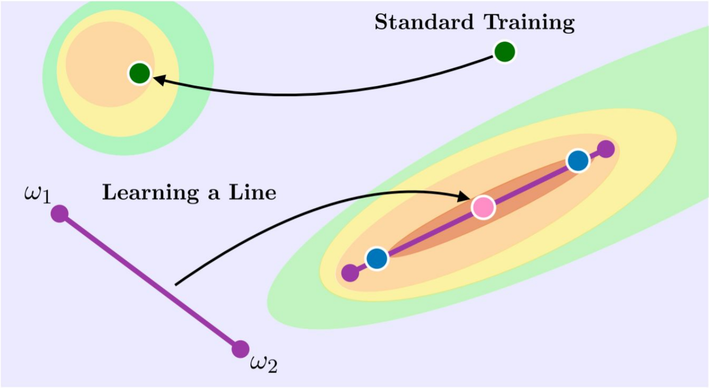
Learning Neural Network Subspaces

Исаев Сергей
Семерова Елена
Ким Михаил

Постановка задачи

При оптимизации нейросети хотим находить не точку (хороший набор весов), а подпространство весов (отрезок, кривая, симлекс), в котором выполняются полезные свойства:



- Модель в из центра подпространства будет работать лучше с точки зрения метрик, калибровки и устойчивости, чем при обычном обучении.
- Из моделей, полученных из подпространства, можно сделать хороший ансамбль.

Немного определений и наблюдений

Ансамбль моделей: $\hat{\mathbf{y}} = \frac{1}{2}\left(f\left(\mathbf{x}, \theta_T^1\right) + f\left(\mathbf{x}, \theta_T^2\right)\right)$ — как правило улучшает метрики, калибровку и устойчивость модели.

Ансамбль весов: $f\left(\mathbf{x}, \frac{1}{2}\left(\theta_T^1 + \theta_T^2\right)\right)$ — как правило так себе идея (но мы будем искать такие веса, чтобы это правило ломалось)

Коннектор это такая непрерывная функция $\mathsf{P} : [0,1] \to \mathbb{R}^n$, возвращающая веса модели $\mathsf{P}(0) = \psi_1,\ \mathsf{P}(1) = \psi_2$, для которой выполняется свойство:

$$\inf_{\alpha \in [0,1]} \mathsf{Acc}(\mathsf{P}(\alpha)) \gtrapprox \frac{1}{2}(\mathsf{Acc}(\psi_1) + \mathsf{Acc}(\psi_2))$$

Что утверждает свойство про ансамбль весов на языке коннекоторов?

Ещё немного наблюдений и определений

Наблюдение: линейный коннектор всё-таки можно построить, например, если уже немного немного предобученную модель начать обучать по разными траектониям. Например, за счет рандомизации семплирования примеров в батче.

Определение: коннектор можно обобщить и на m-мерное пространство. Введём m-мерное пространство для коннектора:

$$\Delta^{m-1} = \{\boldsymbol{\alpha} \in \mathbb{R}^m : \sum_i \boldsymbol{\alpha}_i = 1, \boldsymbol{\alpha}_i \geq 0\}$$

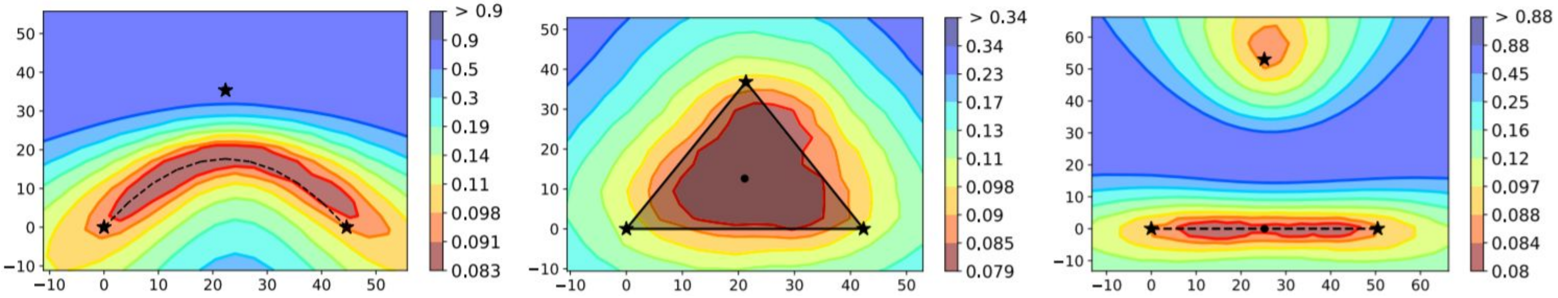
А $\mathbf{e}_i$ — стандартный базисный вектор в Δ^{m-1} из 0 и 1 на позиции i.

Последний слайд с определениями, честно

Тогда m-мерный коннектор на весах модели $\psi_1, \ldots, \psi_m \in \mathbb{R}^n$ — это функция $\mathsf{P} : \Delta^{m-1} \to \mathbb{R}^n$, что $\mathsf{P}(\mathbf{e}_i) = \psi_i$ и

$$\inf_{\boldsymbol{\alpha} \in \Delta^{m-1}} \mathsf{Acc}(\mathsf{P}(\boldsymbol{\alpha})) \gtrapprox \frac{1}{m} \sum_{i=1}^{m} \mathsf{Acc}(\psi_i)$$

В статье описывается эффективный алгоритм поиска коннекторов вида $\mathsf{P}(\boldsymbol{\alpha}) = \sum_i \boldsymbol{\alpha}_i \psi_i$ (симплексы) и одномерных кривых Безье.



А теперь формальная постановка задачи

Пусть определён “коннектор” $\mathsf{P}(\cdot, \{\omega_i\}_{i=1}^m) : \Lambda \to \mathbb{R}^n$

Координаты для коннектора семплируются из распределения $\mathcal{U}(\Lambda)$.

Примеры для модели семплируются из распределения $\mathcal{D}$.

Тогда мы пытаемся минимизировать такое матожидание:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\Big[\mathbb{E}_{\boldsymbol{\alpha}\sim\mathcal{U}(\Lambda)}\big[\ell(f(\mathbf{x}, \mathsf{P}(\boldsymbol{\alpha}, \{\omega_i\}_{i=1}^m)), \mathbf{y})\big]\Big]$$

Algorithm 1 $\mathsf{TrainSubspace}$

Input: P with domain Λ and parameters $\{\omega_i\}_{i=1}^m$, network f, train set $\mathcal{S}$, loss ℓ (*e.g.* a line has $\Lambda = [0, 1]$ and $\mathsf{P}(\alpha; \omega_1, \omega_2) = (1-\alpha)\omega_1 + \alpha\omega_2$).

Initialize each ω_i independently.

for batch $(\mathbf{x}, \mathbf{y}) \subseteq \mathcal{S}$ **do**

  Sample $\boldsymbol{\alpha}$ uniformly from Λ.

  $\theta \leftarrow \mathsf{P}(\boldsymbol{\alpha}; \{\omega_i\}_{i=1}^m)$

  $\hat{\mathbf{y}} \leftarrow f(\mathbf{x}, \theta)$

  $\mathcal{L} \leftarrow \ell(\hat{\mathbf{y}}, \mathbf{y})$

  Backprop $\mathcal{L}$ to each ω_i and update with SGD & momentum using estimate $\frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{\partial \ell}{\partial \theta} \frac{\partial \mathsf{P}}{\partial \omega_i}$

end for

Алгоритм плохой. Нужна регуляризация!

К сожалению, при таком подходе обучаемое подпространство может схлопнуться в одну точку или его точки могут быть недостаточно ортогональными для хорошего ансамбля.

Сделаем регуляризацию, которая будет давать штраф за недостаточно ортогональные точки.

$$\beta \cdot \mathbb{E}_{j \neq k}\left[\cos^2(\omega_j, \omega_k)\right] = \beta \cdot \mathbb{E}_{j \neq k}\left[\frac{\langle \omega_j, \omega_k \rangle^2}{\|\omega_j\|_2^2 \|\omega_k\|_2^2}\right]$$

Algorithm 1 $\mathsf{TrainSubspace}$

Input: P with domain Λ and parameters $\{\omega_i\}_{i=1}^{m}$, network f, train set $\mathcal{S}$, loss ℓ, and scalar β (*e.g.* a line has $\Lambda = [0, 1]$ and $\mathsf{P}(\alpha; \omega_1, \omega_2) = (1 - \alpha)\omega_1 + \alpha\omega_2$).

Initialize each ω_i independently.

for batch $(\mathbf{x}, \mathbf{y}) \subseteq \mathcal{S}$ **do**

  Sample $\boldsymbol{\alpha}$ uniformly from Λ.

  $\theta \leftarrow \mathsf{P}(\boldsymbol{\alpha}; \{\omega_i\}_{i=1}^{m})$

  $\hat{\mathbf{y}} \leftarrow f(\mathbf{x}, \theta)$

  Sample j, k from $\{1, ..., m\}$ without replacement.

  $\mathcal{L} \leftarrow \ell(\hat{\mathbf{y}}, \mathbf{y}) + \beta \cos^2(\omega_j, \omega_k)$

  Backprop $\mathcal{L}$ to each ω_i and update with SGD & momentum using estimate $\frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{\partial \ell}{\partial \theta} \frac{\partial \mathsf{P}}{\partial \omega_i} + \beta \frac{\partial \cos^2(\omega_j, \omega_k)}{\partial \omega_i}$.

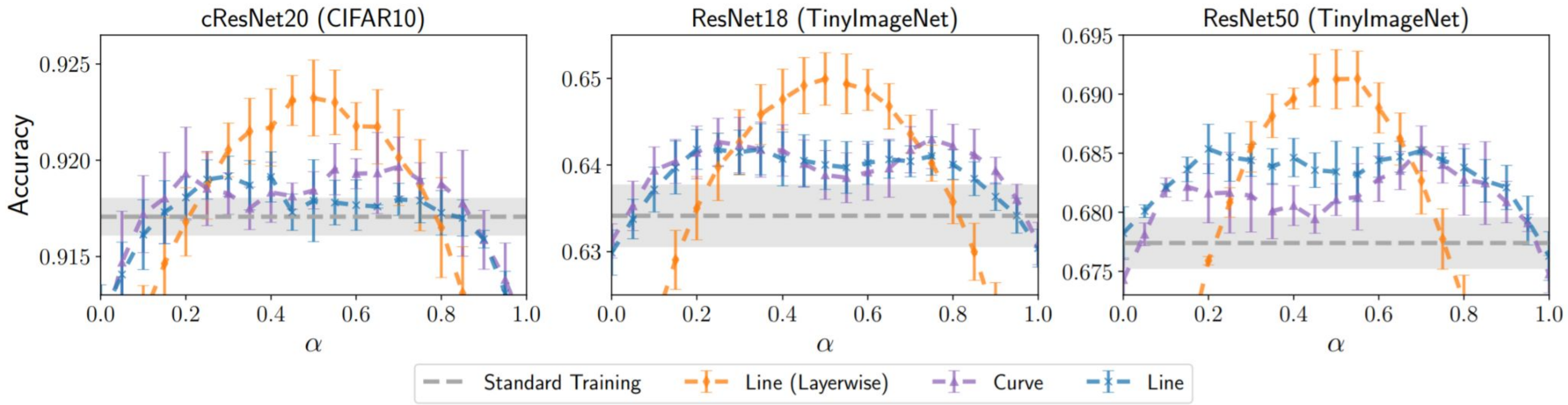
end for

Развитие подхода

Ранее мы считали веса модели просто столбцами.

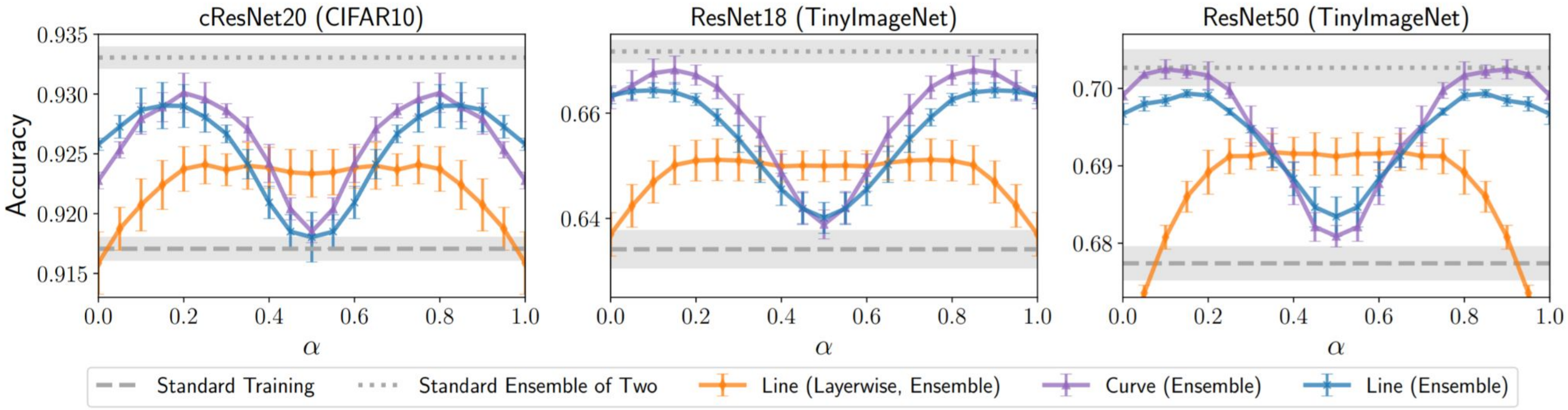
Авторы утверждают, что такой подход для обучения можно использовать отдельно для каждого слоя.

Эксперименты



Исследование качества “средней” модели

Эксперименты



Исследование построения ансамбля.

Тут первая модель берётся в точке P(α), а вторая в P(1-α).

Рецензия

Содержание: предложен и эмпирически изучен метод обучения подпространств в пространстве моделей.

Достоинства:

- Хороший обзор релевантной литературы
- Понятное изложение идеи, информативные графики
- Работы по изучению loss surface важны для всей области
- Доступен понятный код

Недостатки:

- Больше proof-of-concept, чем реальный метод
- Одна задача для тестирования
- Эмпирическая работа, нет теоретических обоснований

Оценка: 8. Уверенность: 4.

Исследование

- Статья была выпущена в 2021 году
- Первая версия на arxiv.org – февраль 2021 года
- Первый коммит на GitHub – март 2021 года
- Представлена на ICML 2021 в формате Poster

ICML
International Conference On Machine Learning

Tue Jul 20 09:00 PM -- 11:00 PM (PDT) Poster

Learning Neural Network Subspaces

Mitchell Wortsman · Maxwell Horton · Carlos Guestrin · Ali Farhadi · Mohammad Rastegari In Poster Session 2

[Visit Poster at Spot C1 in Virtual World]

Mitchell Wortsman » Maxwell Horton » Carlos Guestrin » Ali Farhadi » Mohammad Rastegari »

Recent observations have advanced our understanding of the neural network optimization landscape, revealing the existence of (1) paths of high accuracy containing diverse solutions and (2) wider minima offering improved performance. Previous methods observing diverse paths require multiple training runs. In contrast we aim to leverage both property (1) and (2) with a single method and in a single training run. With a similar computational cost as training one model, we learn lines, curves, and simplexes of high-accuracy neural networks. These neural network subspaces contain diverse solutions that can be ensembled, approaching the ensemble performance of independently trained networks without the training cost. Moreover, using the subspace midpoint boosts accuracy, calibration, and robustness to label noise, outperforming Stochastic Weight Averaging.

Исследование

Mitchell Wortsman

- BA Прикладная математика и компьютерные науки в Brown University in Providence, RI
- PhD студент в University of Washington
- Статья во время стажировки в Apple

Maxwell Horton

- BA Компьютерные науки и физика в The California Institute of Technology
- PhD в University of Washington
- Является сотрудником Apple
- Основная область -- Computer Vision

Исследование

Carlos Guestrin

- Senior Director of AI and Machine Learning в Apple
- Профессор в University of Washington и Stanford University
- Один из соавторов XGBoost

Ali Farhadi

- Науч.рук. Митчелла и Максвелла
- Исследователь AI и ML в Apple
- Профессор в University of Washington
- Один из соавторов YOLO

Mohammad Rastegari

- Senior Technival Leader of AI and Machine Learninng в Apple
- Профессор в University of Washington
- Ментор Митчелла в AI2

Исследование

Frankle et. al, 2020 [1], [2]; Garipov et al., 2018 [3]; Draxler et al., 2018 [4], Fort et al., 2019/20 [5], [6], [7]; Izmailov et al., 2018 [8].

Исследование

Цитирования:

- **[1] LCS: Learning Compressible Subspaces for Adaptive Network Compression at Inference Time**
  Elvis Nunez, Maxwell Horton, Anish Prabhu, Anurag Ranjan, Ali Farhadi, Mohammad Rastegari; *2021.*

- **[2] Subspace Learning for Personalized Federated Optimization**
  Seok-Ju Hahn, Minwoo Jeong, Junghye Lee; *2021.*

- **[3] Exploring the Power of Lightweight YOLOv4**
  Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Yung-Yu Chuang, Youn-Long Lin; *ICCV, 2021.*

- **[4] Towards Better Plasticity-Stability Trade-off in Incremental Learning: A simple Linear Connector**
  Guoliang Lin, Hanglu Chu, Hanjiang Lai; *2021.*

Исследование

Дальнейшее развитие

- Расширить область покрытия экспериментов
- Исследовать разные способы регуляризации
- Применение к другим задачам и областям DL