Local Aggregation for Unsupervised Learning of Visual Embeddings

Zhuang, C., Zhai, A. L., & Yamins, D. (2019). Local Aggregation for Unsupervised Learning of Visual Embeddings. *arXiv preprint arXiv:1903.12355v2*.

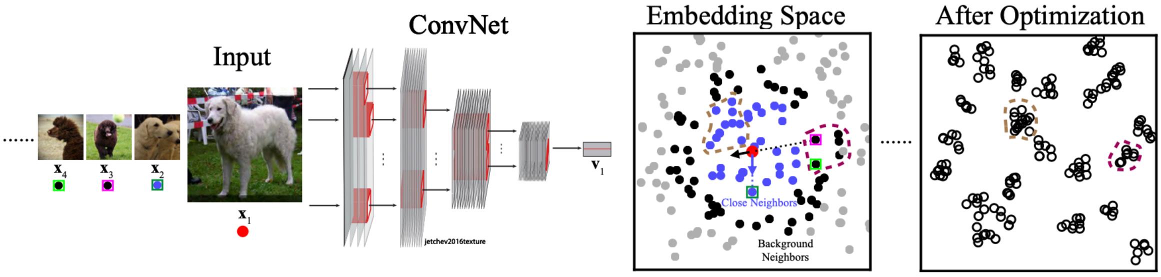


Figure 1. Illustration of the Local Aggregation (LA) method. For each input image, we use a deep neural network to embed it into a lower dimension space (”Embedding Space” panel). We then identify its close neighbors (blue dots) and background neighbors (black dots). The optimization seeks to push the current embedding vector (red dot) closer to its close neighbors and further from its background neighbors. The blue arrow and black arrow are examples of influences from different neighbors on the current embedding during optimization. The ”After Optimization” panel illustrates the typical structure of the final embedding after training.

Primary Problem

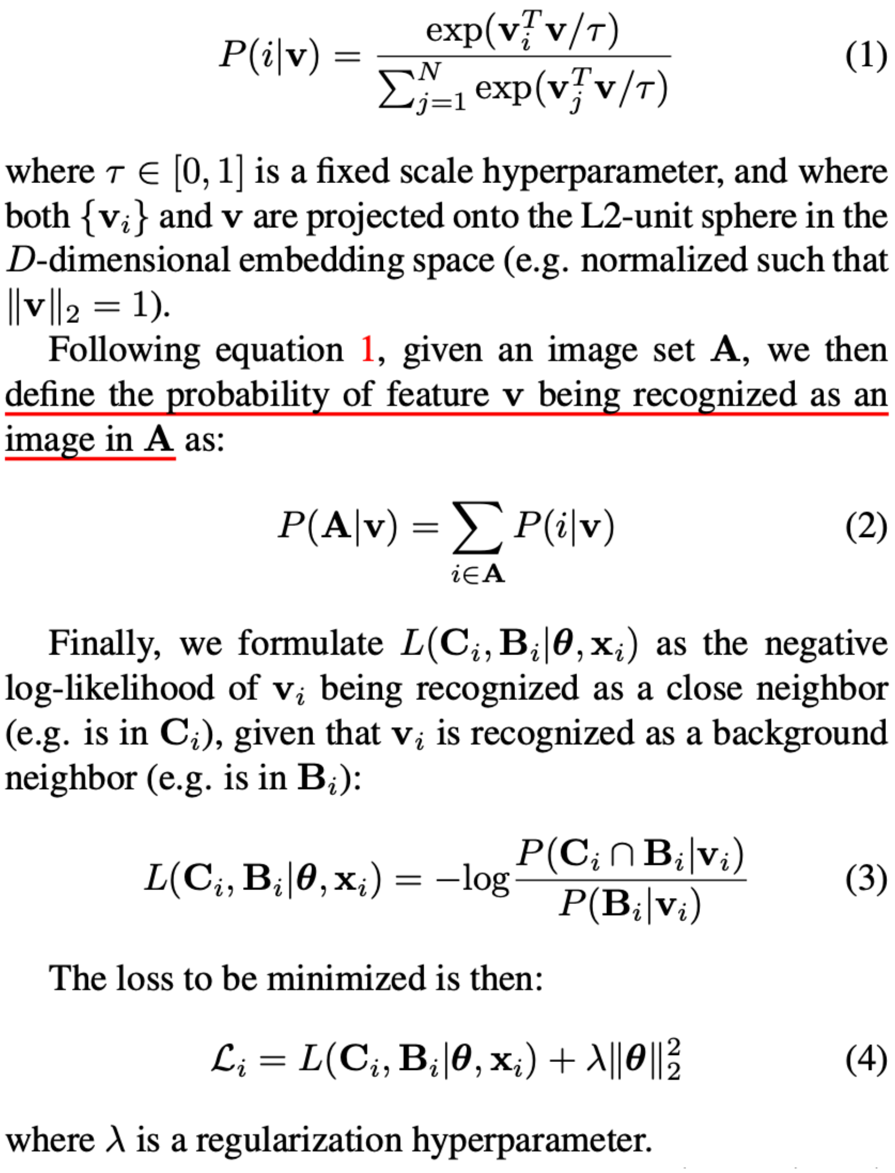
However, unsupervised networks have long **lagged** behind the performance of their supervised counterparts, especially in the domain of large-scale visual recognition.

Key Findings

- **embedding function** to maximize a metric of local aggregation
- causing similar data instances to **move together** in the embedding space, while allowing dissimilar instances to **separate**.
- we propose a novel unsupervised learning algorithm through local non-parametric aggregation in a latent feature space

Methods and Measures used

- **Neighbor Identification**
    - **close neighbors C_i:**At any given step of optimization, the background neighbors for a given embedded point v_i are simply defined as the k closest embedded points Nk(vi) within V, where distance is judged using the cosine distance on the embedding space.
    - **background neighbors B_i**:To identify close neighbors, we first apply an unsupervised clustering algorithm on all embedded points V to cluster the representations into m groups $G = G_1, G_2, \ldots, G_m$.
- **Local Aggregation Metric**

  - 

$$P(i|\mathbf{v}) = \frac{\exp(\mathbf{v}_i^T \mathbf{v}/\tau)}{\sum_{j=1}^N \exp(\mathbf{v}_j^T \mathbf{v}/\tau)} \tag{1}$$

where $\tau \in [0, 1]$ is a fixed scale hyperparameter, and where both $\{\mathbf{v}_i\}$ and $\mathbf{v}$ are projected onto the L2-unit sphere in the D-dimensional embedding space (e.g. normalized such that $\|\mathbf{v}\|_2 = 1$).

Following equation 1, given an image set $\mathbf{A}$, we then define the probability of feature $\mathbf{v}$ being recognized as an image in $\mathbf{A}$ as:

$$P(\mathbf{A}|\mathbf{v}) = \sum_{i \in \mathbf{A}} P(i|\mathbf{v}) \tag{2}$$

Finally, we formulate $L(\mathbf{C}_i, \mathbf{B}_i|\boldsymbol{\theta}, \mathbf{x}_i)$ as the negative log-likelihood of $\mathbf{v}_i$ being recognized as a close neighbor (e.g. is in $\mathbf{C}_i$), given that $\mathbf{v}_i$ is recognized as a background neighbor (e.g. is in $\mathbf{B}_i$):

$$L(\mathbf{C}_i, \mathbf{B}_i|\boldsymbol{\theta}, \mathbf{x}_i) = -\log \frac{P(\mathbf{C}_i \cap \mathbf{B}_i|\mathbf{v}_i)}{P(\mathbf{B}_i|\mathbf{v}_i)} \tag{3}$$

The loss to be minimized is then:

$$\mathcal{L}_i = L(\mathbf{C}_i, \mathbf{B}_i|\boldsymbol{\theta}, \mathbf{x}_i) + \lambda \|\boldsymbol{\theta}\|_2^2 \tag{4}$$

where λ is a regularization hyperparameter.

- **Memory Bank**: $\bar{v}_i \leftarrow (1-t)\bar{v}_i + tv_i$

Hypotheses

- researchers reliably report that infants as young as three months can group perceptually similar stimuli [Categorization in infancy. Trends in cognitive sciences], even for stimulus types that the infants have never seen before
- These findings suggest that biological unsupervised learning may take advantage of inherent visual similarity, **without requiring sharp boundaries between stimulus categories**.

Dataset

- ImageNet
- Places205
- PASCAL VOC 2007