On the Importance of Separation and Labelling on the Hypersphere

Martin Lindström, Ragnar Thobaben and Mikael Skoglund

2026-03-03

Outline of the Talk

- Title: “On the Importance of Separation and Labelling on the Hypersphere”
  - Part I: Information Theory in Representation Learning
  - Part II: Empirical and Geometrical Aspects
  - Part III: Our Contributions

- Publications
  - [1] M. Lindström, B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, ‘A Coding-Theoretic Analysis of Hyperspherical Prototypical Learning Geometry’, in *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, PMLR vol. 251, pp. 78–91.
  - [2] M. Lindström, R. Thobaben, and M. Skoglund, ‘On the Importance of Separation and Labelling on the Hypersphere’, in preparation.

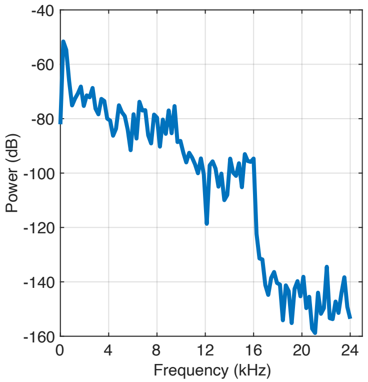
Part I:
Information Theory in Representation Learning

Real-World Data

Images

“So remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist.”

Text



Audio

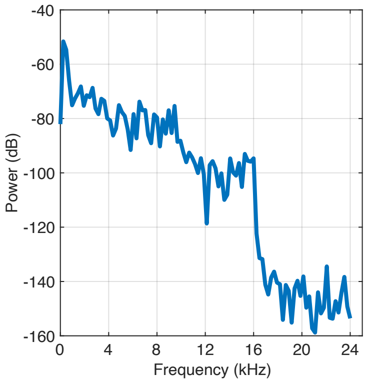
Real-World Data

Lossy JPEG compression

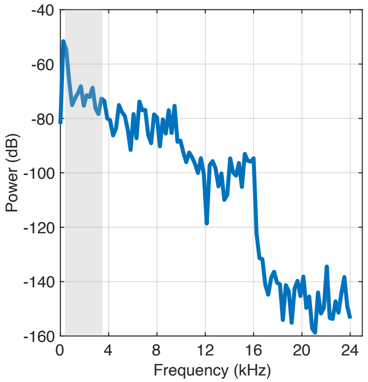
“So remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist.”

↓ Lossy compression

“S rmmbr t lk p t th strs nd nt dwn t r ft. Tr t mk sns f wht s nd wndr bt wht mks th unvrs xst.”

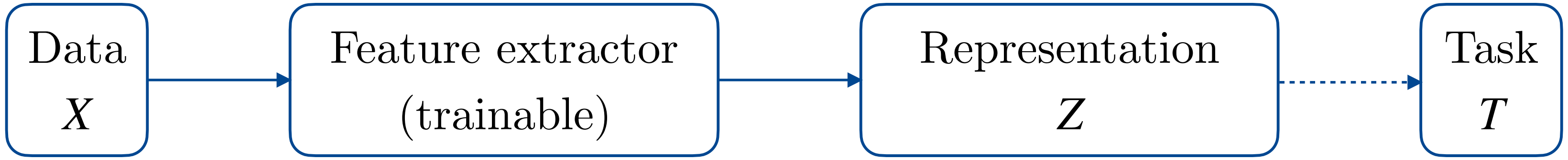


↓ Voice band coding



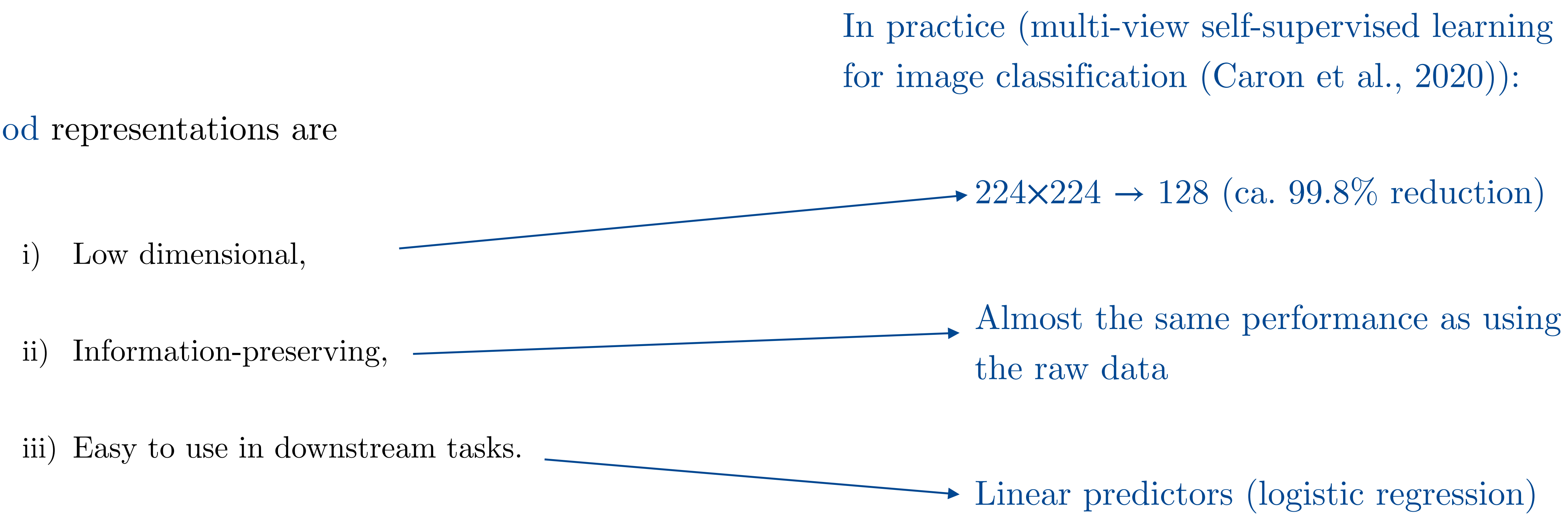
Real-world data can be heavily compressed while still being usable

The Aim



- Some “definitions” of good representations:
  - Bengio et al. (2013): “learning representations of the data that make it easier to extract useful information when building classifiers or other predictors”
  - Goodfellow et al. (2016): “a good representation is one that makes a subsequent learning task easier”
  - Rodríguez Gálvez et al. (2023): “learning lower dimensional representations of data which capture the data’s semantic information”

Good Representations: My Definition



- Good representations are

  i) Low dimensional,

  ii) Information-preserving,

  iii) Easy to use in downstream tasks.

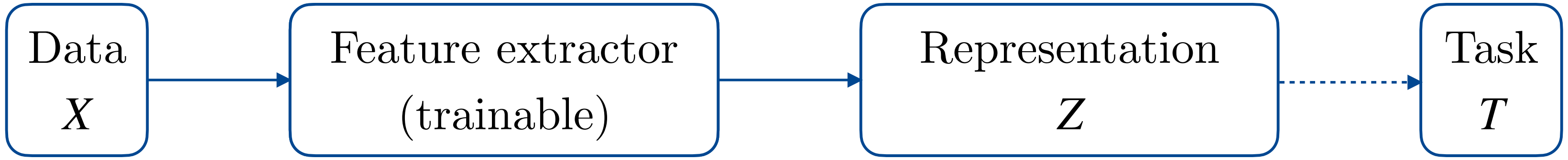
In practice (multi-view self-supervised learning for image classification (Caron et al., 2020)):

224×224 → 128 (ca. 99.8% reduction)

Almost the same performance as using the raw data

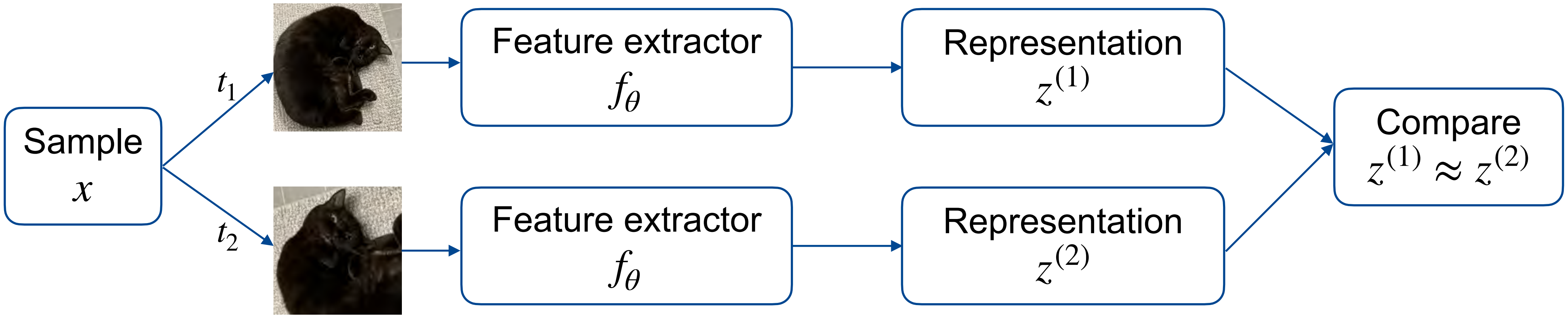
Linear predictors (logistic regression)

Information-Theoretic Analogies



- Information-theoretic analogies:
  - “Keep all relevant information in the data” – the InfoMax principle
  - Approximately lossless source coding
  - Rate-distortion theory
  - Sufficient statistic for most realistic tasks

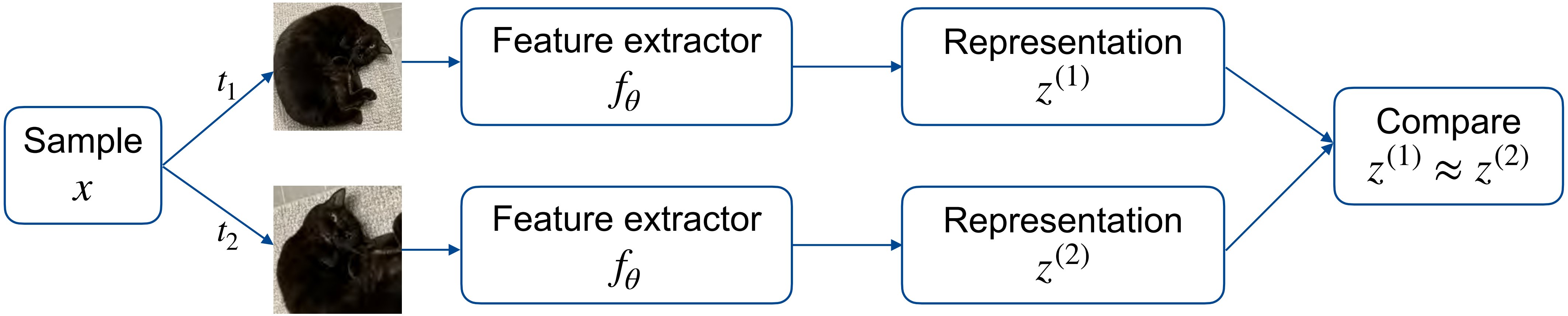
Towards a General Task T: Self-Supervision



Sample x
t_1
t_2
Feature extractor f_θ
Feature extractor f_θ
Representation $z^{(1)}$
Representation $z^{(2)}$
Compare $z^{(1)} \approx z^{(2)}$

- Self-supervised learning: train representations for self-similarity
  - i) Define a class of transforms $\mathcal{T}$ which preserve "enough" information (translations, rotations, flips, crops, colour distortions, blurs ...)
  - ii) Draw $t_1, t_2 \sim \mathcal{T}$ and create the artificial samples $x^{(1)} = t_1(x)$ and $x^{(2)} = t_2(x)$
  - iii) Train the encoder to optimise $\langle z^{(1)}, z^{(2)} \rangle$
- This is the basis of multi-view self-supervised learning (Poole et al., 2019; Chen et al., 2020; Caron et al., 2020; Wang & Isola, 2020; Rodríguez Gálvez et al., 2023; Oquab et al., 2024; Siméoni et al., 2025)

Information-Theoretic Loss Functions



- Broad range of methods, but many are connected to information theory (Rodríguez Gálvez et al., 2023)
- With a tractable density $Q_{Z_2|Z_1}$, we have

$$I(Z_1; Z_2) \geq \underbrace{H(Z_2)}_{\text{Diverse representations}} + \underbrace{\mathbb{E}[\log Q_{Z_2|Z_1}(Z_2)]}_{z^{(1)} \approx z^{(2)}}$$

Information theory is a powerful formalisation of representation learning

Part I Summary

- Representation learning is about learning to extract good features:

    i) Low dimensional,

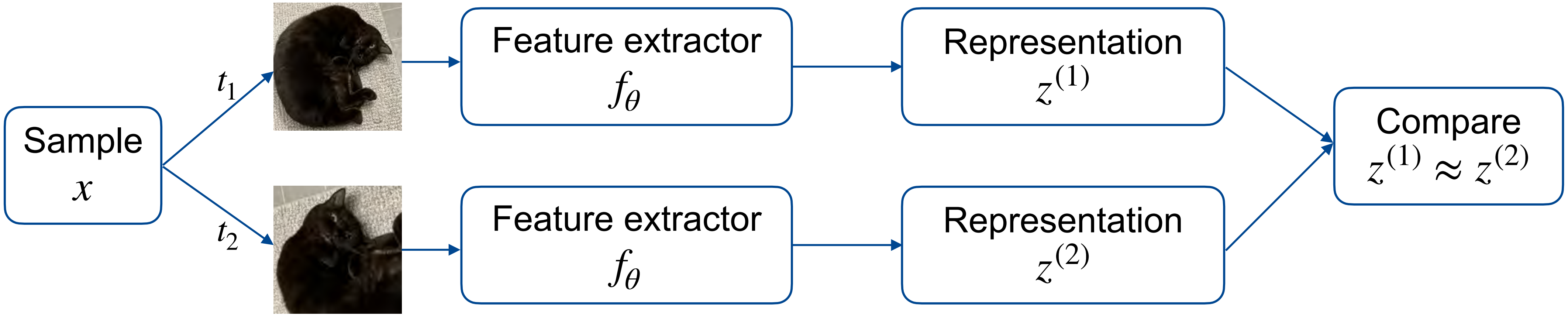
    ii) Information-preserving,

    iii) Easy to use in downstream tasks.

- State-of-the-art methods are typically self-supervised and perform well:
    - Competitive performance on large-scale benchmarks (e.g. ImageNet-1k)
    - Improves performance on smaller benchmarks (e.g. CIFAR-10/100)
- Motivations or losses often are often information-theoretic

Part II: Empirical and Geometrical Aspects

Two Important Practical Details



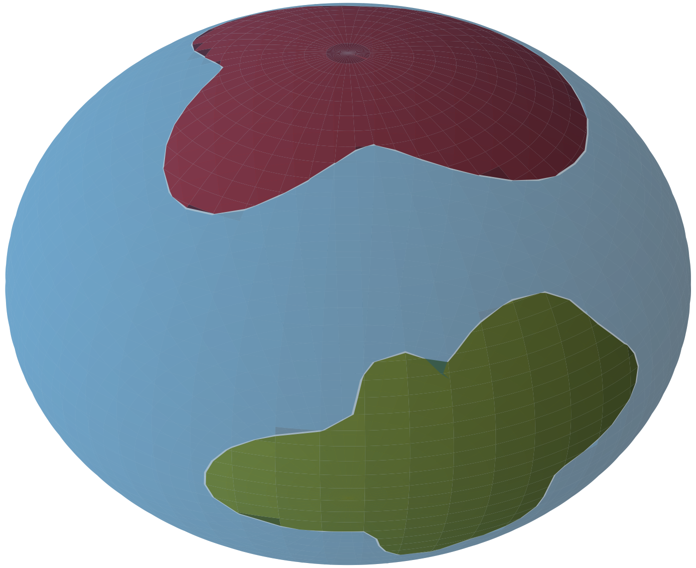
- State-of-the-art methods use hyperspherical latent spaces (Davidson et al., 2018; Chen et al., 2020; Wang & Isola, 2020)
- Neural collapse must be avoided

Hyperspherical Latent Spaces

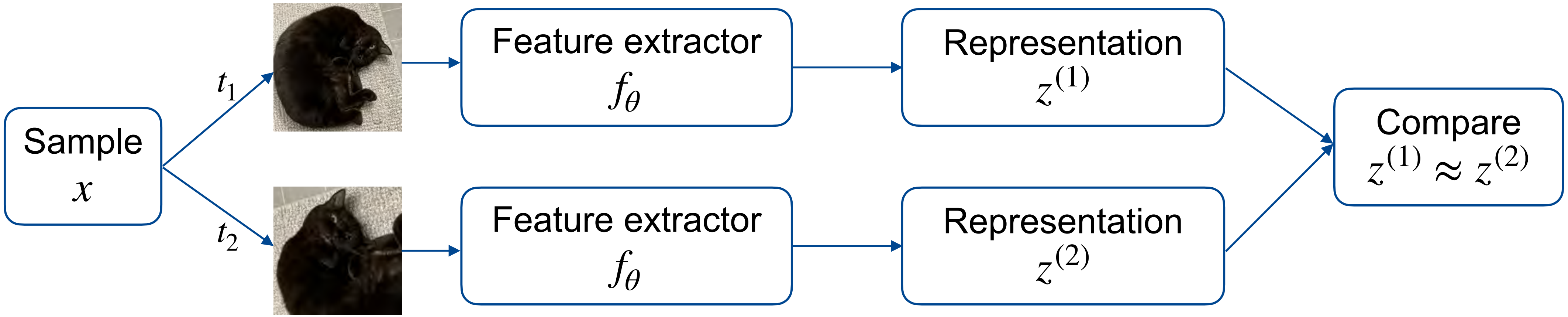
- Normalising representations is helpful empirically (Chen et al., 2020; Wang & Isola, 2020), but the exact reason why is not entirely clear.
- Two commonly cited factors:

1. Most networks output a probability or log-probability. Bounding the output prevents overconfidence and overfitting (Wang et al., 2017)

2. Linear predictors work well (Wang & Isola, 2020):
   If you have tight clustering AND well-separated clusters, then the clusters are linearly separable



Neural Collapse



- Remember: we train to maximise $\langle z^{(1)}, z^{(2)} \rangle$
- An unwanted solution is $z = \text{const.}$ (the representations collapse to one point)

Many differences between methods stem from the need to avoid collapse

Prototypical Learning Avoids Collapse

- Prototypical learning defines a class of methods using the idea:

  i) Pick a codebook $\mathcal{C}$ of $N_{\mathcal{C}}$ prototypes

  ii) Around each prototype, define a prior $P_{c_i|Z} \propto \exp(\langle z, c_i \rangle / \tau)$

  iii) Regularise so that $P_{c_i} = \dfrac{1}{N_{\mathcal{C}}}$ ← Each cluster is equally likely on average

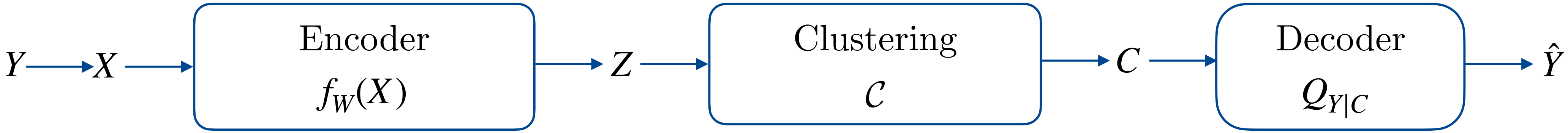
- Key component in self-supervised methods: SwAV (Caron et al., 2020), DINO-v2 (Oquab et al., 2024), DINO-v3 (Siméoni et al., 2025), and many more
- Key idea in supervised methods: hyperspherical prototypical learning (Mettes et al., 2019; Kasarla et al., 2022)

Hyperspherical Prototypical Learning as an Analysis Framework

- Prototypical learning is
  - A state-of-the-art method
  - Amenable to analysis due to the prototypes

- In our work, we consider the supervised setting since:
  - We know exactly how many prototypes are needed
  - It is easy to relate clustering to performance

The Hyperspherical Prototypical Learning Framework



1. Prior to training, pick fixed hyperspherical prototypes $\mathcal{C}$
2. Obtain a hyperspherical representation $z = f_W(x)$
3. Perform a soft clustering through $\operatorname{softmax}(z, \mathcal{C}) = \left[\frac{\exp(\langle z, c_i \rangle / \tau)}{\sum_j \exp(\langle z, c_j \rangle / \tau)} \right]_{i=1}^{N_{\mathcal{C}}}$

   Von Mises-Fisher prior

4. Classify through nearest-neighbour decoding: $\hat{y}$ = index of the closest prototype

Picking Prototypes

- Assumption: we want maximally separated prototypes

$$(P) \quad \min_{\mathcal{C}} \max_{i \neq j} \langle c_i, c_j \rangle$$

Optimise the worst-case separation

- Intuition: channel coding; maximise SNR under a power constraint, ...

Assumption: Larger prototype separation gives better linear separability between classes

Solving (P)

- A closer look at (P)

$$(P) \quad \min_{\mathcal{C}} \max_{i \neq j} \langle c_i, c_j \rangle$$

- Unfortunately, (P) is unsolved even in $\mathbb{R}^3$... (Conway & Sloane, 1999; Musin & Tarasov, 2015)
- Two approximate solutions (Mettes et al., 2019; Kasarla et al., 2022)

$$\begin{aligned}(\mathrm{P}_{\mathrm{Mettes}}) \quad \min_{\mathrm{C}} \quad & \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{N_{\mathcal{C}}} \max_{j} M_{i,j}, \\ \text{s.t.} \quad & M = C^{\top} C - 2I, \\ & \|c_i\| = 1\end{aligned}$$

Average worst case?

Avoid selecting $\langle c_i, c_i \rangle = 1$

$$\begin{aligned}(\mathrm{P}_{\mathrm{Kasarla}}) \quad \min_{\mathcal{C}} \quad & a, \\ \text{s.t.} \quad & \langle c_i, c_j \rangle = a, \quad \forall i \neq j.\end{aligned}$$

Restrict to constant pair-wise separation

Open Problems

- Prototypical learning is a key component of many state-of-the-art methods

- How do we pick the prototypes?

$$(P) \qquad \min_{\mathcal{C}} \max_{i \neq j} \langle c_i, c_j \rangle$$

  - Can we solve (P), or if not, get good approximate solutions?
  - Are the state-of-the-art solutions good?

- What is the connection between separation and downstream performance?
  - Is “better prototype separation ⇒ better linear separability” true?

Part III:
Our Contributions

Warmup: A Convex Relaxation to (P)

- Recall the non-convex and combinatorial problem (P)

$$(P) \quad \min_{\mathcal{C}} \max_{i \neq j} \langle c_i, c_j \rangle$$

- Recall the log-sum-exp approximation:

$$\max_i x_i \leq \frac{1}{t} \log \left(\sum_{i=1}^{n} \exp(t x_i) \right) \leq \max_i x_i + \frac{\log n}{t}, \text{for any } t > 0$$

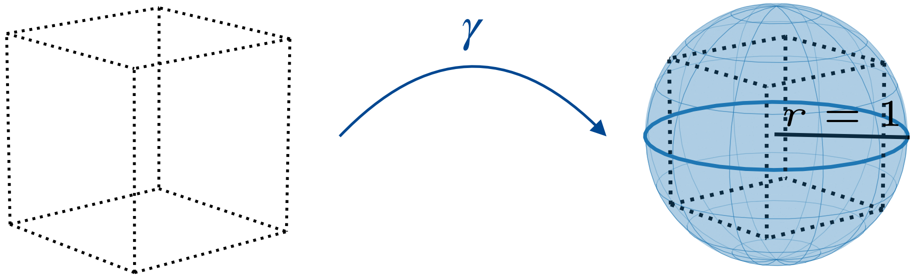
- This gives a relaxed, tractable problem (solvable by projected gradient descent)

$$(\mathrm{P}_{\mathrm{LSE}}) \quad \min_{\mathcal{C}} \frac{1}{t} \log \sum_{i \neq j} \exp(t \langle c_i, c_j \rangle)$$

Coding-Theoretic Relaxation

- Consider the transformation $\gamma : \{0,1\}^n \to \mathbb{S}^{n-1}$ and create prototypes through

$$c = \gamma(b) := \frac{2(b - 1/2)}{\sqrt{n}}$$



- Then $\|c\| = 1$ and $c(k) \in \{-1/\sqrt{n}, +1/\sqrt{n}\}$

- Moreover, if $c_i = \gamma(b_i)$ and $c_j = \gamma(b_j)$ for $b_i \neq b_j$

$$\langle c_i, c_j \rangle = 1 - \frac{2d_{\mathrm{H}}(b_i, b_j)}{n} \underset{\uparrow}{\leq} 1 - \frac{2d_{\min}}{n}$$

For a binary codebook with $d_{\mathrm{H}}(b_i, b_j) \geq d_{\min}$

Good Codebooks

- We have an unusual setting:
  - Often very low rates: $N_{\mathcal{C}} \approx n \Rightarrow R \approx \dfrac{\log n}{n}$
  - Small block lengths
  - Decoding complexity is negligible
  - The aim is not achieving capacity, but a large $d_{\min}$
- Therefore we consider
  - BCH codes: known to be good at low block lengths (MacWilliams & Sloane, 1977)
  - Reed-Muller (RM) codes: good at low rates (Abbe et al., 2021)

Full Characterisation of Solutions to (P)

- The Rankin bound from spherical coding theory states that $\max_{i \neq j} \langle c_i, c_j \rangle \geq \dfrac{-1}{N_{\mathcal{C}} - 1}$

- Main theoretical results (Theorem 7): a tight characterisation of solutions to (P)

$$\frac{-1}{N_{\mathcal{C}} - 1} \leq \max_{i \neq j} \langle c_i, c_j \rangle \leq 1 - \frac{d_{\mathrm{GV}}}{n}$$

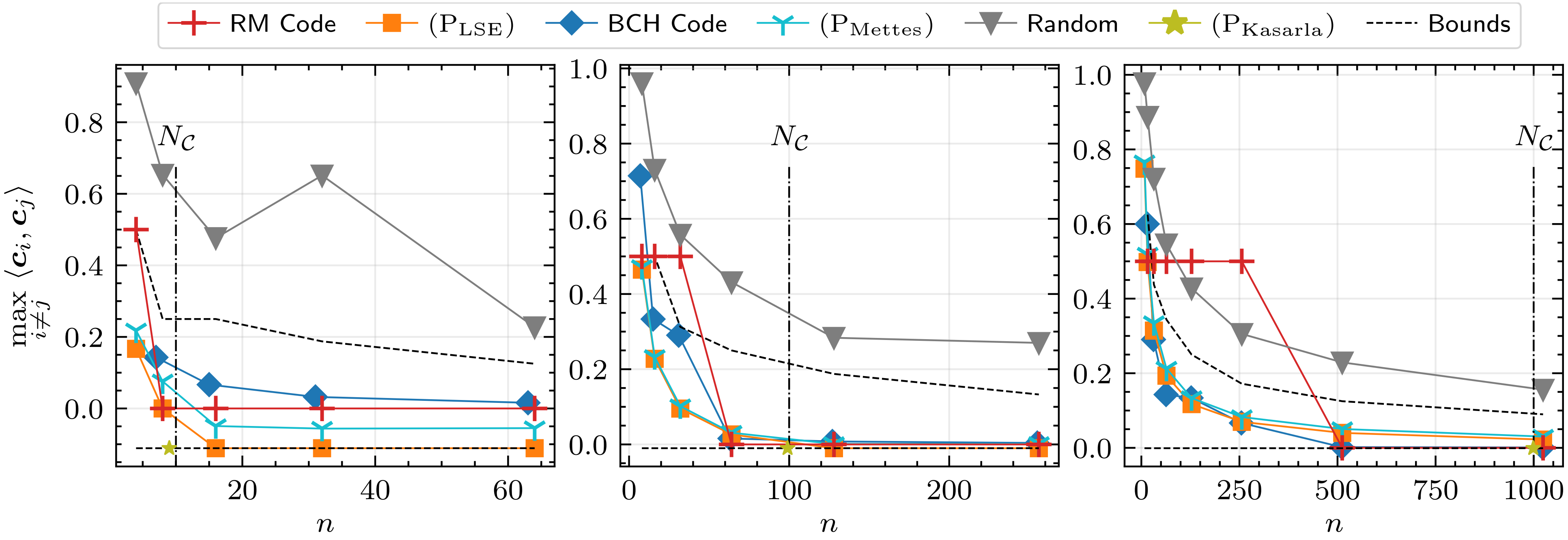
From the Gilbert-Varshamov bound: there are good binary codes

For moderate problem sizes ($N_{\mathcal{C}} \gtrsim 100$) this is close to 0 (worst-case angle $\leq 90.58°$)

In low dimension $n \approx \dfrac{N_{\mathcal{C}}}{2}$, worst-case similarity 0 is achievable (worst-case angle 90°) (next slide)

Orthogonal prototypes are achievable in low dimension,
and there exist no set of prototypes much better than orthogonal!

Numerical Examples on Solutions to (P)

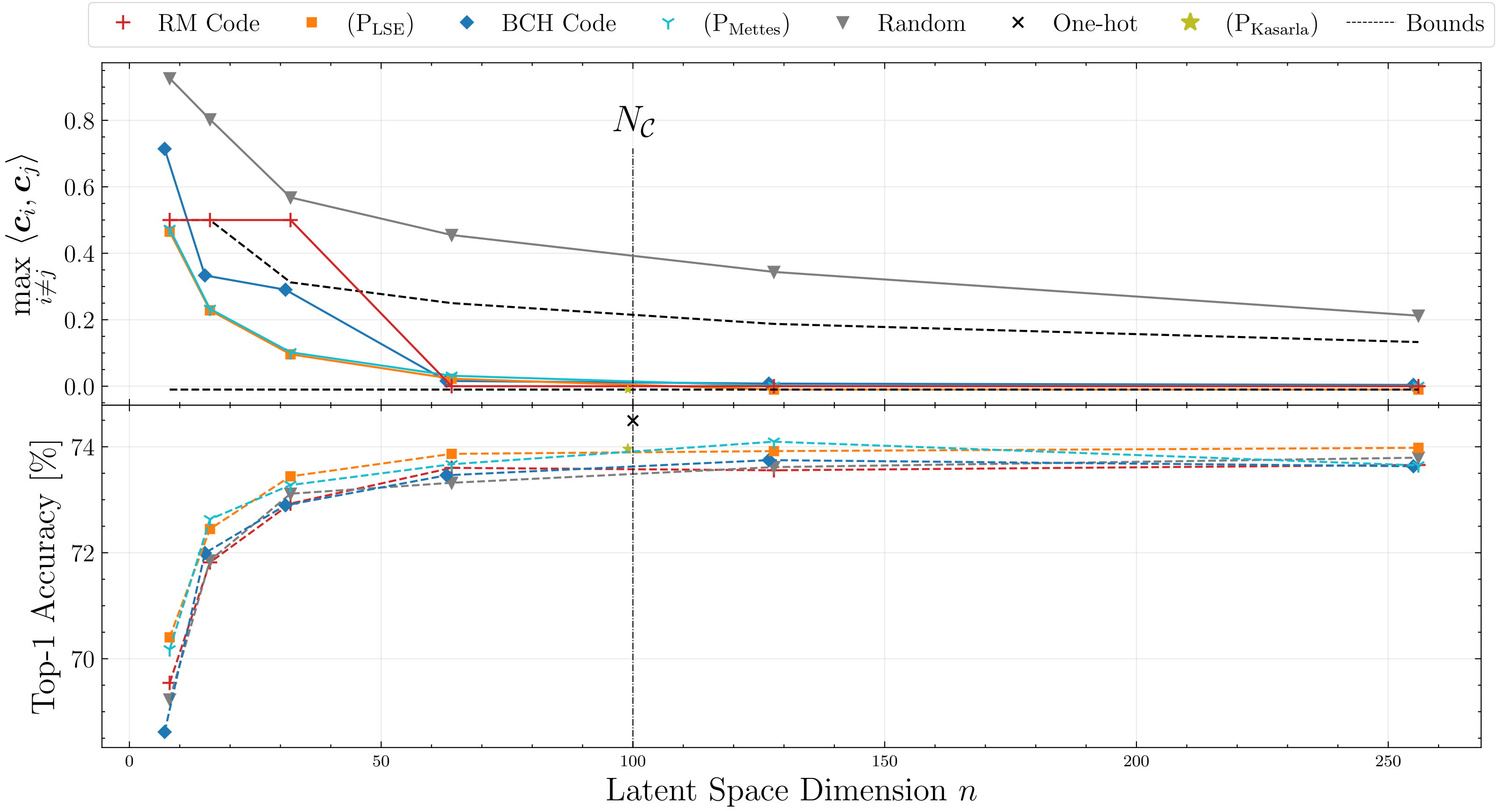


10 prototypes

100 prototypes

1000 prototypes

Connecting Separation to Performance



Better separation does not give better performance!

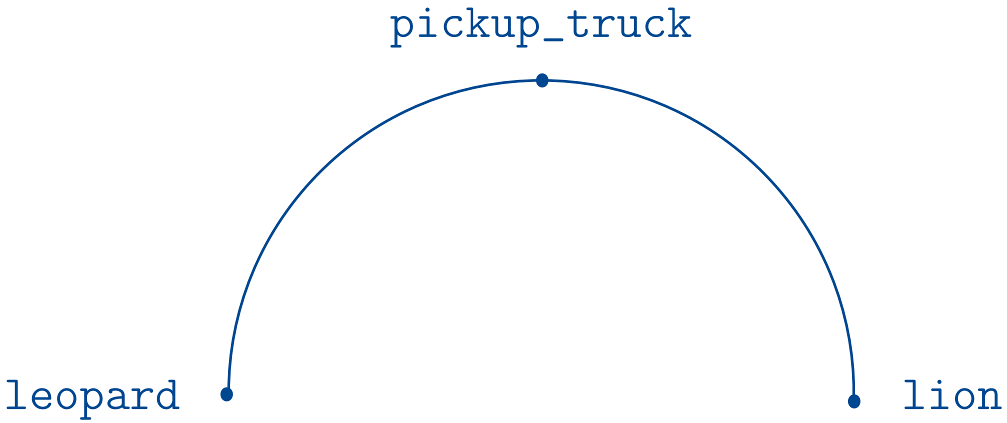
Relabelling is Important: Empirical Motivation

- There is an underlying assumption about the decoder: $z \approx c_i$ is decoded to $\hat{y} = i$
- This gives unwanted behaviour in certain settings, for example CIFAR-100 with RM codes
  - CIFAR-100 is ordered alphabetically, for instance

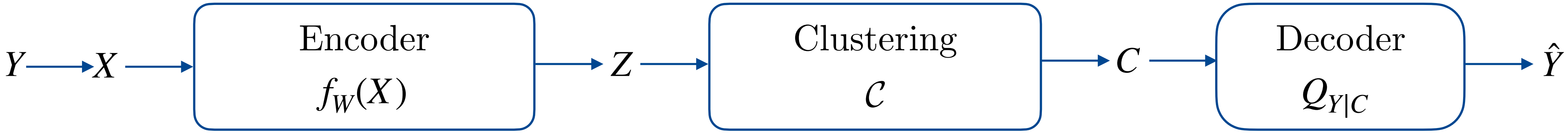
```
CIFAR-100:
…
Class 42: leopard
Class 43: lion
…
```

  - RM codes makes sequential prototypes diametrically opposed, $\langle c_i, c_{i+1} \rangle = -1$

Input data structure is not preserved!



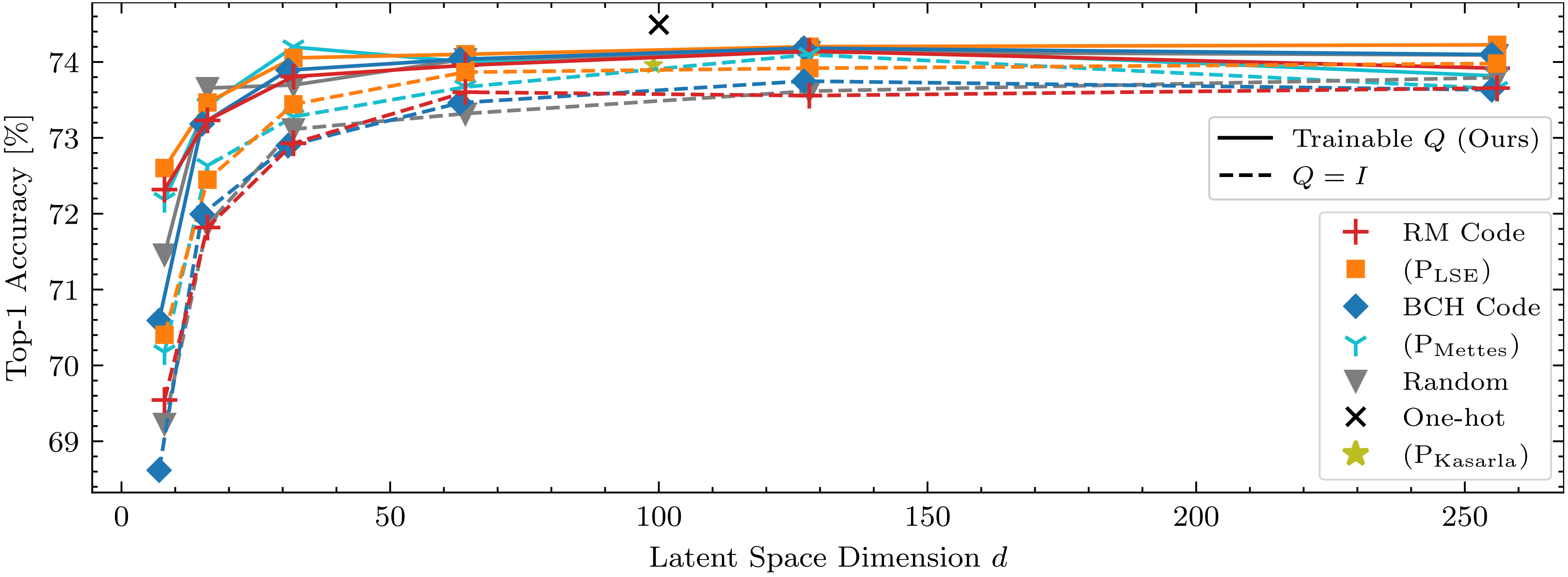
Relabelling is Important: Theoretical Motivation



- A theoretical motivation in three steps:
  1. A large mutual information $I(Z;Y)$ is necessary for good classification accuracy
  2. Our cross-entropy-based loss optimises mutual information
  3. However, mutual information is invariant to shifts and relabellings

The decoder needs to be optimised separately!

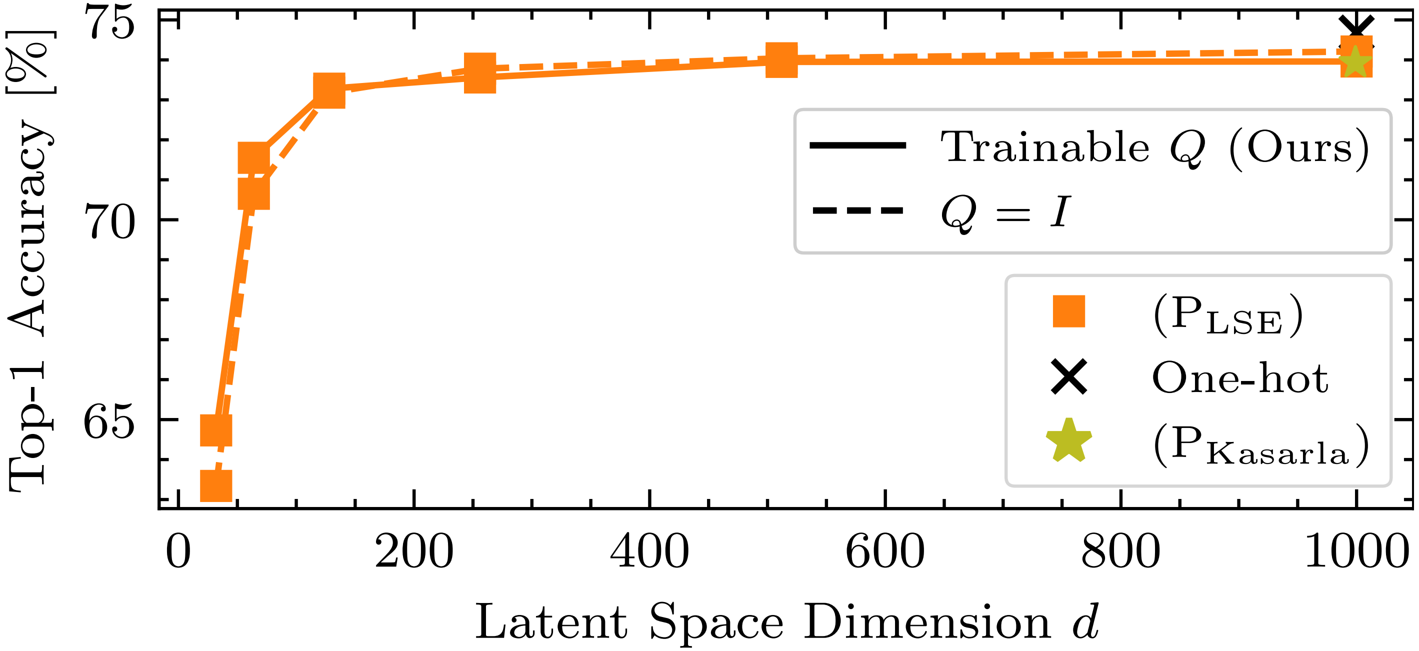
Improvements Through Relabelling



You can compensate poor separation with a trainable relabelling!

Scaling Up to ImageNet

- We now have to pick between 1000! permutations



For large problem sizes, good relabellings are hard to find

Conclusion

- We have analysed hyperspherical clustering, a key component in many methods

- Our contributions include
    - A full characterisation of the optimal hyperspherical prototypes, and efficient algorithms that achieve near-optimal separation in practice
    - Empirical and theoretical motivations why separation is not enough
    - Performance evaluations on CIFAR-100 and ImageNet which show that separation has a surprisingly small impact on performance

References

A. A. Alemi, I. Fischer, J. V. Dillon, and K. P. Murphy, “Deep Variational Information Bottleneck,” in International Conference on Learning Representations, 2017

E. Abbe, A. Shpilka, and M. Ye, “Reed–Muller Codes: Theory and Algorithms,” in IEEE Transactions on Information Theory, vol. 67, no. 6, pp. 3251–3277, 2021

Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013

M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments,” in Advances in Neural Information Processing Systems, 2020

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in Proceedings of the International Conference on Machine Learning, 2020

D.C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, “Flexible, High Performance Convolutional Neural Networks for Image Classification,” in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, vol. 2, pp. 1237–1242, Jul. 2011

J. Conway and N. J. A. Sloane, “Sphere Packings, Lattices, and Groups,” 3rd ed. Springer, 1999

T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, “Hyperspherical Variational Auto-Encoders,” in Conference on Uncertainty in Artificial Intelligence, 2018

N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, “Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders,” 2017, arXiv:1611.02648

I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” MIT Press, 2016

T. Kasarla, G. Burghouts, M. van Spengler, E. van der Pol, R. Cucchiara, and P. Mettes, “Maximum Class Separation as Inductive Bias in One Matrix,” in Advances in Neural Information Processing Systems, 2022

D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in International Conference on Learning Representations, 2014

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information Processing Systems, 2012

J. Lyu, S. Zhang, Y. Qi, and J. Xin, “AutoShuffleNet: Learning Permutation Matrices via an Exact Lipschitz Continuous Penalty in Deep Convolutional Neural Networks,” in Proceedings of the International Conference on Knowledge Discovery & Data Mining, pp. 608–616, 2020

F. J. MacWilliams and N. J. A. Sloane, “The Theory of Error-Correcting Codes”. North-Holland, 1977

P. Mettes, E. van der Pol, and C. Snoek, “Hyperspherical Prototype Networks,” in Advances in Neural Information Processing Systems, 2019

O. R. Musin and A. S. Tarasov, “The Tammes Problem for N=14,” in Experimental Mathematics, vol. 24, no. 4, pp. 460-468, 2015

M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” in Transactions of Machine Learning Research, 2024

B. Poole, S. Ozair, A. Van Den Oord, A. A. Alemi, and G. Tucker, “On Variational Bounds of Mutual Information,” in Proceedings of the International Conference on Machine Learning, 2019

References

B. Rodríguez Gálvez, A. Blaas, P. Rodríguez, A. Golinski, X. Suau, J. Ramapuram, D. Busbridge, and L. Zappella, “The Role of Entropy and Reconstruction in Multi-View Self-Supervised Learning,” in Proceedings of the International Conference on Machine Learning, 2023

M. Sefidgaran, A. Zaidi, and P. Krasnowski, “Minimum Description Length and Generalization Guarantees for Representation Learning,” in Advances in Neural Information Processing Systems, 2023

M. Sefidgaran, A. Zaidi, and P. Krasnowski, “Generalization Guarantees for Multi-View Representation Learning and Application to Regularization via Gaussian Product Mixture Prior,” 2025, arXiv:2504.18455

O. Siméoni et al., “DINOv3,” 2025, arXiv:2508.10104

N. Tishby, F. Pereira, and W. Bialek, “The Information Bottleneck Method,” in Proceedings of the Allerton Conference on Communication, Control and Computing, pp. 368-377, 1999

Y. Uğur, G. Arvanitakis, and A. Zaidi, “Variational Information Bottleneck for Unsupervised Clustering: Deep Gaussian Mixture Embedding,” in Entropy, vol. 22, 2020

F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “NormFace: L2 Hypersphere Embedding for Face Verification,” in Proceedings of the 25th ACM international conference on Multimedia, pp. 1041–1049, 2017

T. Wang and P. Isola, “Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere,” in Proceedings of the International Conference on Machine Learning, 2020

KTH
VETENSKAP
OCH KONST