Mysterious Success of Contrastive Unsupervised Learning

Unsupervised methods for representation learning, reminiscent of word2vec for word embeddings, have been very successful in NLP [1] and to some extent in vision [2]. With access to semantically similar points and random negative samples from unlabeled data, they minimize objectives that look like:

\[ L_{\text{unsup}}(f) = \mathbb{E}_{(x, x') \sim D_{\text{un}}} \left[ \log \left( 1 + e^{f(x)^T f(x') - f(x)^T f(x')} \right) \right] \]

Why are these representations successful on future linear classification tasks? We attempt to demystify this by providing

- Framework connecting unlabeled data with downstream tasks
- Provable guarantees for such algorithms under the framework: Unsupervised loss is surrogate for average supervised loss

Framework

Semantic similarity \( \approx \) membership in same latent class.

Connection

\( \mathcal{X} \): Set of inputs, \( \mathcal{C} \): Set of classes, \( \rho \): Distribution over \( \mathcal{C} \\
\mathcal{D}_{\text{c}} \): Universal distribution over \( \mathcal{X} \) conditioned on class \( c \).

Unlabeled Data

Similarity data: \( (x, x') \sim D_{\text{sim}} \)

\( c^+ \sim \rho \)

\( (x, x') \sim D_{\text{neg}} \)

Negative samples: \( x^+ \sim D_{\text{neg}} \)

\( c^- \sim \rho \)

\( x^- \sim \rho \)

Labeled data: \( (x, c) \sim D_{\mathcal{T}} \)

\( c \sim \mathcal{T} \)

\( x \sim D_{\mathcal{T}} \)

\( c \sim D_{\mathcal{T}} \)

Supervised Tasks

Task: Subset of latent classes \( \mathcal{T} = \{c_1, \ldots, c_k\} \subseteq \mathcal{C} \)

Labeled samples: \( (x, c) \sim D_{\mathcal{T}} \)

Unsupervised Loss Bounds Supervised Loss

\[ \mathcal{F} \subseteq \{ f : \mathcal{X} \rightarrow \mathbb{R}^d, \| f(\cdot) \| \leq R \} : \text{Function class of interest.} \]

\( \tau \): Probability that two classes sampled from \( \rho \) are the same.

Minimizer from \( \mathcal{F} \) of empirical unsupervised loss:

\[ L_{\text{unsup}}(f) = \mathbb{E}_{(x, x') \sim D_{\text{un}}} \left[ \log \left( 1 + e^{f(x)^T f(x') - f(x)^T f(x')} \right) \right] \]

Why are these representations successful on future linear classification tasks? We attempt to demystify this by providing

- Framework connecting unlabeled data with downstream tasks
- Provable guarantees for such algorithms under the framework: Unsupervised loss is surrogate for average supervised loss

Framework

Semantic similarity \( \approx \) membership in same latent class.

Connection

\( \mathcal{X} \): Set of inputs, \( \mathcal{C} \): Set of classes, \( \rho \): Distribution over \( \mathcal{C} \\
\mathcal{D}_{\text{c}} \): Universal distribution over \( \mathcal{X} \) conditioned on class \( c \).

Unlabeled Data

Similarity data: \( (x, x') \sim D_{\text{sim}} \)

\( c^+ \sim \rho \)

\( (x, x') \sim D_{\text{neg}} \)

Negative samples: \( x^+ \sim D_{\text{neg}} \)

\( c^- \sim \rho \)

\( x^- \sim \rho \)

Labeled data: \( (x, c) \sim D_{\mathcal{T}} \)

\( c \sim \mathcal{T} \)

\( x \sim D_{\mathcal{T}} \)

\( c \sim D_{\mathcal{T}} \)

Supervised Tasks

Task: Subset of latent classes \( \mathcal{T} = \{c_1, \ldots, c_k\} \subseteq \mathcal{C} \)

Labeled samples: \( (x, c) \sim D_{\mathcal{T}} \)

Theorem 2: Sufficient Conditions on \( \mathcal{F} \)

Let \( x \sim D_{\mathcal{T}}, \lambda : \text{maximum standard deviation of } f(x) \text{ in a direction, } R_c : \text{mean norm of } f(x) \text{. Let } s(f) = 2 \mathbb{E}_{c \sim \rho} \lambda_c R_c \)

\[ L_{\text{sup}}(f) = \mathbb{E}_{(c_1, c_2) \sim \rho^2} \left[ L_{\text{sup}}(\{c_1, c_2\}, f) \mid c_1 \neq c_2 \right] \]