Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Олейник Михаил Сергеевич

# ДИСТИЛЛЯЦИЯ ЗНАНИЙ В ГЛУБОКИХ СЕТЯХ С ПРИМЕНЕНИЕМ МЕТОДОВ ВЫРАВНИВАНИЯ СТРУКТУР МОДЕЛЕЙ

01.03.02 — Прикладная математика и информатика

Выпускная квалификационная работа бакалавра

**Научный руководитель:**
Бахтеев Олег Юрьевич
канд. физ.-мат. наук

Москва — 2024

## Аннотация

Дистилляция знаний позволяет повысить качество модели, называемой учеником, не увеличивая её число параметров, а используя модель большего размера, называемой учителем. Однако, в случае разных архитектур и несовпадения количества слоев у учителя и ученика, распространённые методы не применимы. Одним из подходов, который позволяет решать задачу для разных архитектур, является максимизация взаимной информации. Мы предлагаем улучшение этого подхода, которое позволит проводить дистилляцию и для моделей с разным количеством слоёв. Мы сравниваем наш метод с остальными с помощью вычислительного эксперимента. Также проводим анализ гиперпараметров и выводим ограничения на них, при которых достигается наибольшее качество.

## Содержание

# 1 Введение

Глубокие нейронные достигли больших успехов в задачах машинного зрения, обработки естественного языка и других. Однако, лучшие результаты достигают модели с большим количеством параметров, из-за этого их трудно встроить в системы с небольшими вычислительными мощностями, например, мобильные телефоны. Если подобрать размер модели под целевую платформу, уменьшив количество параметров, то сильно потеряем и в качестве.

Одним из подходов, которые позволяют не теряя сильно в качестве, получить модель с меньшим количеством параметров, является дистилляция знаний. Этот подход использует большую предобученную на необходимой задаче модель, называемую учителем, данные о слоях которой переносятся определенным образом в модель меньшего размера, называемую учеником. Перенос чаще всего выражается в дополнительном слагаемом в функции потерь ученика.

Так, в работе [1] предлагается переносить знания с последнего слоя модели. К недостаткам этого метода можно отнести то, что мы игнорируем информацию из остальных слоев учителя, а она может быть ценной. В работах ...

Однако, большинство подходов либо неэффективно работают, либо не могут быть применимы к случаям, когда модели имеют разное количество слоёв или разную архитектуру. Также возникают сложности в случае, когда количество параметров в слое учителя сильно больше, чем в соответствующем слое ученика, как показано в работе [2].

Больший интерес представляют подходы, которые можно применить, если учитель и ученик имеют разные архитектуры. В работе [3] моделируется информационный поток в учителе, который имитирует ученик. В работе [4] используется максимизация взаимной информации между парами соответствующих слоёв. В основе этого метода используется вариационный подход [5]. Наша работа предлагает улучшение данного метода, давая возможность проведения дистилляции при разном количестве слоев у учителя и ученика. Также мы проводим анализ гиперпараметров алгоритма, на

примере работы [6].

# 2 Постановка задачи

## 2.1 Задача дистилляции с взаимной информацией

Рассмотрим задачу классификации на $K$ классов:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m, \mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{Y} = \{1, \ldots, K\},$$

где $\mathfrak{D}$ это доступная выборка объектов, $y_i$ — целевая переменная, а $\mathbf{x}_i$ - данные, описывающие объект, взятые из распределения $p(\mathbf{x})$.

В задаче дистилляции нам необходима кроме обучаемой модели, так называемого ученика, еще модель учителя, которая уже предобучена на такой же задаче, и параметры которой не меняются в процессе обучения. Обозначим $i$-й слой модели учителя как $\mathcal{T}^{(i)}$, и $j$-й слой модели ученика как $\mathcal{S}^{(j)}$. Пропуская же входные данные $\mathbf{x}$ через модель, мы можем после каждого слоя получить его активации. Обозначим активации после $i$-го слоя модели учителя как $\mathbf{t}^(i)$, а активации после $j$-го слоя модели ученика как $\mathbf{s}^(j)$.

Взаимная информация пары $(\mathbf{t}, \mathbf{s})$ определена как:

$$I(\mathbf{t}; \mathbf{s}) = H(\mathbf{t}) - H(\mathbf{t}|\mathbf{s}) = -\mathbb{E}_{\mathbf{t}}[\log p(\mathbf{t})] + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log p(\mathbf{t}|\mathbf{s})],$$

где энтропия $H(\mathbf{t})$ и условная энтропия $H(\mathbf{t}|\mathbf{s})$ получены из совместного распределения $p(\mathbf{t},\mathbf{s})$. Определение $I(\mathbf{t}; \mathbf{s})$ можно понимать, как уменьшение неопределенности в знаниях учителя, которые закодированны в его слое $\mathbf{t}$, когда известен слой $\mathbf{s}$ ученика.

Теперь, можем определить функцию потерь для модели ученика, минимизируя которую ученик будет не только обучаться для задачи классификации, но и будет максимизироваться взаимная информация между слоями учителя и ученика:

$$\mathcal{L} = \beta \mathcal{L}_{\text{CE}} - (1 - \beta) \sum_{i=1}^{T} \sum_{j=1}^{S} \lambda_{i,j} I(\mathbf{t}_i, \mathbf{s}_j),$$

где $\mathcal{L}_{\text{CE}}$ – кросс-энтропия, $T$ — количество слоёв учителя, $S$ — количество слоёв ученика, $\beta \in (0; 1)$ — гиперпараметр, отвечающий за баланс между минимизацией кросс-энтропии и максимизацией взаимной информации

между слоями учителя и ученика, $\lambda_{i,j} > 0$ — гиперпараметр, отвечающий за важность связи $i$-го слоя учителя и $j$-го слоя ученика.

### 2.2 Вариационная максимизация информации

Функция потерь должна быть минимизирована относительно параметров модели ученика. Однако, сделать это будет сложно, так как трудно вычислить взаимную информацию.

Вместо этого используется вариационная нижняя граница для каждого члена взаимной информации $I(\mathbf{t};\mathbf{s})$, в которой определяется вариационное распределение $q(\mathbf{t}|\mathbf{s})$, которое аппроксимирует $p(\mathbf{t}|\mathbf{s})$:

$$\begin{aligned} I(\mathbf{t};\mathbf{s}) = H(\mathbf{t}) - H(\mathbf{t}|\mathbf{s}) = H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log p(\mathbf{t}|\mathbf{s})] \\ + H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})] + \mathbb{E}_{\mathbf{s}}[D_{\mathrm{KL}}(p(\mathbf{t}|\mathbf{s})||q(\mathbf{t}|\mathbf{s}))] \\ \geq H(\mathbf{t}) + \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})]. \quad (1) \end{aligned}$$

Данная техника известна как вариационная максимизация информации. Применяя её к каждому члену взаимной информации в функции потерь, получим:

$$\mathcal{L} = \beta \mathcal{L}_{\mathrm{CE}} - (1-\beta) \sum_{i=1}^{T} \sum_{j=1}^{S} \lambda_{i,j} \mathbb{E}_{\mathbf{t},\mathbf{s}}[\log q(\mathbf{t}^{(i)}|\mathbf{s}^{(j)})].$$

Эта новая функция потерь минимизируется относительно параметров модели ученика и вариационного распределения $q(\mathbf{t}|\mathbf{s})$. Стоит обратить внимание, что энтропия $H(\mathbf{t})$ убрана из выражения, так как не зависит от оптимизируемых параметров.

ДОПИСАТЬ О виде $q(\mathbf{t}|\mathbf{s})$.

### 2.3 Двухуровневая задача оптимизации

До этого момента мы рассматривали задачу как обычную задачу оптимизации с гиперпараметрами. Однако, в таком подходе возникают сложности с подбором значений гиперпараметров. Они могут быть заданы какими-то наивными соображениями, быть подобраны с помощью полного перебора или с использованием вероятностных моделей. Однако, лучшее качество

показывают методы, которые основаны на градиентном спуске, при этом с использованием разностной аппроксимации возможно сохранить качество, сильно уменьшив сложность метода.

Разобьём нашу выборку на тренировочную и валидационную: $\mathfrak{D} = \mathfrak{D}_{\text{train}} \bigsqcup \mathfrak{D}_{\text{test}}$. Обозначим $\mathcal{L}_{\text{train}}$ как функцию потерь $\mathcal{L}$, вычисляемую на выборке $\mathfrak{D}_{\text{train}}$, а $\mathcal{L}_{\text{val}}$ — как функцию потерь $\mathcal{L}$, вычисляемую на выборке $\mathfrak{D}_{\text{val}}$.

Определим вектор $\boldsymbol{\lambda}$ из всех гиперпараметров задачи:

$$\boldsymbol{\lambda} = [\lambda_{0,0}, \ldots, \lambda_{i,j}, \ldots, \beta].$$

Определим все обучаемые параметры модели ученика и обучаемые параметры взаимной информации как $\mathbf{w}$.

Заметим, что функции потерь $\mathcal{L}_{\text{train}}$ и $\mathcal{L}_{\text{val}}$ зависят и от $\boldsymbol{\lambda}$, и от $\mathbf{w}$. Цель — найти $\hat{\boldsymbol{\lambda}}$, которое минимизирует функцию потерь на валидационной выборке $\mathcal{L}_{\text{val}}(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}})$, где параметры $\mathbf{w}$ получаются в результате минимизации функции потерь на тренировочной выборке $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \hat{\boldsymbol{\lambda}})$.

И это определяет задачу двухуровневой оптимизации:

$$\begin{aligned} &\min_{\boldsymbol{\lambda}} \quad \mathcal{L}_{\text{val}}(\hat{\mathbf{w}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \\ &\text{s.t.} \quad \hat{\mathbf{w}}(\boldsymbol{\lambda}) = \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}) \end{aligned}$$

Используем аппроксимацию (РАСПИСАТЬ ПОДРОБНЕЕ):

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\text{val}}(\hat{\mathbf{w}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \approx \mathcal{L}_{\text{val}}(\mathbf{w} - \xi \nabla_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\lambda}), \boldsymbol{\lambda}),$$

где .... Идея заключается в ....

## 3 Вычислительный эксперимент

Эксперименты

# 4 Заключение

Заключение

## Список литературы


[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[3] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2020.

[4] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

[6] M Gorpinich, O Yu Bakhteev, and VV Strijov. Gradient methods for optimizing metaparameters in the knowledge distillation problem. *Automation and Remote Control*, 83(10):1544–1554, 2022.