Состязательный метод дообучения нейронной сети в задаче переноса информации

Колесов А.С.

Московский Физико-Технический интитут
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем
Научный руководитель: к.ф.-м.н. Бахтеев О.Ю.

22.04.2022

Задача переноса информации

Цель

Предложить метод оптимизации параметров модели нейронной сети при помощи информации с другой модели глубокого обучения, обученной на схожей выборке.

Исследуемая проблема

Современные алгоритмы переноса информации нацелены на биективное соответствие параметров в моделях, тем самым теряя гибкость модели для обучения ее на новой выборке.

Метод Решения

Предлагается метод переноса информации , основанный на вероятностном подходе. Он обладает более быстрой сходимостью и использует меньший объем информации.

Постановка задачи переноса информации

Определение

Моделью глубокого обучения является $\mathbf{f}(\mathbf{x}, \mathbf{w})$ функция дифференцируемая по параметрам из множества признакового описания объектов во множество меток $\mathbf{f} : \mathbb{X} \times \mathbb{W} \to \mathbb{Y}$, где $\mathbb{W}$ - пространство параметров функции $\mathbf{f}$.

Определение

Множество объектов и их меток $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n : \mathbf{x}_i \in \mathbb{X}_s, \mathbf{y}_i \in \mathbb{Y}_s$ назовем выборкой-источником , данные которого доступны только при оптимизации модели глубокого обучения с некоторого произвольного начального положения $\mathbf{w}_0 \in \mathbb{W}$.

Определение

Множество объектов и их меток $\mathcal{T} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n : \mathbf{x}_i \in \mathbb{X}_t, \mathbf{y}_i \in \mathbb{Y}_t$ назовем целевой выборкой, данные которой доступны только при оптимизации модели глубокого обучения с фиксированного начального положения $\mathbf{w}_{fix} \in \mathbb{W}$.

Постановка задачи переноса информации

В качестве модели глубокого обучения рассматривается суперпозиция :

- $\mathbf{f}_{enc}(\mathbf{x}, \mathbf{w}) : \mathbb{X} \times \mathbb{W} \to \mathbb{Q}$,где $\mathbb{Q}$- скрытое пространство признаков модели-энкодера.
- $\mathbf{f}_{cl}(\mathbf{q}, \mathbf{w}) : \mathbb{W} \times \mathbb{Q} \to \mathbb{Y}$ - модель-классификатор.

На источнике и целевой выборках модель представима в виде суперпозиции:

$$\mathbf{f}^{src} = \mathbf{f}^{src}_{cl} \odot \mathbf{f}^{src}_{enc}, \qquad \mathbf{f}^{tgt} = \mathbf{f}^{tgt}_{cl} \odot \mathbf{f}^{tgt}_{enc}$$

Общий метод переноса информации

- Обучить $\mathbf{f}^{src}_{enc}$ на $\mathcal{S}$ выборке-источнике с произвольного начального положения $\mathbf{w}_0$ до фиксированного положения $\mathbf{w}_{fix}$.
- Провести оптимизацию по параметрам модели $\mathbf{f}^{tgt}_{enc}$ на $\mathcal{T}$ целевой выборке, взяв в качестве начального фиксированного положения $\mathbf{w}_{fix}$.
- Обучить $\mathbf{f}^{tgt}_{cl}$ на $\mathcal{T}$ с начального фиксированного положения $\mathbf{w}'_0 \in \mathbb{W}$.

Обзор существующих методов

Современные методы переноса информации в общем виде могут быть сформулированны как задача минимизации следующего функционала:

$$\min_{\mathbf{w}\in\mathbb{W}} \sum_{i=1}^{n} \mathcal{L}(\mathbf{f}_{cl}^{tgt}(\mathbf{f}_{enc}^{tgt}(\mathbf{x}_i)), \mathbf{y}_i) + \Omega(\cdot)$$

$\mathbf{w}$ -параметры модели, $\mathcal{L}(\cdot,\cdot)$ - функция потерь и $\Omega(\cdot)$ -регуляризация на параметры или выходы слоев модели.

Методы

- **L2-penalty :**

  $$\Omega(\mathbf{w}) = \alpha||\mathbf{w}^{tgt}||_2^2$$

  где α - гиперпараметр, контролирующий силу регуляризации.
- **L2-SP :** Метод регуляризации стремит параметры модели $\mathbf{f}_{enc}^{tgt}$ приблизить к параметрам $\mathbf{f}_{enc}^{src}$ по $L2$ метрике,

  $$\Omega(\mathbf{w}) = \beta||\mathbf{w}_{enc}^{tgt} - \mathbf{w}_{enc}^{src}||_2^2 + \alpha||\mathbf{w}_{cl}^{tgt}||_2^2.$$
- **DELTA :** Обозначим выходы слоев модели как $\mathbf{FM}_{enc}$:

  $$\Omega(\mathbf{w}) = \beta||\mathbf{FM}_{enc}^{tgt}(\mathbf{w}_{enc}^{tgt}) - \mathbf{FM}_{enc}^{src}(\mathbf{w}_{enc}^{src})||_2^2 + \alpha||\mathbf{w}_{cl}^{tgt}||_2^2.$$

Вероятностный подход

Цель вероятностного подхода

Оценка апостериорного распределения $p(\mathbf{w}|\mathbf{x})$ при помощи заданного априорного распределения $p(\mathbf{w})$:

$$p(\mathbf{w}|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w},\mathbf{x})p(\mathbf{w})}{\int_{\mathbb{W}} p(\mathbf{y},\mathbf{w}|\mathbf{x})d\mathbf{w}}.$$

Вариационный вывод

Пусть $q_{\phi^*}(\mathbf{w})$ - вариационное распределение, параметры которой $\phi \in \Phi$ минимизируют:

$$\phi^* = \arg\min_{\phi} KL(\mathbf{q}_\phi(\mathbf{w})||p(\mathbf{w}|\mathbf{y},\mathbf{x})).$$

Для вычисления оптимизируется вариационная нижняя оценка $\mathcal{L}(\phi)$:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\mathbf{w})} \log p(\mathbf{y}|\mathbf{w},\mathbf{x}) - KL(q_\phi(\mathbf{w})||p(\mathbf{w})) \to \max_{\phi}.$$

Альтернативная функция потерь

$KL(\mathbb{P}||\mathbb{Q})$ не является метрикой и запрашивает $\mathbb{P},\mathbb{Q} \in \mathcal{P}$, $\mathcal{P}$ - вероятностное пространство, в отличие от $\mathbb{W}_1(\mathbb{P},\mathbb{Q})$.

Расстояние Вассерштайна

Определение

Рассмотрим пространство $\mathbb{R}^D$ с метрикой $||\cdot||_2$. Пусть $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_1(\mathbb{R}^D)$, где $\mathcal{P}_1(\mathbb{R}^D)$ — множество вероятностных мер измеримых по Борелю с коненым первым моментом. Расстояние Вассерштйна-1 ($\mathbb{W}_1(\mathbb{P}, \mathbb{Q})$):

$$\mathbb{W}_1(\mathbb{P}, \mathbb{Q}) \overset{def}{=} \inf_{T\sharp\mathbb{P}=\mathbb{Q}} \int ||\mathbf{x} - T(\mathbf{x})||_2 d\mathbb{P}(\mathbf{x}),$$

где $T : \mathbb{R}^D \to \mathbb{R}^D$ измеримая функция (детерминистичный план).

Теорема

Пусть $\mathbb{P}$ и $\mathbb{Q} \in \mathcal{P}_1(\mathbb{R}^D)$. Пусть биективное соответствие в задаче переноса информации при L_2 регуляризации соответсвует детерминистичному плану $\tilde{\gamma}$. Тогда для оценки расстояния Вассерштайна $\tilde{W}_1(\mathbb{P}, \mathbb{Q})$ по плану $\tilde{\gamma}$ справедливо следующее соотношение

$$W_1(\mathbb{P}, \mathbb{Q}) \leq \tilde{W}_1(\mathbb{P}, \mathbb{Q}).$$

Двойственность Канторовича

Определение

Функцию $\mathbf{f}(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ будем называть строго 1-Липшицевой функцией и обозначать $||\mathbf{f}||_L = 1$, если

$$\forall \mathbf{x} \in \mathbb{R}^D \Rightarrow ||\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x})|| = 1$$

Теорема

Пусть $\mathbb{P}$ и $\mathbb{Q} \in \mathcal{P}_1(\mathbb{R}^D)$. Траснпортными лучами назовем прямые определяемые оптимальным планом $T(\boldsymbol{x})$ вида:

$$r = \boldsymbol{x}t + (1-t)T(\boldsymbol{x}), \quad t \in [0,1].$$

Тогда двойственная форма записи для $\mathbb{W}_1$:

$$\mathbb{W}_1(\mathbb{P}, \mathbb{Q}) = \sup_{||\boldsymbol{f}||_L = 1} [\int \boldsymbol{f}(x) d\mathbb{P}(x) - \int \boldsymbol{f}(y) d\mathbb{Q}(y)],$$

где двойственный потенциал $\boldsymbol{f}$ удовлетворяет условию $||\boldsymbol{f}||_L = 1$ и $\sup$ берется по классу строго 1-Липшицевых функций $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}$.

Предлагаемый метод

Введем следующие обозначения :

- $\mathbf{w}_{enc_j}^{src} \sim p_j(\mathbf{w})$ параметры в j-ом слое модели $\mathbf{f}_{enc}^{src}$.
- $\mathbf{w}_{enc_j}^{tgt} \sim q_j(\mathbf{w})$ параметры в j-ом слое модели $\mathbf{f}_{enc}^{tgt}$.
- $||\mathbf{f}_{j_{\phi_j}}(\mathbf{w}_j)||_L = 1$ — модель глубокого обучения на j-ом слое с оптимизируемыми параметрами ϕ_j, именуемая дискриминатором (критиком).
- Пара $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{T}$ — это пара объект и метка на целевой выборке.

Тогда задача оптимизации параметов модели $\mathbf{f}^{tgt}$ ставится как следующая мини-максная задача:

$$\max_{\phi} \min_{\mathbf{w}^{tgt}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}^{tgt}) + \sum_{j=1}^{J} \lambda_j [\mathbb{E}_{q_j(\mathbf{w})} \mathbf{f}_{j_{\phi_j}}(\mathbf{w}_{enc_j}^{tgt}) - \mathbb{E}_{p_j(\mathbf{w})} \mathbf{f}_{j_{\phi_j}}(\mathbf{w}_{enc_j}^{src})],$$

где λ_j является настраиваемым гиперпараметром для каждого слоя модели $\mathbf{f}$. Первое слагаемое соответствует функции потерь для обучения $\mathbf{f}^{tgt}$, второе слагаемое — двойственная задача Канторовича.

Строго 1-Липшицевы нейронные сети

Теорема

Рассмотрим функцию $\boldsymbol{f}: \mathbb{R}^D \to \mathbb{R}$*, представимую в виде* $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$*, где* $\boldsymbol{W}$ *некоторая матрица преобразования. Тогда* $||\nabla_x \boldsymbol{f}(\boldsymbol{x})||_2 = 1$*, если* $||\boldsymbol{W}||_2 = 1$.

Теорема

Рассмотрим функции $\boldsymbol{f}, \boldsymbol{g}: \mathbb{R}^D \to \mathbb{R}$ *такие, что :* $||\boldsymbol{f}||_L = 1, ||\boldsymbol{g}||_L = 1$. *Тогда следующие функции будут строго 1-Липшицевы :*

$$\max(\boldsymbol{f}, \boldsymbol{g}), \quad \min(\boldsymbol{f}, \boldsymbol{g}).$$

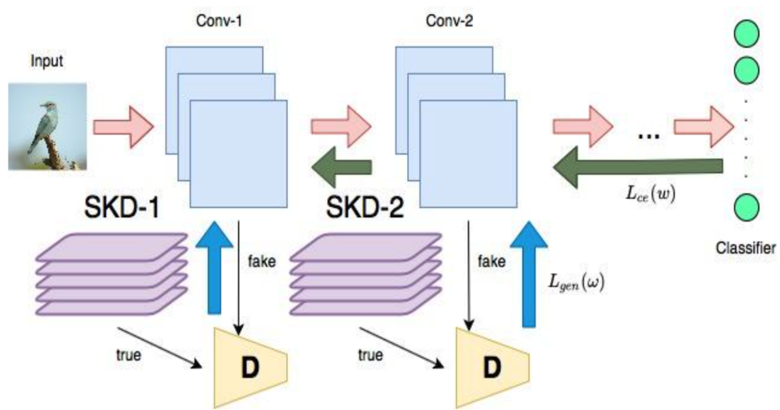
Теорема

Рассмотрим функции $\boldsymbol{f}, \boldsymbol{g}: \mathbb{R}^D \to \mathbb{R}$ *такие , что:* $||\boldsymbol{f}||_L = 1, ||\boldsymbol{g}||_L = 1$. *Тогда функция* $\boldsymbol{t}: \mathbb{R}^{2D} \to \mathbb{R}$*, определяемую как*

$$\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{y}) = \alpha \boldsymbol{f}(\boldsymbol{x}) + \beta \boldsymbol{g}(\boldsymbol{y}),$$

является строго 1-Липшицевой с коэффициентом $\alpha = \sqrt{1 - \beta^2}$.

Схема предлагаемого метода



Вычислительный эксперимент

Цель

Исследовать поведение модели глубокого обучения при переносе информации с другой модели. Сравнить предложенный метод с различными существующими подходами переноса информации.

Проведенно сравнение со следующими методами переноса информации:

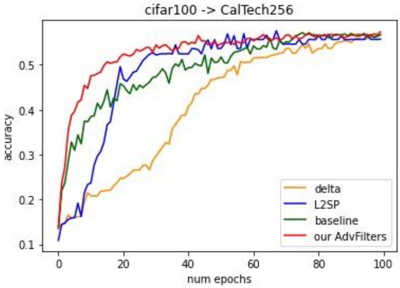
- L2-Penalty(baseline)
- L2-SP
- DELTA.

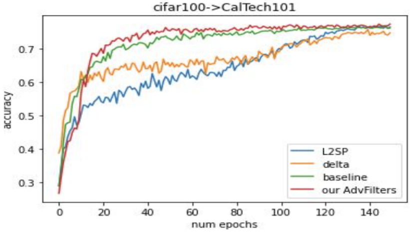
$\mathbf{f}^{src}$ обученная сверточная модель архитектуры ResNet-18 на выборке-источнике CIFAR-100 . Информация переносится на $\mathbf{f}^{tgt}$ сверточную нейронную сеть той же архитектуры обучаемую на целевых выборках CalTech-256 и CalTech-101.

Критерий качества модели

$$Accuracy(\mathbf{x}) = 1 - \frac{1}{m}\sum_{i=1}^{m}[\mathbf{f}^{tgt}(\mathbf{x}_i, \mathbf{w}^{tgt}) \neq \mathbf{y}_i].$$

Эксперименты на CalTech-256 и CalTech-101





Выносится на защиту

- Предложен и обоснован метод переноса информации, реегуляризатор которого является точной нижней оценкой между распределениями параметров моделей.
- Предложенны и теоретически обоснованны модели глубокого обучения, имеющие константу Липшица ровно 1.
- Проведены эксперименты для моделей глубокого обучения для различных целевых выборок, подтверждающие работоспособность предложенного метода.