How much do Label Representations Matter for Image Classification?

Ishaan Chandratreya, Katon Luaces
Columbia University

What's in a Label?

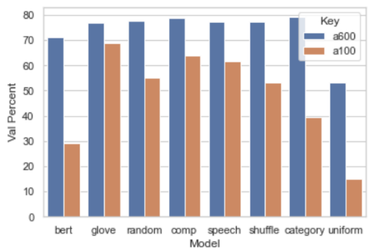
Recent work (Chen et al, 2021) argues that the ***representation of the label space*** plays a critical role in a supervised learning problem in addition to the representation of the input space. They investigate this in context of image classification and endorse (I) **high dimensional** and (II) **high entropy** labels for:

1. Strong *adversarial robustness* of representation
2. Quicker convergence of models with *fewer samples*

We systematically attempt to reproduce their work and reach the conclusions:

1. Label space manipulation seems to have *tangible* impact on training dynamics
2. Claimed dependence is inconsistent as multiple parameters are co-dependent

As such we note this area as a promising direction for further investigation, and propose a systematic study of label representations to stress-test and expand the claims of (Chen et al, 2021) on the quicker convergence with fewer samples claim. As such we will train only on 10% of the CIFAR-10 dataset throughout.



Parameter \ Work	Chen et al. 2021	Our Work (UML!)
Dimension	Directly compare 64*64 labels to 10 dimension one-hot	Compare impact of dim. while preserving distances
Entropy	Entropy is conflated with dimensions ('oddly defined')	Entropy of underlying label space used
Label-Label Relationships	NA	Organize dependence on ordinality, dissimilarity and misc. semantics

Gaussian Random Projections

Theorem: Johnson-Lindenstrauss: Given $0 < \varepsilon < 1$, a set X of m points in $\mathbb{R}^N$, and a number $n > 8\ln(m)/\varepsilon^2$, there is a linear map $f : \mathbb{R}^N \to \mathbb{R}^n$ such that:

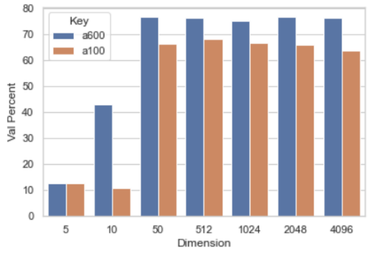
$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2$$

for all $u, v \in X$.

Dimensionality: We sample from ~N(0, 0.1* I) Gaussians of various dimensions.

Rotations: We apply gaussian projections to the standard one-hot label system to get random rotations. We use both the same dimension and the min. JL dimension.

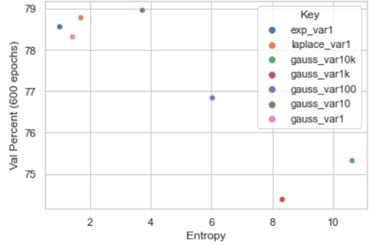
Type	Best Val. 600	Best Val. 100
Category (Regression)	40.64	12.52
Same Dim.	53.16	13.15
Min. JL Dim.	12.93	11.97

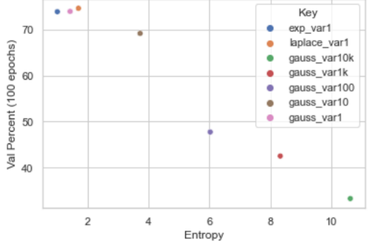


Label Distribution Entropies

We draw 512 dimensional labels with each element sampled independently to measure the impact of changing **differential entropy.** Contrary to [Chen et. al], we notice a negative trend**:**

Distribution	D.E. Formula
Exp.	$1 - \ln(\lambda)$
Laplace	$1 + \ln(2\sigma)$
Gaussian	$\ln(\sigma\sqrt{2\pi e})$





Language Models for Labels

If the label associated with a classification task carries a semantic meaning, then embeddings trained on this semantic meaning can be used as a potential label space. **Chen et al, 2021** experiment with BERT and Glove embeddings.

We briefly consider predicting a set of such self-supervised pre-text task labels as a vector and note that they tend to have strong performance. Please note that the BERT and CLIP embeddings are zero-centered and scaled with std. deviation.

Embedding	Pretext details	Dimensions	Best Val. 600
Glove	Matrix Fact. on Word-Context	50	76.92
Word2Vec	CBOW, Skip-Gram	100	76.84
CLIP	Multimodal Retrieval	512	77.92
BERT	Language *cloze*, Sentence Prediction	768	78.14

Contrastive (Metric) Learning

A more general framework to preserve label-label relationships will make use of the implicit semantic distance between features rather than rely on the label to have a reasonable language realization.

Here we use contrastive learning to do "double work" to test this: **in order to see if it helps for a label space to be organized as such,** we:

1. Learn a image representation where image embeddings are separated by label
2. Form a label embedding for each label using the corresponding image embeddings and solve a supervised learning problem from the image itself

To learn the image representation, we can simply use data augmentation to get positives and treat all others as negatives, or we can use the labels to mine positives and negatives. We choose to use ***SimCLR*** and ***Supervised Contrastive Learning (SupCon),*** each from the respective family of methods.

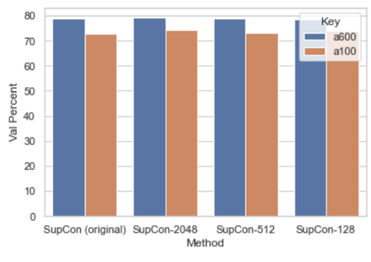
$$l(i,j) = -\log \frac{\exp(z_i, z_j/\tau)}{\sum_{k=1,k\neq j}^{2N} \exp(z_i, z_k/\tau)} \qquad l^{out}(i) = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_k \in A(i) \exp(z_i \cdot z_k/\tau)}$$

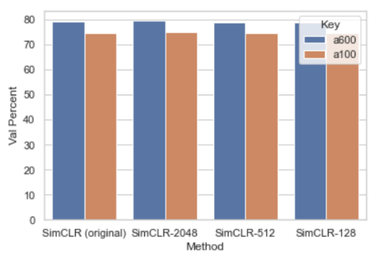
Then with the learnt encoder *f*, the corresponding label representation becomes:

$$l_i = \frac{1}{\sum_{j=1}^{N} 1[y_j = i]} \sum_{j=1, y_j=i}^{N} f(x_j)$$

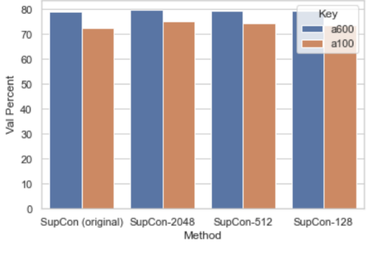
We then calculate a label-label representation matrix and further use this with MDS (metric and non-metric) to learn embeddings of the labels in lower dimensions while preserving the relative distances between the labels. These sets of labels are then used for 10% data CIFAR-10 prediction as before.

Original v. Non Metric MDS





Original v. Metric MDS



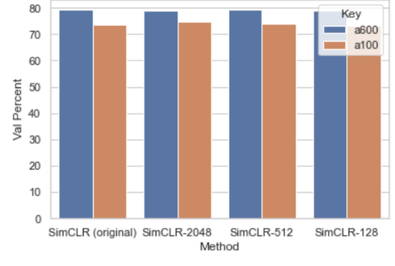
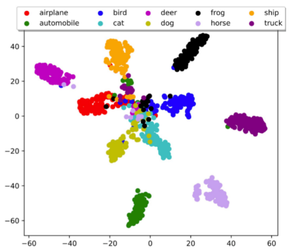
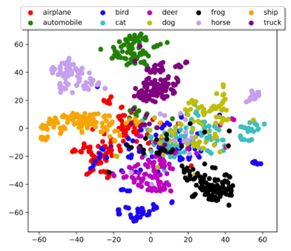


Figure from Khosla et al. (2020)

SupCon



SimCLR



Order Embeddings

The outcome of the previous experiment seems to emphasize that order matters more than distances. Such ordinal constraints might be realized via many forms of hierarchy (and hence be more general):

1. **Hypernymy** (E.g. mammal < canine < husky)
2. **Entailment** (E.g. "sphere drop" < "it rolled").

In an embedding, an ordinal constraint can be enforced via a max-margin push-pull loss:

$$\sum_{u,v \in \mathcal{P}} E(u,v) + \sum_{u',v' \in \mathcal{N}} \max(0, \alpha - E(u',v'))$$

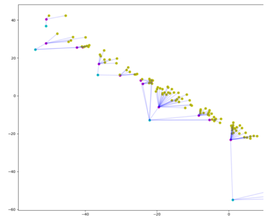
Here, the energy measures the violation of the relationship. Some methods we implement include

Method	Energy
OE (Vendrov et al.2016)	$\|\|\max(0, v-u)\|\|_2^2$
EC (Ganea et al. 2018)	$\max(0, \epsilon(u,v) - \psi(u))$

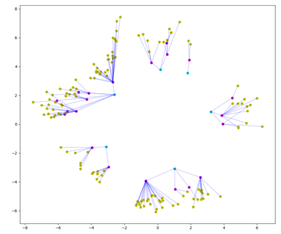
We trained OE and EC using the paper spec and use them for 10% CIFAR Prediction. The results plot the val. accuracy against the number of epochs trained for OE / EC

Figure from Dhall et al. (2020)

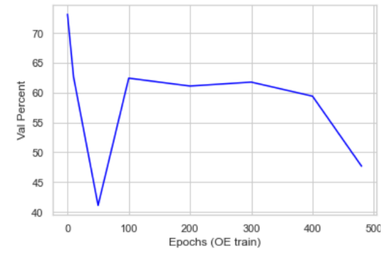
Order Embeddings (OE)

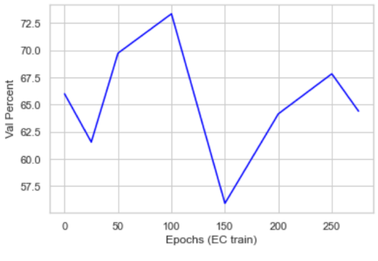


Euclidean Cones (EC)

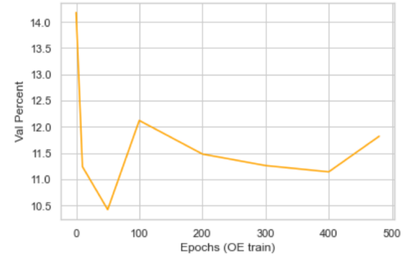


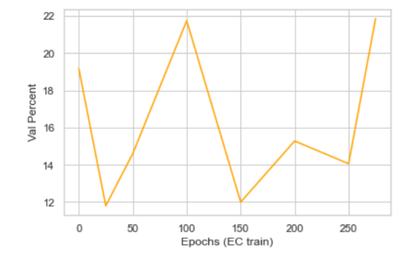
After 600 epochs





After 100 epochs





Further Directions

Direction	Current Deficit	Solution
Seeding dependence	High variance in low-dim. exp.	More seeds with more compute
More Labels	CIFAR 10 Restriction	CIFAR100/Mini-ImageNet
Geometry	Euclidean restrictive for hierarchy	Extension Survey in Paper
Probabilistic Embeddings	Space "Crowding"	Extension Survey in Paper
Metric Learning on labels	Require access to labels	Learn from label space