ChatGPT

Deep Learning Feature Extraction with Batch Context

Introduction

Modern deep learning models have leveraged batch-level interactions – relationships between samples within the same mini-batch – to learn more powerful feature representations for images. In particular, recent advances in **representation learning for image classification** (especially since 2020) show that considering other samples in a batch (as negatives, positives, or contextual references) can significantly improve learned features. This report surveys several **notable research papers (2020–present)** that use such batch-context techniques. For each paper, we list the title, authors, year, a brief summary (with a focus on the method's main idea), how it leverages batch-level interactions, and a link to the paper. The papers are grouped by the type of approach (contrastive learning, clustering-based methods, negative-free methods, and attention-based methods).

Contrastive Learning Approaches

Contrastive learning methods train an encoder by **pulling together representations of similar images and pushing apart representations of different images**. Typically, each image instance is treated as its own class (instance discrimination), and other images in the batch serve as negative examples to repel in feature space. These methods benefit strongly from large batch sizes to provide many negatives [1]. Below are key papers in this category:

A Simple Framework for Contrastive Learning of Visual Representations (Ting Chen et al., 2020)

Authors: Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton
Summary: Introduces **SimCLR**, a simple yet effective framework for self-supervised visual representation learning [2]. The method learns image features by maximizing agreement between two augmented views of the same image (a positive pair) while **discriminating against other images in the batch as negatives**. Key findings of this work include the importance of strong data augmentations and a learnable projection head for better representations [3]. Notably, SimCLR demonstrates that performance improves with **larger batch sizes** (providing more negative samples) and more training steps [1], achieving near supervised-level accuracy on ImageNet with a ResNet-50.
Batch Interaction: SimCLR uses an in-batch contrastive loss (NT-Xent) where each image's representation is pushed away from all other images' representations in the **same batch** (except its augmented duplicate) [3]. In other words, every other sample in the batch acts as a negative example, directly influencing the feature learning of a given sample. This batch-wise comparison is central to SimCLR's success and obviates the need for a memory bank.
Link: arXiv:2002.05709 (ICML 2020)

Momentum Contrast for Unsupervised Visual Representation Learning (Kaiming He et al., 2020)

Authors: Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick
Summary: This work presents **MoCo (Momentum Contrast)**, an unsupervised framework that builds a dynamic dictionary of feature keys using a momentum-updated encoder [4]. MoCo maintains a queue of past encoded examples to serve as a large set of negative samples, enabling contrastive learning with a manageable batch size. The dictionary keys are updated via a moving-average (momentum) mechanism to keep them consistent even as the model evolves [4]. MoCo achieves competitive ImageNet results under linear evaluation and shows that unsupervised pretraining can rival or surpass supervised pretraining on various detection and segmentation tasks [5].
Batch Interaction: MoCo's contrastive loss treats other images' features as negatives similar to SimCLR, but the pool of negatives is extended beyond the current batch via the memory queue [4]. In each iteration, the current batch's samples are encoded (with one encoder) and matched against a dictionary of keys (encoded by a momentum encoder) to compute an InfoNCE loss. The context of other batch elements (and queued samples from recent batches) influences representation learning by encouraging the model to differentiate the current image from a wide array of other images.
Link: arXiv:1911.05722 (CVPR 2020)

Supervised Contrastive Learning (Prannay Khosla et al., 2020)

Authors: Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan
Summary: This paper extends contrastive learning to the fully supervised setting, introducing the **SupCon (Supervised Contrastive)** loss [6]. Rather than treating every instance as its own class, SupCon uses label information: embeddings of images with the same class label are pulled together, while those of different classes are pushed apart [7]. The authors show that this approach can **outperform standard cross-entropy** classification loss, achieving higher ImageNet accuracy (81.4% top-1 with ResNet-200) and improved robustness to data corruptions [8]. SupCon also proves more stable across various hyperparameters compared to cross-entropy.
Batch Interaction: SupCon leverages batch sample relationships by grouping positives by class. In each batch, all images of the same class form a positive cluster that is contrasted against images of other classes. Thus, each sample's representation is influenced by **other same-class samples (pulled closer)** and **different-class samples (pushed away)** present in the batch [7]. This multi-sample contrast within a batch effectively uses the batch as a mini-dataset of both positive and negative pairs beyond the one-to-one instance level.
Link: arXiv:2004.11362 (NeurIPS 2020)

Clustering-Based Representation Learning

Clustering-based methods incorporate grouping of similar images during training, often assigning pseudo-labels or prototype codes to images and using these labels to learn representations. These approaches exploit batch context by forming on-the-fly clusters or prototypes from the batch and enforcing consistency within those groups. Below are key papers in this category:

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments (Mathilde Caron et al., 2020)

Authors: Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin
Summary: Proposes **SwAV** (Swapped Assignments), an online clustering-based self-supervised method

[9]. SwAV avoids explicit pairwise feature comparison; instead, it **clusters images in the batch** and enforces consistency between cluster assignments of different augmentations of the same image [10]. Concretely, each batch of image features is clustered into a set of prototype vectors, and a "swapped prediction" task is used: the model predicts the cluster code of one augmentation from the feature of another augmentation [11]. SwAV also introduces a multi-crop augmentation strategy (using several small-crop views) to improve training efficiency [12]. This method achieves 75.3% top-1 ImageNet accuracy with a ResNet-50 (linear eval), rivaling previous contrastive methods, without requiring large memory banks or momentum encoders [12] [13].
Batch Interaction: SwAV explicitly **clusters the batch's image features** at each iteration [10]. The cluster assignments (essentially pseudo-labels) for the batch provide context: images that fall into the same cluster are pulled together by making their different augmented views predict one another's cluster code. The presence of other samples in the batch affects the cluster prototypes – each sample's representation is learned in relation to these prototype "codes" which are computed from the entire batch's feature distribution [10]. Thus, the representation for an image is influenced by how other batch images group with it or not, combining contrastive learning benefits with clustering [9].
Link: arXiv:2006.09882 (NeurIPS 2020)

Contrastive Clustering (Yunfan Li et al., 2021)

Authors: Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey T. Zhou, Xi Peng
Summary: This paper presents **Contrastive Clustering (CC)**, an approach that **bridges instance-level and cluster-level representation learning** in an online fashion [14]. CC simultaneously optimizes two contrastive objectives: one at the instance level (like instance discrimination via data augmentations) and one at the cluster level. The key idea is to treat the **rows of the batch feature matrix as "soft labels" (probability of assignment to clusters for each instance) and the columns as cluster prototypes** [15] [16]. By maximizing agreement for positive pairs and disagreement for negatives at both the instance and cluster assignment levels, the model learns representations and cluster assignments jointly [17]. CC's end-to-end training yields significantly improved clustering performance (e.g., much higher NMI on CIFAR-10/100) compared to prior deep clustering methods [18].
Batch Interaction: Contrastive Clustering relies on batch-wise computation of cluster assignments. In each batch, features are used to compute "soft" cluster assignments for all samples, and a **cluster-level contrastive loss** pulls features towards their batch cluster's centroid while pushing them from other clusters [14] [19]. Simultaneously, the usual instance-level contrastive loss considers other images in the batch as negatives. The cluster assignments themselves are derived from the batch's feature distribution (often via k-means or an assignment algorithm on the batch), meaning each sample's learning signal is directly influenced by how other batch samples group together.
Link: AAAI 2021 Paper (AAAI 2021)

(Note: Another notable method in this vein is **Prototypical Contrastive Learning (PCL)** by Li et al. (ICLR 2021), which also combines contrastive learning with clustering by assigning multiple prototype centroids to each image and contrasting images with their nearest prototypes [20] [21]. This further highlights the trend of using batch-level clustering or prototypes to guide feature learning.)*

Negative-Free Siamese Representation Learning

While contrastive methods typically rely on comparing each sample to others (negatives) in the batch, recent **"negative-free" Siamese networks** show that meaningful representations can be learned without explicit negative pairs. These methods still leverage multiple samples (usually two augmented views of the same image) and often incorporate batch statistics or architectural tricks to avoid trivial solutions. We highlight a few influential papers:

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning (Jean-Bastien Grill et al., 2020)

Authors: Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, et al. (13 others)
Summary: This paper introduces **BYOL**, a pioneering self-supervised method that surprisingly **does not use negative samples or contrastive pairs** [22]. BYOL employs two networks: an online network and a target network. Given two augmented views of the same image, the online network is trained to predict the target network's representation of the other view [23]. The target network's parameters are a slow-moving average of the online network's (a form of momentum encoder). Through this bootstrapping mechanism, BYOL avoids collapse and learns high-quality features, achieving 74.3% top-1 ImageNet accuracy with a ResNet-50 (and 79.6% with a larger ResNet) without any explicit negatives [24]. It matches or exceeds state-of-the-art on transfer and semi-supervised benchmarks, indicating that contrastive negatives are not strictly necessary for representation learning.
Batch Interaction: BYOL's training objective focuses on pairs of augmentations from the same image, so it does not directly pull or push against other images in the batch. However, **batch normalization** and other implicit effects mean the batch still provides context (e.g., the BN layers compute statistics across the batch). Importantly, BYOL demonstrates a case where feature representations can be learned without comparing to other batch samples – the presence of other images is only incidental for normalization. The heavy reliance on the **momentum-updated target network** and the prediction head on the online network are what prevent collapse, instead of negative sample comparisons [24]. This method highlights an alternative to batchwise contrast: using two networks and iterative knowledge distillation between them.
Link: arXiv:2006.07733 (NeurIPS 2020)

Barlow Twins: Self-Supervised Learning via Redundancy Reduction (Jure Zbontar et al., 2021)

Authors: Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, Stéphane Deny
Summary: Barlow Twins proposes a new loss function based on redundancy reduction to learn invariant features without negatives [25]. The method takes two augmented views of each image (processed by two identical networks) and computes the cross-correlation matrix of the two sets of embeddings across a batch. The objective then forces this cross-correlation to be as close to the identity matrix as possible [26]. Intuitively, this means: (1) the diagonal terms (same image, same feature component across views) are pushed to 1, encouraging similarity between the two views of the same image, and (2) the off-diagonal terms are pushed to 0, meaning different feature dimensions are decorrelated, reducing redundancy [27]. This avoids the trivial constant solution by ensuring the network doesn't encode all features the same. Barlow Twins achieved on-par performance with the best methods of the time on ImageNet linear evaluation, and did so **without requiring large batches or special momentum encoders or stop-gradients** [28].
Batch Interaction: The core of Barlow Twins is the **cross-correlation matrix computed over the batch** – this matrix captures relationships between representations of all samples in the batch. The loss explicitly relies on batch statistics: if any two different images in the batch have a correlated feature dimension, the off-diagonal entries capture that and the loss will penalize it. Thus, even though there are no explicit negative pairs, the presence of other samples in the batch influences the loss via the off-diagonal terms. Each image's features are optimized not only to match its augmentation twin, but also to ensure they are uncorrelated with other images' features along each dimension (at least in expectation over batches) [25]. In summary, Barlow Twins uses batch-level covariance information to enrich representations.
Link: arXiv:2103.03230 (ICML 2021)

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning (Adrien Bardes et al., 2022)

Authors: Adrien Bardes, Jean Ponce, Yann LeCun
Summary: **VICReg** is another negative-free self-supervised method designed to avoid collapsed representations via explicit regularizers [29]. It introduces three terms: (1) an invariance term (mean squared error between two augmentations' embeddings, encouraging similarity like BYOL), (2) a variance term that forces the per-dimension variance of embeddings in a batch to exceed a threshold (preventing all outputs from collapsing to a constant) [29], and (3) a covariance term which, similar to Barlow Twins, penalizes off-diagonal covariance between different feature dimensions (decorrelation) [30]. By combining these, VICReg achieves strong results on ImageNet and other tasks, comparable to state-of-the-art, with stable training. The authors also show that adding the variance regularization to other approaches (like BYOL/SimCLR) can improve their stability and performance [31]. (VICReg was accepted at ICLR 2022).
Batch Interaction: VICReg explicitly uses batch statistics in its loss. The **variance term** looks at the batch of embeddings and makes sure that along each feature dimension, the batch has sufficient variance (i.e. images aren't all producing the same value) [29]. The **covariance term** computes the covariance matrix of the batch's embeddings (after centering) and penalizes off-diagonals, meaning if any two different dimensions have correlated responses across the batch of samples, it pushes them toward zero [29]. These mechanisms ensure informational diversity and implicitly utilize relationships across the batch (since only with multiple samples can one measure variance or covariance). Thus, although VICReg does not compare different image instances via contrasting, it leans on batch-level dynamics (variance and covariance) to sculpt the feature space.
Link: arXiv:2105.04906 (ICLR 2022)

Note: Another important negative-free method is **SimSiam (Xinlei Chen & Kaiming He, 2021)**, which found that a simple Siamese network can learn good features without negatives, large batches, or momentum encoders – using just two views and a **stop-gradient** trick [32]. SimSiam's success further confirms that batch context can sometimes be minimized; however, most methods above leverage batch interactions more explicitly.

Attention-Based Batch Interaction

Beyond loss functions and clustering, researchers have also explored architectural modules that directly allow **attention across samples in a batch** to enhance feature extraction. These methods aim to identify important samples or features by looking at the batch as a whole. Below is a notable example:

BA^2M: A Batch Aware Attention Module for Image Classification (Qishang Cheng et al., 2021)

Authors: Qishang Cheng, Hongliang Li, Qingbo Wu, King Ngi Ngan
Summary: BA^2M introduces a novel attention mechanism that operates **across a batch of images** rather than only within individual images [33]. Traditional attention modules (SE block, CBAM, etc.) focus on enhancing features within one sample. In contrast, BA^2M adds a batch-aware stage: First, it computes a **Sample-wise Attention Representation (SAR)** for each image by fusing multi-scale attention maps (channel-wise, local spatial, global spatial) within that image [34]. Then, it takes all SARs in the batch and feeds them into a normalization function to produce a weight for each sample [35]. These weights indicate the relative importance or complexity of each sample's features in the batch context [36]. The network can use these weights to re-weight or modulate features, effectively paying more attention to harder or more informative examples in the batch. Experiments on CIFAR-100 and

ImageNet-1K showed consistent performance improvements across various CNN architectures, outperforming classical attention modules and sample re-weighting schemes [37].
Batch Interaction: BA^2M explicitly makes different samples in the batch attend to each other. By computing a batch-level normalization of the per-sample attention features, it **distinguishes the importance of features between samples in the training batch** [35]. For example, if one image in the batch is more complex or contains less common features, BA^2M can assign it a higher weight, causing the network to focus more on it during that forward pass. This means the representation of a given image is adjusted not in isolation, but in the context of what other images are in the same batch. In essence, the module learns to highlight or dampen features relative to the batch's overall feature distribution, introducing an attention-based interplay among batch samples [33].
Link: arXiv:2103.15099 (2021 preprint)

Conclusion

In summary, recent advances in representation learning for images increasingly exploit **batch-level relationships** to extract richer feature representations. Contrastive learning methods use other batch samples as explicit negatives (or positives in supervised cases) to shape the embedding space. Clustering-based approaches form on-the-fly groups within each batch to provide a broader context than instance-only views. Newer negative-free techniques avoid direct instance comparison but still harness batch statistics (like covariance) or Siamese architectures to learn invariances. Lastly, attention mechanisms like BA^2M show the potential of directly incorporating batch context in the network's forward pass to focus on important samples. All these works demonstrate that considering the **context of multiple samples concurrently** during training can lead to more discriminative and robust image representations, which in turn improves performance on image classification benchmarks. The field continues to evolve, with ongoing research exploring even more ways to leverage relationships across training samples for representation learning.

---

[1] [2] [3] [2002.05709] A Simple Framework for Contrastive Learning of Visual Representations
https://arxiv.org/abs/2002.05709

[4] [5] [1911.05722] Momentum Contrast for Unsupervised Visual Representation Learning
https://arxiv.org/abs/1911.05722

[6] [7] [8] [2004.11362] Supervised Contrastive Learning
https://arxiv.org/abs/2004.11362

[9] [10] [11] [12] [13] papers.neurips.cc
https://papers.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf

[14] [15] [16] [17] [18] [19] Contrastive Clustering
https://yunfan-li.github.io/assets/pdf/Contrastive%20Clustering.pdf

[20] [21] openreview.net
https://openreview.net/references/pdf?id=8Gyq2JakVU

[22] [23] [24] [2006.07733] Bootstrap your own latent: A new approach to self-supervised Learning
https://arxiv.org/abs/2006.07733

[25] [26] [27] [28] [2103.03230] Barlow Twins: Self-Supervised Learning via Redundancy Reduction
https://arxiv.org/abs/2103.03230

29 30 31 [2105.04906] VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning
https://arxiv.org/abs/2105.04906

32 [2011.10566] Exploring Simple Siamese Representation Learning
https://arxiv.org/abs/2011.10566

33 34 35 36 37 [2103.15099] BA^2M: A Batch Aware Attention Module for Image Classification
https://arxiv.org/abs/2103.15099