits $\rightarrow$ comparable results to simple discriminate neural networks.
- No need to train discriminatively?

Inverse Problems in Signal Processing

NTNU

Norwegian University of Science and Technology

$$v = Au + \sigma w$$

- $v \in \mathcal{V}$ observed signal
- $A : \mathcal{U} \to \mathcal{V}$ (linear) forward operator. Often ill-posed/ill-conditioned.
- $u \in \mathcal{U}$ signal of interest
- $w \in \mathcal{W}$ noise
- $\sigma > 0$ noise level.
- Goal: Recover u given observation v.
- Statisticians: GLM with high correlation and/or $n < k$.

Denoising Gaussian noise

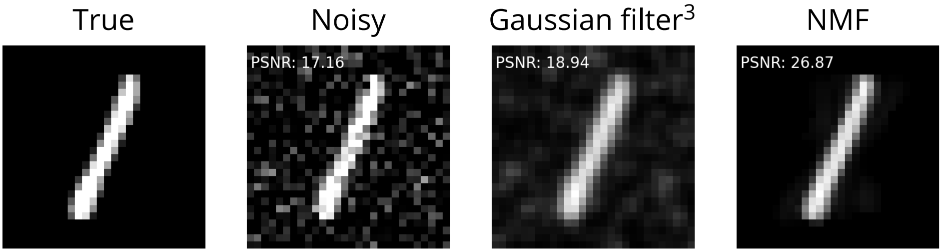


Figure: Observe noisy $v = u + \sigma w$ where w is Gaussian noise. Recover $u \approx P_{C(W)}(v)$. Core Idea: Generative models are able to separate the signal from the noise.

[3] Equivalently: Solve Heat equation.

Denoising Salt and pepper noise

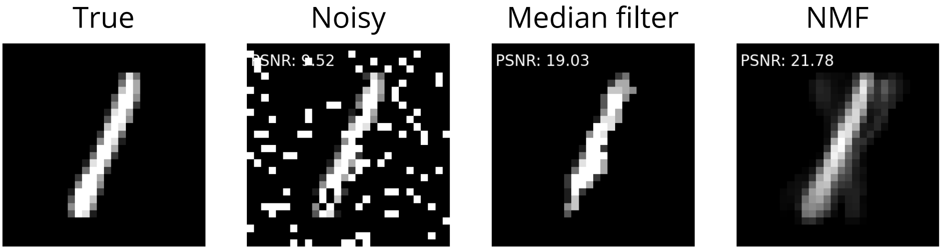


Figure: Observe S+P noise. Recover $u \approx P_{C(W)}(v)$. [4]

[4] Should solve in 1-norm for better results: $h^* = \arg\min_{h \geq 0} \|Wh - v\|_1$

Denoising periodic noise

NTNU
Norwegian University of Science and Technology

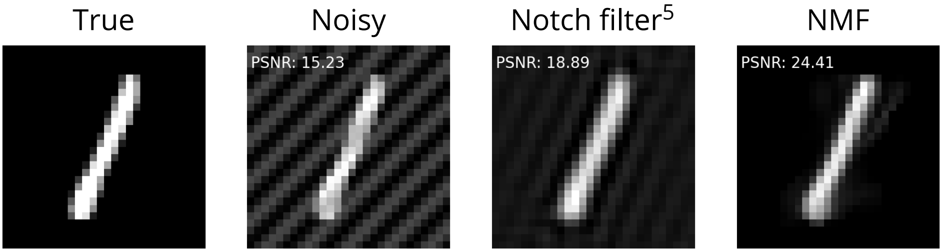
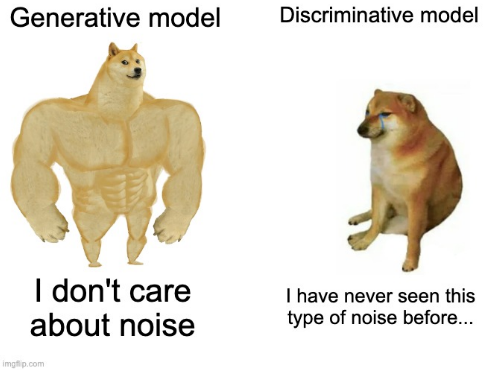


Figure: Observe periodic noise. Recover $u \approx P_{C(W)}(v)$.

[5] Remove exact frequencies in Fourier space. Requires a lot of tuning and knowledge about the noise!

Discriminative denoising

- Train different models for each noise type? (Wasteful!)
- Train model using data with all noise types? (Redundant!)
- Discriminative models require strong supervision data, generative models only use weak/semi supervision data!
- Discriminative models need a generative model at their core!



Solving Inverse Problems With Generative Regularization

NTNU

Norwegian University of Science and Technology

We can solve more general inverse problems with projections!

$$\min_{u \in \mathcal{U}} \|Au - v\|_2^2 \approx \min_{h \geq 0} \|AWh - v\|_2^2 = \min_{h \geq 0} \|\tilde{W}h - v\|_2^2$$

If $\tilde{W} = AW$ is still a non-negative dictionary $\rightarrow$ Can solve inverse problem via projection $u \approx P_{C(\tilde{W})}(v)$!

Can reuse dictionary W for many different problems!

Inpainting/Imputation

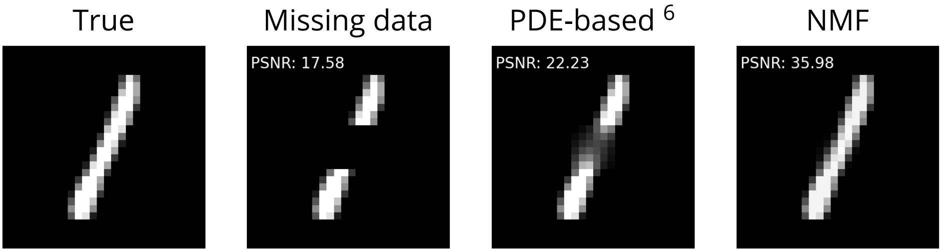


Figure: Inpaint/impute missing data.

[6] Biharmonic equation with Neumann boundary conditions

Deblurring

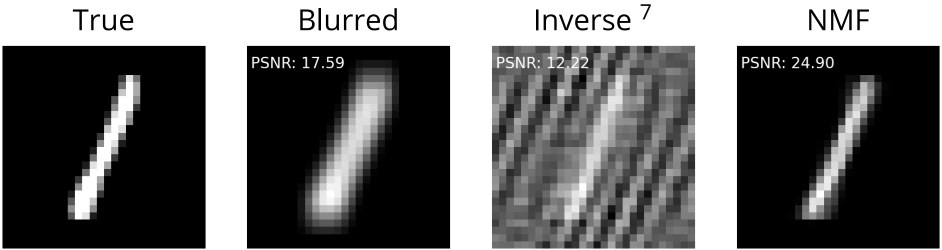
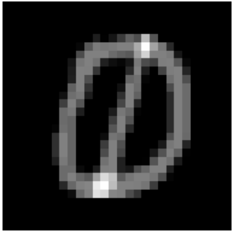


Figure: Deblurring where A is a convolution with a constant kernel, and we add a tiny bit of noise.

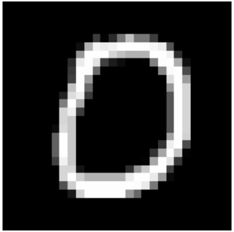
[7] With L_2-regularization in Fourier space

Source Separation



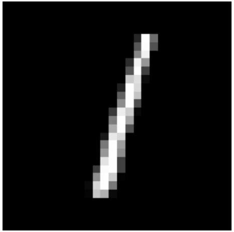
v

$=$



u_0

$+$



u_1

Given a mixed image $v = u_0 + u_1$, can we recover the individual images u_0 and u_1?

Source Separation Method

NTNU

Norwegian University of Science and Technology

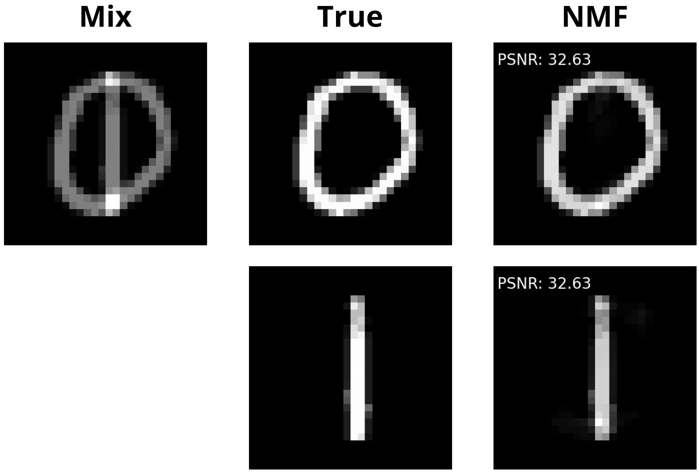
1. Training: Fit bases for each of the S signal types $W_1, \dots, W_S$.
2. Testing: Given mixed data v, solve

$$(h_1^*, \dots, h_S^*) = \underset{h_i \geq 0, i=1,\dots,S}{\arg\min} \|v - \sum_{i=1}^{S} W_i h_i\|_2^2 = \underset{h \geq 0}{\arg\min} \|v - Wh\|_2^2,$$

   where W and h are concatenated bases and latent variables.
3. Separate: $u_1 \approx W_1 h_1^*, \dots, u_s \approx W_S h_S^*$
4. Do some post-processing …

Separate via projection onto convex cones!

Source Separation Results



NTNU

Norwegian University of Science and Technology

NTNU

Norwegian University of Science and Technology

Can we go further?

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com

Karthik Narasimhan OpenAI karthikn@openai.com

Tim Salimans OpenAI tim@openai.com

Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

Figure: Famous 2018 paper from OpenAI that lead to ChatGPT (Generative Pre-trained Transformer). Reuse generative functions with discriminative fine-tuning!

Questions?

MA

I am doing a talk about reusing Generative models for different tasks like classification and denoising. What would be a good question fro the audience to ask after the presentation?

Here's a potential question that could spark an interesting discussion:

- Can you talk more about the limitations or challenges of reusing generative models for tasks other than the ones they were originally trained on, and how can these challenges be addressed?