

Дистилляция знаний

Московский Физико-Технический Институт

2022

Что такое дистилляция [Hinton et al., 2015]

Given a teacher model. Knowledge distillation is a transfer of the knowledge from the teacher model to a small model that is more suitable for deployment.



Teacher



Student

Дистилляция и привилегированная информация

Общий подход к дистилляции [Lopez-Paz et al., 2015]:

- ① обучить учителя на выборке $\mathbf{X}^*, \mathbf{y}^*$;
- ② Обучить ученика на ответах учителя и выборке \mathbf{X}, \mathbf{y} .

Если $\mathbf{X} = \mathbf{X}^*$ — это дистилляция по Hinton.

В качестве \mathbf{X}^* могут выступать не только фичи объектов, но и информация о схожености объектов (например).

Подход Hinton

$$\alpha \log p(\mathbf{y} | \text{SM}(\mathbf{f}_{\text{student}}(\mathbf{X}))) + (1 - \alpha) p \left(\text{SM} \left(\frac{\mathbf{f}_{\text{teacher}}(\mathbf{X})}{T} \right) \middle| \text{SM} \left(\frac{\mathbf{f}_{\text{student}}(\mathbf{X})}{T} \right) \right) \rightarrow \max_{\mathbf{w}_{\text{student}}},$$

где $\mathbf{f}_{\text{teacher}}$, $\mathbf{f}_{\text{student}}$ — модель учителя и ученика (до softmax), SM — softmax, T — гиперпараметр температуры.

При $T \rightarrow \infty$ второе слагаемое эквивалентно минимизации l_2 расстояния между логитами.

При $T \rightarrow 0$ логиты превращаются в one-hot векторы.

Дистилляция однородных моделей



Teacher layer 1



Student layer 1



Teacher layer 2



Student layer 2



Teacher layer 3



Student layer 3

TinyBERT: пример

$$\mathcal{L}_{\text{attn}} = \frac{1}{h} \sum_{i=1}^h \text{MSE}(\mathbf{A}_i^S, \mathbf{A}_i^T),$$

$$\mathcal{L}_{\text{hidn}} = \text{MSE}(\mathbf{H}^S \mathbf{W}_h, \mathbf{H}^T),$$

$$\mathcal{L}_{\text{embd}} = \text{MSE}(\mathbf{E}^S \mathbf{W}_e, \mathbf{E}^T),$$

$$\mathcal{L}_{\text{pred}} = \text{CE}(\mathbf{z}^T / t, \mathbf{z}^S / t),$$

Дистилляция неоднородных моделей

- Слоев может быть разное количество
- Слои могут иметь разное функциональное назначение



Teacher (Transformer)

???



Student (LSTM)

TextBrewer: пример наивной дистилляции

Model	MNLI		SQuAD		CoNLL-2003
	m	mm	EM	F1	F1
BERT _{BASE}	83.7	84.0	81.5	88.6	91.1
<i>Public</i>					
DistilBERT	81.6	81.1	79.1	86.9	-
TinyBERT	80.5	81.0	-	-	-
+DA	82.8	82.9	72.7	82.1	-
<i>TextBrewer</i>					
BiGRU	-	-	-	-	85.3
T6	83.6	84.0	80.8	88.1	90.7
T3	81.6	82.5	76.3	84.8	87.5
T3-small	81.3	81.7	72.3	81.4	78.6
T4-tiny	82.0	82.6	73.7	82.5	77.5
+DA	-	-	75.2	84.0	89.1

Model	# Layers	Hidden size	Feed-forward size	# Parameters	Relative size
BERT _{BASE} (teacher)	12	768	3072	108M	100%
T6	6	768	3072	65M	60%
T3	3	768	3072	44M	41%
T3-small	3	384	1536	17M	16%
T4-tiny	4	312	1200	14M	13%
BiGRU	1	768	-	31M	29%

Распределение косинусов: наивная дистилляция

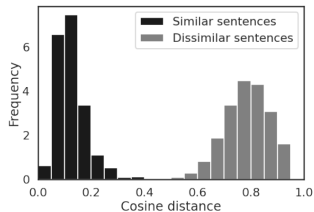


Figure 2a. The distribution of distances between pairs of similar and dissimilar sentences with different phrase embedding models: LaBSE model.

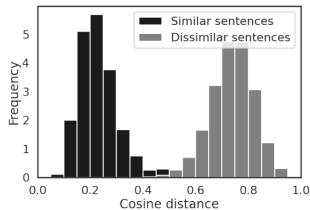


Figure 2b. The distribution of distances between pairs of similar and dissimilar sentences with different phrase embedding models: LSTM model.

Probabilistic knowledge transfer [Passalis et al., 2018]

Идея метода: Пусть заданы две однородные модели с одним скрытым слоем $\mathbf{h}_t, \mathbf{h}_s$. Задача дистилляции:

$$KL(p(\mathbf{H}_s), p(\mathbf{H}_t)) \rightarrow \min.$$

Альтернативная постановка:

$$I(\mathbf{H}_s, \mathbf{y}) = I(\mathbf{H}_t, \mathbf{y}),$$

$$I(\mathbf{H}_s, \mathbf{y}) = KL(p(\mathbf{H}_t, \mathbf{y}) || p(\mathbf{H}_t)p(\mathbf{y})).$$

Аппроксимация:

$$IQ(\mathbf{h}_s, \mathbf{y}) = E_{\mathbf{H}, \mathbf{y}} || p(\mathbf{H}_t, \mathbf{y}) - p(\mathbf{H}_t)p(\mathbf{y}) ||_2^2.$$

Как считать вероятности?

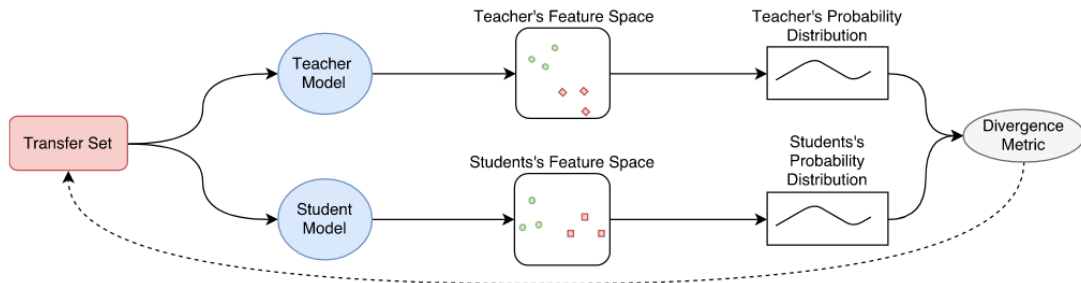
$$IQ(\mathbf{h}_s, y) = \sum_y \int_{\mathbf{h}} d\mathbf{h} [p(\mathbf{H}, y)^2 + p(\mathbf{H})p(y)^2 - p(\mathbf{H}, y)p(\mathbf{H})p(y)] .$$

Вероятность каждого класса $p(y)$ — считается эмпирически по выборке.

$$p(\mathbf{h}) \propto \sum_{\mathbf{h}'} K(\mathbf{h}, \mathbf{h}'),$$

$p(\mathbf{h}, y)$ считается аналогично, суммирование ведется по объектам класса y .

PKT



Что делать, если модели неоднородны? [Passalis et al., 2020]

- Создать “тяжелую” модель ученика, которая
 - ▶ Однородна с обычным учеником
 - ▶ Имеет бОльшую сложность, чем ученик (сопоставимую со сложностью учителя)
- Продистиллировать знания из учителя в “тяжелого” ученика, допустима дистилляция нескольких слоев учителя в один слой “тяжелого ученика”
- Продистиллировать знания из “тяжелого” ученика в ученика



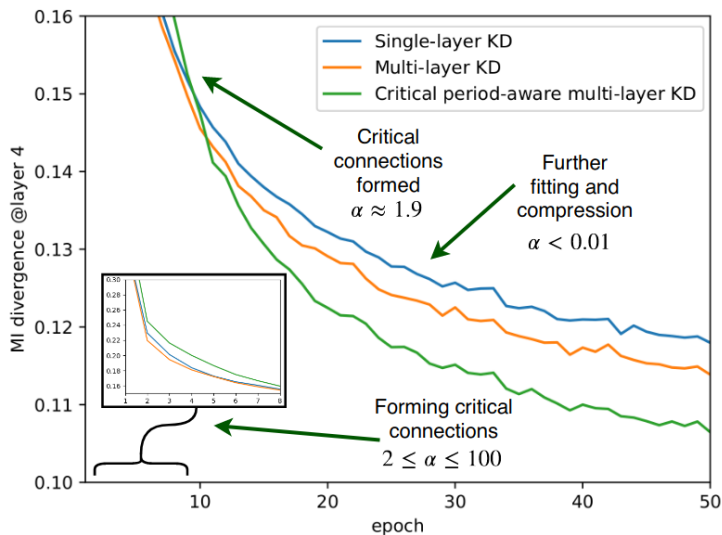
Результаты

Table 1. Metric Learning Evaluation: CIFAR-10

Method	mAP (e)	mAP (c)	top-100 (e)	top-100 (c)
Baseline Models				
Teacher (ResNet-18)	87.18	90.47	92.15	92.26
Aux. (CNN1-A)	62.12	66.78	73.72	75.91
With Constrastive Supervision				
Student (CNN1)	47.69	48.72	57.46	58.50
Hint.	43.56	48.73	60.44	62.43
MKT	45.34	46.84	55.89	57.10
PKT	48.87	49.95	58.44	59.48
Hint-H	43.24	47.46	58.97	61.07
MKT-H	44.83	47.12	56.28	57.90
PKT-H	48.69	50.09	58.71	60.20
Proposed	49.55	50.82	59.50	60.79
Without Constrastive Supervision				
Student (CNN1)	35.30	39.00	55.87	58.77
Distill.	37.39	40.53	56.17	58.56
Hint.	43.99	48.99	60.69	62.42
MKT	36.26	38.20	50.55	52.72
PKT	48.07	51.56	60.02	62.50
Hint-H	42.65	46.46	58.51	60.59
MKT-H	41.16	43.99	55.10	57.63
PKT-H	48.05	51.73	60.39	63.01
Proposed	49.20	53.06	61.54	64.24

* В Proposed вместо KL-дивергенции используется симметризация KL: $KL(a||b) + KL(b||a)$.

Важность переноса знаний о скрытых слоях



Где тут Байес?

Variational Information Distillation for Knowledge Transfer (Ahn et al., 2019)

$$\begin{aligned} I(\mathbf{t}; \mathbf{s}) &= H(\mathbf{t}) - H(\mathbf{t}|\mathbf{s}) \\ &= H(\mathbf{t}) + \mathbb{E}_{\mathbf{t}, \mathbf{s}}[\log p(\mathbf{t}|\mathbf{s})] \\ &= H(\mathbf{t}) + \mathbb{E}_{\mathbf{t}, \mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})] + \mathbb{E}_{\mathbf{s}}[D_{\text{KL}}(p(\mathbf{t}|\mathbf{s})||q(\mathbf{t}|\mathbf{s}))] \\ &\geq H(\mathbf{t}) + \mathbb{E}_{\mathbf{t}, \mathbf{s}}[\log q(\mathbf{t}|\mathbf{s})], \end{aligned} \tag{3}$$

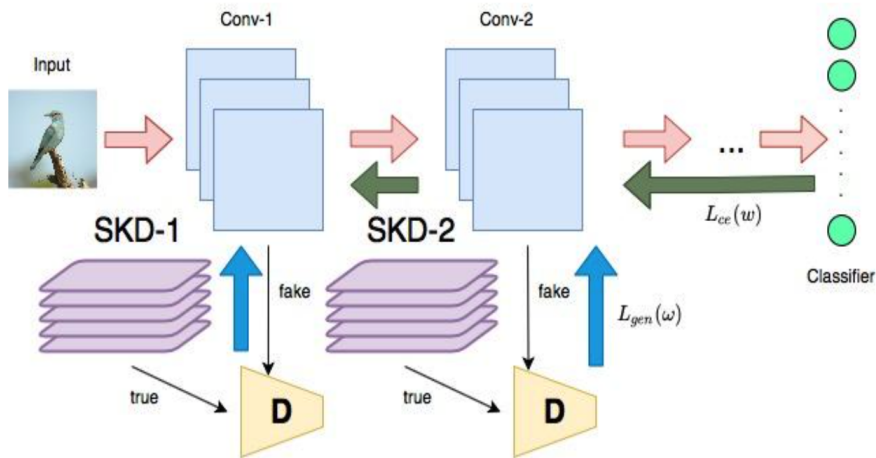
$$\begin{aligned} -\log q(\mathbf{t}|\mathbf{s}) &= -\sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W \log q(t_{c,h,w}|\mathbf{s}) \\ &= \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W \log \sigma_c + \frac{(t_{c,h,w} - \mu_{c,h,w}(\mathbf{s}))^2}{2\sigma_c^2} + \text{constant}, \end{aligned} \tag{5}$$

Что будет, если переусложнить вариационное распределение?

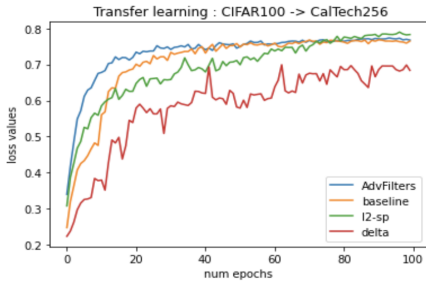
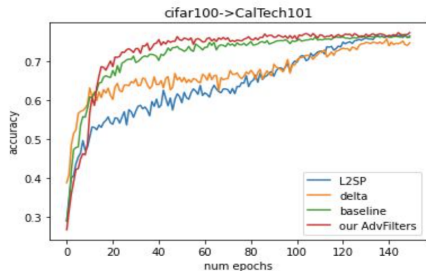
Байесовская дистилляция моделей глубокого обучения (Грабовой, Стрижов, 2021)

- Модель-учитель получена при помощи стандартного вариационного вывода
- Модель-ученик получается при помощи вариационного вывода с априорным распределением $p(\mathbf{w}_{\text{st}}) = q(\mathbf{w}_{\text{teacher}})$
- Если ученик и учитель разнородны: решается задача выравнивания двух распределений, последовательным удалением слабо выравнивающихся параметров.

Адверсариальный перенос информации, (Колесов, 2022)

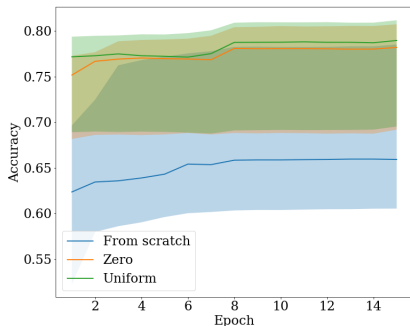


Адверсариальный перенос информации, (Колесов, 2022)



Anti-Distillation from a simple model to a complex one.

Goal: Adapting the model to more complex data.



Increasing complexity:

1. More classes
2. More features

Our solution: Weight initialization using the parameters of the pre-trained model.

$$\mathbf{w}_1 = \varphi(\mathbf{u}^*)$$

$$\mathbf{w}_1 = (\mathbf{u}^*, \mathbf{w}'_1), \quad \mathbf{w}'_1 = \mathbf{0}, \quad \mathbf{w}'_1 \sim U\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]$$

Дистилляция без выборки (Lopes et al., 2017)

Проблема: чтобы продистиллировать модель, нужно хранить выборку.
Что делать, если выборку занимает терабайты данных?













Решение: будем запоминать статистики выборок и сэмплировать объекты на лету.

Дистилляция без выборки (Lopes et al., 2017)

Простой вариант:

- Запоминаем у учителя распределение логитов (среднее и ковариацию для каждого класса)
- На стадии дистилляции: $\mathbf{b} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$;
- Для ученика \mathbf{f} : $\mathbf{x}_0 = \arg \min_{\mathbf{x}} \|\mathbf{f}(\mathbf{x}) - \mathbf{b}\|^2$
- Проводим дистилляцию на \mathbf{x}_0 .

Дистилляция без выборки (Lopes et al., 2017)

Activation Record	Means	Randomly sampled example
MNIST		
Top Layer Statistics		
All Layers Statistics		
All Layers + Dropout		
Spectral All Layers		
Spectral Layer Pairs		

Дистилляция без выборки (Lopes et al., 2017)

Table 5: Accuracies of the ALEXNET model and CELEBA dataset for each procedure.

Model Name	Procedure	Accuracy on test set
ALEXNET	Train on CelebA	80.82%
ALEXNET-HALF	Train on CelebA	81.59%
ALEXNET-HALF	Knowledge Distillation [8]	69.53%
ALEXNET-HALF	Top Layer Statistics	54.12%
	All-Layers Spectral	77.56%
	Layer-Pairs Spectral	76.94%

Литература

- Hinton G. et al. Distilling the knowledge in a neural network //arXiv preprint arXiv:1503.02531. – 2015. – Т. 2. – №. 7.
- Lopez-Paz D. et al. Unifying distillation and privileged information //arXiv preprint arXiv:1511.03643. – 2015.
- Jiao X. et al. Tinybert: Distilling bert for natural language understanding //arXiv preprint arXiv:1909.10351. – 2019.
- Yang Z. et al. Textbrewer: An open-source knowledge distillation toolkit for natural language processing //arXiv preprint arXiv:2002.12620. – 2020.
- Bakhteev Oleg et al. Cross-language plagiarism detection: a case study of European universities academic works // ENAI. - 2021.
- Passalis N., Tefas A. Learning deep representations with probabilistic knowledge transfer //Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – С. 268-284.
- Ahn S. et al. Variational information distillation for knowledge transfer //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 9163-9171.
- Грабовой А. В., Стрижов В. В. Байесовская дистилляция моделей глубокого обучения //Автоматика и телемеханика. – 2021. – №. 11. – С. 16-29.