Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning

Sorokin Dmitrii

2024

Феномены ансамблей

- Ансамбль из моделей, отличающихся только сидом, дает прирост в качестве

Феномены ансамблей

- Ансамбль из моделей, отличающихся только сидом, дает прирост в качестве
- Ансамбль можно дистиллировать в одну модель и не потерять в качестве

Феномены ансамблей

- Ансамбль из моделей, отличающихся только сидом, дает прирост в качестве
- Ансамбль можно дистиллировать в одну модель и не потерять в качестве
- Можно дистиллировать модель в саму себя и все равно получить прирост в качестве

💡 *Исследования в этой области поможет понять тонкости подготовки датасета перед ансамблированием моделей*

Гипотезы: Random Features Mapping

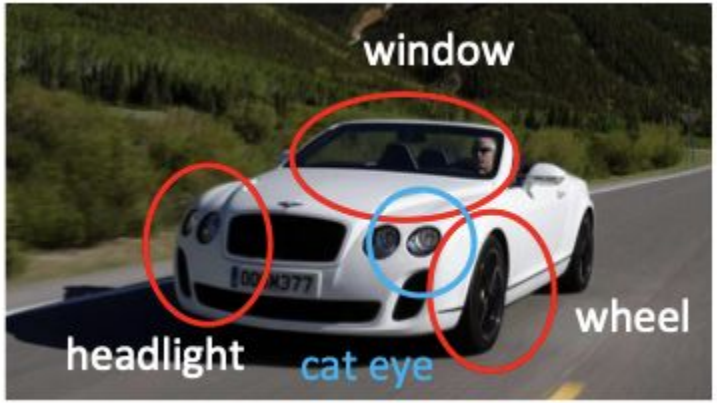
NTK: $$f(W, x) \approx f(W_0, x) + \langle W - W_0, \nabla_W f(W_0, x) \rangle$$

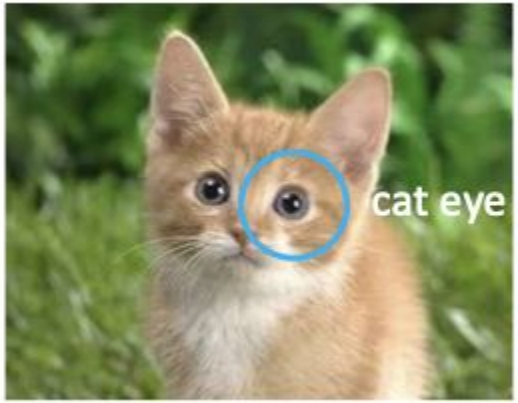
$$\Phi_{W_0}(x) := \nabla_W f(W_0, x)$$

- Усреднение давало лучшее качество
- Дистилляция не работает с NTK

*Почему модель может через дистилляцию выучить признаки, которая не может выучить непосредственно при использовании NTK? Все дело в **Dark Knowledges***

Dark Knowledge





💡 *Dark Knowledge – причина по которой работает дистилляция*

Гипотезы: Multi View

- When the label is class 1, then:[2]

$$\begin{cases} \text{both } v_1, v_2 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 80\%; \\ \text{only } v_1 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 10\%; \\ \text{only } v_2 \text{ appears with weight 1, one of } v_3, v_4 \text{ appears with weight 0.1} & \text{w.p. } 10\%. \end{cases}$$

- When the label is class 2, then

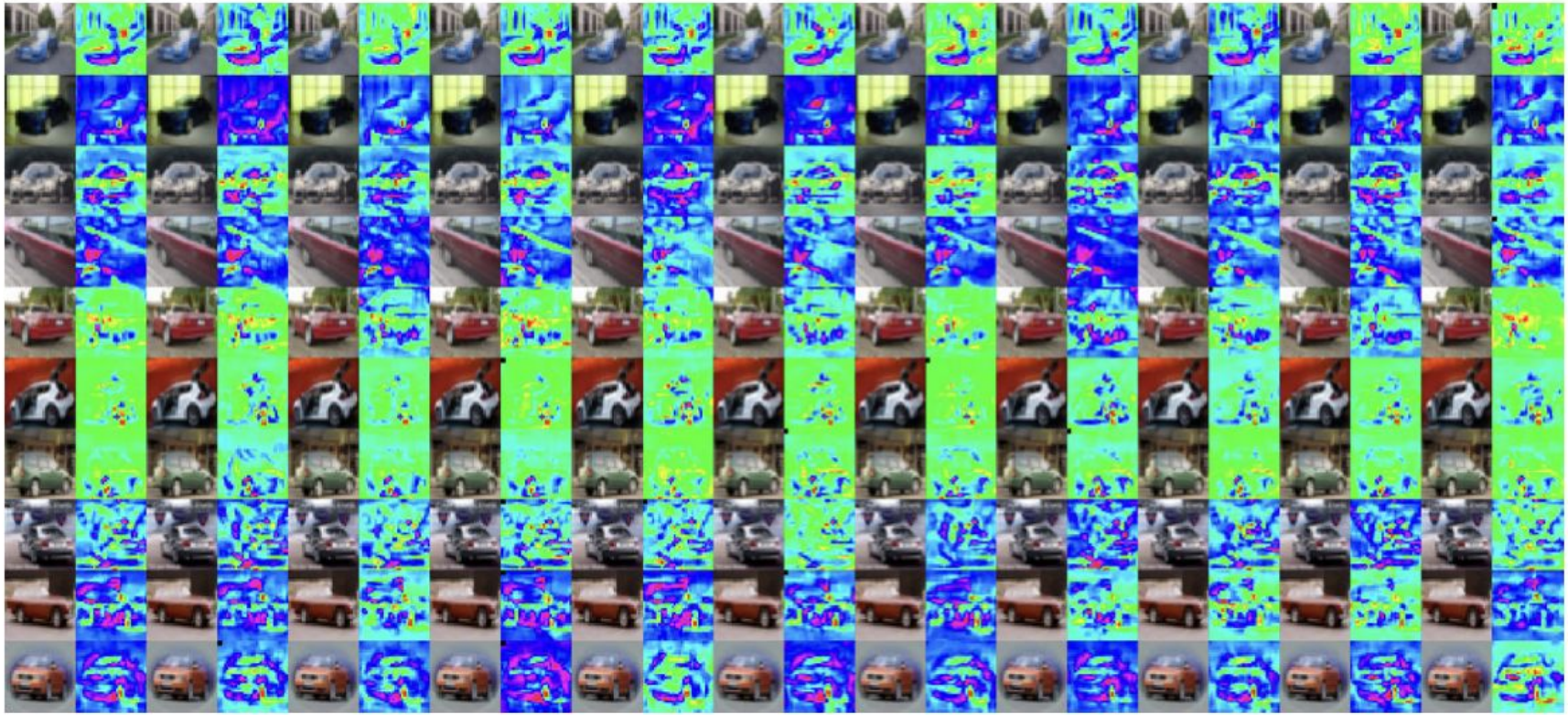
$$\begin{cases} \text{both } v_3, v_4 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 80\%; \\ \text{only } v_3 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 10\%; \\ \text{only } v_4 \text{ appears with weight 1, one of } v_1, v_2 \text{ appears with weight 0.1} & \text{w.p. } 10\%. \end{cases}$$

Гипотезы: Multi View

ResNet-34 learns three features (views) of a car:
(1) front wheel (2) front window (3) side window

ResNet-34 learns three features (views) of a horse:
(1) tail (2) legs (3) head

Ансамбли: откуда прирост в качестве?



Теоретическое обоснование

1. В нашей постановке задачи, при обучении одиночной модели точность на трейне будет 100%, точность на тесте от 49% до 51%
2. При обучении ансамбля точность на обеих выборках около 100%*
3. При обучении дистилляции точность на обеих выборках около 100%*
4. При обучении селф дистилляции точность на трейне 100%, точность на тесте строго лучше чем в одиночной модели

* При достаточно большом количестве моделей в ансамбле

Эксперименты

	CIFAR10 test accuracy				CIFAR100 test accuracy			
	single model (over 10)	ensemble (over 10)	10 runs of knowledge distill	ensemble over knowledge distill	single model (over 10)	ensemble (over 10)	10 runs of knowledge distill	ensemble over knowledge distill
ResNet-28-2	95.22±*0.14%*	96.33%	95.89±0.07%	96.21%	76.38±*0.23%*	81.13%	78.94±0.21%	80.35%
ResNet-34	93.65±*0.19%*	94.97%	94.37±0.13%	94.88%	71.66±*0.43%*	76.85%	73.57±0.34%	75.60%
ResNet-34-2	95.45±*0.14%*	96.55%	96.00±0.12%	96.42%	77.01±*0.35%*	81.48%	79.43±0.23%	81.56%
ResNet-16-10	96.08±*0.16%*	96.80%	96.73±0.07%	96.76%	80.03±*0.17%*	83.18%	82.51±0.14%	83.36%
ResNet-22-10	96.44±*0.09%*	97.12%	97.01±0.09%	97.09%	81.17±*0.23%*	84.33%	83.54±0.19%	84.27%
ResNet-28-10	96.70±*0.21%*	97.20%	97.06±0.08%	97.24%	81.51±*0.16%*	84.69%	83.75±0.16%	84.87%

Message ①: an ensemble over *single models* (independently trained) can be distilled into a single model with moderate accuracy loss.

Message ②: an ensemble over models *after knowledge distillation* *does not improve accuracy by much* – in fact, not exceeding the ensemble accuracy of the original single models ③ – despite the training objective is still non-convex and different random seeds are used. This means, knowledge distillation models (i.e. simply matching the soft labels) *have learned most of the features* from the ensemble, and have less variety comparing to the original single models. This also means that "(huge) non-convexity" in neural networks and SGD with "different random seeds" *even together do not guarantee ensemble advantage unconditionally*; the structure of the data (and hard labels) is extremely important for ensemble to work as we mainly focus on in this paper.

Эксперименты

neural networks	single model (over 10) ⑦	ensemble (over 10)	train $\sum_\ell f_\ell$ (over 10)	knowledge distillation	self-distill	single model (over 10) ⑦	ensemble (over 10)	train $\sum_\ell f_\ell$ (over 10)	knowledge distillation	self-distill
ResNet-28-2	95.22±*0.14%*	96.33%	95.02%	96.16%	95.78%	76.38±*0.23%*	81.13%	73.18%	79.03%	78.12%
ResNet-34	93.65±*0.19%*	94.97%	93.12%	94.59%	94.21%	71.66±*0.43%*	76.85%	68.88%	73.74%	73.14%
ResNet-34-2	95.45±*0.14%*	96.55%	95.00%	96.08%	95.86%	77.01±*0.35%*	81.48%	72.99%	79.23%	79.07%
ResNet-16-10	96.08±*0.16%*	96.80%	95.88% (over 6)◊	96.81%	96.62%	80.03±*0.17%*	83.18%	80.53% (over 6)◊	82.67%	82.25%
ResNet-22-10	96.44±*0.09%*	97.12%	96.41% (over 5)◊	97.09%	97.05%	81.17±*0.23%*	84.33%	81.59% (over 5)◊	83.71%	83.26%
ResNet-28-10	96.70±*0.21%*	97.20%	96.46% (over 4)◊	97.22%	97.13%	81.51±*0.16%*	84.69%	81.83% (over 4)◊	83.81%	83.56%

Message ④: for neural nets, ensemble helps on improving test accuracies, **and** this accuracy gain *cannot* be matched by training the sum of the individuals directly. In other words, the benefit of using ensemble comes from somewhere other than enlarging the model.

Message ⑤: for neural nets, the superior test performance of ensemble *can be distilled* into single model by a large extent.

Message ⑥: for neural nets, self-distillation *clearly improves* the test performance of single models.

Message ⑦: for neural nets, the superior performance of ensemble *does not* come from the variance of test accuracies in single models.