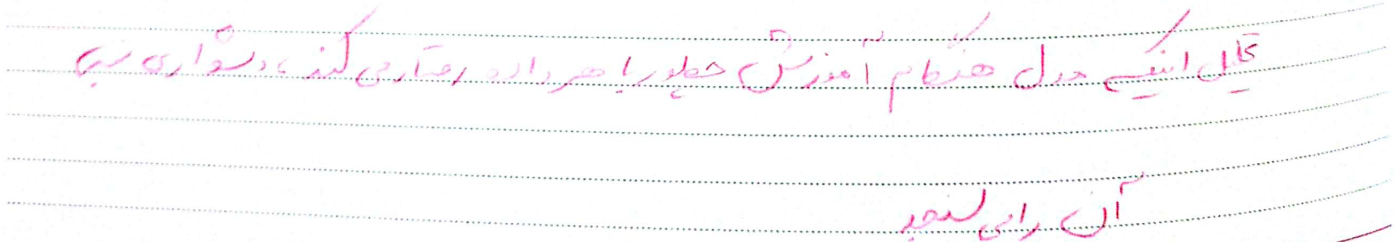
نکات مقاله

«ELFS: LABLE-FRee Coreset selection with Proxy Training Dynamics»

1 ELFS utilizes deep clustering to estimate training dynamics-based data difficulty scores without ground truth labels.

2 Pseduo-labels introduce a distribution shift in the data difficulty scores, and we propose a simple but effective double-end pruning method to mitigate bias on calculated scores.

منظورش از training dynamics چیه؟



مهر ۳۰ | 22 Oct 2023
یکشنبه | ۶ ربیع الثانی ۱۴۴۵

۲ چالش اصلی که مقاله مطرح و سعی در پاسخ دهی به آن‌ها دارد ؟

1 How to estimate training dynamics without ground truth labels ?

2 How to select high-quality coresets with proxy training dynamics ?

برای برطرف کردن چالش ۱ :

* Pseudo-labels provide a good estimation for training dynamics based difficulty scores.

* ELFS utilizes deep clustering to assign a pseudo-label to each data point.

★ For the second challenge, we observe that directly applying existing selection methods like CCS results in a performance gap when compared to coreset selection with ground truth labels.

تفاوت بین توزیع داده‌های واقعی و توزیع تخمینی (یا یادگرفته شده) که باعث می‌شود روش‌های انتخاب شده دیگر نماینده‌ی واقعی داده‌ها نباشند.

★ this distribution shift leads traditional sampling methods to select more easy examples, resulting in inferior performance

→ To mitigate this distribution shift issue, we propose a simple yet effective double-end pruning method for the pseudo-label-based difficulty scores.

★ This double-end pruning significantly reduces the number of selected easy examples and improves the coreset selection performance.

آبان | ۴ | 26 Oct 2023
پنجشنبه | ۱۰ ربیع الثانی ۱۴۴۵

* CCS introduces a method that combines hard example pruning with stratified sampling to jointly consider data difficulty and coverage, significantly improving coreset selection performance.

روش پیشنهادی مقاله

وفات حضرت معصومه (س) (۲۰۱ هـ ق)

آبان | ۵ | 27 Oct
جمعه | ۱۱ ربیع الثانی

* Deep clustering is a technique that combines deep learning methods to group data into clusters without explicit supervision.

* we utilize Teacher Ensemble_weighted pointwise Mutual Information (TEMI) to generate pseudo_labels for the unlabeled dataset.

روش TEMI:

* calculates the k nearest neighbor set N_x.
* After getting the nearest neighbor sets for all examples,
* TEMI trains an ensemble of classification heads to assign pseudo labels.
* TEMI uses an ensemble of student head $h_s(.)$ and a teacher head $h_t(.)$ that share the same architecture but differ in their updates.
* Each instance x is processed through both types of heads, producing probabilistic classifications $q_{f_s}(c|x)$ and $q_{f_t}(c|x)$, respectively, where $c \in \{1, \ldots, C\}$ denotes the class label.

* The student head's parameters θ_s are updated via backpropagation, while the teacher head's parameters θ_t are updated via an exponential moving average (EMA) of the student's parameters, ensuring more stable target distributions

Loss Function

* Given H clustering heads, TEMI aims to minimize the following accumulated weighted pointwise mutual information (PMI) objective to encourage the model to align instances that are likely to belong to the same cluster while penalizing those that are less likely to be related:

* $$\mathcal{L}_{TEMI}(x) = -\frac{1}{2H}\sum_{h=1}^{H}\sum_{x' \in \mathcal{N}_x} w_h(x,x') \cdot \left(pmi^h(x,x') + pmi^h(x',x)\right)$$

* $$pmi(x,x') = \log\left(\sum_{c=1}^{C}\frac{\left(q_s(c|x)\, q_t(c|x')\right)^{\beta}}{q_t(c)}\right)$$

pmi → point wise mutual information را تعریف می‌زند

دارد جون پاس شدم

بین جفت‌های (x, x')

و

$$w(x,x') = \sum_{c=1}^{C} q_t(c|x)\, q_t(c|x')$$

وزن‌های بین جفت‌هایی که احتمال تعلق به same cluster را نشان می‌دهد.

* TEMI assigns the pseudo label $\tilde{y}$ by aggregating these predictions and selecting the class with the highest combined probability:

$$\tilde{y} = \arg\max_{c \in \{1, \dots, C\}} \frac{1}{H} \sum_{h=1}^{H} q_h(c|x)$$

* Pseudo-labels generated by deep clustering make it feasible to calculate training dynamics scores with supervised training.

* However, the clustering-based pseudo-labeling can introduce many label noises, which causes the distribution shift of difficulty scores. ?

توضیح ← یک روش انتخاب داده است که ابتدا داده‌های خیلی سخت را (به‌ترتیب) و سپس داده‌های خیلی آسان (نیز هم) را حذف می‌کند تا بخش‌های متعادل از داده‌های آموزشی باقی بماند

* Double-end pruning consists of two steps: 1 Prune β hard examples first, where β is a hyperparameter

2 Continue pruning easy examples until the budget is met.

* A common practice is to perform a grid search to select the optimal β.

* We show that pseudo-labels generated in the deep clustering step provide a good estimation for this β search: After generating the pseudo-labeled dataset, we split it into 90% for training and 10% for validation.

* We use the validation set to determine the optimal β. Once the optimal β is identified, ELFS selects coresets from the

* entire training dataset. Our evaluation shows that β grid search with pseudo-labels provides a good estimation for the optimal β

* ELFS guarantees that ELFS does not use any ground truth labels in the coreset selection process.

* We report ELFS results using DINO and SwAV as the pretrained vision model to extract embeddings.

* We use the area under the margin (AUM) as the training dynamics metric for all datasets.