
CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP

Andreas Fürst ^{*1} Elisabeth Rumetshofer ^{*1} Johannes Lehner ¹ Viet Tran ¹
Fei Tang ³ Hubert Ramsauer ¹ David Kreil ² Michael Kopp ²
Günter Klambauer ¹ Angela Bitto-Nemling ¹ Sepp Hochreiter ^{1 2}

¹ ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University, Linz, Austria

² Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria

³ HERE Technologies, Zurich, Switzerland

* Equal contribution

Abstract

CLIP yielded impressive results on zero-shot transfer learning tasks and is considered as a foundation model like BERT or GPT3. CLIP vision models that have a rich representation are pre-trained using the InfoNCE objective and natural language supervision before they are fine-tuned on particular tasks. Though CLIP excels at zero-shot transfer learning, it suffers from an explaining away problem, that is, it focuses on one or few features, while neglecting other relevant features. This problem is caused by insufficiently extracting the covariance structure in the original multi-modal data. We suggest to use modern Hopfield networks to tackle the problem of explaining away. Their retrieved embeddings have an enriched covariance structure derived from co-occurrences of features in the stored embeddings. However, modern Hopfield networks increase the saturation effect of the InfoNCE objective which hampers learning. We propose to use the InfoLOOB objective to mitigate this saturation effect. We introduce the novel “Contrastive Leave One Out Boost” (CLOOB), which uses modern Hopfield networks for covariance enrichment together with the InfoLOOB objective. In experiments we compare CLOOB to CLIP after pre-training on the Conceptual Captions and the YFCC dataset with respect to their zero-shot transfer learning performance on other datasets. CLOOB consistently outperforms CLIP at zero-shot transfer learning across all considered architectures and datasets.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) showed spectacular performance at zero-shot transfer learning (Radford et al., 2021). CLIP learns expressive image embeddings directly from raw text, thereby leverages a much richer source of supervision than just labels. The CLIP model is considered as an important foundation model (Bommasani et al., 2021), therefore a plethora of follow-up work has been published (see Appendix Section A.4). CLIP as a contrastive learning method has two simultaneous goals, namely (i) increasing the similarity of matched language-image pairs and (ii) decreasing the similarity of unmatched language-image pairs. Though CLIP yielded

Code is available at: <https://github.com/ml-jku/cloob>

striking zero-shot transfer learning results, it still suffers from “explaining away”. Explaining away is known in reasoning as the concept that the confirmation of one cause of an observed event dismisses alternative causes (Pearl, 1988; Wellman & Henrion, 1993). CLIP’s explaining away problem is its focus on one or few features while neglecting other relevant features. This problem is caused by insufficiently extracting feature co-occurrences and covariance structures in the original multi-modal data. Humans extract co-occurrences and covariances by associating current perceptions with memories (Bonner & Epstein, 2021; Potter, 2012). In analogy to these human cognitive processes, we suggest to use modern Hopfield networks to amplify co-occurrences and covariance structures of the original data.

Hopfield networks are energy-based, binary associative memories, which popularized artificial neural networks in the 1980s (Amari, 1972; Hopfield, 1982, 1984). Associative memory networks have been designed to store and retrieve samples. Their storage capacity can be considerably increased by polynomial terms in the energy function (Chen et al., 1986; Psaltis & Cheol, 1986; Baldi & Venkatesh, 1987; Gardner, 1987; Abbott & Arian, 1987; Horn & Usher, 1988; Caputo & Niemann, 2002; Krotov & Hopfield, 2016). In contrast to these binary memory networks, we use continuous associative memory networks with very high storage capacity. These modern Hopfield networks for deep learning architectures have an energy function with continuous states and can retrieve samples with only one update (Ramsauer et al., 2021). Modern Hopfield networks have already been successfully applied to immune repertoire classification (Widrich et al., 2020), chemical reaction prediction (Seidl et al., 2022) and reinforcement learning (Widrich et al., 2021; Paischer et al., 2022). Modern Hopfield networks are a novel concept for contrastive learning to extract more covariance structure.

However, modern Hopfield networks lead to a higher similarity of retrieved samples. The increased similarity exacerbates the saturation of CLIP’s InfoNCE objective (van den Oord et al., 2018). InfoNCE saturates because it contains terms of the form $a/(a + b)$. In analogy to Wang & Isola (2020), a is called the “alignment score” that measures the similarity of matched pairs and b is called the “uniformity penalty” that measures the similarity of unmatched pairs. The saturation problem becomes more severe for retrieved samples of the modern Hopfield network since the alignment score a increases. Saturation of InfoNCE hampers the decrease of the uniformity penalty b (see also Yeh et al. (2021)). Contrary to InfoNCE, the “InfoLOOB” (LOOB for “Leave One Out Bound”) objective (Poole et al., 2019) contains only terms of the form a/b which do not saturate. Thus, even for a large alignment score a , learning still decreases the uniformity penalty b by distributing samples more uniformly.

We introduce “Contrastive Leave One Out Boost” (CLOOB) which combines modern Hopfield networks with the “InfoLOOB” objective. Our contributions are:

- (a) we propose CLOOB, a new contrastive learning method,
- (b) we propose modern Hopfield networks to reinforce covariance structures,
- (c) we propose InfoLOOB as an objective to avoid saturation as observed with InfoNCE, and provide theoretical justifications for optimizing InfoLOOB.

2 CLOOB: Modern Hopfield Networks with InfoLOOB

Our novel contrastive learning method CLOOB can be seen as a replacement of CLIP and therefore be used in any method which builds upon CLIP. Figure 1 sketches the CLOOB architecture for image-text pairs. The training set consists of N pairs of embeddings $\{(x_1, y_1), \dots, (x_N, y_N)\}$ with $\mathbf{X} = (x_1, \dots, x_N)$ and $\mathbf{Y} = (y_1, \dots, y_N)$, M stored embeddings $\mathbf{U} = (u_1, \dots, u_M)$, and K stored embeddings $\mathbf{V} = (v_1, \dots, v_K)$. The state or query embeddings x_i and y_i retrieve U_{x_i} and U_{y_i} , respectively, from \mathbf{U} — analog for retrievals from \mathbf{V} . The samples are normalized: $\|x_i\| = \|y_i\| = \|u_i\| = \|v_i\| = 1$. U_{x_i} denotes an image-retrieved image embedding, U_{y_i} a text-retrieved image embedding, V_{x_i} an image-retrieved text embedding and V_{y_i} a text-retrieved text embedding. These retrievals from modern Hopfield networks are computed as follows (Ramsauer et al., 2021):

$$U_{x_i} = \mathbf{U} \operatorname{softmax}(\beta \mathbf{U}^T x_i), \quad (1) \quad V_{x_i} = \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T x_i), \quad (3)$$

$$U_{y_i} = \mathbf{U} \operatorname{softmax}(\beta \mathbf{U}^T y_i), \quad (2) \quad V_{y_i} = \mathbf{V} \operatorname{softmax}(\beta \mathbf{V}^T y_i). \quad (4)$$

The hyperparameter β corresponds to the inverse temperature: $\beta = 0$ retrieves the average of the stored pattern, while large β retrieves the stored pattern that is most similar to the state pattern (query).

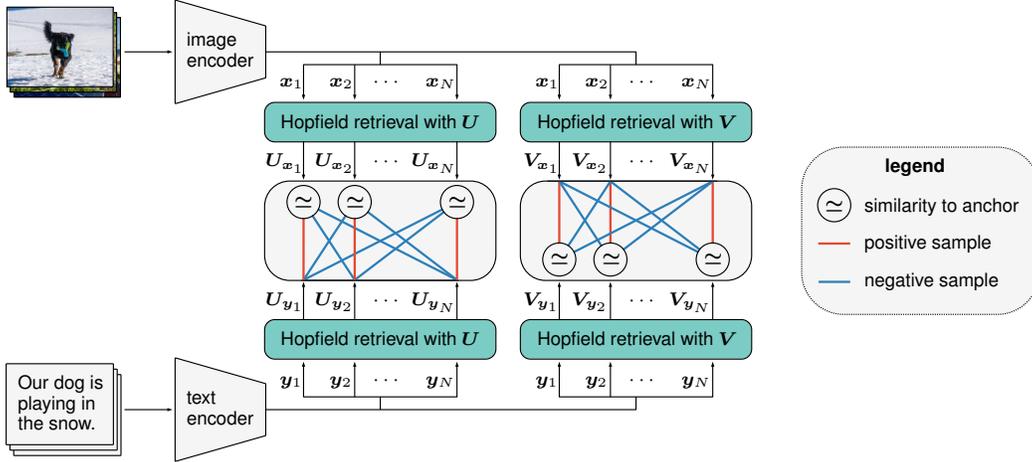


Figure 1: The CLOOB architecture for image-text pairs. The image embedding x_i and the text embedding y_i retrieve the embeddings U_{x_i} and U_{y_i} , respectively, from a modern Hopfield network that stores image embeddings $U = (u_1, \dots, u_M)$ (green boxes at the left). The image-retrieved image embedding U_{x_i} serves as anchor in order to contrast the positive text-retrieved image embedding U_{y_i} with the negative text-retrieved image embedding U_{y_j} for $j \neq i$. Analogously, for the second modern Hopfield network that stores text embeddings $V = (v_1, \dots, v_K)$ (green boxes at the right).

In the InfoLOOB loss Eq. (8), CLOOB substitutes the embedded samples x_i and y_i by the normalized retrieved embedded samples. In the first term, x_i and y_i are substituted by U_{x_i} and U_{y_i} , respectively, while in the second term they are substituted by V_{x_i} and V_{y_i} . After retrieval, the samples are re-normalized to ensure $\|U_{x_i}\| = \|U_{y_i}\| = \|V_{x_i}\| = \|V_{y_i}\| = 1$.

We obtain the CLOOB loss function:

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} U_{x_i}^T U_{y_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} U_{x_i}^T U_{y_j})} - \frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} V_{x_i}^T V_{y_i})}{\sum_{j \neq i}^N \exp(\tau^{-1} V_{x_i}^T V_{y_j})}. \quad (5)$$

By default, we store the minibatch in the modern Hopfield networks, that is, $U = X$ and $V = Y$. Thus, in Eq. (1) x_i can retrieve itself from $U = X$, but in Eq. (3) it can not retrieve itself from $V = Y$. Analogously, in Eq. (4) y_i can retrieve itself from $V = Y$, but in Eq. (2) it can not retrieve itself from $U = X$. By storing the embeddings of the mini-batch examples in the Hopfield memory, we do not require to compute the embeddings of additional samples via text and image encoders. However, the modern Hopfield networks can also store prototypes, templates, or proprietary samples to amplify particular embedding features via the stored samples. Either the original embeddings x and y or the retrieved embeddings U_x , U_y , V_x , and V_y may serve for the downstream tasks, e.g. for zero-shot transfer learning.

Pseudocode 1 shows CLOOB in a PyTorch-like style. CLOOB has two major components: (i) modern Hopfield networks that alleviate CLIP’s problem of insufficiently exploiting the covariance structure in the data and (ii) the InfoLOOB objective that does not suffer from InfoNCE’s saturation problem. The next two sections analyze CLOOB’s major components.

3 Modern Hopfield Networks for Enriching the Covariance Structure

We use modern Hopfield networks to amplify co-occurrences and the covariance structure. Replacing the original embeddings by retrieved embeddings reinforces features that frequently occur together in stored embeddings. Additionally, spurious co-occurrences that are peculiar to a sample are averaged out. By this means, the covariance structure is reinforced by the retrieved embeddings $U_{x_i}^T U_{y_i}$ and $V_{x_i}^T V_{y_i}$. The Jacobian J of the softmax $p = \text{softmax}(\beta a)$ is $J(\beta a) = \beta (\text{diag}(p) - pp^T)$. We define the *weighted covariance* $\text{Cov}(U)$, where sample u_i is drawn with probability p_i , as $[\text{Cov}(U)]_{kl} = [UJ(\beta a)U^T]_{kl} = \beta (\sum_{i=1}^M p_i u_{ik} u_{il} - \sum_{i=1}^M p_i u_{ik} \sum_{i=1}^M p_i u_{il})$. The formula of the

Pseudocode 1 CLOOB in a PyTorch-like style.

```

1 # image_encoder - ResNet
2 # text_encoder - Text Transformer
3 # I[n, h, w, c] - minibatch of images
4 # T[n, l] - minibatch of texts
5 # W_i[d_i, d_e] - image projection
6 # W_t[d_t, d_e] - text projection
7 # beta - inverse temperature Hopfield retrieval
8 # tau - temperature InfoLOOB
9
10 # extract feature representations
11 I_f = image_encoder(I) #[n, d_i]
12 T_f = text_encoder(T) #[n, d_t]
13
14 # joint multimodal embedding
15 x = l2_normalize(I_f @ W_i) #[n, d_e]
16 y = l2_normalize(T_f @ W_t) #[n, d_e]
17
18 # Hopfield retrieval H with batch stored
19 # H(beta, A, B) = B.T @ softmax(beta * A @ B.T)
20 U_x = H(beta, x, x).T #[n, d_e]
21 U_y = H(beta, y, y).T #[n, d_e]
22 V_x = H(beta, x, y).T #[n, d_e]
23 V_y = H(beta, y, x).T #[n, d_e]
24
25 # normalize retrievals
26 U_x = l2_normalize(U_x) #[n, d_e]
27 U_y = l2_normalize(U_y) #[n, d_e]
28 V_x = l2_normalize(V_x) #[n, d_e]
29 V_y = l2_normalize(V_y) #[n, d_e]
30
31 # loss: info_loob(tau, anchors, samples)
32 # samples contain pos. and neg. embeddings
33 loss_i = info_loob(tau, U_x, U_y)
34 loss_t = info_loob(tau, V_y, V_x)
35 loss = (loss_i + loss_t) * tau

```

weighted covariance differs from the standard empirical covariance, since the factor $1/M$ is replaced by p_i . Thus, \mathbf{u}_i is sampled with probability p_i instead with probability $1/M$ (uniformly).

We apply the mean value theorem to the softmax function with mean Jacobian matrix $\mathbf{J}^m(\beta\mathbf{a}) = \int_0^1 \mathbf{J}(\lambda\beta\mathbf{a}) d\lambda$. The mean Jacobian $\mathbf{J}^m(\beta\mathbf{a})$ is a symmetric, diagonally dominant, positive semi-definite matrix with one eigenvalue of zero for eigenvector $\mathbf{1}$ and spectral norm bounded by $\|\mathbf{J}^m\|_2 \leq 0.5\beta$ (see Appendix Lemma A1). According to Appendix Theorem A3, we can express $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ as:

$$(\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{x}_i) \mathbf{x}_i)^T (\bar{\mathbf{u}} + \text{Cov}(\mathbf{U}, \mathbf{y}_i) \mathbf{y}_i), \quad (6)$$

where the mean is $\bar{\mathbf{u}} = 1/M \mathbf{U} \mathbf{1}$ and the weighted covariances are $\text{Cov}(\mathbf{U}, \mathbf{x}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{x}_i) \mathbf{U}^T$ and $\text{Cov}(\mathbf{U}, \mathbf{y}_i) = \mathbf{U} \mathbf{J}^m(\beta \mathbf{U}^T \mathbf{y}_i) \mathbf{U}^T$. The weighted covariance $\text{Cov}(\mathbf{U}, \cdot)$ is the covariance if the stored pattern \mathbf{u}_i is drawn according to an averaged p_i given by $\mathbf{J}^m(\cdot)$. Maximizing the dot product $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ forces the normalized vectors \mathbf{x}_i and \mathbf{y}_i to agree on drawing the patterns \mathbf{u}_i with the same probability p_i in order to generate similar weighted covariance matrices $\text{Cov}(\mathbf{U}, \cdot)$. If subsets of \mathbf{U} have a strong covariance structure, then it can be exploited to produce large weighted covariances and, in turn, large dot products of $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$. Furthermore, for a large dot product $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$, \mathbf{x}_i and \mathbf{y}_i have to be similar to each other to extract the same direction from the covariance matrices. The above considerations for $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ analogously apply to $\mathbf{V}_{x_i}^T \mathbf{V}_{y_i}$.

We did not use a loss function that contains dot products like $\mathbf{U}_{x_i}^T \mathbf{V}_{y_i}$, because they have higher variance than the ones we have used. The dot product $\mathbf{U}_{x_i}^T \mathbf{V}_{y_i}$ has higher variance, since it uses $M + K$ stored patterns, whereas $\mathbf{U}_{x_i}^T \mathbf{U}_{y_i}$ and $\mathbf{V}_{x_i}^T \mathbf{V}_{y_i}$ use M and K stored patterns, respectively.

Modern Hopfield networks enable to extract more covariance structure. To demonstrate the effect of modern Hopfield networks, we computed the eigenvalues of the covariance matrix of the image and text embeddings. We counted the number of effective eigenvalues, that is, the number of eigenvalues needed to obtain 99% of the total sum of eigenvalues. Figure 2 shows the relative change of the number of effective eigenvalues compared to the respective reference epoch (the epoch before the first learning rate restart). Modern Hopfield networks consistently increase the number of effective eigenvalues during learning. Consequently, modern Hopfield networks enable to extract more covariance structure during learning, i.e. enrich the embeddings by covariances that are already in the raw multi-modal data. This enrichment of embeddings mitigates explaining away. Further details can be found in Appendix Section A.2.7.

4 InfoLOOB for Contrastive Learning

Modern Hopfield networks lead to a higher similarity of retrieved samples. The increased similarity exacerbates the saturation of the InfoNCE objective. To avoid the saturation of InfoNCE, CLOOB uses the ‘‘InfoLOOB’’ objective. The ‘‘InfoLOOB’’ objective is called ‘‘Leave one out upper bound’’ in Poole et al. (2019) and ‘‘L1Out’’ in Cheng et al. (2020). InfoLOOB is not established as a contrastive

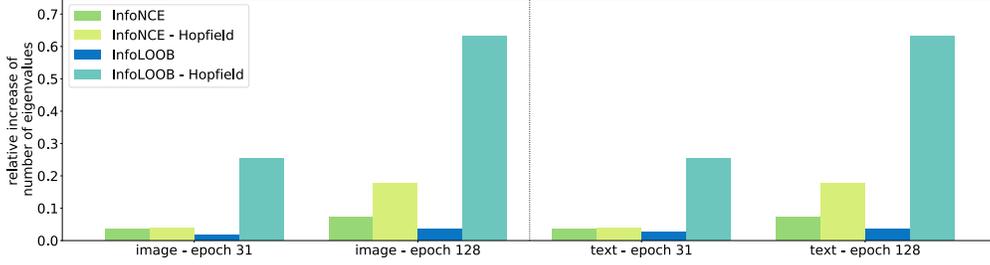


Figure 2: Relative change in the number of the effective eigenvalues of the embedding covariance matrices, which were obtained from image and text encoders at two different training points. Models with modern Hopfield networks steadily extract more covariance structure during learning.

objective, although it is a known bound. Recently, InfoLOOB was independently introduced as objective for image-to-image contrastive learning (Yeh et al., 2021).

InfoNCE and InfoLOOB loss functions. N samples are drawn iid from $p(\mathbf{x}, \mathbf{y})$ giving the training set $\{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$. For the sample \mathbf{y}_1 , InfoNCE uses for the matrix of negative samples $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, while InfoLOOB uses $\tilde{\mathbf{X}} = (\mathbf{x}_2, \dots, \mathbf{x}_N)$. The matrices differ by the positive sample \mathbf{x}_1 . For the score function $f(\mathbf{x}, \mathbf{y})$, we use $f(\mathbf{x}, \mathbf{y}) = \exp(\tau^{-1} \text{sim}(\mathbf{x}, \mathbf{y}))$ with the similarity $\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T \mathbf{x}$ and τ as the temperature. We have the InfoNCE and InfoLOOB loss functions:

$$L_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)} - \frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}, \quad (7)$$

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)} - \frac{1}{N} \sum_{i=1}^N \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}. \quad (8)$$

We abbreviate $\mathbf{y} = \mathbf{y}_1$ leading to the pair $(\mathbf{x}_1, \mathbf{y})$ and the negatives $\tilde{\mathbf{X}} = (\mathbf{x}_2, \dots, \mathbf{x}_N)$. In the second sum of the losses in Eq. 7 and Eq. 8, we consider only the first term, respectively:

$$L_{\text{InfoNCE}}(\mathbf{y}) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y})}^a}{\underbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y})}_a + \underbrace{\sum_{j=2}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y})}_b}, \quad (9)$$

$$L_{\text{InfoLOOB}}(\mathbf{y}) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y})}^a}{\underbrace{\sum_{j=2}^N \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y})}_b}. \quad (10)$$

In analogy to Wang & Isola (2020), a is called the ‘‘alignment score’’ that measures the similarity of matched pairs and b the ‘‘uniformity penalty’’ that measures the similarity of unmatched pairs.

Gradients of InfoNCE and InfoLOOB loss functions. Eq. (9) and Eq. (10) are equal to

$$-\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \text{lse}(\tau^{-1}, \mathbf{X}^T \mathbf{y}), \quad -\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \text{lse}(\tau^{-1}, \tilde{\mathbf{X}}^T \mathbf{y}),$$

where lse is the log-sum-exp function (see Eq. (A73) in the Appendix).

The gradients of Eq. (9) and Eq. (10) with respect to \mathbf{y} are

$$-\tau^{-1} \mathbf{x}_1 + \tau^{-1} \mathbf{X} \text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y}), \quad -\tau^{-1} \mathbf{x}_1 + \tau^{-1} \tilde{\mathbf{X}} \text{softmax}(\tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y}).$$

Using $\mathbf{p} = (p_1, \dots, p_N)^T = \text{softmax}(\tau^{-1} \mathbf{X}^T \mathbf{y})$, the gradient of InfoNCE with respect to \mathbf{y} is

$$\frac{\partial L_{\text{InfoNCE}}(\mathbf{y})}{\partial \mathbf{y}} = -\tau^{-1} (1 - p_1) (\mathbf{x}_1 - \tilde{\mathbf{X}} \text{softmax}(\tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y})) = (1 - p_1) \frac{\partial L_{\text{InfoLOOB}}(\mathbf{y})}{\partial \mathbf{y}}.$$

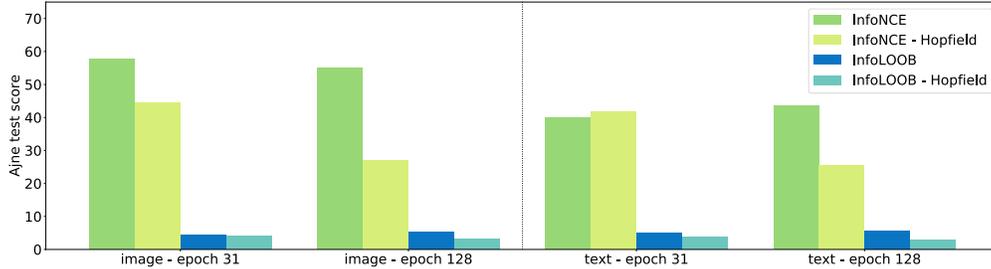


Figure 3: Ajne uniformity test statistics for image and text embeddings for two different epochs during training. A high test statistic indicates low uniformity of an embedding. Models trained with the InfoLOOB objective develop more uniform image and text embeddings on the hypersphere.

By and large, the gradient of InfoNCE is scaled by $(1 - p_1)$ compared to the gradient of InfoLOOB, where p_1 is the softmax similarity between the anchor \mathbf{y} and the positive sample \mathbf{x}_1 . Consequently, InfoNCE is saturating with increasing similarity between the anchor and the positive sample. For more details we refer to Appendix Section A.1.4.

This saturation of InfoNCE motivated the use of the InfoLOOB objective in order to decrease the uniformity penalty even for large alignment scores. The uniformity penalty decreases since learning does not stall and the most prominent features become down-scaled which makes negative examples less similar to the anchor sample. The InfoNCE objective Eq. 9 has the form $a/(a + b)$, while the InfoLOOB objective Eq. 10 has the form a/b . InfoLOOB does not saturate and keeps decreasing the uniformity penalty b . Figure 3 shows how InfoLOOB leads to an increase in the uniformity of image and text embeddings on the sphere, which is described by the statistics of the uniformity test of Ajne that was extended by Prentice (Ajne, 1968; Prentice, 1978). Higher uniformity on the sphere correlates with a lower uniformity penalty b . For more details we refer to Appendix Section A.2.7.

Theoretical justification for optimizing InfoLOOB. The InfoNCE information is a lower bound on the mutual information, which was proven by Poole et al. (2019). In the Appendix Section A.1, we elaborate more on theoretical properties of the bounds and properties of the objective functions. Specifically, we show that InfoLOOB with neural networks is not an upper bound on the mutual information. Thus, unlike hitherto approaches to contrastive learning we use InfoLOOB as an objective, since it does not suffer from saturation effects as InfoNCE.

5 Experiments

CLOOB is compared to CLIP with respect to zero-shot transfer learning performance on two pre-training datasets. The first dataset, Conceptual Captions (CC) (Sharma et al., 2018), has a very rich textual description of images but only three million image-text pairs. The second dataset, a subset of YFCC100M (Thomee et al., 2016), has 15 million image-text pairs but the textual description is less rich than for CC and often vacuous. For both pre-training datasets, the downstream zero-shot transfer learning performance is tested on seven image classification datasets.

5.1 Conceptual Captions Pre-training

Pre-training dataset. The Conceptual Captions (CC) (Sharma et al., 2018) dataset contains 2.9 million images with high-quality captions. Images and their captions have been gathered from the web via an automated process and have a wide variety of content. Raw descriptions of images are from the *alt-text* HTML attribute.

Methods. The CLOOB implementation is based on OpenCLIP (Ilharco et al., 2021), which achieves results equivalent to CLIP on the YFCC dataset (see Section 5.2). OpenCLIP also reports results on the CC dataset. As CLIP does not train models on CC, we report results from this reimplementation as baseline. Analogously to Radford et al. (2021, Section 2.4), we used the modified ResNet (He et al., 2016) and BERT (Devlin et al., 2018, 2019) architectures to encode image and text input. We used the ResNet encoder ResNet-50 for experiments on CC.

Table 1: Zero-shot results for models trained on CC with ResNet-50 vision encoders for two different checkpoints. Results are given as mean accuracy over 5 runs. Statistically significant results are shown in bold. CLIP and CLOOB were trained for 31 epochs while CLIP* and CLOOB* were trained for 128 epochs. In the majority of tasks CLOOB significantly outperforms CLIP.

Dataset	CLIP RN-50	CLOOB RN-50	CLIP* RN-50	CLOOB* RN-50
Birdsnap	2.26 ± 0.20	3.06 ± 0.30	2.8 ± 0.16	3.24 ± 0.31
Country211	0.67 ± 0.11	0.67 ± 0.05	0.7 ± 0.04	0.73 ± 0.05
Flowers102	12.56 ± 0.38	13.45 ± 1.19	13.32 ± 0.43	14.36 ± 1.17
GTSRB	7.66 ± 1.07	6.38 ± 2.11	8.96 ± 1.70	7.03 ± 1.22
UCF101	20.98 ± 1.55	22.26 ± 0.72	21.63 ± 0.65	23.03 ± 0.85
Stanford Cars	0.91 ± 0.10	1.23 ± 0.10	0.99 ± 0.16	1.41 ± 0.32
ImageNet	20.33 ± 0.28	23.97 ± 0.15	21.3 ± 0.42	25.67 ± 0.22
ImageNet V2	20.24 ± 0.50	23.59 ± 0.15	21.24 ± 0.22	25.49 ± 0.11

Hyperparameter selection and learning schedule. The hyperparameter values of OpenCLIP were used as default, concretely, a learning rate of 1×10^{-3} and a weight decay of 0.1 for the Adam optimizer (Kingma et al., 2014) with decoupled weight decay regularization (Loshchilov & Hutter, 2019). Deviating from OpenCLIP, we used a batch size of 512 due to computational restraints, which did not change the performance. The learning rate scheduler for all experiments was cosine annealing with warmup and hard restarts (Loshchilov & Hutter, 2017). We report the hyperparameter τ (default 0.07) from CLIP as τ^{-1} of 14.3 to be in the same regime as the hyperparameter β for the modern Hopfield networks. The main hyperparameter search for CLOOB (also for YFCC pre-training in the next section) was done with ResNet-50 as the vision encoder. Learnable τ^{-1} in combination with the InfoLOOB loss results in undesired learning behavior (see Appendix Section A.1.4). Therefore, we set τ^{-1} to a fixed value of 30, which was determined via a hyperparameter search (see Appendix Section A.2.2). For modern Hopfield networks, the hyperparameter β was set to 8. Further we scaled the loss L_{InfoLOOB} with τ to remove the factor τ^{-1} from the gradients (see Appendix Section A.1.4) resulting in the loss function τL_{InfoLOOB} .

Evaluation metrics: Zero-shot transfer learning. We evaluated and compared both CLIP and CLOOB on their zero-shot transfer learning capabilities on the following downstream image classification tasks. Birdsnap (Berg et al., 2014) contains images of 500 different North American bird species. The Country211 (Radford et al., 2021) dataset consists of photos across 211 countries and is designed to test the geolocalization capability of visual representations. Flowers102 (Nilsback & Zisserman, 2008) is a dataset containing images of 102 flower species. GTSRB (Stallkamp et al., 2011) contains images for classification of German traffic signs. UCF101 (Soomro et al., 2012) is a video dataset with short clips for action recognition. For UCF101 we followed the procedure reported in CLIP and extract the middle frame of every video to assemble the dataset. Stanford Cars (Krause et al., 2013) contains images of 196 types of cars. ImageNet (Deng et al., 2009) is a large scale image classification dataset with images across 1,000 classes. ImageNetv2 (Recht et al., 2019) consists of three new test sets with 10,000 images each for the ImageNet benchmark. For further details see Appendix Section A.2.3.

Results. We employed the same evaluation strategy and used the prompts as published in CLIP (see Appendix Section A.2.3). We report zero-shot results from two checkpoints in Table 1. CLIP and CLOOB were trained for a comparable number of epochs used in CLIP (see Appendix Section A.2.2) while CLIP* and CLOOB* were trained until evaluation performance plateaued (epoch 128). In both cases CLOOB significantly outperforms CLIP on the majority of tasks or matches its performance. Statistical significance of these results was assessed by an unpaired Wilcoxon test on a 5% level.

5.2 YFCC Pre-training

Pre-training dataset. To be comparable to the CLIP results, we used the same subset of 15 million samples from the YFCC100M dataset (Thomee et al., 2016) as in Radford et al. (2021), which we refer to as YFCC. YFCC was created by filtering YFCC100M for images which contain natural language descriptions and/or titles in English. It was not filtered by quality of the captions, therefore the textual descriptions are less rich and contain superfluous information. The dataset with 400

million samples used to train the CLIP models in Radford et al. (2021) has not been released and, thus, is not available for comparison. Due to limited computational resources we were unable to compare CLOOB to CLIP on other datasets of this size.

Methods. Besides experiments with a ResNet-50 image encoder, we additionally conducted experiments with the larger ResNet variants ResNet-101 and ResNet-50x4. In addition to the comparison of CLOOB and CLIP based on the OpenCLIP reimplementation (Ilharco et al., 2021), we include the original CLIP results (Radford et al., 2021, Table 12).

Hyperparameter selection. Hyperparameters were the same as for the Conceptual Captions dataset, except learning rate, batch size, and β . For modern Hopfield networks, the hyperparameter β was set to 14.3, which is default for τ^{-1} in Radford et al. (2021). Furthermore, the learning rate was set to 5×10^{-4} and the batch size to 1024 as used in OpenCLIP of Ilharco et al. (2021). All models were trained for 28 epochs. For further details see Appendix Section A.2.2.

Evaluation metrics. As in the previous experiment, methods were again evaluated at their zero-shot transfer learning capabilities on downstream tasks.

Table 2: Results of CLIP and CLOOB trained on YFCC with ResNet-50 encoder. Except for one linear probing dataset, CLOOB consistently outperforms CLIP at all tasks.

Dataset	Linear Probing		Zero-Shot	
	CLIP (OpenAI)	CLOOB (ours)	CLIP (OpenAI)	CLOOB (ours)
Birdsnap	47.4	56.2	19.9	28.9
Country211	23.1	20.6	5.2	7.9
Flowers102	94.4	96.1	48.6	55.1
GTSRB	66.8	78.9	6.9	8.1
UCF101	69.2	72.3	22.9	25.3
Stanford Cars	31.4	37.7	3.8	4.1
ImageNet	62.0	65.7	31.3	35.7
ImageNet V2	-	58.7	-	34.6

Table 3: Zero-shot results for the CLIP reimplementation and CLOOB using different ResNet architectures trained on YFCC. CLOOB outperforms CLIP in 7 out of 8 tasks using ResNet-50 encoders. With larger ResNet encoders CLOOB outperforms CLIP on all tasks. The performance of CLOOB scales with increased encoder size.

Dataset	RN-50		RN-101		RN-50x4	
	CLIP	CLOOB	CLIP	CLOOB	CLIP	CLOOB
Birdsnap	21.8	28.9	22.6	30.3	20.8	32.0
Country211	6.9	7.9	7.8	8.5	8.1	9.3
Flowers102	48.0	55.1	48.0	55.3	50.1	54.3
GTSRB	7.9	8.1	7.4	11.6	9.4	11.8
UCF101	27.2	25.3	28.6	28.8	31.0	31.9
Stanford Cars	3.7	4.1	3.8	5.5	3.5	6.1
ImageNet	34.6	35.7	35.3	37.1	37.7	39.0
ImageNet V2	33.4	34.6	34.1	35.6	35.9	37.3

Results. Table 2 provides results of the original CLIP and CLOOB trained on YFCC. Results on zero-shot downstream tasks show that CLOOB outperforms CLIP on all 7 tasks (ImageNet V2 results have not been reported in Radford et al. (2021)). Similarly, CLOOB outperforms CLIP on 6 out of 7 tasks for linear probing. Results of CLOOB and the CLIP reimplementation of OpenCLIP are given in Table 3. CLOOB exceeds the CLIP reimplementation in 7 out of 8 tasks for zero-shot classification using ResNet-50 encoders. With larger ResNet encoders, CLOOB outperforms CLIP on all tasks. Furthermore, the experiments with larger vision encoder networks show that CLOOB performance increases with network size. Results of CLOOB zero-shot classification on all datasets are shown in Appendix Section A.2.4.

5.3 Ablation studies

CLOOB has two new major components compared to CLIP: (1) modern Hopfield networks and (2) the InfoLOOB objective instead of the InfoNCE objective. To assess effects of the new major components of CLOOB, we performed ablation studies.

Modern Hopfield networks. Modern Hopfield networks amplify the covariance structure in the data via the retrievals. Ablation studies confirm this amplification as modern Hopfield networks consistently increase the number of effective eigenvalues of the embedding covariance matrices during learning. Figure 2 shows the relative change of the number of effective eigenvalues compared to the respective reference epoch, which is the epoch before the first learning rate restart. These results indicate that modern Hopfield networks steadily extract more covariance structure during learning. Modern Hopfield networks induce higher similarity of retrieved samples, which in turn leads to stronger saturation of the InfoNCE objective. As a result, we observe low uniformity (see Figure 3) and a small number of effective eigenvalues (see Appendix Figure A1).

Modern Hopfield networks with InfoLOOB. CLOOB counters the saturation of InfoNCE by using the InfoLOOB objective. The effectiveness of InfoLOOB manifests in an increased uniformity measure of image and text embeddings on the sphere, as shown in Figure 3. The ablation study verifies that modern Hopfield networks together with InfoLOOB have a strong synergistic effect.

InfoLOOB. However, using solely InfoLOOB results in overfitting of the alignment score. This overfitting occasionally leads to high similarities of unmatched pairs (see Figure 4), which may decrease the zero-shot downstream performance. The reason for this is that the top-1 evaluation metric is very sensitive to occasionally high similarities of prompts of the incorrect class. Yeh et al. (2021) and Zhang et al. (2022) reported similar observations of overfitting.

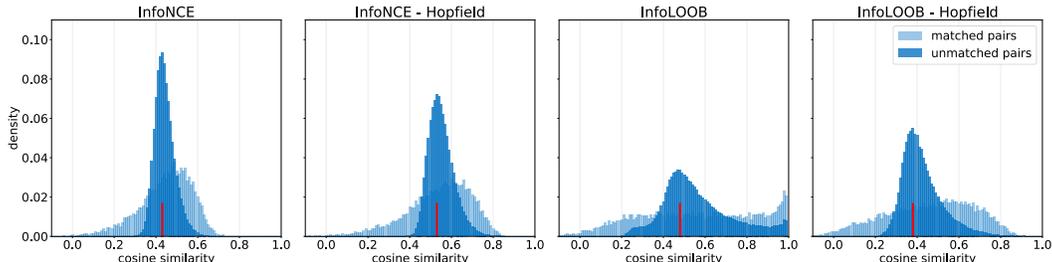


Figure 4: Distribution of the cosine similarity of matched pairs and the cosine similarity of the 10 unmatched pairs that have the highest similarity score with the anchor. Modern Hopfield networks lead to higher values of both matched and unmatched pairs. InfoLOOB without Hopfield has high similarity scores of the matched pairs which correlate with high similarity scores of the top-10 unmatched pairs. In contrast, InfoLOOB with Hopfield does not suffer from this overfitting problem.

CLOOB balances the overfitting of InfoLOOB with the underfitting of modern Hopfield networks and remains in effective learning regimes. For more details and further ablation studies see Appendix Section A.2.1.

6 Conclusion

We have introduced “Contrastive Leave One Out Boost” (CLOOB), which combines modern Hopfield networks with the InfoLOOB objective. Modern Hopfield networks enable CLOOB to extract additional covariance structure in the data. This allows for building more relevant features in the embedding space, mitigating the explaining away problem. We show that InfoLOOB avoids the saturation problem of InfoNCE. Additionally, we theoretically justify the use of the InfoLOOB objective for contrastive learning and suggest it as an alternative to InfoNCE. At seven zero-shot transfer learning tasks, the novel CLOOB was compared to CLIP after pre-training on the Conceptual Captions and the YFCC dataset. CLOOB consistently outperforms CLIP at zero-shot transfer learning across all considered architectures and datasets.

Acknowledgments

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. IARAI is supported by Here Technologies. We thank the projects AIMOTION (LIT-2018-6-YOU-212), AI-SNN (LIT-2018-6-YOU-214), DeepFlood (LIT-2019-8-YOU-213), Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), AIRI FG 9-N (FWF-36284, FWF-36235), ELISE (H2020-ICT-2019-3 ID: 951847). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, Software Competence Center Hagenberg GmbH, TÜV Austria, Frauscher Sonsonic and the NVIDIA Corporation.

References

- Abbott, L. F. and Arian, Y. Storage capacity of generalized networks. *Physical Review A*, 36: 5091–5094, 1987. doi: 10.1103/PhysRevA.36.5091.
- Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. Evaluating CLIP: Towards characterization of broader capabilities and downstream implications. *ArXiv*, 2108.02818, 2021.
- Ajne, B. A simple test for uniformity of a circular distribution. *Biometrika*, 55(2):343–354, 1968. doi: 10.1093/biomet/55.2.343.
- Amari, S.-I. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.
- Arbel, J., Marchal, O., and Nguyen, H. D. On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ArXiv*, 1901.09188, 2019.
- Baldi, P. and Venkatesh, S. S. Number of stable points for spin-glasses and neural networks of higher orders. *Physical Review Letters*, 58:913–916, 1987. doi: 10.1103/PhysRevLett.58.913.
- Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., and Torralba, A. Paint by word. *ArXiv*, 2103.10951, 2021.
- Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of International Conference on Machine Learning (ICML)*, pp. 531–540, 2018.
- Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W., and Belhumeur, P. N. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2019–2026, 2014. doi: 10.1109/CVPR.2014.259.
- Biewald, L. Experiment tracking with weights and biases, 2020. URL <https://www.wandb.com/>. Software available from wandb.com.
- Bommasani, R. et al. On the opportunities and risks of foundation models. *ArXiv*, 2108.07258, 2021.
- Bonner, M. F. and Epstein, R. A. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(4081), 2021. doi: 10.1038/s41467-021-24368-2.
- Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- Cai, Q., Wang, Y., Pan, Y., Yao, T., and Mei, T. Joint contrastive learning with infinite possibilities. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12638–12648, 2020.
- Caputo, B. and Niemann, H. Storage capacity of kernel associative memories. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pp. 51–56, Berlin, Heidelberg, 2002. Springer-Verlag.
- Carlini, N. and Terzis, A. Poisoning and backdooring contrastive learning. *ArXiv*, 2106.09667, 2021.
- Chen, H. H., Lee, Y. C., Sun, G. Z., Lee, H. Y., Maxwell, T., and Giles, C. L. High order correlation model for associative memory. *AIP Conference Proceedings*, 151(1):86–99, 1986. doi: 10.1063/1.36224.
- Chen, J., Gan, Z., Li, X., Guo, Q., Chen, L., Gao, S., Chung, T., Xu, Y., Zeng, B., Lu, W., Li, F., Carin, L., and Tao, C. Simpler, faster, stronger: Breaking the log-K curse on contrastive learners with FlatNCE. *arXiv*, 2107.01152, 2021.
- Chen, T., Sun, Y., Shi, Y., and Hong, L. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 767–776, New York, NY, USA, 2017. Association for Computing Machinery. doi: 10.1145/3097983.3098202.

- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In Daumé, H. and Singh, A. (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.
- Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021.
- Cheng, G., Han, J., and Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017. ISSN 1558-2256. doi: 10.1109/jproc.2017.2675998.
- Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. CLUB: A contrastive log-ratio upper bound of mutual information. In Daume, H. and Singh, A. (eds.), *International Conference on Machine Learning (ICLR)*, pp. 1779–1788, 2020.
- Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- Coates, A., Ng, A., and Lee, H. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *AISTATS*, 2011.
- D’Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *ArXiv*, 2011.03395, 2020.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. Ieee, 2009.
- Devillers, B., Bielawski, R., Choski, B., and VanRullen, R. Does language help generalization in vision models? *ArXiv*, 2104.08313, 2021.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, 2018.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.
- Fang, H., Xiong, P., Xu, L., and Chen, Y. CLIP2Video: Mastering video-text retrieval via image CLIP. *ArXiv*, 2106.11097, 2021.
- Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.
- Frans, K., Soros, L. B., and Witkowski, O. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. *ArXiv*, 2106.14843, 2021.
- Galatolo, F. A., Cimino, M. G. C. A., and Vaglini, G. Generating images from caption and vice versa via CLIP-guided generative latent space search. *ArXiv*, 2102.01645, 2021.
- Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *ArXiv*, 2017.
- Gao, T., Yao, X., and Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. *ArXiv*, 2104.08821, 2021.
- Gardner, E. Multiconnected neural network models. *Journal of Physics A*, 20(11):3453–3464, 1987. doi: 10.1088/0305-4470/20/11/046.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *ArXiv*, 2004.07780, 2020.

- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shave-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. Challenges in representation learning: A report on three machine learning contests, 2013.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21271–21284, 2020.
- Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterton, M. (eds.), *International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.
- Han, T., Xie, W., and Zisserman, A. Self-supervised co-training for video representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5679–5690, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Helber, P., Bischke, B., Dengel, A., and Borth, D. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207. IEEE, 2018.
- Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosatsat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.
- Hénaff, O. J., Srinivas, A., DeFauw, J., Razavi, A., Doersch, C., Eslami, S. M. A., and vanDenOord, A. Data-efficient image recognition with contrastive predictive coding. *ArXiv*, 1905.09272, 2019.
- Henderson, M. L., Al-Rfou, R., Strobe, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. Efficient natural language response suggestion for smart reply. *ArXiv*, 1705.00652, 2017.
- Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pp. 2554–2558, 1982.
- Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. In *Proceedings of the National Academy of Sciences*, volume 81, pp. 3088–3092. National Academy of Sciences, 1984. doi: 10.1073/pnas.81.10.3088.
- Horn, D. and Usher, M. Capacities of multiconnected memory models. *Journal of Physics France*, 49(3):389–395, 1988. doi: 10.1051/jphys:01988004903038900.
- Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. OpenCLIP, 2021.
- Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*. Open-Review, 2022.
- Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3581–3589. 2014.

- Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3D object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013.
- Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.
- Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1172–1180, 2016.
- Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 951–958, 2009.
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10, 2019. doi: 10.1038/s41467-019-08987-4.
- Li, J., Zhou, P., Xiong, C., Socher, R., and Hoi, S. C. H. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2021.
- Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2018.
- Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2017.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2019.
- Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval. *ArXiv*, 2104.08860, 2021.
- Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.
- McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. *ArXiv*, 1811.04251, 2018.
- McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 875–884, 26–28 Aug 2020.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3111–3119, 2013.
- Milbich, T., Roth, K., Sinha, S., Schmidt, L., Ghassemi, M., and Ommer, B. Characterizing generalization under out-of-distribution shifts in deep metric learning. *ArXiv*, 2107.09562, 2021.
- Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. *ArXiv*, 2107.04649, 2021.
- Misra, I. and vanDerMaaten, L. Self-supervised learning of pretext-invariant representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- Narasimhan, M., Rohrbach, A., and Darrell, T. CLIP-It! Language-guided video summarization. *ArXiv*, 2107.00650, 2021.
- Nguyen, X., Wainwright, M. J., and Jordan, M. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010. doi: 10.1109/tit.2010.2068870.

- Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47.
- Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W. *NIST handbook of mathematical functions*. Cambridge University Press, 1 pap/cdr edition, 2010. ISBN 9780521192255.
- Paischer, F., Adler, T., Patil, V., Bitto-Nemling, A., Holzleitner, M., Lehner, S., Eghbal-Zadeh, H., and Hochreiter, S. History compression via language models in reinforcement learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, pp. 17156–17185, 2022.
- Pakhomov, D., Hira, S., Wagle, N., Green, K. E., and Navab, N. Segmentation in style: Unsupervised semantic image segmentation with stylegan and CLIP. *ArXiv*, 2107.12518, 2021.
- Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.
- Pearl, J. Embracing causality in default reasoning. *Artificial Intelligence*, 35(2):259–271, 1988.
- Poole, B., Ozair, S., vanDenOord, A., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5171–5180, 2019.
- Potter, M. Conceptual short term memory in perception and thought. *Frontiers in Psychology*, 3:113, 2012. doi: 10.3389/fpsyg.2012.00113.
- Prentice, M. J. On invariant tests of uniformity for directions and orientations. *The Annals of Statistics*, 6(1):169–176, 1978. doi: 10.1214/aos/1176344075.
- Psaltis, D. and Cheol, H. P. Nonlinear discriminant functions and associative memories. *AIP Conference Proceedings*, 151(1):370–375, 1986. doi: 10.1063/1.36241.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.
- Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2021.
- Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.
- Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Hochreiter, S., and Klambauer, G. Modern Hopfield networks for few- and zero-shot reaction prediction. *ArXiv*, 2104.03279, 2021.
- Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Segler, M., Hochreiter, S., and Klambauer, G. Improving few- and zero-shot reaction template prediction using modern Hopfield networks. *Journal of Chemical Information and Modeling*, 2022.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2018.
- Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can CLIP benefit vision-and-language tasks? *ArXiv*, 2107.06383, 2021.

- Soomro, K., Zamir, A. R., and Shah, M. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.
- Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. The German traffic sign recognition benchmark: A multi-class classification competition. *The International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.
- Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18583–18599, 2020.
- Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802.
- Tsai, Y.-H. H., Ma, M. Q., Zhao, H., Zhang, K., Morency, L.-P., and Salakhutdinov, R. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *ArXiv*, 2106.02866, 2021.
- Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2019.
- van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, 1807.03748, 2018.
- Wainwright, M. J. *Basic tail and concentration bounds*, pp. 21–57. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.002.
- Wang, F. and Liu, H. Understanding the behaviour of contrastive loss. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 2021.
- Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- Wellman, M. P. and Henrion, M. Explaining ‘explaining away’. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993. doi: 10.1109/34.204911.
- Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., and Klambauer, G. Modern Hopfield networks and attention for immune repertoire classification. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18832–18845, 2020.
- Widrich, M., Hofmarcher, M., Patil, V. P., Bitto-Nemling, A., and Hochreiter, S. Modern hopfield networks for return decomposition for delayed rewards. *Deep Reinforcement Learning Workshop, Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. *ArXiv*, 2109.01903, 2021.
- Wu, M., Mosse, M., Zhuang, C., Yamins, D., and Goodman, N. Conditional negative sampling for contrastive learning of visual representations. In *International Conference on Learning Representations (ICLR)*. OpenReview, 2021.
- Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742, Los Alamitos, CA, USA, 2018. doi: 10.1109/CVPR.2018.00393.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.

- Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 547–558, 2020.
- Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., and LeCun, Y. Decoupled contrastive learning. *ArXiv*, 2110.06848, 2021.
- Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. *ArXiv*, 2203.17248, 2022.
- Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *ArXiv*, 2109.01134, 2021.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
 - (b) Did you describe the limitations of your work? [Yes] Our method is currently limited to natural images and short text prompts as inputs, and, thus its use for other types of images, such as medical or biological images, is unexplored. While we hypothesize that our approach could also be useful for similar data in other domains, this has not been assessed.
 - (c) Did you discuss any potential negative societal impacts of your work? [Yes] One potential danger arises from users that overly rely on systems built on our method. For example in the domain of self-driving cars, users might start paying less attention to the traffic because of the AI-based driving system. Finally, our method might also be used to automate various simple tasks, which might lead to reduced need for particular jobs in production systems. As for almost all machine learning methods, our proposed method relies on human-annotated training data and thereby human decisions, which are usually strongly biased. Therefore, the responsible use of our method requires the careful selection of the training data and awareness of potential biases within those.
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [Yes]
 - (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the URL to a GitHub repository that contains the code.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.1, Section 5.2 and Appendix Section A.2.2.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We added error bars for all experiments for which this was computationally feasible (see Table 1).
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We used several different servers equipped with GPUs of different types, such as V100 and A100. The total amount of compute is roughly 11, 000 GPU hours (with A100).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [Yes] For the model implementations we used PyTorch (Paszke et al., 2017, BSD license) and for monitoring the runs we used Weights & Biases (Biewald, 2020, MIT license).
 - (b) Did you mention the license of the assets? [Yes] See above.
 - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide the code as supplementary material.
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] We only use public datasets that can be used for research purposes, such as the YFCC dataset which was published under the Creative Commons licence.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] As almost all computer vision and natural language datasets, the data suffers from many biases including social biases. We refer to Yang et al. (2020) for a detailed analysis of biases in such datasets.
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

- (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]