Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution

Докладчик: Фролова Анна

Содержание

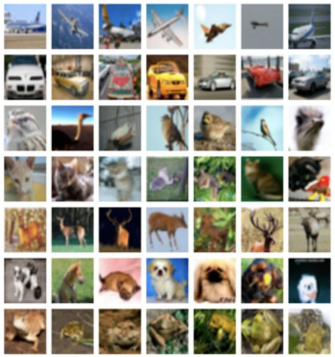
- Введение
- Теория
- Эксперименты
- Заключение

Введение

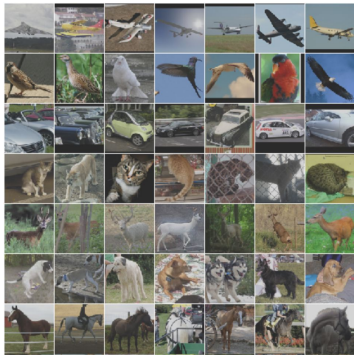
Fine-Tuning может привести к искажению предобученных функций и снижению производительности в OOD

Out-of-distribution (OOD)

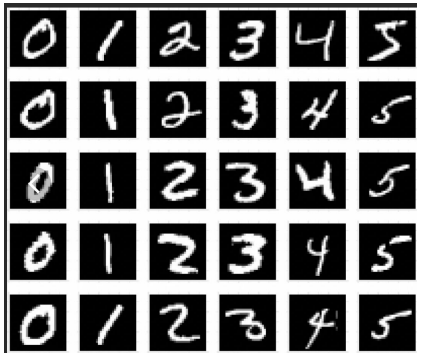
- В статье Out-of-distribution используется как данные которые не использовались для обучения модели
- К примеру, ImageNet была обучена на датасете CIFAR-10, тогда
  - CIFAR-10: In-Distribution (ID)
  - STL : Out-of-Distribution (OOD)



CIFAR-10



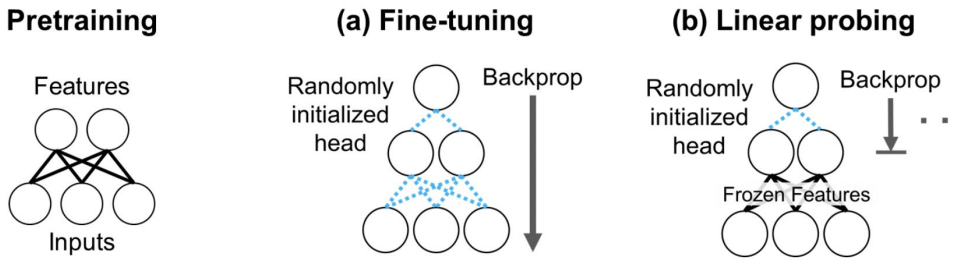
STL



MNIST

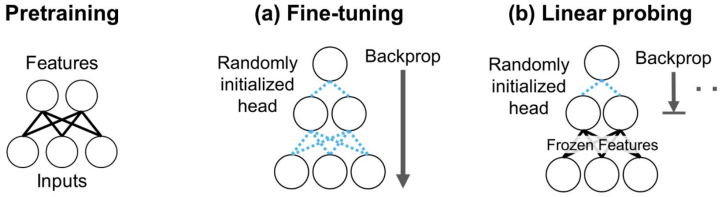
Transfer Learning

- Transfer Learning - это исследовательская задача в области машинного обучения, которая фокусируется на хранении знаний, полученных при решении одной проблемы, и применении их к другой, но связанной проблеме.
- Существуют два основных метода TL:
  - Fine-Tuning
    - ✓ Обновляет все параметры модели
  - Linear Probing
    - ✓ Обновляет параметры только последнего линейного слоя



Fine-Tuning VS Linear Probing

- Хорошо известно, что Fine-Tuning почти всегда дает больше accuracy, чем Linear Probing.
- При некоторых обстоятельствах, однако, Fine Tuning может показывать accuracy ниже Linear Probing.
  - ✓ Когда?
    - Когда сдвиг распределения между ID и OOD большой
  - ✓ Почему?
    - Потому что Fine Tuning искажает признаки



ID test	85.1%	82.9%
OOD test	59.3%	66.2%

Теория

- Обучить предиктор l, чтобы сопоставьте входные данные с выходными
- Оценить предикторы на данных **In-Distribution** и **Out-of-Distribution**

$$L_{\text{id}}(f) = \mathop{\mathbb{E}}_{(x,y)\sim P_{\text{id}}}[\ell(f(x),y)] \text{ and } L_{\text{ood}}(f) = \mathop{\mathbb{E}}_{(x,y)\sim P_{\text{ood}}}[\ell(f(x),y)]$$

- Предиктор параметризован следующим образом: $f_{v,B}(x) = v^\top g_B(x)$

  B - feature extractor, $g_B(x) \in \mathbb{R}^k$, v - linear head

- Предположим, что feature extractor B_0 получен при обучении на большом количестве данных.
  - ✓ Linear Probing минимизирует loss, сохранив исходный feature extractor
  - ✓ Fine-tuning минимизирует loss обновляя и feature extractor, и linear head

Теория

- Для анализа, статья фокусируется на задаче регрессии:

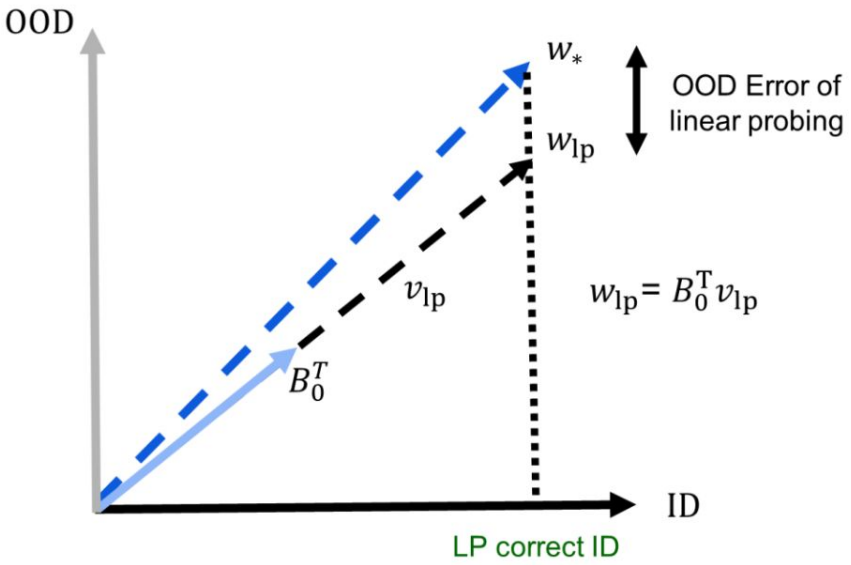
$$\mathcal{Y} = \mathbb{R} \text{ and } \ell(\widehat{y}, y) = (\widehat{y} - y)^2$$

- Предположим, что feature extractor - линейный: $f_{v,B}(x) = v^\top Bx$

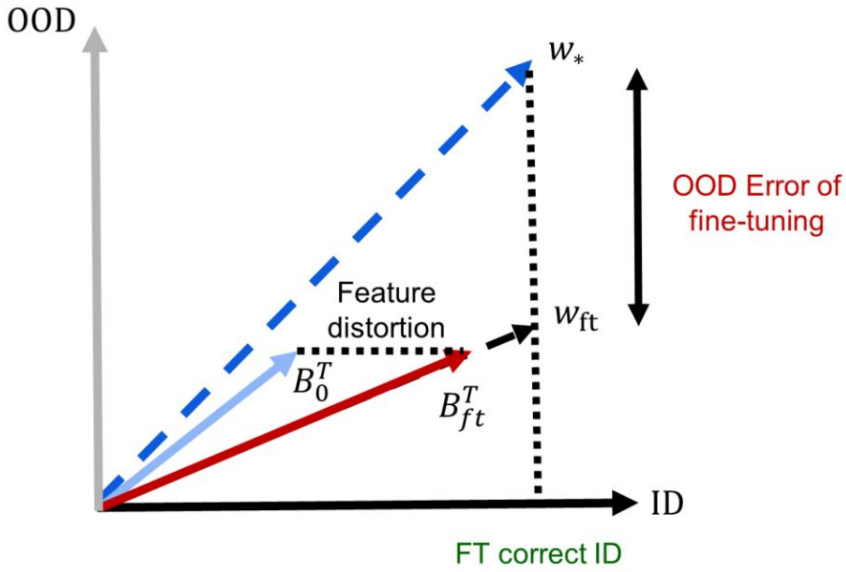
$B \in \mathcal{B} = \mathbb{R}^{k \times d}$ v - linear head

- Good pretrained model
  - ✓ Для простоты мы предполагаем, что модели хорошо определены $y = v_\star^\top B_\star x$
  - ✓ Для любой матрицы вращения U, $(Uv_\star)^T(UB_\star)x = v_\star^T B_\star x$
  - ✓ Предположим, что у нас есть предобученный feature extractor B_0 очень близкий к B_*, так чт $d(B_0, B_\star) \leq \epsilon$, гд $d(B, B') = \min_U \|B - UB'\|_2$
  - ✓ Предобучение было на большом количестве данных, таким образом B_0 становится хорошей предобученной моделью

Когда почему происходит искажение признаков?



(a) Toy example (Linear probing)



(b) Toy example (fine-tuning)

Эксперименты

- Данные и архитектура

ID	DomainNet	Living-17	FMoW Geo-Shift (North America)	CIFAR-10	CIFAR-10	ImageNet-1K
OOD	DomainNet	Entity-30	FMoW Geo-Shift (Africa ,Europe)	STL	CIFAR-10.1	ImageNet V2 ImageNet-R ImageNet-A ImageNet-Sketch
Architecture	ResNet-50	ResNet-50	ResNet-50	ResNet-50	ResNet-50	CLIP pretrained ViT-B/16

- Метод обучения:
  1. Fine-Tuning
     - ✓ Cosine learning rate schedule
     - ✓ Batch size of 64
     - ✓ Early stop and choose the best learning rate using ID validation accuracy

Эксперименты

- Метод обучения:
  2. Linear probing
     - ✓ Обучаем логистическую регрессию с L_2- регуляризатором frozen features from the penultimate layer
     - ✓ Выбираем лучший L_2- регуляризатор как гиперпараметр основанный на ID validation accuracy
- Для всех датасетов гиперпараметры подбирались исходя из 3 запусков эксперимента
- ImageNet - слишком большой датасет, поэтому эксперимент запускался 1 раз
- OOD данные использовались только для оценивания

Результаты

- Fine-Tuning дал лучшие результаты чем Linear Probing на 5/6 ID датасетах
- Linear Probing дал лучшие результаты чем Fine-Tuning на 8/10 OOD датасетах
- Результаты подтверждают предположения статьи почти во всех случаях
  - Исключение в случае датасетов CIFAR-10.1 и ImageNetV2
  - Это как раз может происходить из-за небольшого различия между ID и OOD данными (CIFAR-10 ↔ CIFAR-10.1, ImageNet ↔ ImageNetV2)
- LP-FT инициализирует голову нейронной сети, используя линейное Linear Probing, а затем Fine-Tuning модели
- LP-FT подсчитывается аналогично тому, как это делалось для Fine-Tuning
- LP-FT дал лучшие результаты на 5/6 ID датасетах и на 10/10 OOD датасетах

Результаты

ID Accuracy

	CIFAR-10	Ent-30	Liv-17	DomainNet	FMoW	ImageNet	Average
FT	**97.3 (0.2)**	**93.6 (0.2)**	97.1 (0.2)	84.5 (0.6)	**56.5 (0.3)**	**81.7 (-)**	85.1
LP	91.8 (0.0)	90.6 (0.2)	96.5 (0.2)	89.4 (0.1)	49.1 (0.0)	79.7 (-)	82.9
LP-FT	**97.5 (0.1)**	**93.7 (0.1)**	**97.8 (0.2)**	**91.6 (0.0)**	51.8 (0.2)	**81.7 (-)**	**85.7**

OOD Accuracy

	STL	CIFAR-10.1	Ent-30	Liv-17	DomainNet	FMoW
FT	82.4 (0.4)	92.3 (0.4)	60.7 (0.2)	77.8 (0.7)	55.5 (2.2)	32.0 (3.5)
LP	85.1 (0.2)	82.7 (0.2)	**63.2 (1.3)**	82.2 (0.2)	79.7 (0.6)	**36.6 (0.0)**
LP-FT	**90.7 (0.3)**	**93.5 (0.1)**	**62.3 (0.9)**	**82.6 (0.3)**	**80.7 (0.9)**	**36.8 (1.3)**

	ImNetV2	ImNet-R	ImNet-Sk	ImNet-A	Average
FT	**71.5 (-)**	52.4 (-)	40.5 (-)	27.8 (-)	59.3
LP	69.7 (-)	70.6 (-)	46.4 (-)	45.7 (-)	66.2
LP-FT	**71.6 (-)**	**72.9 (-)**	**48.4 (-)**	**49.1 (-)**	**68.9**

Изучение теории искажения признаков

- Ранняя остановка не уменьшает искажения функций
  - Можно подумать что дообучаясь на ID данных, модель сильно подстраивается под них, поэтому дает плохие результаты, поэтому ранняя остановка может помочь не переобучиться
  - Но Fine-Tuning все равно показывает худшие результаты, чем Linear Probing
  - Даже если модель выбирать исходя из OOD accuracy то это все равно не ничего не меняет
- ID-OOD features искажаются из-за fine-tuning
- Предобученные функции должны быть хорошими, но ID-OOD должны быть далеко друг от друга

Резюме

- Искажение настроек, вызванное Fine-Tuning-ом модели снижает точность OOD
- Сохранение функций может быть важно для надежности предсказаний
  - ✓ Показано теоретически и экспериментально в статье
- LP-FT может уменьшить разницу между ID и OOD accuracy