Representation Learning & Unsupervised Transfer for Alzheimer's Disease in MRI

Javier Antonio Salazar Cavazos

September 18, 2024

Summer:

- Internship @ KLA
- A little bit of research on the side…

Summer:

- Internship @ KLA
- A little bit of research on the side…

Presentation:

- Very little MRI here, sorry for misleading name 😔
- Previously, I spoke about classification of Alzheimer's in sMRI
- Due to potential confounders, labels can be misleading…
  - Chronic stress, depression, PTSD can play a role
- Let's change focus and pose the problem without the use of labels
  - Train model to extract relevant image features

Representation Learning

2017 *2020* *2024*

β-**VAE** [1]
Reconstruction Loss
"Nonlinear PCA"

SimCLR [2]
Contrastive Loss
Attract & Repel

SimSiam [4]
Similarity Loss
Attract Only

AUC-CL [3]
Contrastive Loss
Attract & Repel

Contrastive Learning (SimCLR)

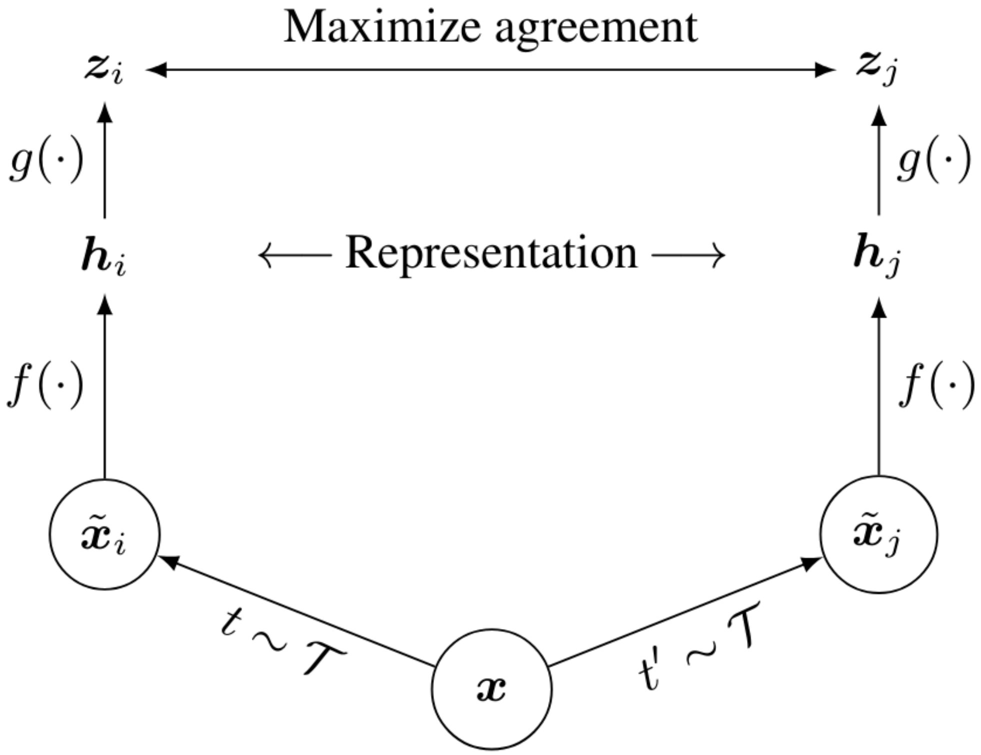


Figure 1: SimCLR framework.

$$\mathrm{sim}\big(z_i, z_j\big) = \frac{z_i^T}{\|z_i\|_2}\frac{z_j}{\|z_j\|_2}$$

$$l(i,j) = -\log\left(\frac{\exp\big(\mathrm{sim}\big(z_i, z_j\big)\cdot\tau^{-1}\big)}{\sum_{k=1}^{2N}\mathbb{I}_{[k\neq i]}\exp\big(\mathrm{sim}\big(z_i, z_j\big)\cdot\tau^{-1}\big)}\right)$$

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[l(2k-1, 2k) + l(2k, 2k-1)]$$

“Contrastive” Learning (SimSiam)

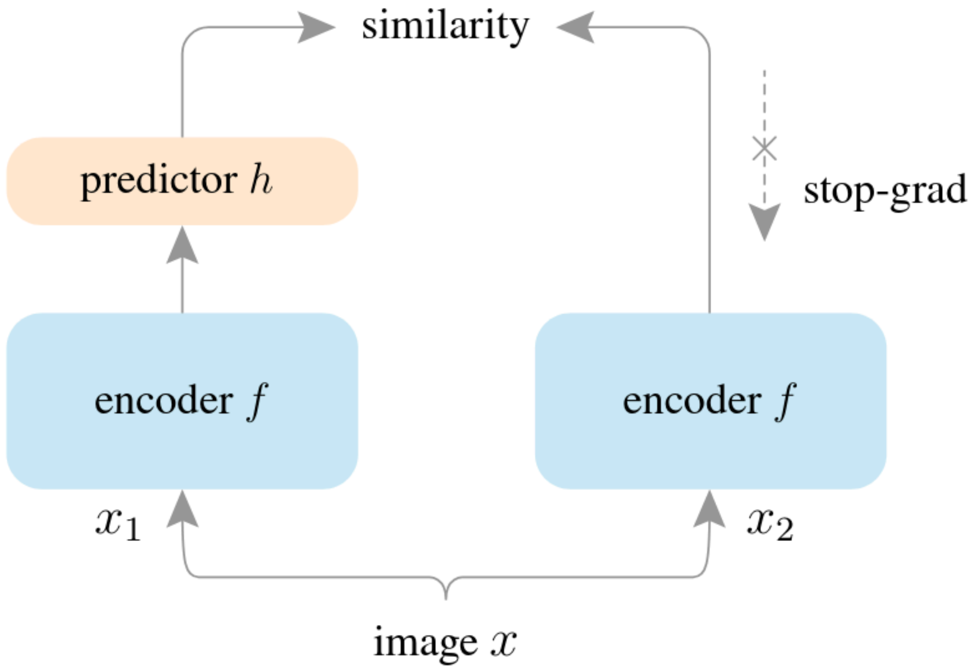


Figure 2: SimSiam framework.

$$p_1 = h(f(x_1)), z_2 = f(x_2)$$

$$\mathcal{D}(p_1, z_2) = -\frac{p_1^T}{\|p_1\|_2} \frac{z_2}{\|z_2\|_2}$$

$$\mathcal{L} = \mathcal{D}(p_1, \text{stopgrad}(z_2))$$

Contrastive Learning vs. Instance Discrimination

- Contrastive approach = pushing away from samples
- Instance discrimination = pull only

- If positive sample = class dog, what happens if negative samples have dog?
  - Is latent space worse? From my observations, this is not the case!

Contrastive Learning vs. Instance Discrimination

- Contrastive approach = pushing away from samples
- Instance discrimination = pull only

- If positive sample = class dog, what happens if negative samples have dog?
  - Is latent space worse? From my observations, this is not the case!

As a side note, both methods benefit from higher batch size so one aspect to explore is small batch size effects.

Small Batch Size CL

I worked with SimSiam (pull only method)

I suspect small batches lead to cost function "instability"

Idea 1: Retain a memory bank and add kNN such as:

$$\mathcal{L} = \mathcal{D}(p_1, \mathrm{sg}(z_2)) + \sum_{k=1}^{K} \mathcal{D}(p_1, \mathrm{sg}(\mathrm{NN}(z_2, k)))$$

Small Batch Size CL

I worked with SimSiam (pull only method)

I suspect small batches lead to cost function "instability"

Idea 1: Retain a memory bank and add kNN such as:

$$\mathcal{L} = \mathcal{D}(p_1, \mathrm{sg}(z_2)) + \sum_{k=1}^{K} \mathcal{D}(p_1, \mathrm{sg}(\mathrm{NN}(z_2, k)))$$

Idea 2: Create a mixed model of CL and AE:

$$\mathcal{L} = \mathcal{D}(p_1, \mathrm{sg}(z_2)) + \lambda \| \, x_1 - \mathrm{AE}_{\mathrm{DECODE}}(p_1) \, \|_2^2$$

Small Batch Size CL

I worked with SimSiam (pull only method)
I suspect small batches lead to cost function "instability"

Idea 1: Retain a memory bank and add kNN such as:
$\mathcal{L} = \mathcal{D}(p_1, \mathrm{sg}(z_2)) + \sum_{k=1}^{K} \mathcal{D}(p_1, \mathrm{sg}(\mathrm{NN}(z_2, k)))$

Idea 2: Create a mixed model of CL and AE:
$\mathcal{L} = \mathcal{D}(p_1, \mathrm{sg}(z_2)) + \lambda \| \, x_1 - \mathrm{AE}_{\mathrm{DECODE}}(p_1) \, \|_2^2$

Both ideas failed 😔 so I turned to the literature

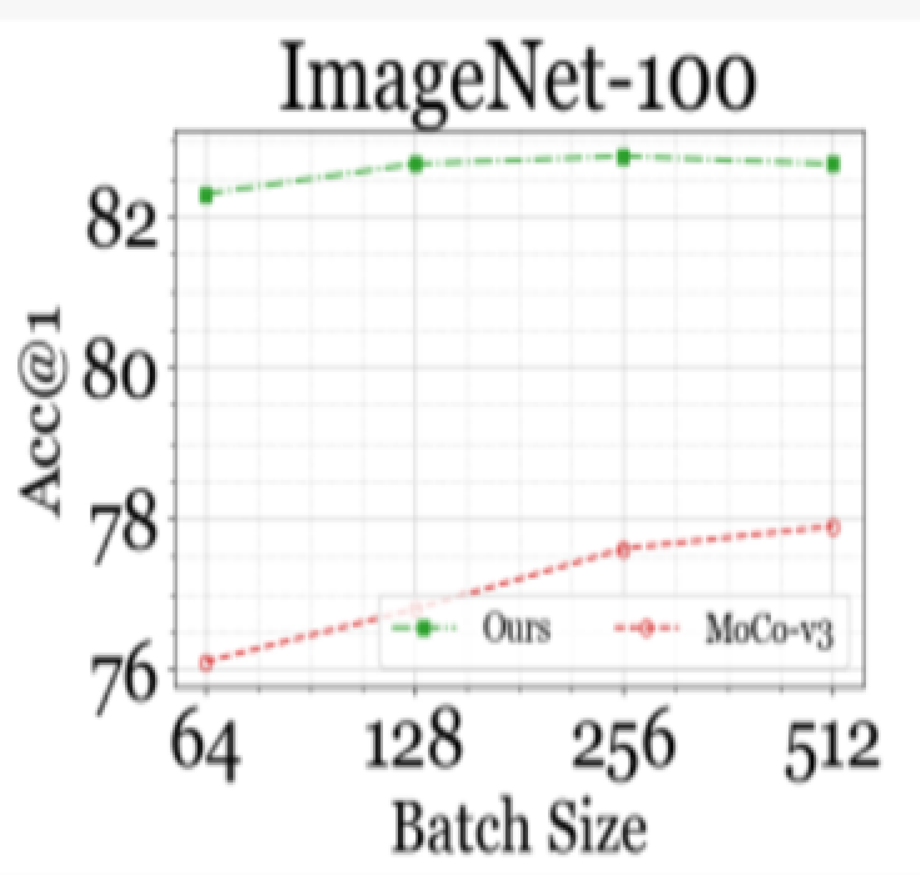
AUC-CL (Small Batch Size CL)

AUC-CL (ICLR2024) is a new method that outperforms others that attempt to mitigate small batch size

AUC-CL is batch size robust. For parameters v_t, lr η, batch size B, and timestep t sampled from T,

$$E\|\nabla L_{AUC}(v_t)\|^2 \leq O\left(\frac{1}{\eta T} + \eta + \frac{1}{\boxed{B}}\right) \quad \textbf{Biased}$$

$$E\|\nabla L_{NTXent}(v_t)\|^2 \leq O\left(\frac{1}{\eta T} + \eta\right) \quad \textbf{Unbiased}$$



sMRI Comparisons (2 class AD/CN)

For balanced dataset of equal AD/CN ADNI subjects:

Method	AUC-CL	VAE
kNN	81%	67%
Linear Probe	79%	53%

Both models trained from scratch using only ADNI AD/CN sMRI data.

For comparisons, standard classification CNN gives 86% ACC.

Could be improved depending on data augmentations?

Latent Visualizations

No results here, 2D space is too restrictive.

No good separation for VAE or AUC-CL (256 dimensional space)

Clustering
How do we find labels in our data?

Unsupervised Learning - Clustering

1950s-2000s

Kmeans [5]
Centroid method
Clustering only

Spectral Clustering [6]
Distance metric method
Clustering only

2015-2020

DeepCluster [7]
Cross Entropy Loss
Backbone + Cluster

DeepKMeans [8]
Joint Loss
VAE + Kmeans

2021-2024

Contrastive Clustering [9]
Contrastive Loss
Backbone + Cluster

TURTLE [10]
Unsupervised SVM
Clustering only

TURTLE

Publication

Let Go of Your Labels with Unsupervised Transfer
Artyom Gadetsky*, Yulun Jiang*, Maria Brbić.
International Conference on Machine Learning (ICML), 2024.

For data x, let $\phi(x)$ be latent variables. Solve the following:

$$\mathcal{L}_{\text{TURTLE}}(\theta) = \sum_{x \in \mathcal{D}} \mathcal{L}_{ce}(w_\theta \cdot \phi(x); \tau_\theta(\phi(x))) \;\text{ s.t. } w_\theta = \Xi(w_\theta, \mathcal{D})$$

$$\min_\theta \mathcal{L}_{\text{TURTLE}}(\theta) - \underbrace{\gamma \mathbb{H}(\overline{\tau_\theta})}_{\text{reg.}} \;\text{ where } \overline{\tau_\theta} = |\mathcal{D}|^{-1} \sum_{x \in \mathcal{D}} \tau_\theta(\phi(x))$$

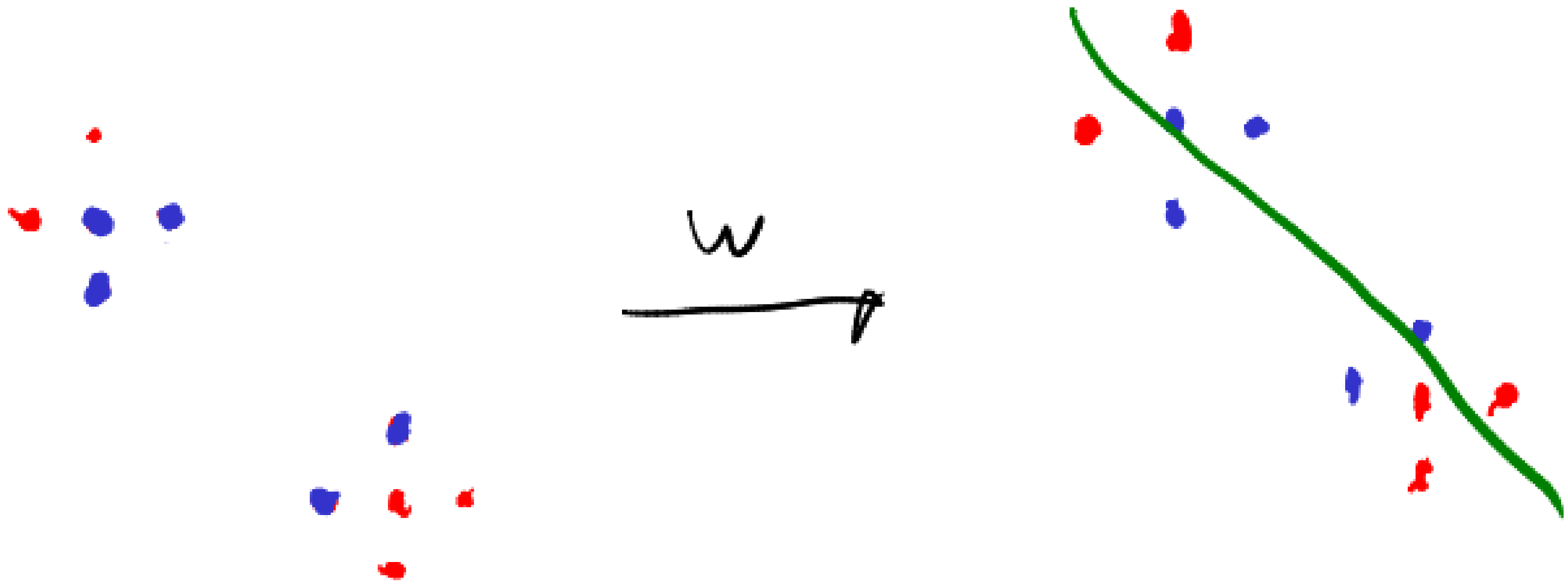
w_θ = learnable hyperplane, τ_θ = learnable "classifier", $\overline{\tau_\theta}$ = empirical dist.

TURTLE

Given random τ, first "classify" data as follows:

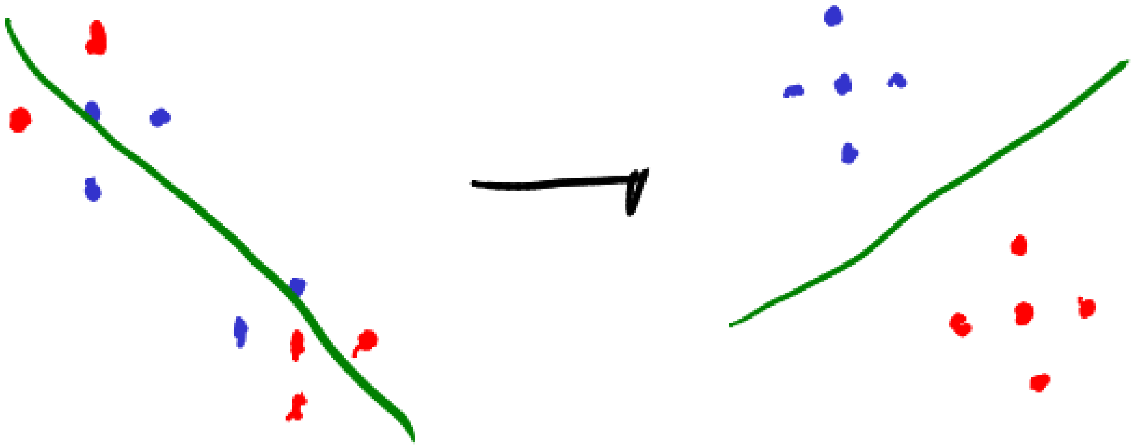
TURTLE

Now given fixed τ, find optimal w as follows:



TURTLE

Alternate between the two until optimal solution reached:



TURTLE = Unsupervised SVM method reminiscent of K-means (**opinion**)

PET-TURTLE (Prior Enforcement Term TURTLE)

TURTLE objective:

$$\min_\theta \mathcal{L}_{\text{TURTLE}}(\theta) - \gamma \underbrace{\mathbb{H}(\overline{\tau_\theta})}_{\text{reg.}} \text{ where } \overline{\tau_\theta} = |\mathcal{D}|^{-1} \sum_{x \in \mathcal{D}} \tau_\theta(\phi(x))$$

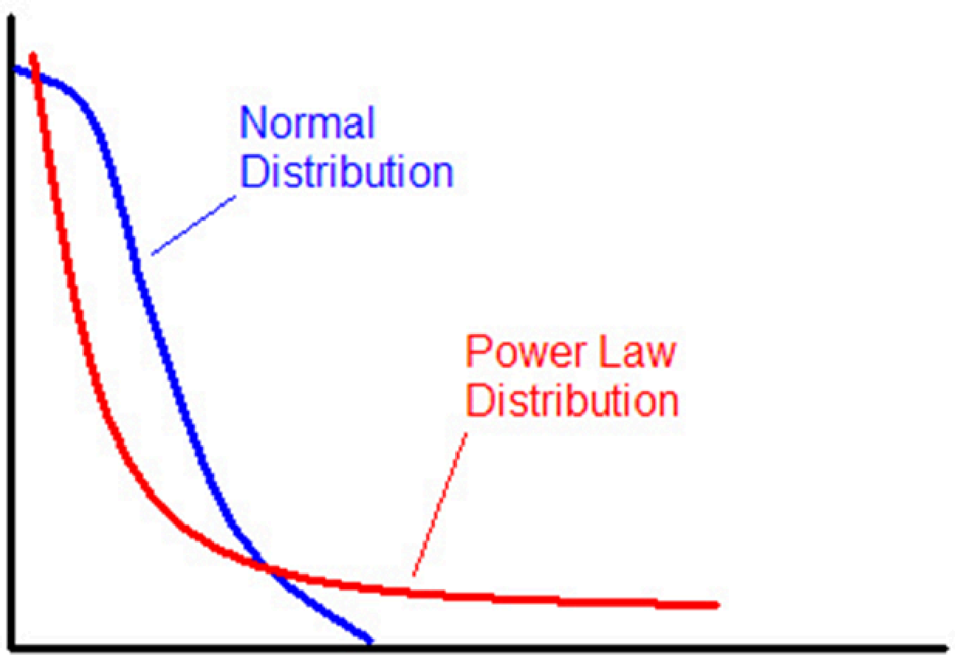
I propose the following generalization for imbalanced data:

$$\min_\theta \mathcal{L}_{\text{TURTLE}}(\theta) + \gamma \mathcal{R}(\overline{\tau_\theta}) \text{ where } \mathcal{R}(\overline{\tau_\theta}) = D_{\text{KL}}\left(\underbrace{s}_{\text{sort}}(\overline{\tau_\theta}) \parallel \Pi\right)$$

Π can be many things such as:

- Uniform if balanced
- Known for some applications
- Surrogate priors can be utilized…

Powerlaw Distribution



Powerlaw = heavy tail distribution

Powerlaw Distribution: Applications

Astronomy [edit]

- Kepler's third law
- The initial mass function of stars
- The differential energy spectrum of cosmic-ray nuclei
- The M–sigma relation
- Solar flares

Biology [edit]

- Kleiber's law relating animal metabolism to size, and allometric laws in general
- The two-thirds power law, relating speed to curvature in the human motor system.[17]
- The Taylor's law relating mean population size and variance of populations sizes in ecology
- Neuronal avalanches[5]
- The species richness (number of species) in clades of freshwater fishes[18]
- The Harlow Knapp effect, where a subset of the kinases found in the human body compose a majority of published research[19]
- The size of forest patches globally follows a power law [20]
- The species–area relationship relating the number of species found in an area as a function of the size of the area

Chemistry [edit]

- Rate law

Climate science [edit]

- Sizes of cloud areas and perimeters, as viewed from space[3]
- The size of rain-shower cells[21]
- Energy dissipation in cyclones[22]
- Diameters of dust devils on Earth and Mars [23]

General science [edit]

- Exponential growth and random observation (or killing)[24]
- Progress through exponential growth and exponential diffusion of innovations[25]
- Highly optimized tolerance
- Proposed form of experience curve effects
- Pink noise
- The law of stream numbers, and the law of stream lengths (Horton's laws describing river systems)[26]
- Populations of cities (Gibrat's law)[27]
- Bibliograms, and frequencies of words in a text (Zipf's law)[28]
- 90–9–1 principle on wikis (also referred to as the 1% rule)[29][30]
- Richardson's Law for the severity of violent conflicts (wars and terrorism)[31][32]
- The relationship between a CPU's cache size and the number of cache misses follows the power law of cache misses.
- The spectral density of the weight matrices of deep neural networks[33]

Economics [edit]

- Population sizes of cities in a region or urban network, Zipf's law.
- Distribution of artists by the average price of their artworks.[34]
- Income distribution in a market economy.
- Distribution of degrees in banking networks.[35]
- Firm-size distributions.[36]

Finance [edit]

- Returns for high-risk venture capital investments[37]
- The mean absolute change of the logarithmic mid-prices[38]
- Large price changes, volatility, and transaction volume on stock exchanges[39]
- Average waiting time of a directional change[40]
- Average waiting time of an overshoot

Mathematics [edit]

- Fractals
- Pareto distribution and the Pareto principle also called the "80–20 rule"
- Zipf's law in corpus analysis and population distributions amongst others, where frequency of an item or event is inversely proportional to its frequency rank (i.e. the second most frequent item/event occurs half as often as the most frequent item, the third most frequent item/event occurs one third as often as the most frequent item, and so on).
- Zeta distribution (discrete)
- Yule–Simon distribution (discrete)
- Student's t-distribution (continuous), of which the Cauchy distribution is a special case
- Lotka's law
- The scale-free network model

Physics [edit]

- The Angstrom exponent in aerosol optics
- The frequency-dependency of acoustic attenuation in complex media
- The Stefan–Boltzmann law
- The input-voltage–output-current curves of field-effect transistors and vacuum tubes approximate a square-law relationship, a factor in "tube sound".
- Square–cube law (ratio of surface area to volume)
- A 3/2-power law can be found in the plate characteristic curves of triodes.
- The inverse-square laws of Newtonian gravity and electrostatics, as evidenced by the gravitational potential and Electrostatic potential, respectively.
- Self-organized criticality with a critical point as an attractor
- Model of van der Waals force
- Force and potential in simple harmonic motion
- Gamma correction relating light intensity with voltage
- Behaviour near second-order phase transitions involving critical exponents
- The safe operating area relating to maximum simultaneous current and voltage in power semiconductors.
- Supercritical state of matter and supercritical fluids, such as supercritical exponents of heat capacity and viscosity.[41]
- The Curie–von Schweidler law in dielectric responses to step DC voltage input.
- The damping force over speed relation in antiseismic dampers calculus
- Folded solvent-exposed surface areas of centered amino acids in protein structure segments[42]

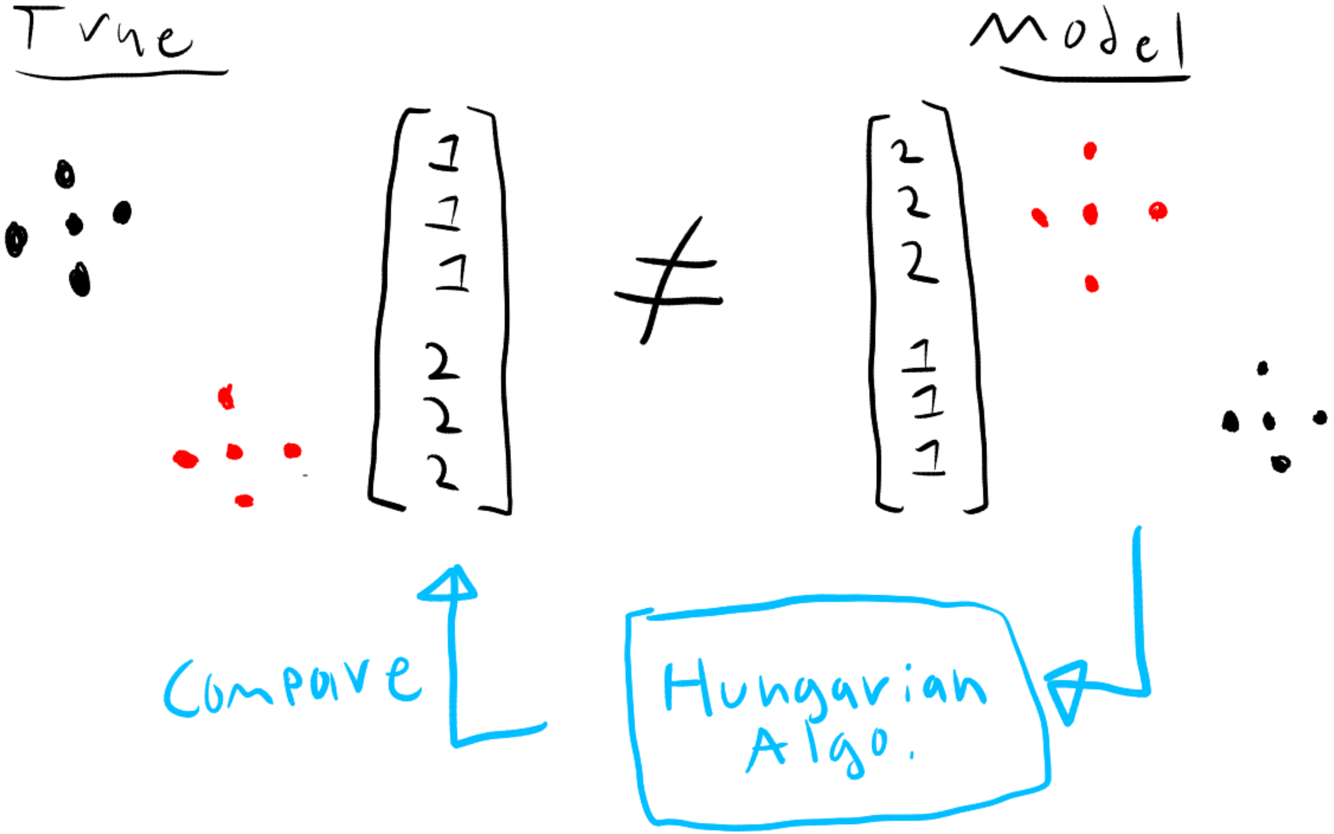
Political Science [edit]

- Cube root law of assembly sizes

Psychology [edit]

- Stevens's power law of psychophysics (challenged with demonstrations that it may be logarithmic[43][44])
- The power law of forgetting[45]

Permutation Issue



Experiments

- Balanced datasets: CIFAR10 & FOOD101
- CIFAR10 test size for each cluster: 1000
- FOOD101 test size for each cluster: 750

Experiments

- Balanced datasets: CIFAR10 & FOOD101
- CIFAR10 test size for each cluster: 1000
- FOOD101 test size for each cluster: 750

- α = decay factor for cluster size
- $\alpha = 2 \Rightarrow$ cluster sizes: 1000, 500, 250, …
- CIFAR10-LT ($\alpha = 2$) & FOOD101-LT ($\alpha = 1.05$)

Experiments

- Balanced datasets: CIFAR10 & FOOD101
- CIFAR10 test size for each cluster: 1000
- FOOD101 test size for each cluster: 750

- α = decay factor for cluster size
- $\alpha = 2 \Rightarrow$ cluster sizes: 1000, 500, 250, ...
- CIFAR10-LT ($\alpha = 2$) & FOOD101-LT ($\alpha = 1.05$)

- I use CLIP trained Resnet50 to extract latent variables for this data
- I explore having/not having a bottom floor for cluster size

Prior Distribution Comparisons

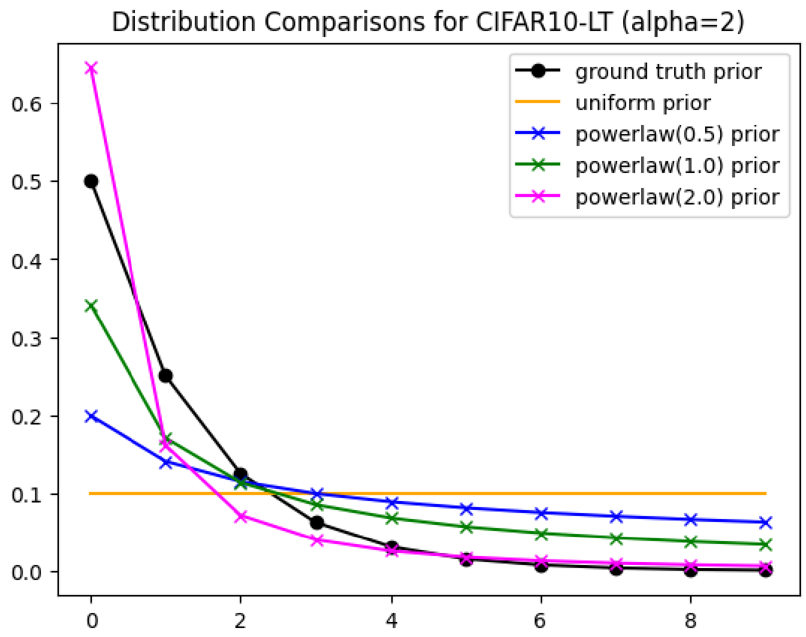


Figure 13: No clip on cluster size.

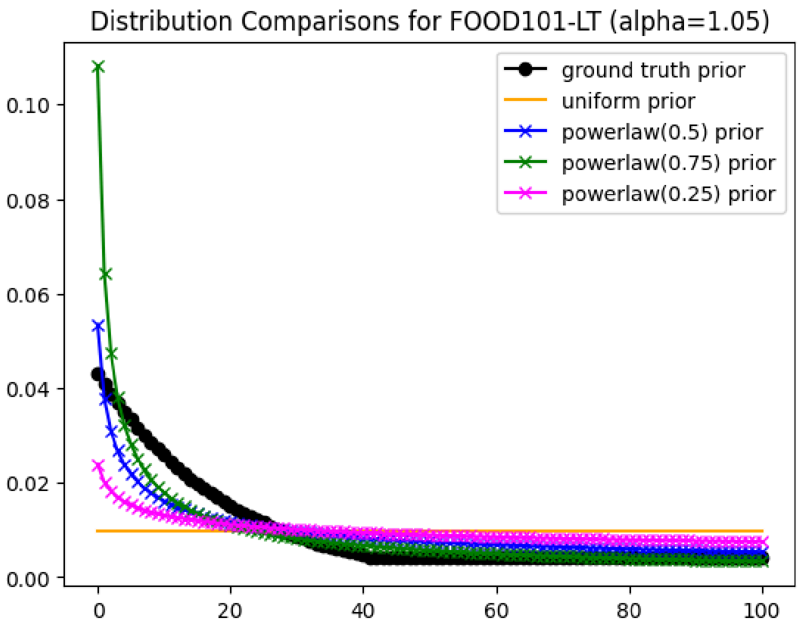


Figure 14: Cluster size clipped.

Accuracy Comparisons

Distribution Type	CIFAR10-LT	FOOD101
Paper	66.0%	53.4%
Powerlaw (0.5)	70.5%	67.8%
Powerlaw (1.0)	72.8%	-
Powerlaw (2.0)	71.9%	-
Truth Prior	73.5%	69.6%

Confusion Matrices (truth on y-axis)

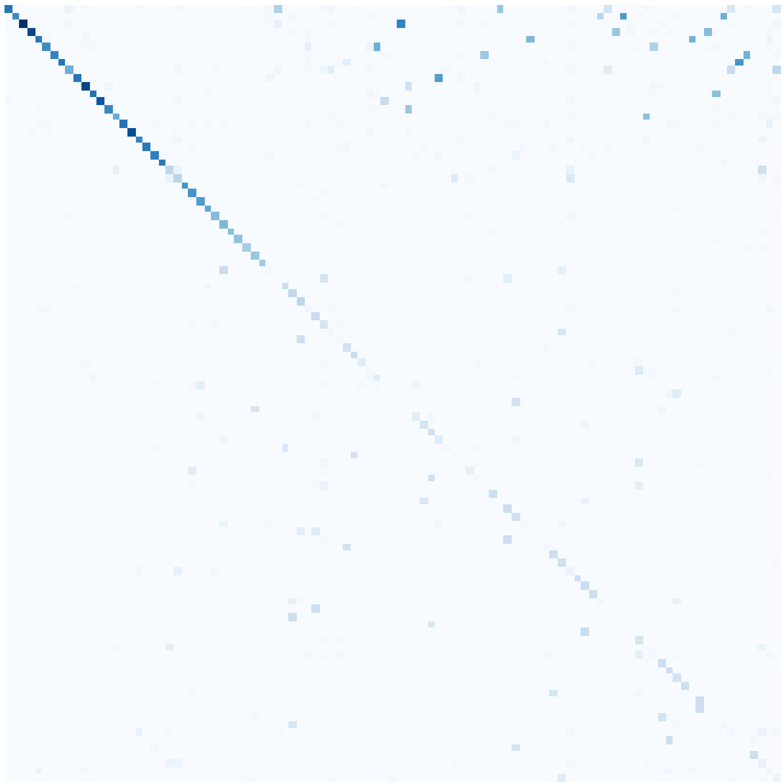


Figure 15: TURTLE on Food101.

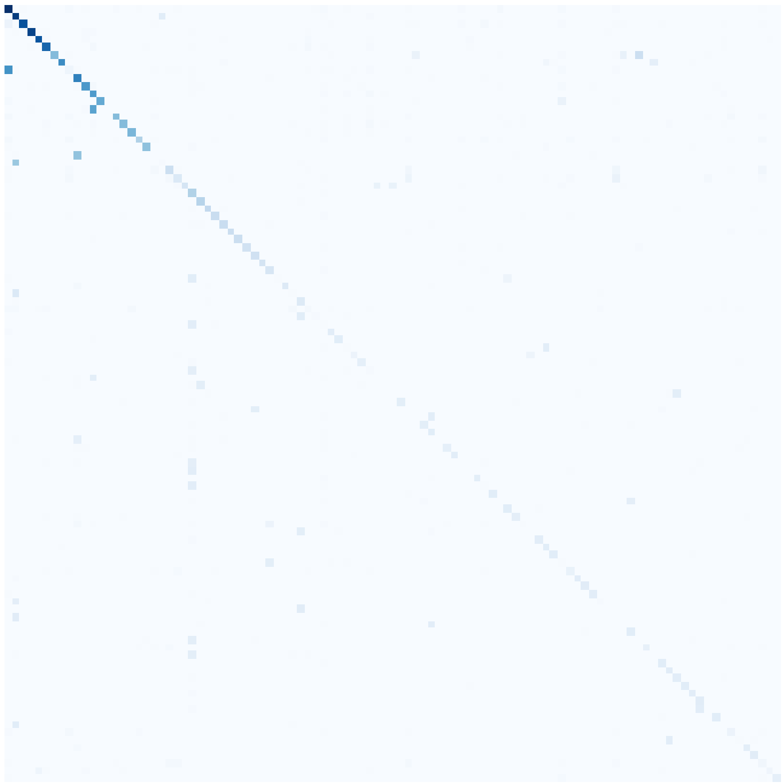


Figure 16: PET-TURTLE on Food101.

PET-TURTLE: Potential Next Steps

- Explore other terms besides KL divergence such as Wasserstein distance

PET-TURTLE: Potential Next Steps

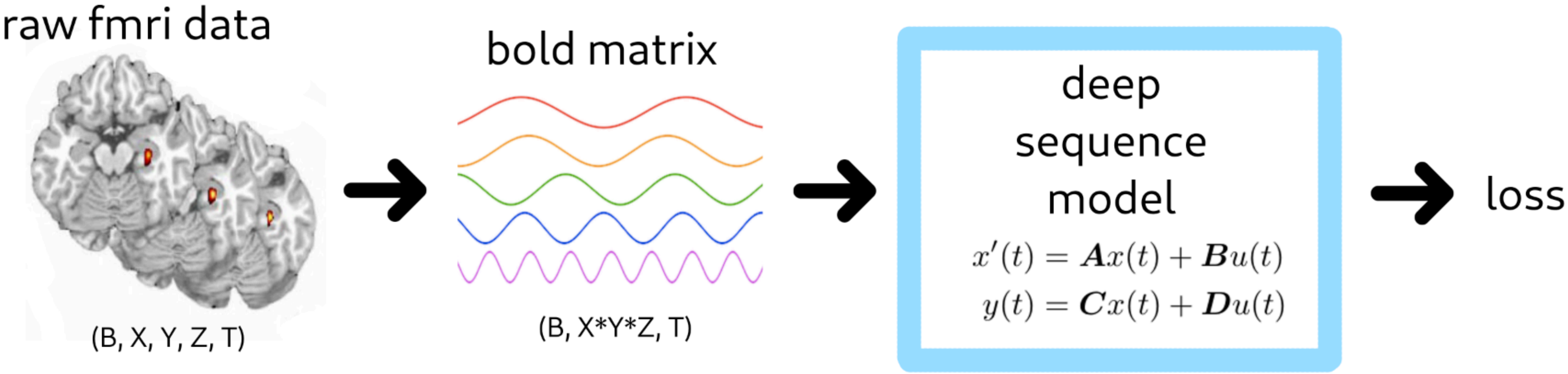
- Explore other terms besides KL divergence such as Wasserstein distance
- Estimate prior distribution Π along the way with τ, w ?
  - May need reformulation, otherwise could estimate empirical dist.

PET-TURTLE: Potential Next Steps

- Explore other terms besides KL divergence such as Wasserstein distance
- Estimate prior distribution Π along the way with τ, w ?
  - May need reformulation, otherwise could estimate empirical dist.
- Generalize to do joint learning from sMRI & fMRI as shown:

$$\sum_{x \in \mathcal{D}} \sum_{k=1}^{2} \mathcal{L}_{ce}\left(w_\theta^k \cdot \phi^k\left(x^k\right); \tau_\theta\left(\phi^k\left(x^k\right)\right)\right) \text{ s.t. } w_\theta^k = \Xi\left(w_\theta^k, \mathcal{D}\right) \ \forall k$$

Next Steps (4D fMRI Temporal Encoder)



Goal:

Create 4D encoder to temporally distill patient state from entire session

Could be used in CL, supervised CL, or standard classification CE loss

Thesis Proposal
Dimensionality Reduction & Clustering

Classical ML Methods

Subspace Learning

🦙 ALPCAH: 🦙
PCA + Heteroscedastic Data

Subspace Clustering

🦙 🦙 ALPCAHUS: 🦙 🦙
UoS + Heteroscedastic Data

Deep Learning Methods

Representation Learning

METHOD:
4D fMRI Temporal Encoder

Unsupervised Learning

🐢 PET-TURTLE: 🐢
Joint Clustering + Data Imbalance

Bibliography

[1] I. Higgins *et al.*, "beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR (Poster)*, vol. 3, 2017.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, 2020, pp. 1597–1607.

[3] R. Sharma, K. Ji, zhiqiang xu, and C. Chen, "AUC-CL: A Batchsize-Robust Framework for Self-Supervised Contrastive Representation Learning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=YgMdDQB09U

[4] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15750–15758.

[5] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *biometrics*, vol. 21, pp. 768–769, 1965.

[6] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[7] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[8] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, “Towards k-means-friendly spaces: Simultaneous deep learning and clustering,” in *international conference on machine learning*, 2017, pp. 3861–3870.

[9] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, “Contrastive clustering,” in *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 8547–8555.

[10] A. Gadetsky, Y. Jiang, and M. Brbic, “Let Go of Your Labels with Unsupervised Transfer,” *arXiv preprint arXiv:2406.07236*, 2024.