Exercise Session 1

Learning at Scale: Supervised, Self-Supervised, and Beyond

Prepared by Xiuying Wei, Johann Wenckstern, Petr Grinberg

Overview

Additional Reading Materials. We recommend following three papers:
[1] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "**Representation Learning with Contrastive Predictive Coding**." arXiv preprint arXiv:1807.03748 (2018).
[2] Chen, Ting, et al. "**A Simple Framework for Contrastive Learning of Visual Representations**." International Conference on Machine Learning (ICML). PMLR, 2020.
[3] Radford, Alec, et al. "**Learning Transferable Visual Models from Natural Language Supervision**." International Conference on Machine Learning. PMLR, 2021.

Task 1. InfoNCE in Contrastive Learning.

In contrastive learning, the goal is to learn useful representations without labels. For each condition c, the model is trained to identify the single positive sample among K distractors. For example, in instance discrimination for images, c is one augmented view of an image, x^+ is another independent augmentation of the same image, and the negatives $\{x_i^-\}$ are views of other images (e.g., from the same minibatch or from a memory queue). In vision language tasks, c is an image, x denotes captions, x^+ is the matched caption, and the negatives are captions of other images. A common objective for this is the InfoNCE loss. We will investigate it from a probabilistic perspective in this task.

Suppose that, given a condition c, we form a candidate set $X = \{x_0, \ldots, x_K\}$ where $x_0 = x^+$ is the positive sample and $x_1, \ldots, x_K$ are K negative samples drawn i.i.d. from $q_{\text{noise}}(x)$. The InfoNCE objective is defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{(X,c)\in D} \left[\log \frac{\frac{p_{\text{data}}(x^+ \mid c)}{q_{\text{noise}}(x^+)}}{\sum_{j=0}^{K} \frac{p_{\text{data}}(x_j \mid c)}{q_{\text{noise}}(x_j)}} \right],$$

where the term inside the logarithm represents the probability of correctly identifying x^+ as the positive sample among all candidates.

In practice, we typically set $q_{\text{noise}}(x) = p_{\text{data}}(x)$, so the numerator simplifies to $\frac{p_{\text{data}}(x^+|c)}{p_{\text{data}}(x^+)}$, whose logarithm equals the pointwise mutual information. Then, we can encode x and c with neural networks to obtain embeddings, define a score to approximate this density ratio. Consequently, the InfoNCE loss encourages the model to learn effective representations that captures mutual information between x and c, by learning a higher estimated ratio (or score) for true positive pairs than for negative samples.

(a) **Connection to cross-entropy loss.** Choose logits whose exponentials match the relative weights of candidates in the InfoNCE numerator/denominator; then build connection between InfoNCE and cross-entropy loss.

(b) **Relation to NCE.** The local NCE loss is defined as

$$\mathcal{L}_{\text{NCE}} = \sum_{(X,c)\in D} \left[-\log \frac{p_{\text{data}}(x^+ \mid c)}{p_{\text{data}}(x^+ \mid c) + K\, q_{\text{noise}}(x^+)} - \sum_{j=1}^{K} \log \frac{K\, q_{\text{noise}}(x_j^-)}{p_{\text{data}}(x_j^- \mid c) + K\, q_{\text{noise}}(x_j^-)} \right].$$

Show that this objective can be seen as optimizing the same logits as the InfoNCE but with a binary cross-entropy loss.

(c) **Effect of K.** In the case $q_{\text{noise}}(x) = p_{\text{data}}(x)$, analyze the effect of the number of negative samples K on the InfoNCE loss.

Hint: In this case, the logarithm of the numerator equals the pointwise mutual information. Optimizing the InfoNCE loss corresponds to maximizing a lower bound on the mutual information.

Task 2. Masked Language Modeling as Pseudo-Likelihood and -Perplexity.

Consider a sequence $x = (x_1, \ldots, x_T)$ from a data distribution $\mathcal{D}$. For masked language modeling (MLM), let us draw a random mask set $M \subseteq \{1, \ldots, T\}$ by sampling each position independently with probability $q \in (0, 1)$. Let $x_{\setminus t}$ denote x with the token in position t hidden (or replaced) and other tokens visible. Let x_t be an actual token in the position t. The MLM training objective is

$$\mathcal{L}_{\text{MLM}}(\theta) \;=\; -\,\mathbb{E}_{x\sim\mathcal{D}}\, \mathbb{E}_M \left[\sum_{t\in M} \log p_\theta\big(x_t \mid x_{\setminus t}\big) \right]. \tag{1}$$

(a) **Connection to pseudo-likelihood.** Define the (negative) pseudo log-likelihood (NPLL):

$$\text{NPLL}(\theta) = -\,\mathbb{E}_{x\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log p_\theta\big(x_t \mid x_{\setminus t}\big) \right]. \tag{2}$$

Show that under independent Bernoulli masking at rate q, i.e., $M \overset{iid}{\sim} \text{Bern}(q)$:

$$\mathbb{E}_M \left[\sum_{t\in M} \log p_\theta(x_t \mid x_{\setminus t}) \right] = q \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{\setminus t}), \tag{3}$$

and hence $\mathcal{L}_{\text{MLM}}(\theta) = q \cdot \text{NPLL}(\theta)$.

Hint: Use indicators $\mathbf{1}_{\{t\in M\}}$ and linearity of expectation.

(b) **Pseudo-perplexity and an unbiased estimator.** Define the pseudo-perplexity (PPPL) for a sequence x by

$$\mathrm{PPPL}(x) \;=\; \exp\left(\frac{1}{T}\sum_{t=1}^{T} -\log p_\theta\big(x_t \mid x_{\setminus t}\big)\right). \tag{4}$$

(i) Show that $\log \mathrm{PPPL}(x)$ equals the average token-wise NPLL.

(ii) Propose a practical *unbiased* single-pass estimator of $S(x) = \sum_{t=1}^{T} -\log p_\theta(x_t \mid x_{\setminus t})$ by sampling one index $U \sim \mathrm{Unif}\{1, \ldots, T\}$ and evaluating only $-\log p_\theta(x_U \mid x_{\setminus U})$. Then, prove that it is indeed unbiased.

(c) **Relation to autoregressive maximum likelihood.** An autoregressive (AR) model maximizes

$$\log p_\theta(x) = \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}). \tag{5}$$

(i) Explain why AR likelihood can be evaluated exactly, whereas MLM/pseudo-likelihood generally cannot yield a normalized joint $p_\theta(x)$.

(ii) *Bonus*: State a modeling assumption(s) under which minimizing $\mathrm{NPLL}(\theta)$ is statistically consistent. That is, if θ_n denotes the estimator from n samples, then $\mathrm{NPLL}(\theta_n) \to \mathrm{NPLL}(\theta^\star)$ and $\theta_n \to \theta^\star$ as $n \to \infty$. Sketch the proof of consistency.
Hint: Consider applying results from M-estimation theory.

(iii) Give one advantage and one limitation of MLM vs. AR for downstream tasks.

Bonus (BERT 80/10/10). In BERT, of the selected tokens, 80% are replaced by `[MASK]`, 10% by a random token, and 10% are left unchanged. Argue how this reduces train–test mismatch and prevents over-reliance on `[MASK]`; predict qualitative effects of using 100% or 0% `[MASK]`.

Task 3. Exploring Contrastive Learning with SimCLR.

See Task A in the Jupyter notebook.

Task 4. Exploring the Scaling Behaviour of LMs with a Series of Pythia Models.

See Task B in the Jupyter notebook.
