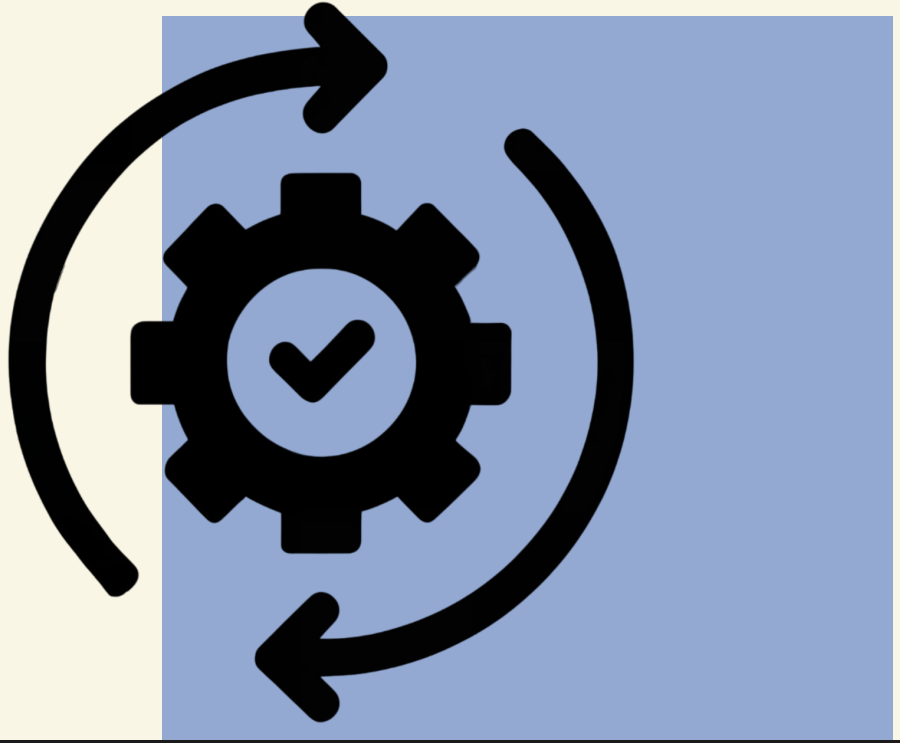
Эффективные системы машинного обучения



Лекция 1

“Правила игры и дистилляция знаний”

Преподаватель

Феоктистов Дмитрий

20 сентября 2024

ВМК МГУ

Команда курса

- **Оганов Александр** (ВМК МГУ)
- **Феоктистов Дмитрий** (ВМК МГУ, Yandex Research)
- **Алексеев Илья** (ВМК МГУ, Лаборатория нейронных систем и глубокого обучения МФТИ)

Что такое эффективность?

Эффективность – соотношение между достигнутым результатом и использованными ресурсами

Что такое эффективность?

Эффективность – соотношение между достигнутым результатом и использованными ресурсами

В рамках курса мы **оптимизируем** такие ресурсы:
–**Компьют**
–Умственную активность

А также увеличиваем результаты, правильно **учитывая структуру задачи**

Программа курса

01

Методы сжатия нейронных сетей

-Дистилляция знаний

-Квантизаци и спарсификация

-Parameter efficient fine tuning

Феоктистов Дмитрий

02

Работа с неклассическими типами данных в глубоком обучении

-Neural Architecture Search
-Временные ряды
-Рекомендательные системы
-Сверточные сети в речи
-Машинное обучение на графах
-Машинное обучение на табличных данных

Алексеев Илья

Феоктистов Дмитрий

03

Современные генеративные модели

-Введение в диффузионные модели
-Дистилляция диффузионных моделей
-Сэмплирование в диффузионных моделях - Применение генеративных моделей

Оганов Александр

Правила игры

Оценка ставится по следующим правилам:

Одна зачтенная домашка – хорошо. Две зачтенные домашки – отлично. Домашка считается зачтенной, если работа проведена качественно (мы понимаем, что не все может получиться), и **мы понимаем, каков ваш вклад в работу.**

Домашка 1

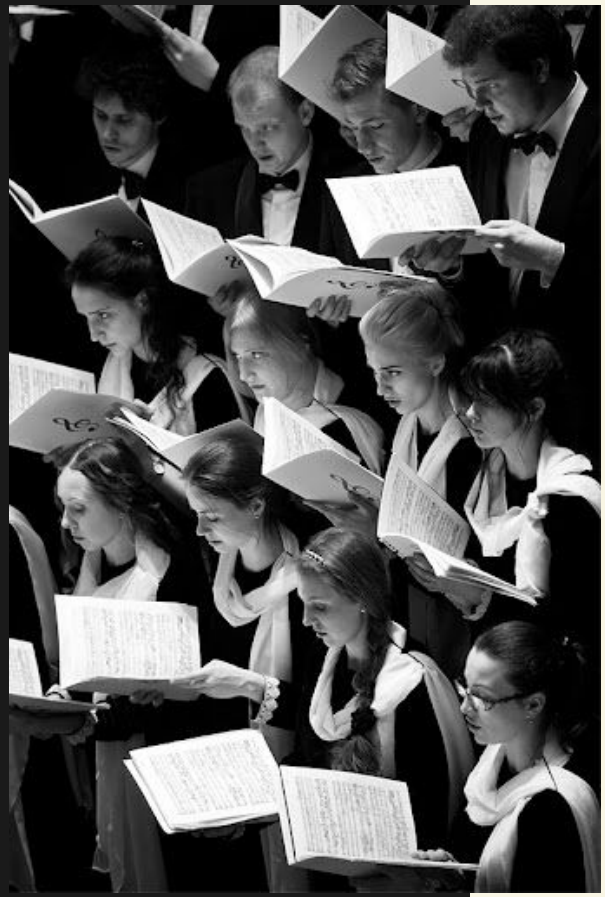
Выполняется в командах размером до 4х человек. Тема домашки: методы сжатия нейронных сетей. Можно брать чужой код, если указывать источники. Желательно брать код из статей, а не рандомный. **Нельзя списывать друг у друга**. Формат сдачи: гитхаб + отчет до 6 страниц в формате статьи. Время на выполнение – 3 недели

Примерная дата выдачи: 18.10.2024

Домашка 2

Комбинирование различных подходов к обработке особенностей данных для решения прикладной задачи. Условия аналогичные

Примерная дата выдачи: 15.11.2024



Пояснения

Указанные ниже пункты не являются строгими критериями, а лишь примерами, на зачет работы влияет общее впечатление. То есть если каждый участник выполнил ровно один пункт и больше ничего, то это не гарантирует получение зачета за домашку. Однако **жестить мы не будем, если хорошо постараться, то работа точно будет зачтена**

Red flags

- Все сделал один участник (ему зачет, остальным нет)

-Было выполнено мало пунктов задания (меньше пункта на человека)

-В отчете числа не соответствую действительности (эксперименты не воспроизводятся)

Green flags

-Каждый участник выполнил один или больше пункт задания, и это видно по отчету и коммитам

-Работа целостная, сделаны выводы по экспериментам, приложены графики/числа

-Если что-то не получилось сделать, но была выполнена попытка, то об этом написано

Дистилляция знаний

1. Мотивация
2. Вывод формул
3. Разбор кода
4. Какие знания дистиллируем?
5. Современные результаты

Мотивация

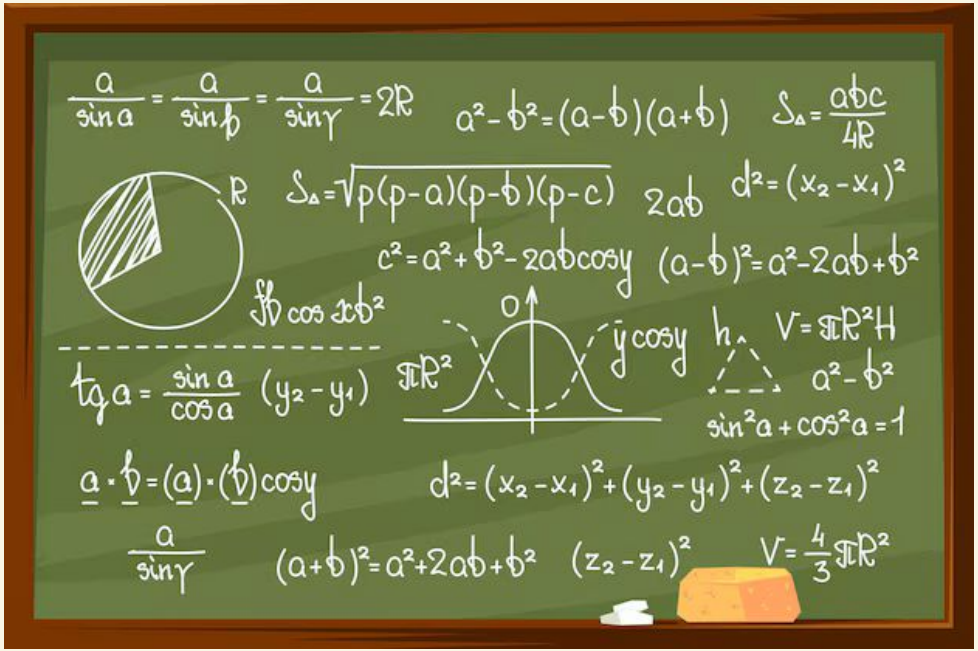
–Мотивация из 2015 – ансамбли это круто, но вот DL модельки большие, тут одну учить по полгода, вот бы выучить одну большую, а из нее получить много маленьких, которые сможем заансамблировать

–Мотивация из 2024 – нейронные сети везде, хочется даже самую последнюю ChatGPT запускать не телефоне, нужно как-то сжимать. Но вот ChatGPT закрытая, поэтому всякие квантизации нам не подходят. Вот бы просто учиться на ее выходах!

–Даже если моделька открытая и есть доступ к весам, то иногда может быть желание не сжать веса модельки, а перетащить знания из нее в модельку поменьше

Hinton et al, Distilling the knowledge in a Neural Network, 2015

Формулы



Смотрите на доске

Код

Ноутбук с кодом тут ↑

Эффективные системы машинного обучения, ВМК МГУ

Занятие 1. Дистилляция знаний

Феоктистов Дмитрий

20 сентября 2024

Какие знания дистиллируются?

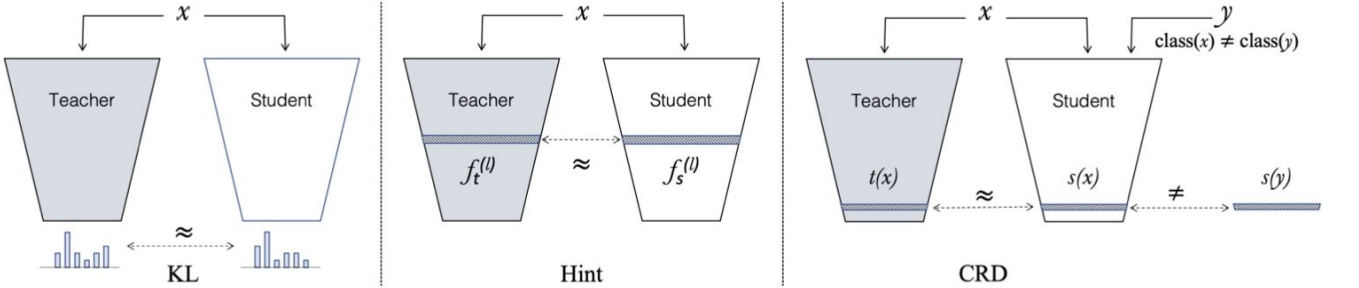


Figure 1: Methods used in this work. (i) KL [12]: mimicking of class probabilities. (ii) $Hint$ [26]: mimicking of features at an intermediate layer. (iii) CRD [30]: features from the student and teacher for the same image constitute a positive pair, and those from different classes make up a negative pair.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Какие знания дистиллируются?

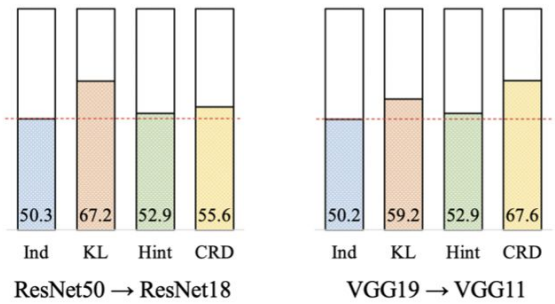


Figure 2: **Left:** An example of how the distilled student can focus on similar regions as the teacher while classifying an image. **Right:** % where teacher's CAM is more similar to the distilled student's CAM than to the independent student's CAM. The red line indicates chance performance (50%).

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Какие знания дистиллируются?

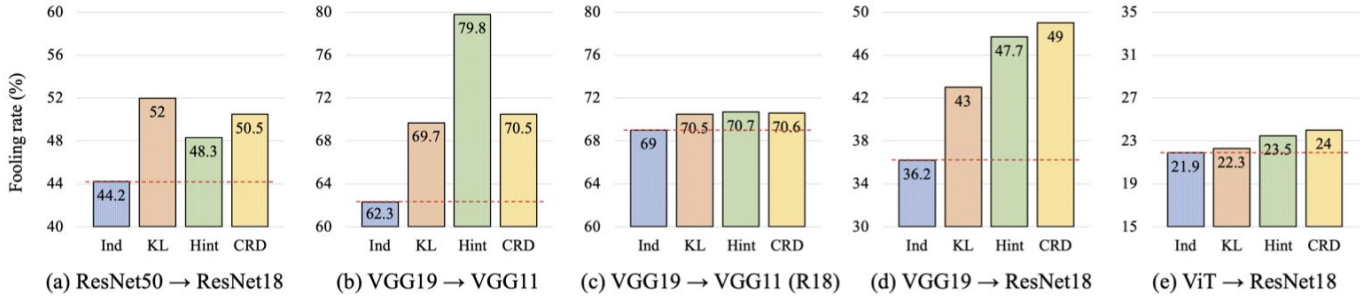


Figure 3: The images which fool the teacher fool the distilled student more than the independent student in **(a)**, **(b)** and **(d)**, but not in **(e)**. If the adversarial attack is generated using a foreign network (ResNet18, **(c)**), it fails to convincingly fool the distilled student more than the independent one.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Какие знания дистиллируются?

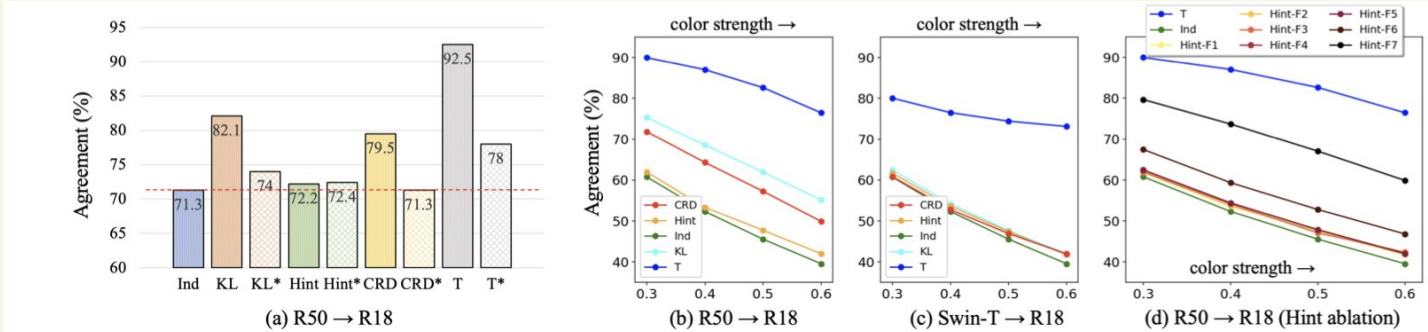


Figure 4: **(a)** Agreement between two images with different color properties. * indicates distillation done by a teacher T* not trained to be color invariant. **(b, c)** Agreement between two images having increasingly different color properties. **(d)** Effect of different layers used in *Hint* distillation.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Эффективные системы машинного обучения, ВМК МГУ

Занятие 1. Дистилляция знаний

Феоктистов Дмитрий

20 сентября 2024

Какие знания дистиллируются?

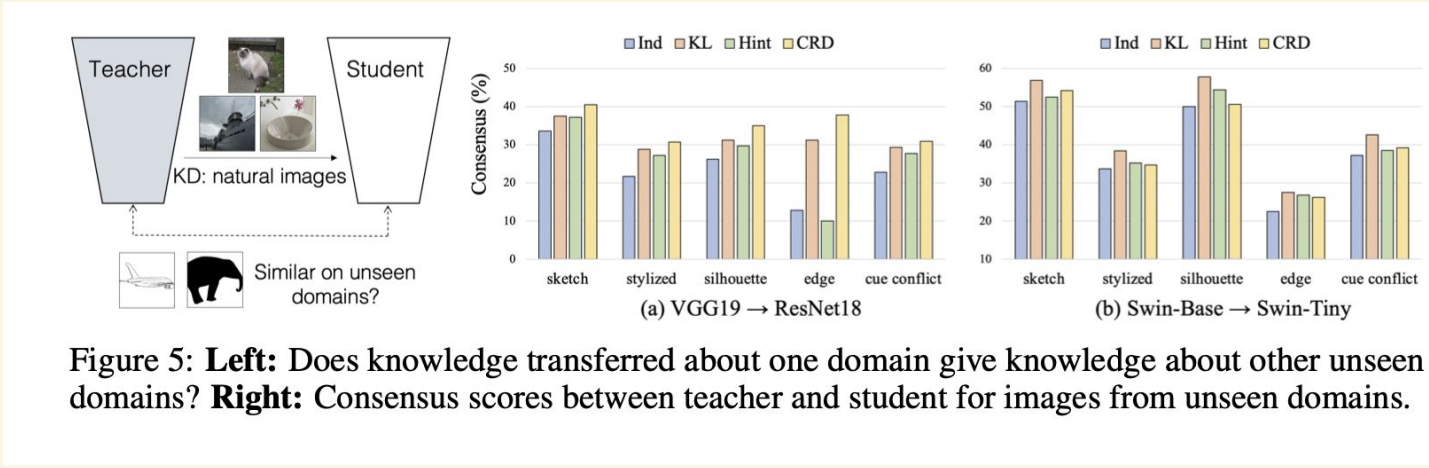


Figure 5: **Left:** Does knowledge transferred about one domain give knowledge about other unseen domains? **Right:** Consensus scores between teacher and student for images from unseen domains.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Какие знания дистиллируются?

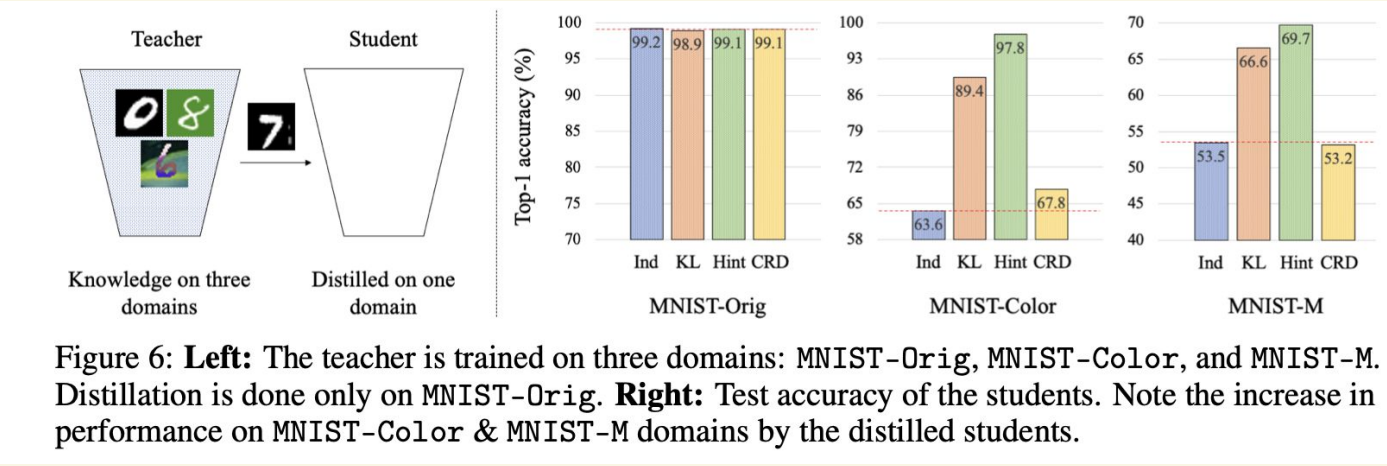


Figure 6: **Left:** The teacher is trained on three domains: `MNIST-Orig`, `MNIST-Color`, and `MNIST-M`. Distillation is done only on `MNIST-Orig`. **Right:** Test accuracy of the students. Note the increase in performance on `MNIST-Color` & `MNIST-M` domains by the distilled students.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Какие знания дистиллируются?

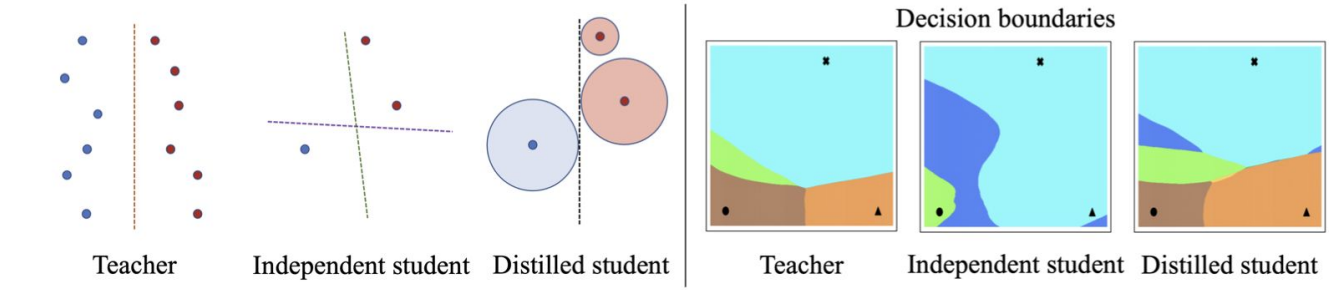
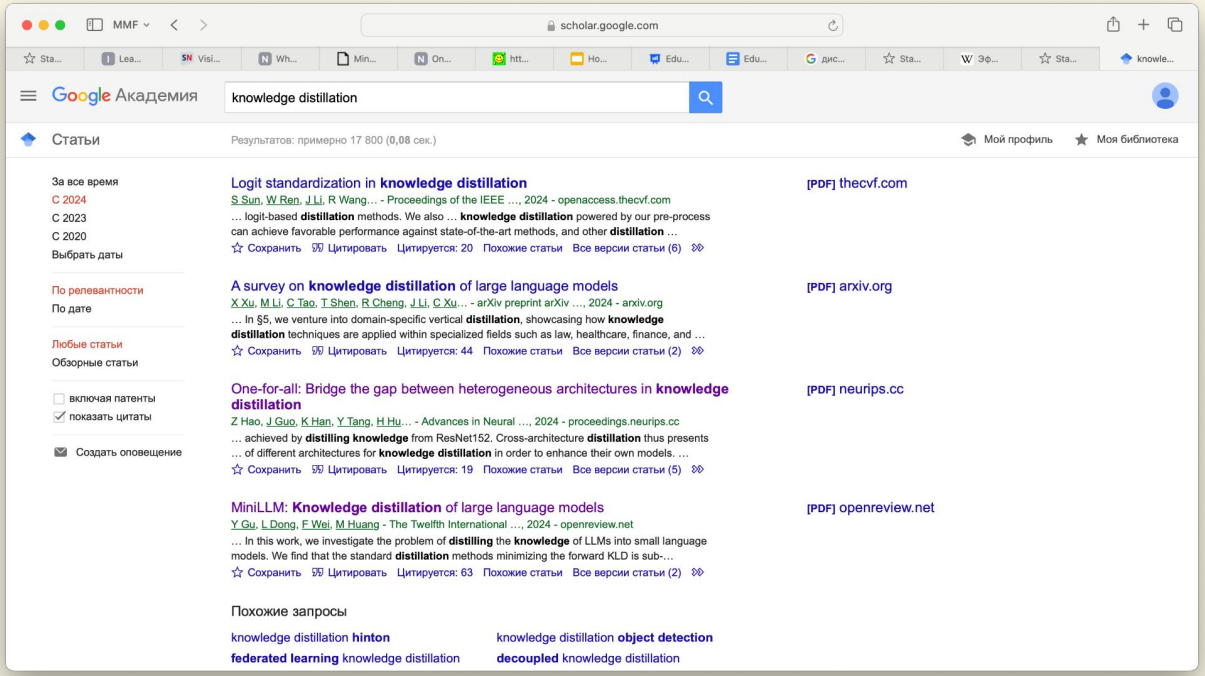


Figure 7: **Left.** Teacher learns on the the entire dataset. Independent student without access to the full dataset might learn some spurious correlations. Distillation provides constraints to reconstruct the teacher's decision boundary. **Right.** Sampled decision boundaries for points from MNIST-Color visualized using [28]. The distilled student is better at reconstructing the teacher's decision boundary.

Utkarsh Ojha et al, What knowledge Gets Distilled in Knowledge Distillation, NeurIPS 2023

Современные результаты



Эффективные системы машинного обучения, ВМК МГУ

Занятие 1. Дистилляция знаний

Феоктистов Дмитрий

20 сентября 2024

MiniLLM

-*Black-box* KD, where only the teacher-generated texts are accessible, and *white-box* KD, where the teacher model's output distribution or intermediate hidden states are also available

-*Black-box* KD has shown promising results in fine-tuning small models on the prompt-response pairs generated by LLM APIs

-However, *white-box* KD approaches are mostly studied for small (< 1B parameters)

Yuxian Gu et al, MiniLLM: Knowledge distillation of large language models, ICLR 2024

MiniLLM

For text classification tasks, KL[p||q_θ] works well because the output space usually consists of a finite number of classes such that both p(y|x) and q_θ(y|x) have few mode. For open-ended text generation tasks, the output spaces are much more complex, and p(y|x) can contain many more modes than what q_θ(y|x) can express due to the limited model capacity. Minimizing *forward* KLD causes q_θ to assign unreasonably high probabilities to the void regions of p and produces very unlikely samples under p during free-run generation. Compared to KL[p||q_θ], minimizing KL[q_θ||p] causes q_θ to seek the major modes of p, and assign low probabilities to p's void regions

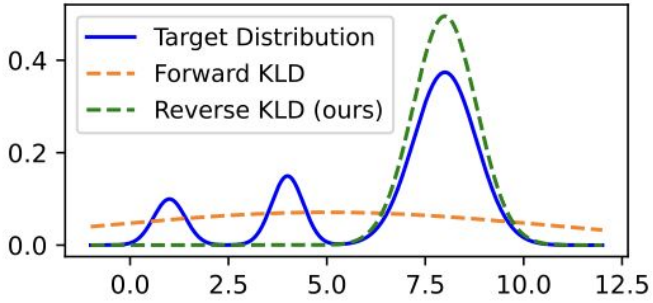


Figure 2: We fit a Gaussian mixture distribution with a single Gaussian distribution using *forward* KLD and *reverse* KLD.

Yuxian Gu et al, MiniLLM: Knowledge distillation of large language models, ICLR 2024

MiniLLM

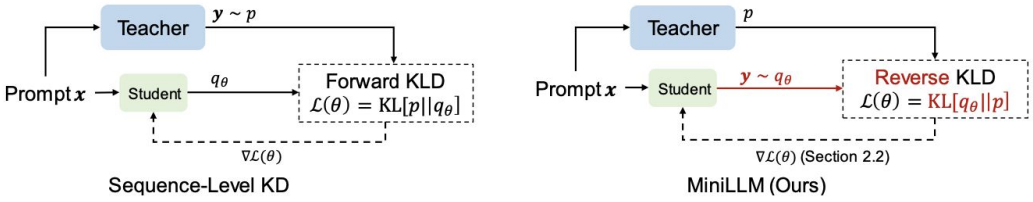


Figure 3: Comparison between sequence-level KD (left) and MINILLM (right). Sequence-level KD forces the student to memorize all samples generated by the teacher model, while MINILLM improves its generated texts with the teacher model's feedback.

To optimize $\min_\theta \mathrm{KL}[q_\theta||p]$ we derive the gradient of the objective with Policy Gradient. To further stabilize and accelerate training, we propose single-step decomposition to reduce variance, teacher-mixed sampling to alleviate reward hacking, and length normalization to eliminate the length bias (We found that long sequences tend to have small R_{t+1}, which encourages the model to produce short response).

Yuxian Gu et al, MiniLLM: Knowledge distillation of large language models, ICLR 2024

MiniLLM

OPT	13B	Teacher	29.2	18.4	17.8	30.4	36.1
	1.3B	SFT w/o KD	26.0	11.4	15.6	23.1	28.4
		KD	25.4	12.2	14.9	21.9	27.0
		SeqKD	26.1	12.7	16.6	21.4	28.2
		MINILLM	**26.7**	**14.8**	**17.9***	**28.6**	**33.4**
	2.7B	SFT w/o KD	27.1	13.9	16.6	24.9	32.3
		KD	25.9	13.8	16.7	26.3	30.2
		SeqKD	27.5	13.3	16.5	25.3	32.3
		MINILLM	**27.4**	**17.2**	**19.1***	**30.7***	**35.1**
	6.7B	SFT w/o KD	27.6	16.4	17.8	30.3	28.6
		KD	28.3	17.0	17.5	30.7*	26.7
		SeqKD	28.5	17.0	17.9*	30.4	28.2
		MINILLM	**29.0**	**17.5**	**18.7***	**32.5***	**36.7***
LLaMA	13B	Teacher	29.7	23.4	19.4	35.8	38.5
	7B	SFT w/o KD	26.3	20.8	17.5	32.4	35.8
		KD	27.4	20.2	18.4	33.7	37.9
		SeqKD	27.5	20.8	18.1	33.7	37.6
		MINILLM	**29.0**	**23.2**	**20.7***	**35.5**	**40.2***

We report the average Rouge-L scores across 5 random seeds on some popular LLM evaluation datasets (Dolly, SelfInst, Vicuna, S-NI, UnNI)

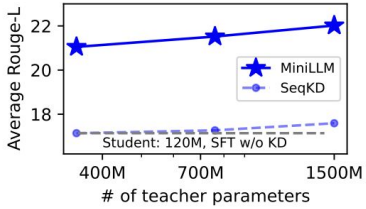


Figure 5: The scaling law of teacher based on the GPT-2 family models. We compare MINILLM and SeqKD with GPT-2-125M as the student and GPT-2 340M, 760M, and 1.5B as teachers.

Yuxian Gu et al, MiniLLM: Knowledge distillation of large language models, ICLR 2024

Эффективные системы машинного обучения, ВМК МГУ

Занятие 1. Дистилляция знаний

Феоктистов Дмитрий

20 сентября 2024

MiniLLM

	SST2		BoolQ	
	ECE	Acc.	ECE	Acc.
Teacher	0.025	93.0	0.356	74.5
KD	0.191	84.7	0.682	63.5
SeqKD	0.243	66.5	0.681	62.8
MINILLM	**0.099**	**89.7**	**0.502**	**67.8**

Table 2: The ECE scores and accuracy scores (Acc.) on SST2 and BoolQ datasets. The best scores among student models are **boldfaced**.

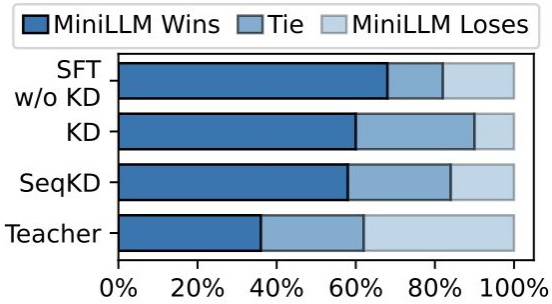


Figure 4: Human evaluation results. We use LLaMA-7B as the student and LLaMA-13B as the teacher.

Yuxian Gu et al, MiniLLM: Knowledge distillation of large language models, ICLR 2024

OFA-KD

Most existing distillation methods are designed under the assumption that the teacher and student models belong to the same model family, particularly the hint-based approaches and we want to distill VITs and CNNs for example

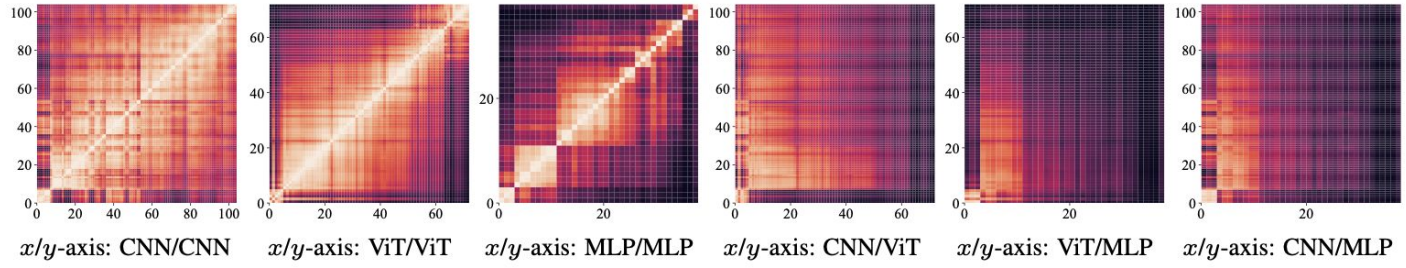


Figure 1: Similarity heatmap of intermediate features measured by CKA. We compare features from **MobileNetV2** (CNN), **ViT-Small** (Transformer) and **Mixer-B/16** (MLP model). The first three figures illustrate similarity between homogeneous models, and the last three illustrate feature similarity between heterogeneous models. Coordinate axes indicate the corresponding layer index.

Zhiwei Hao et al,One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation, NeurIPS 2023

OFA-KD

Centered kernel alignment analysis.[2] To demonstrate the representation gap among heterogeneous architectures, we adopt centered kernel alignment (CKA) [48, 49] to compare features extracted by CNN, ViT, and MLP models. CKA is a feature similarity measurement allowing cross-architecture comparison achievable, as it can work with inputs having different dimensions.

CKA evaluates feature similarity over mini-batch. Suppose $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_2}$ are features of n samples extracted by two different models, where d_1 and d_2 are their dimensions, respectively. CKA measures their similarity by:

$$\mathrm{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\mathcal{D}_{\mathrm{HSIC}}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathcal{D}_{\mathrm{HSIC}}(\mathbf{K}, \mathbf{K}) \mathcal{D}_{\mathrm{HSIC}}(\mathbf{L}, \mathbf{L})}}, \tag{3}$$

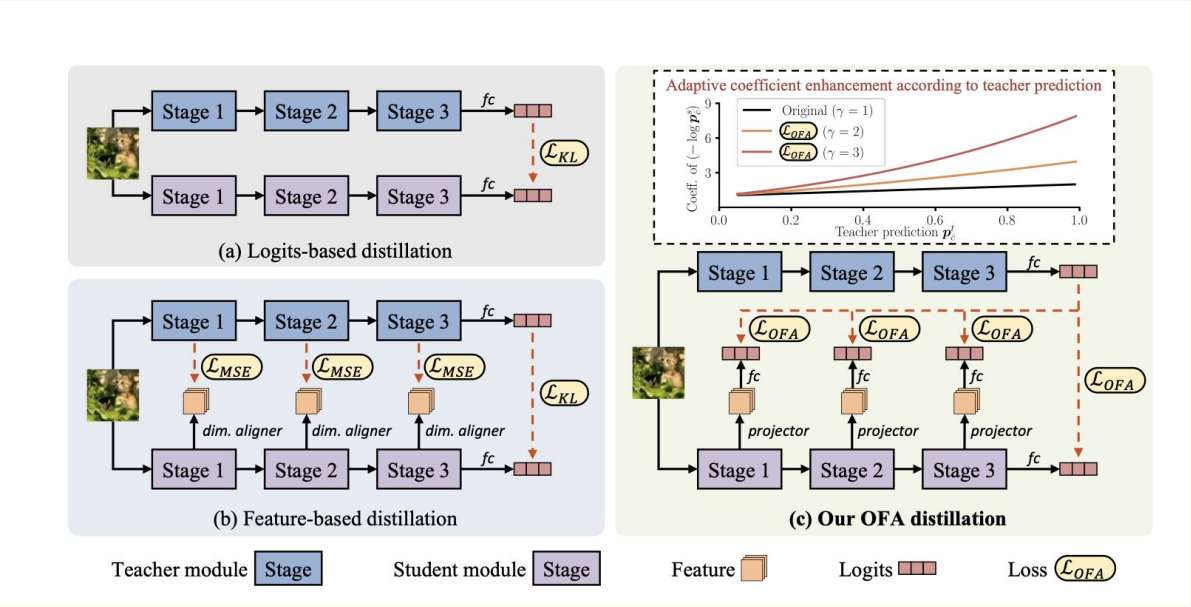
where $\mathbf{L} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{K} = \mathbf{Y}\mathbf{Y}^T$ are Gram matrices of the features, and $\mathcal{D}_{\mathrm{HSIC}}$ is the Hilbert-Schmidt independence criterion [50], a non-parametric independence measure. The empirical estimator of $\mathcal{D}_{\mathrm{HSIC}}$ can be formulated as:

$$\mathcal{D}_{\mathrm{HSIC}}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} \mathrm{tr}(\mathbf{KHLH}), \tag{4}$$

where $\mathbf{H}$ is the centering matrix $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Zhiwei Hao et al,One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation, NeurIPS 2023

OFA-KD



$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} &= -\log \boldsymbol{p}^s_{\hat{c}} - \mathbb{E}_{c\sim\mathcal{Y}}[\boldsymbol{p}^t_c \log \boldsymbol{p}^s_c] \\ &= -(1+\boldsymbol{p}^t_{\hat{c}}) \log \boldsymbol{p}^s_{\hat{c}} - \mathbb{E}_{c\sim\mathcal{Y}/\{\hat{c}\}}[\boldsymbol{p}^t_c \log \boldsymbol{p}^s_c],\end{aligned}$$

$$\mathcal{L}_{\mathrm{OFA}} = -(1+\boldsymbol{p}^t_{\hat{c}})^{\gamma} \log \boldsymbol{p}^s_{\hat{c}} - \mathbb{E}_{c\sim\mathcal{Y}/\{\hat{c}\}}[\boldsymbol{p}^t_c \log \boldsymbol{p}^s_c].$$

Zhiwei Hao et al,One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation, NeurIPS 2023

Эффективные системы машинного обучения, ВМК МГУ

Занятие 1. Дистилляция знаний

Феоктистов Дмитрий

20 сентября 2024

OFA-KD

Teacher	Student	From Scratch		hint-based				Logits-based			
		T.	S.	FitNet	CC	RKD	CRD	KD	DKD	DIST	OFA
CNN-based students											
DeiT-T	ResNet18	72.17	69.75	70.44	69.77	69.47	69.25	70.22	69.39	70.64	**71.34**
Swin-T	ResNet18	81.38	69.75	71.18	70.07	68.89	69.09	71.14	71.10	70.91	**71.85**
Mixer-B/16	ResNet18	76.62	69.75	70.78	70.05	69.46	68.40	70.89	69.89	70.66	**71.38**
DeiT-T	MobileNetV2	72.17	68.87	70.95	70.69	69.72	69.60	70.87	70.14	71.08	**71.39**
Swin-T	MobileNetV2	81.38	68.87	71.75	70.69	67.52	69.58	72.05	71.71	71.76	**72.32**
Mixer-B/16	MobileNetV2	76.62	68.87	71.59	70.79	69.86	68.89	71.92	70.93	71.74	**72.12**
ViT-based students											
ResNet50	DeiT-T	80.38	72.17	75.84	72.56	72.06	68.53	75.10	75.60†	75.13†	**76.55**†
ConvNeXt-T	DeiT-T	82.05	72.17	70.45	73.12	71.47	69.18	74.00	73.95	74.07	**74.41**
Mixer-B/16	DeiT-T	76.62	72.17	74.38	72.82	72.24	68.23	74.16	72.82	74.22	**74.46**
ResNet50	Swin-N	80.38	75.53	78.33	76.05	75.90	73.90	77.58	78.23†	77.95†	**78.64**†
ConvNeXt-T	Swin-N	82.05	75.53	74.81	75.79	75.48	74.15	77.15	77.00	77.25	**77.50**
Mixer-B/16	Swin-N	76.62	75.53	76.17	75.81	75.52	73.38	76.26	75.03	76.54	**76.63**
MLP-based students											
ResNet50	ResMLP-S12	80.38	76.65	78.13	76.21	75.45	73.23	77.41	78.23†	77.71†	**78.53**†
ConvNeXt-T	ResMLP-S12	82.05	76.65	74.69	75.79	75.28	73.57	76.84	77.23	77.24	**77.53**
Swin-T	ResMLP-S12	81.38	76.65	76.48	76.15	75.10	73.40	76.67	76.99	77.25	**77.31**

Zhiwei Hao et al,One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation, NeurIPS 2023

Спасибо за внимание!

Если остались вопросы, то задавайте их в чат или пишите преподавателю в tg: @trandelik