
“Model soups”

Или почему смешивать модели - круто.

Докладчик: Сидоренко Иван Алексеевич БПМИ2110

Стандартный пайплайн обучения

1. Выбор “архитектуры”
2. Пока !надоело
Выбор гиперпараметров
Обучение
3. Выбор лучшей модели

Очевидный минус

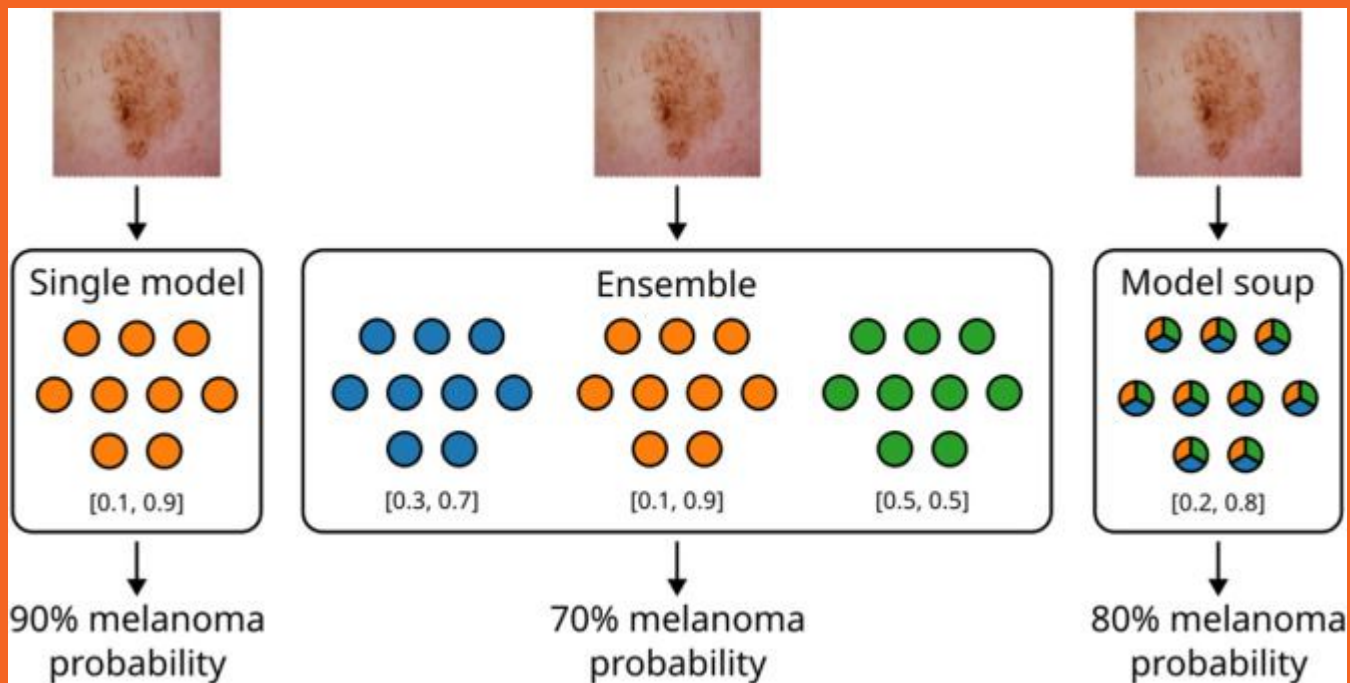
Слишком много работы,
результаты которой не
используются.

Как можно использовать несколько моделей вместе?

1. Детерминированная комбинация ответов.
2. Бэггинг.
3. Бустинг.
4. Стекинг.

Минус этих подходов - весомые затраты на “инференс” ансамбля.

Комбинация моделей без потери производительности



Виды супов

1. Равномерный суп (uniform soup).
2. Жадный суп (greedy soup).
3. Эрудированный суп (learned soup).

Суть равномерного супа в усреднении весов всех имеющихся моделей.

Жадный суп

Добавляем
ингредиенты жадно,
пока качество супа не
начнёт ухудшаться.

Recipe 1 GreedySoup

Input: Potential soup ingredients $\{\theta_1, \dots, \theta_k\}$ (sorted in decreasing order of $\text{ValAcc}(\theta_i)$).

ingredients $\leftarrow \{\}$

for $i = 1$ **to** k **do**

if $\text{ValAcc}(\text{average}(\text{ingredients} \cup \{\theta_i\})) \geq$
 $\text{ValAcc}(\text{average}(\text{ingredients}))$ **then**

 ingredients $\leftarrow \text{ingredients} \cup \{\theta_i\}$

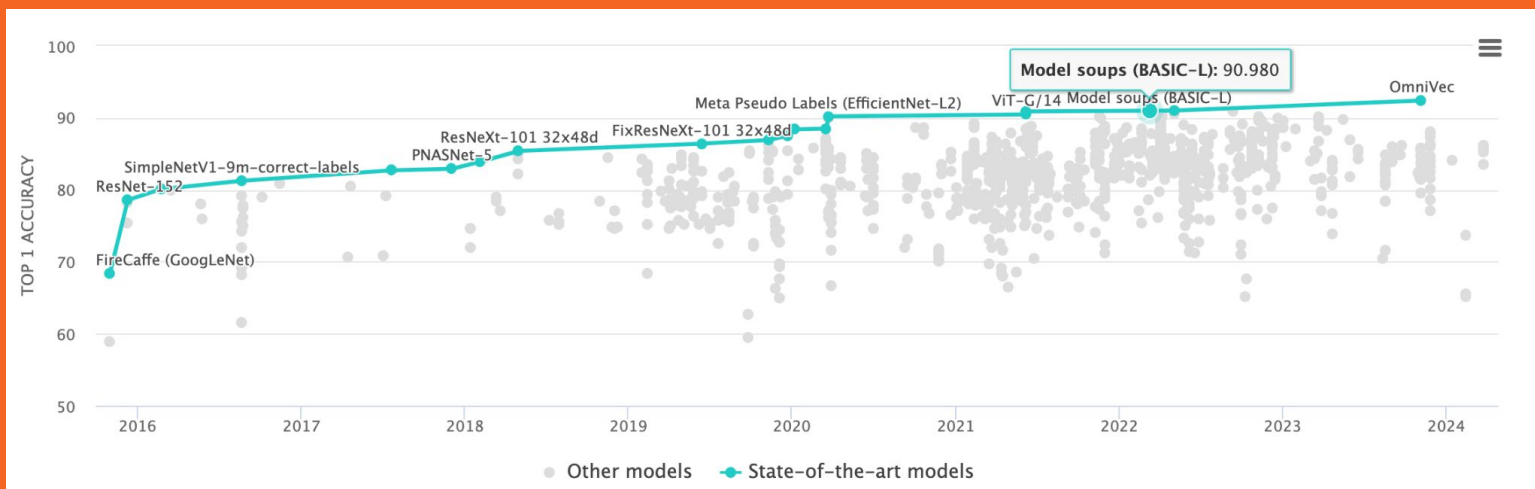
return average(ingredients)

Эрудированный суп





Коэффициенты для весов ингредиентов
обучаются.

$$\sum_{j=1}^n L \left(\beta f \left(x_j, \sum_{i=1}^k \alpha_i \theta_i \right), y_j \right) \rightarrow \min_{\alpha, \beta}$$

<https://paperswithcode.com/>

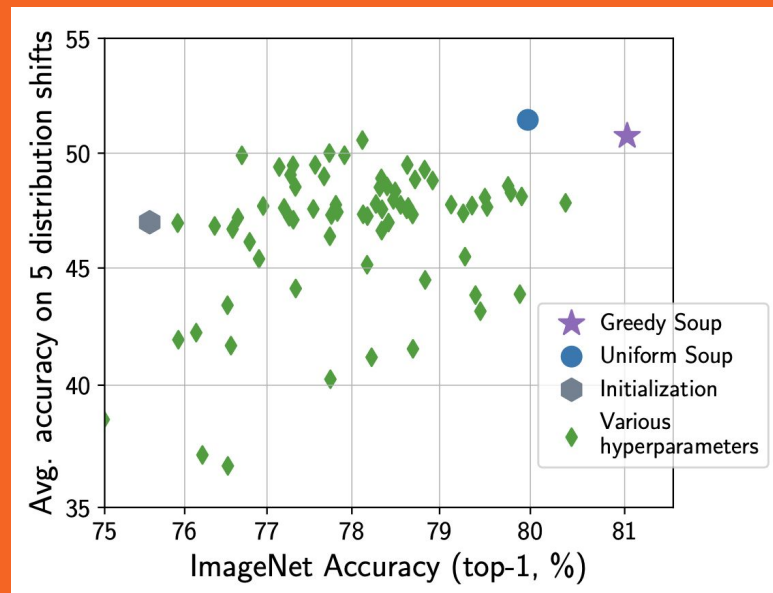


<https://paperswithcode.com/>

4	Model soups (BASIC-L)	90.98%	2440M	×	Model soups: averaging weights of multiple fine- tuned models improves accuracy without increasing inference time			2022	ALIGN	JFT-3B	Conv+Transform
5	Model soups (ViT-G/14)	90.94%	1843M	×	Model soups: averaging weights of multiple fine- tuned models improves accuracy without increasing inference time			2022	JFT-3B	Transf	

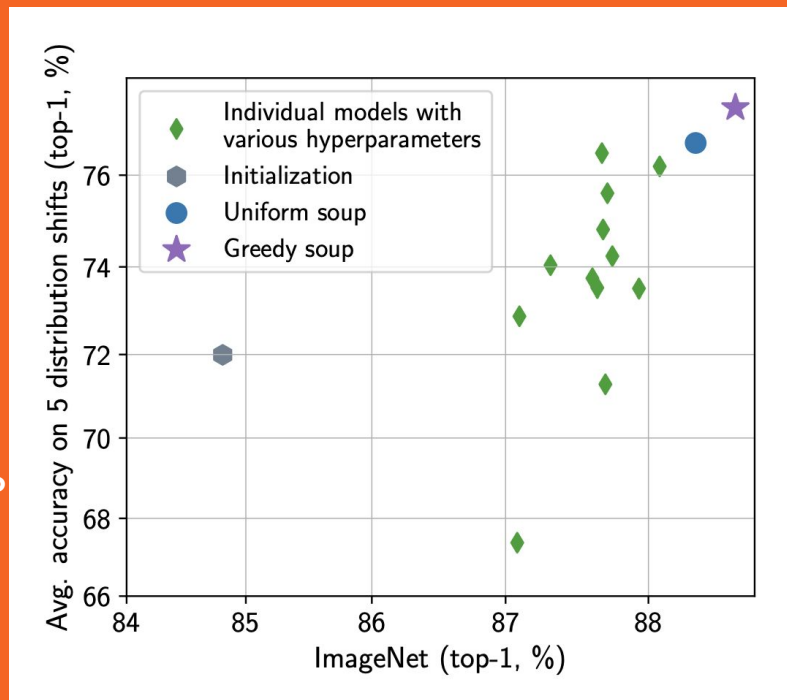
Эксперименты

1. CLIP ViT-B/32
2. ImageNet-A, -R, -Sketch, -V2, а также ObjectNet.
3. Случайным образом выбирались lr, wd, количество эпох, аугментации.



Эксперименты

1. ALIGN
2. ImageNet-A, -R, -Sketch, -V2, а также ObjectNet.
3. lr и количество эпох выбирались по “сетке”.
4. Микс аугментация.

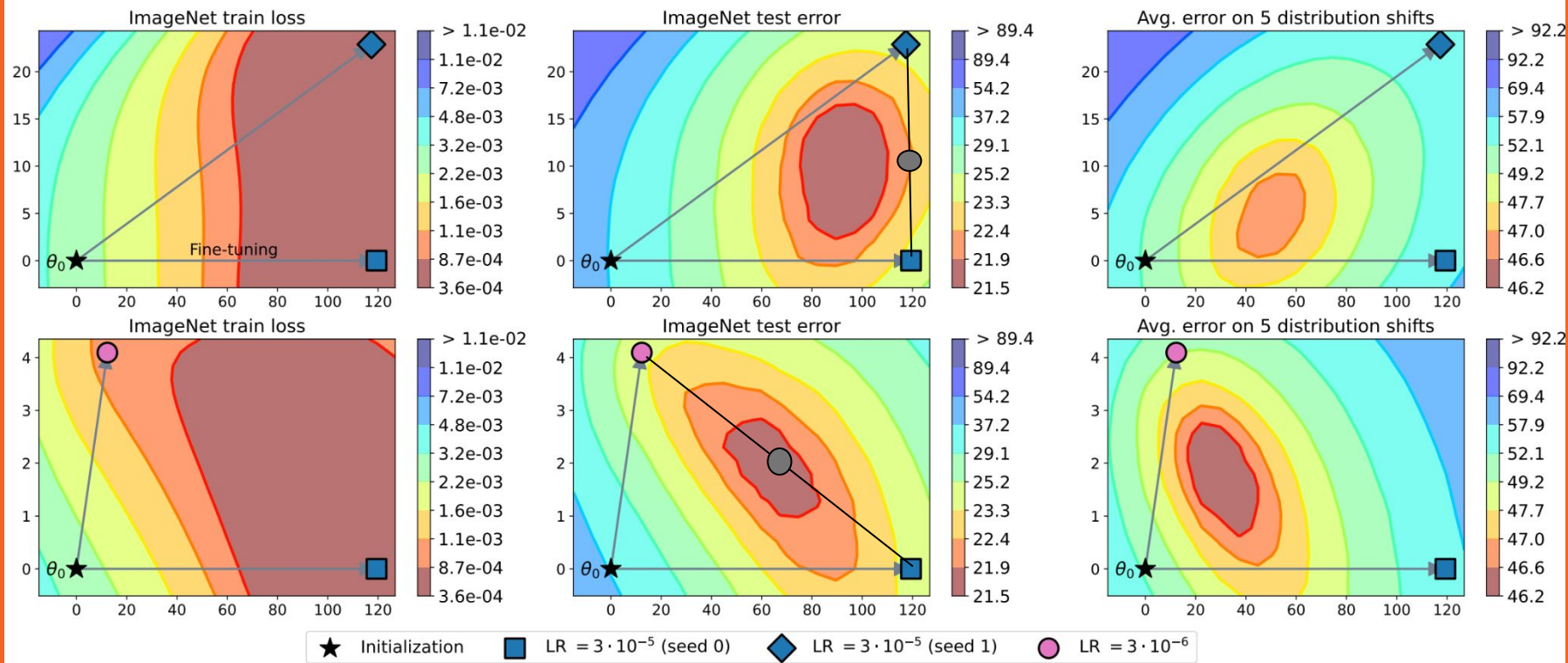


Эксперименты

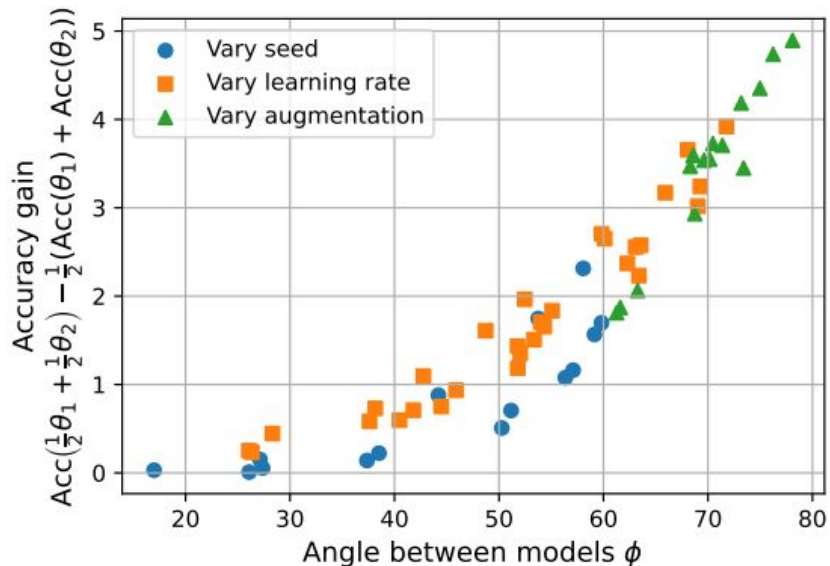
1. Bert и T5.
2. GLUE бенчмарк.
3. 32 модели.
4. avg асс и F1 для MRPC, асс для RTE и SST-2, корреляция Метьюса для CoLA.

Model	Method	MRPC	RTE	CoLA	SST-2
BERT (Devlin et al., 2019b)	Best individual model	88.3	61.0	59.1	92.5
	Greedy soup	88.3 (+0.0)	61.7 (+0.7)	59.1 (+0.0)	93.0 (+0.5)
T5 (Raffel et al., 2020b)	Best individual model	91.8	78.3	58.8	94.6
	Greedy soup	92.4 (+0.6)	79.1 (+0.8)	60.2 (+0.4)	94.7 (+0.1)

Почему супы работают?



Предельная полезность добавления ингредиента в суп



$$\varphi(\theta_1, \theta_2) = \angle(\theta_1 - \theta_0, \theta_2 - \theta_0)$$

Pruned soup

Убираем из равномерного супа ингредиенты по одному, пока качество не падает.

Input: weights of Potential soup ingredients $\theta_1, \dots, \theta_k$ (optionally sorted in decreasing order of $\mathbf{ValAcc}(\theta_i)$)

Parameter: numbers of passes (N)

```
1: soup =  $\frac{1}{k} \sum_{i=1}^k \theta_i$ 
2: baseline = ValAcc(averaged weights).
3: for pass=1 to N do
4:   for i=1 to k do
5:     new soup  $\leftarrow$  Remove a model  $\theta_i$  from the soup.
6:     if ValAcc(new soup)  $\geq$  baseline then
7:       baseline = ValAcc(new soup)
8:       soup  $\leftarrow$  new soup
9:     end if
10:  end for
11: end for
12: return weights of the final soup  $\theta_{soup}$ 
```

Pruned soup

1. CIFAR-100
2. ViT
3. SGD(momentum=0.9)
4. Разные lr и wd

Method	Acc. (%)	Ingredients (avg)
Best individual model	50.3	-
Uniform soup	32.22	22
Greedy soup (random)	51.06	5.1
Greedy soup (sorted)	51.76	5
Pruned soup (random)	52.04	3.2
Pruned soup (Sorted)	52.1	3

Вредный совет

1. Учим ингредиенты с нуля.
2. Смешиваем их в суп.
3. Негодует, что суп испорчен (не сильно лучше рандомного классификатора).

Method (ResNets)	Acc. (%)	Ingredients (avg)
Best individual model	65.78	-
Uniform soup	1.08	33
Greedy soup (random)	65.78	1
Greedy soup (sorted)	65.78	1
Pruned soup (random)	1.19	6.2
Pruned soup (Sorted)	1.19	6.2

Method (EfficientNets)	Acc. (%)	Ingredients (avg)
Best individual model	40.12	-
Uniform soup	0.98	36
Greedy soup (random)	40.12	1
Greedy soup (sorted)	40.12	1
Pruned soup (random)	1.12	5.1
Pruned soup (Sorted)	1.12	5.1

Заключение

1. Предобучение модели (инициализация).
2. Дообучение множества экземпляров с различными гиперпараметрами, аугментациями.
3. Выбор подмножества моделей для усреднения весов.

