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.

15.06.2022

Задача переноса информации

Цель

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

Исследуемая проблема

Современные алгоритмы нацелены на минимизацию информации между соответствующими параметрами моделей, тем самым теряя гибкость для обучения на целевой 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 которой 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.

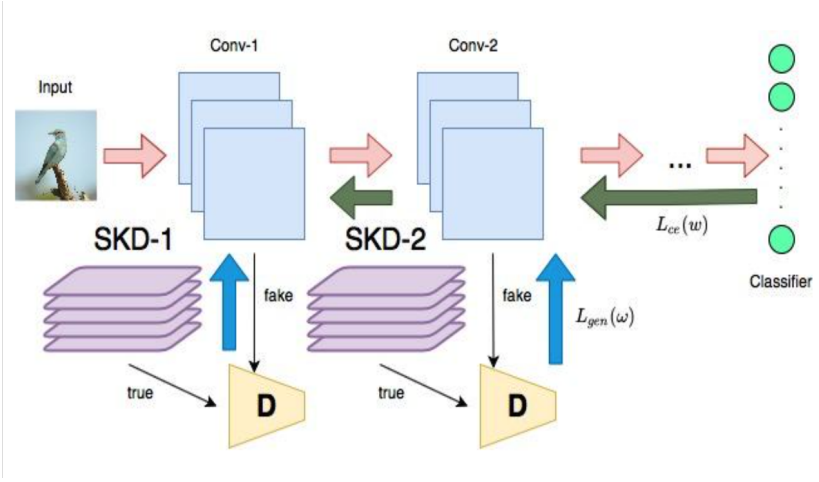
L2-SP : Метод регуляризации стремит параметры модели $\mathbf{f}_{enc}^{tgt}$ приблизить к параметрам $\mathbf{f}_{enc}^{src}$ по $\mathcal{L}^2$ метрике,

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

DELTA : Обозначим выходы слоев моделей-энкодеров как $\mathbf{FM}_{enc}$:

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

Предлагаемый метод



Предлагается **минимизировать расстояние между распределениями параметров** моделей $\mathbf{f}_{enc}^{src}$ и $\mathbf{f}_{enc}^{tgt}$ с примнением **генеративно-состязательного подхода**.

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 требует совпадения носителей вероятностных распределений $\mathbb{P},\mathbb{Q} \in \mathcal{P}$ ($\mathcal{P}$ — вероятностное пространство), в отличие от $\mathbb{W}_1(\mathbb{P},\mathbb{Q})$.

Расстояние Васерштейна

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 конечным первым моментом. Расстояние Васерштейна—1 ($\mathbb{W}_1(\mathbb{P}, \mathbb{Q})$):

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

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

Теорема (Колесов 1.1)

Пусть $\mathbb{P}$ и $\mathbb{Q} \in \mathcal{P}_1(\mathbb{R}^D)$. Пусть соответствие параметров между моделями f_{enc}^{src} и f_{enc}^{tgt} в задаче переноса информации при $\mathcal{L}^2$ регуляризации соответсвует детерминистичному транспортному плану $\tilde{\gamma}$. Тогда для смещенной оценки расстояния Васерштейна $\tilde{W}_1(\mathbb{P}, \mathbb{Q})$ по плану $\tilde{\gamma}$ справедливо следующее соотношение:

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

Двойственность Канторовича

Двойственная форма Канторовича для нахождения расстояния Васерштейна—1 между вероятностными мерами $\mathbb{P}$ и $\mathbb{Q}$ вводится одним из следующим трех способов с двойственными потенциалами $f(\cdot), g(\cdot) : \mathbb{R}^D \to \mathbb{R}$:

- $\mathbb{W}_1(\mathbb{P}, \mathbb{Q}) = \sup_{f \oplus g \leq ||\cdot||_2} \int_{\mathbb{R}^D} f(x) d\mathbb{P}(x) + \int_{\mathbb{R}^D} g(y) d\mathbb{Q}(y)$ (1)
- $\mathbb{W}_1(\mathbb{P}, \mathbb{Q}) = \sup_f \int_{\mathbb{R}^D} f(x) d\mathbb{P}(x) + \int_{\mathbb{R}^D} f^c(y) d\mathbb{Q}(y)$ (2)
- $\mathbb{W}_1(\mathbb{P}, \mathbb{Q}) = \sup_{||f||_L \leq 1} \int_{\mathbb{R}^D} f(x) d\mathbb{P}(x) - \int_{\mathbb{R}^D} f(y) d\mathbb{Q}(y)$ (3)

Указанные ограничения на **f** обеспечивают слабую локальную аппроксимацию расстояния Васерштейна, поскольку не гарантируют строгую 1—Липшицевость глобально. Предлагается разработать модели глубокого обучения, удовлетворяющие заявленному свойству.

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:

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

Строго 1—Липшицевы нейронные сети

Теорема (Колесов 2.1)

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

Теорема (Колесов 2.2)

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 :*

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

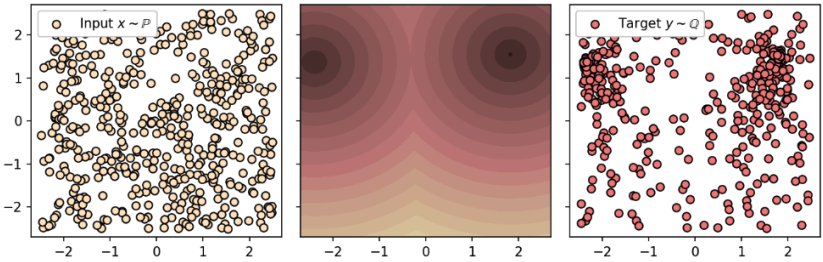
Теорема (Колесов 2.3)

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 $\alpha = \sqrt{1 - \beta^2}$.

Вычисление $\mathbb{W}_1$ расстояние Васерштейна



Движение точек в размерности 2. Слева направо изображенны точки исходного базого распределения $\mathbb{P} = 5\mathcal{U}(0,1) - 2.5$, поверхность строго 1—Липшицевой функции $\mathbf{f}(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ и конечное местоположение точек, описываемое некоторым неизвестным рапспределением $\mathbb{Q}$.

Теорема (Колесов 2.4)

Для вероятностных мер $\mathbb{P}, \mathbb{Q}$ выполняется соотношение $\mathcal{E}^2(\mathbb{P}, \mathbb{Q}) \leq \mathbb{W}_1(\mathbb{P}, \mathbb{Q}) \leq \mathcal{I}(\mathbb{P}, \mathbb{Q})$, где $\mathcal{I}(\mathbb{P}, \mathbb{Q}) \stackrel{def}{=} \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} \|\boldsymbol{x} - \boldsymbol{y}\|_2 d\mathbb{P}(\boldsymbol{x}) d\mathbb{Q}(\boldsymbol{y})$ является усредненным парным расстоянием между $\mathbb{P}, \mathbb{Q}$, и $\mathcal{E}^2(\mathbb{P}, \mathbb{Q}) \stackrel{def}{=} \mathcal{I}(\mathbb{P}, \mathbb{Q}) - \frac{1}{2}\mathcal{I}(\mathbb{P}, \mathbb{P}) - \frac{1}{2}\mathcal{I}(\mathbb{Q}, \mathbb{Q})$ является квадратом так называемого energy-distance.

Вычисление $\mathbb{W}_1$ расстояние Васерштейна

Dim	2	4	8	16	32	64	128
Lip-Clip	6.41	3.55	2.02	1.44	0.51	0.04	0.01
Lip-GP	9.25	7.66	6.69	5.05	3.25	2.45	1.77
Lip-LP	9.01	7.38	6.07	5.78	4.20	3.29	1.56
Lip-SN	11.27	10.03	7.94	6.32	5.28	3.19	1.26
Lip-SO	14.67	13.89	12.67	11.00	9.87	4.03	3.09
Lip-WP	14.89	14.21	13.96	11.22	10.07	7.68	3.35
Reg	12.05	10.31	7.99	4.08	2.13	0.11	0.02
MM:B	14.98	13.43	10.24	5.16	1.28	0.19	0.06
MM	14.21	13.21	12.21	11.08	8.42	7.29	6.26
Lower	5.22	4.98	4.57	4.18	3.21	2.39	2.06
True (Our)	15.11	14.33	13.29	13.18	12.22	11.29	11.26
Upper	15.50	17.31	17.98	19.23	20.44	22.29	25.67

Сравнение оценок $\mathbb{W}_1$ расстояния с методами WGANs и нашим предложенным методом для несмещенного оценивания, а также с верхней и нижней оценками на расстояние Васерштейна согласно теореме 2.4.

Предлагаемый метод переноса информации

Введем следующие обозначения :

$\mathbf{w}^{src}_{enc_j} \sim p_j(\mathbf{w})$ параметры в j-ом слое модели $\mathbf{f}^{src}_{enc}$.

$\mathbf{w}^{tgt}_{enc_j} \sim q_j(\mathbf{w})$ параметры в j-ом слое модели $\mathbf{f}^{tgt}_{enc}$.

$||\mathbf{f}_{j_{\phi_j}}(\mathbf{w}_j)||_L = 1$ — 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:

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

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.

Эксперимент метода переноса информации

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:

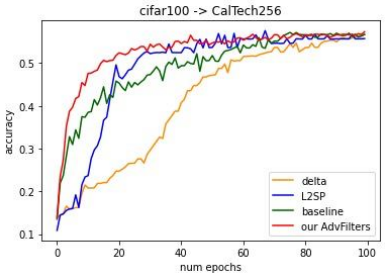
- L2-Penalty (baseline)
- L2-SP
- DELTA

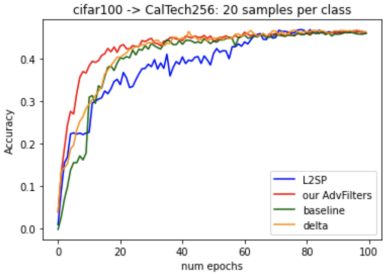
В качетсве выборки-источника $\mathcal{S}$ используется датасет CIFAR-100, на котором обученна модель $\mathbf{f}^{src}$. Информация переносится на модель $\mathbf{f}^{tgt}$ обучаемую на целевых выборках CalTech-256 и CalTech-101 с объемами тренировочных объектов 20 и 60 объектов соответственно.

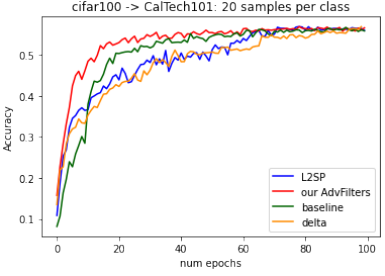
Критерий качества модели

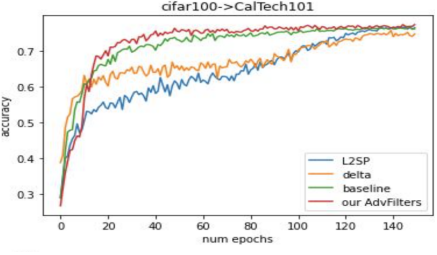
$$Accuracy(\mathbf{x}) = 1 - \frac{1}{m}\sum_{i=1}^{m}[\mathbf{f}^{tgt}(\mathbf{x}_i, \mathbf{w}^{tgt}) \neq \mathbf{y}_i].$$

Эксперименты на CalTech-256 и CalTech-101









Эксперименты на CalTech-256 и CalTech-101

Используется метрика AUC, представляющая собой площадь под кривой точности метода, характеризующая скорость сходимости метода относительно других. Для вышеупомянутых экспериментов вычисляются значения AUC метрики, которые приведенны в таблице ниже.

Experiment	CalTech101:60	CalTech101:20	CalTech256:60	CalTech256:20
L2-SP	67.45	49.74	50.51	38.68
Delta	67.76	49.58	27.77	41.33
Baseline	71.09	51.77	52.94	40.04
AdvFilters (Our)	72.28	53.18	52.94	42.57

Сранение AUC metric для методов переноса информации с нашим предложенным алгоритмом в 4 вышеупомянутых экспериментах.

Выносится на защиту

– Предложен и обоснован метод переноса информации, регуляризатор которого является точной нижней оценкой расстояния Вассерштайна между распределениями параметров моделей.

– Предложенны и теоретически обоснованны модели глубокого обучения, имеющие константу Липшица ровно 1.

– Предложен метод для вычисления несмещенной оценки расстояния Вассерштайна и сравнение его с текущуими методами.

– Проведены эксперименты для моделей глубокого обучения для различных целевых выборок, подтверждающие работоспособность предложенного метода.