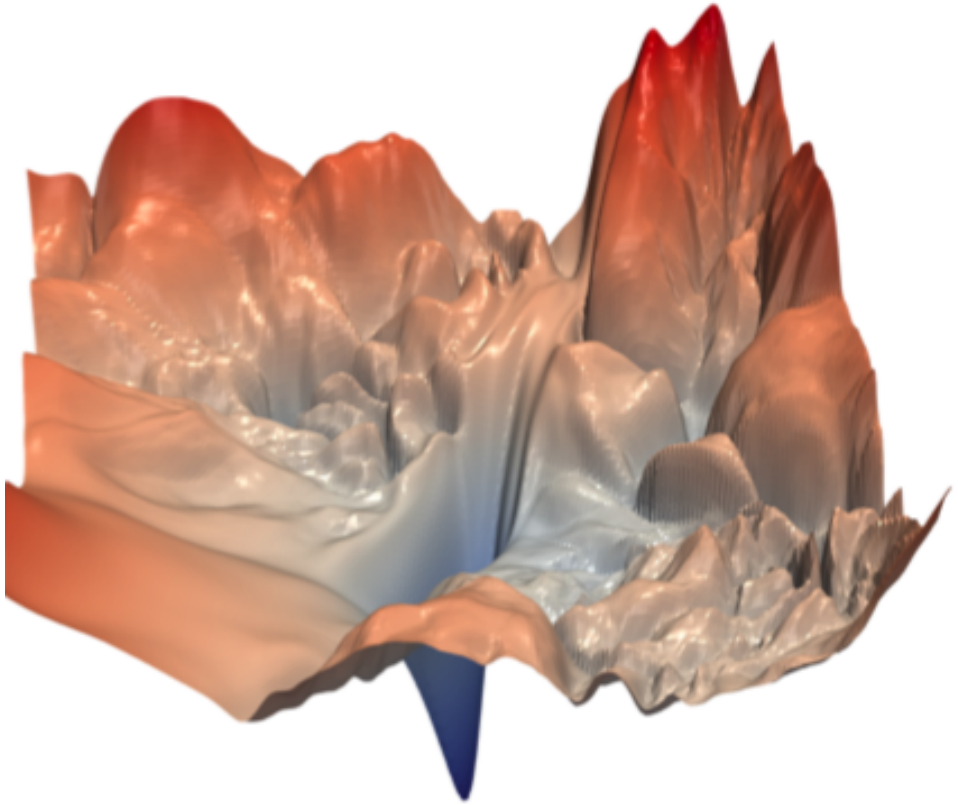
Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs.

Цыбакин Александр, гр. 162.

Вступление.

- Лосс-функции невыпуклые и зависят от большого количества параметров.
- Геометрические свойства поверхностей лосс-функций плохо изучены.
- Огромное количество локальных оптимумов и седловых точек экспоненциально зависит от числа параметров модели.

Вступление.

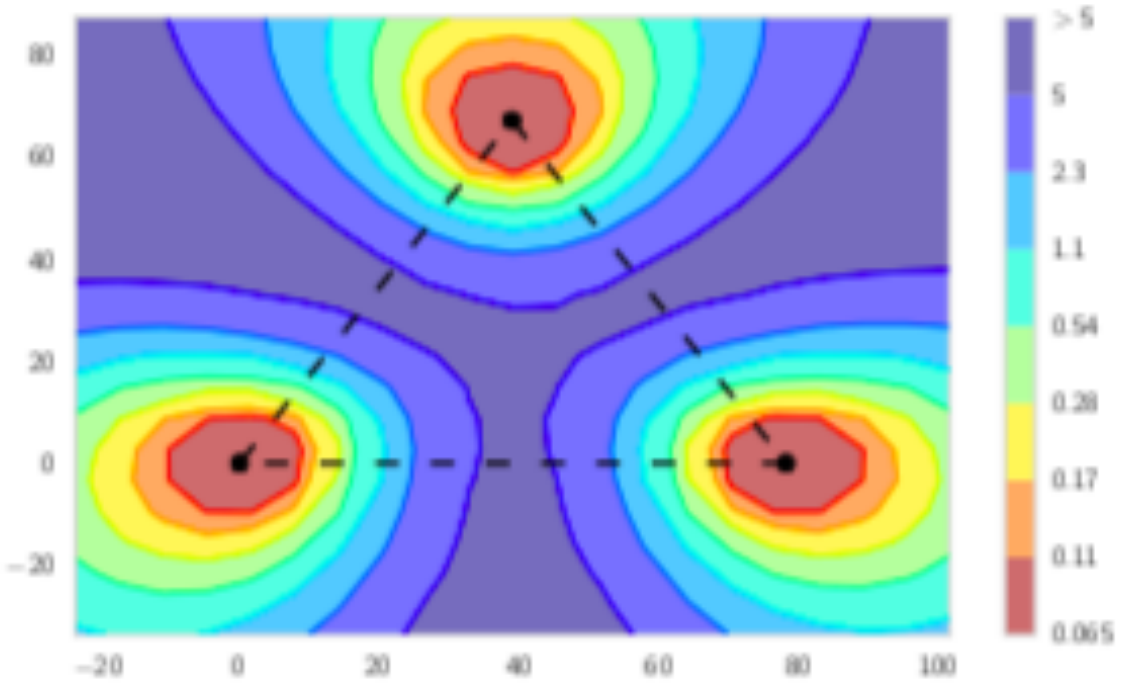


Визуализация поверхности лосс-функции ResNet-56.

[Visualizing the Loss Landscape of Neural Nets, 2018 y.]

Вступление.

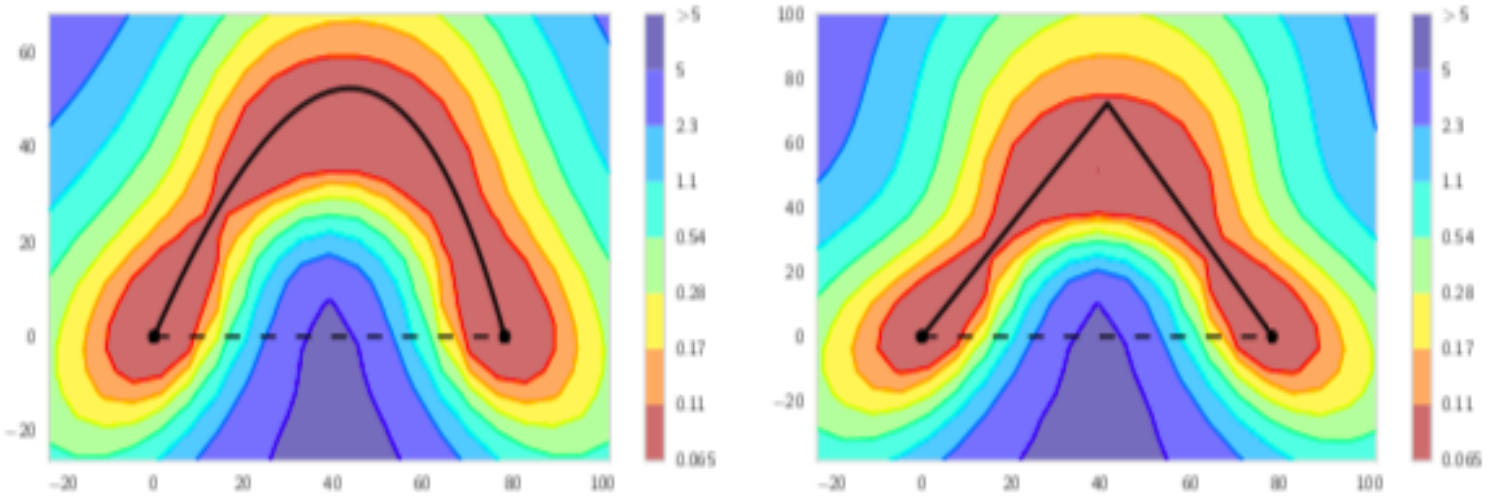
- Оптимумы изолированы друг от друга, например:



Проекция поверхности Кросс-Энтропии с L2-регуляризатором ResNet-164 на датасете CIFAR-100.

Построение кривых между оптимумами.

- Наблюдение: между оптимумами существуют кривые, для которых значение лосс-функции практически всегда низкое (mode connectivity).
- Более того, такие кривые имеют простой вид, например, ломаной кривой.
- Можно использовать для построения ансамблей.



Примеры построения кривых между двумя фиксированными оптимумами в пространстве весов.

Построение кривых между оптимумами.

- Пусть у нас есть w_0 и w_1 из $\mathbb{R}^{|net|}$, где $|net|$ - количество весов модели.
- *L(w)* – некоторая лосс-функция, например, кросс-энтропия.
- Введем функцию $\phi_\theta(t)$: $[0,1] \to \mathbb{R}^{|net|}, \phi_\theta(0) = w_0$ и $\phi_\theta(1) = w_1$.

- Авторы предлагают ввести новую лосс-функцию $l(\theta)$, которая минимизирует мат. ожидание значений исходной лосс-функции от равномерно распределенной на кривой величины.

- Вспомним, что $p(x) = \frac{1}{b-a}$, где $x \sim U[a,b]$.
- В нашем случае $p(\phi_\theta) = \frac{1}{\int d\phi_\theta}$, для равномерно распределенной величины ϕ_θ на кривой.
- $E\big(f(x)\big) = \int f(x)\, p(x) dx$ - Мат. ожидание непрерывной величины.
- Значит, $l(\theta) = \frac{\int L(\phi_\theta) d\phi_\theta}{\int d\phi_\theta} = \frac{\int_0^1 L(\phi_\theta(t))\, ||\phi_\theta'(t)|| dt}{\int_0^1 ||\phi_\theta'(t)|| dt} = \int_0^1 L\big(\phi_\theta(t)\big) q_\theta(t) dt = E_{t \sim q_\theta(t)}[L(\phi_\theta(t))]$,

где $q_\theta(t) = ||\phi_\theta'(t)|| / \int_0^1 ||\phi_\theta'(t)|| dt$, t ∈ $[0,1]$.

Построение кривых между оптимумами.

- В итоге , $l(\theta) = E_{t \sim q_\theta(t)}[L(\phi_\theta(t))]$.

Посчитать градиент $l(\theta)$ сложно, т.к. $q_\theta(t)$ зависит от θ.

- Авторы предлагают использовать $U[0,1]$ вместо $q_\theta(t)$.

Тогда $l(\theta) = E_{t \sim U[0,1]}[L(\phi_\theta(t))]$.

- **Оптимизация: градиентный спуск**.

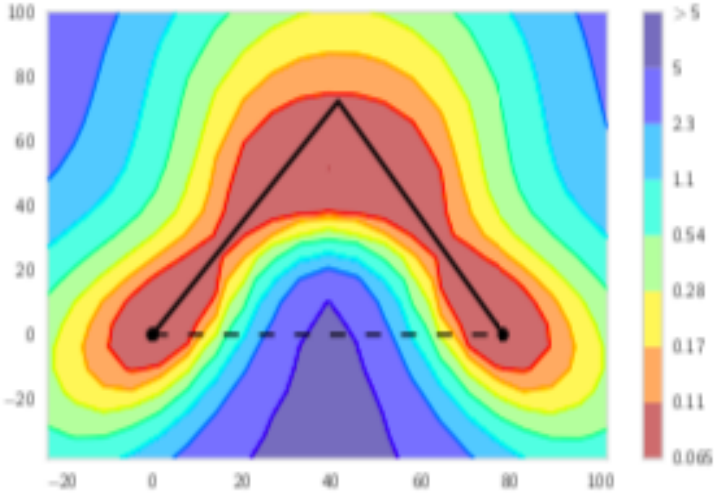
Шаг: генерируется $\tilde{t}$ из $U[0,1]$ и рассчитывается $\nabla_\theta L(\phi_\theta(\tilde{t}))$.

Т.к. $\nabla_\theta l(\theta) = \nabla_\theta E_{t \sim U[0,1]}\left[L\left(\phi_\theta(t)\right)\right] = E_{t \sim U[0,1]}\left[\nabla_\theta L\left(\phi_\theta(t)\right)\right] \cong \nabla_\theta L(\phi_\theta(\tilde{t}))$

Примеры параметризации.

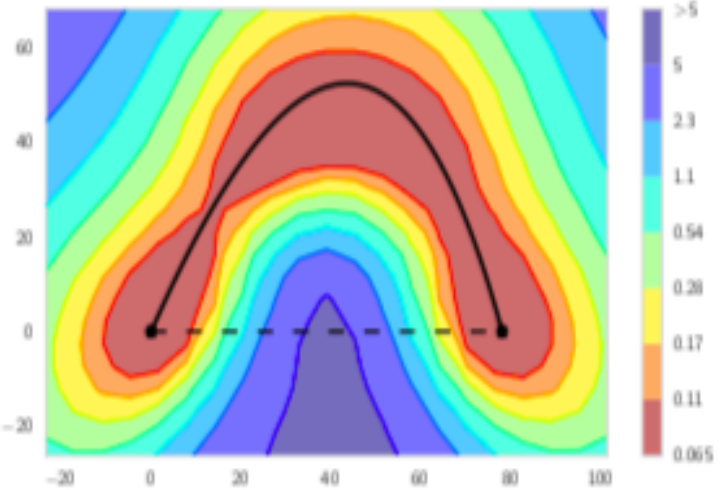
1. Ломаная.

$$\phi_\theta(t) = \begin{cases} 2(t\theta + (0.5 - t)w_1, & 0 \le t \le 0.5 \\ 2\big((t-0.5)w_2 + (1-t)\theta\big), & 0.5 < t \le 1 \end{cases}.$$



2. Кривая Безье.

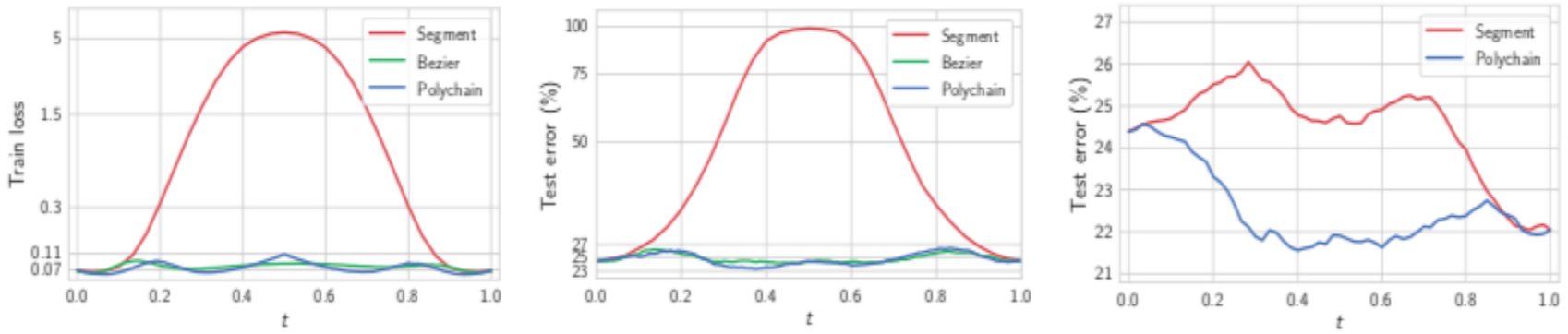
$$\phi_\theta(t) = (1-t)^2 w_1 + 2t(1-t)\theta + t^2 w_2, \quad 0 \le t \le 1.$$



Эксперименты с построением.

- Два раза независимо обучили модель, получили w_0 и w_1.
- Знаем, что можно построить кривую с низким лоссом между двумя оптимумами.
- Попробуем построить простой ансамбль из двух моделей.
- Первая модель: зафиксируем первую точку $\phi_\theta(0) = w_0$.
- Вторая модель: $\phi_\theta(t)$ при переборе t ∈ $[0, 1]$.

Результат:



ResNet-164 на датасете CIFAR-100.
(Слева) Значение кросс-энтропии с L2-регуляризатором на трейне.
(Посередине) Ошибка на тесте.
(Справа) Ошибка ансамбля на тесте при переборе t.

Fast Geometric Ensembling.

- Теперь имеем одну предобученную модель w_0.
- Знаем, что между оптимумами существуют кривые с низким лоссом.
- Как увидели ранее, даже при небольшом отступе от оптимума в пространстве весов модель уже имеет другие предсказания, а значит и другое представление о данных.

Общая идея:

- Необходимо без построения самой кривой перемещаться от предобученной модели(w_0) небольшими шагами (learning rate) без большого увеличения лосса по пространству весов.
- На некоторых шагах “собирать” модели в ансамбль.
- Усреднить предсказания собранных моделей (учитываем разное представление данных).

Fast Geometric Ensembling.

Изменение размера шагов (learning rate).

Авторы вводят циклическое изменение шага.

Пусть есть $\alpha_1 > \alpha_2$ – два learning rate, тогда изменение шага:

$$\alpha(i) = \begin{cases} \big(1 - 2t(i)\big)\alpha_1 + 2t(i)\alpha_2, 0 < t(i) \leq 0.5 \\ \big(2 - 2t(i)\big)\alpha_2 + (2t(i) - 1)\alpha_1, 0.5 < t(i) \leq 1. \end{cases}$$

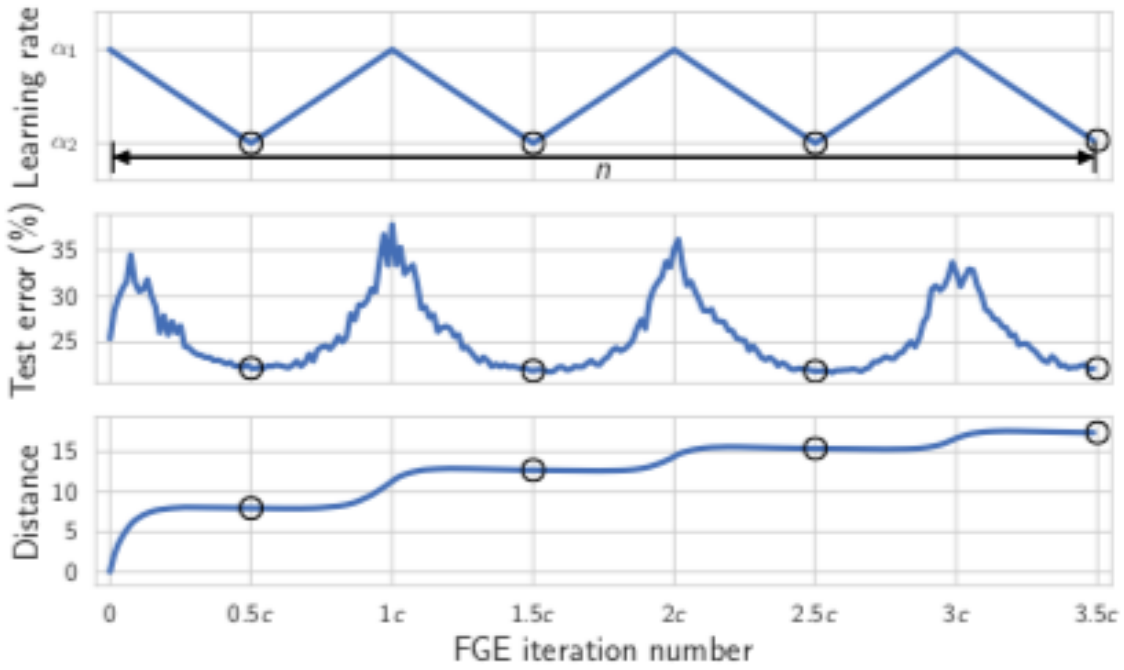
Где *i = 1,2…* - номер итерации.

$t(i) = \frac{1}{c}(mod(i-1, c) + 1)$, c – количество итераций в одном цикле(обычно 2-4 эпохи).

На каждой середине цикла (т.е. при $t(i) = \frac{1}{2}, \alpha(i) = \alpha_2$) модели с текущими весами добавляются в ансамбль.

Такое изменение вызвано балансом между двумя фазами: исследованием (большие шаги - значения близкие к α_1) и уточнением (маленькие шаги – значения близкие к α_2).

Fast Geometric Ensembling.



(Сверху) Циклическое изменение шага.
(Посередине) Ошибка на тесте при циклическом изменении шага.
(Снизу) Расстояние (по Евклиду) между зафиксированными моделями, которые добавляются в ансамбль.

Fast Geometric Ensembling.

- Общее описание алгоритма.

Algorithm 1 Fast Geometric Ensembling

Require:
weights $\hat{w}$, LR bounds α_1, α_2,
cycle length c (even), number of iterations n
Ensure: ensemble
$w \leftarrow \hat{w}$ {Initialize weight with $\hat{w}$}
ensemble $\leftarrow [\,]$
for $i \leftarrow 1, 2, \ldots, n$ **do**
  $\alpha \leftarrow \alpha(i)$ {Calculate LR for the iteration}
  $w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$ {Stochastic gradient update}
  if $\mathrm{mod}(i, c) = c/2$ **then**
    ensemble $\leftarrow$ ensemble $+ [w]$ {Collect weights}
  end if
end for

Эксперименты с FGE.

DNN (Budget)	method	CIFAR-100			CIFAR-10		
		$1B$	$2B$	$3B$	$1B$	$2B$	$3B$
VGG-16 (200)	Ind	27.4 ± 0.1	25.28	24.45	6.75 ± 0.16	5.89	5.9
	SSE	26.4 ± 0.1	25.16	24.69	6.57 ± 0.12	6.19	5.95
	FGE	$\mathbf{25.7 \pm 0.1}$	**24.11**	**23.54**	$\mathbf{6.48 \pm 0.09}$	**5.82**	**5.66**
ResNet-164 (150)	Ind	21.5 ± 0.4	19.04	18.59	4.72 ± 0.1	**4.1**	**3.77**
	SSE	20.9 ± 0.2	19.28	18.91	4.66 ± 0.02	4.37	4.3
	FGE	$\mathbf{20.2 \pm 0.1}$	**18.67**	**18.21**	$\mathbf{4.54 \pm 0.05}$	4.21	3.98
WRN-28-10 (200)	Ind	19.2 ± 0.2	17.48	17.01	3.82 ± 0.1	3.4	**3.31**
	SSE	17.9 ± 0.2	17.3	16.97	3.73 ± 0.04	3.54	3.55
	FGE	$\mathbf{17.7 \pm 0.2}$	**16.95**	**16.88**	$\mathbf{3.65 \pm 0.1}$	**3.38**	3.52

Сравнительная таблица ошибок (в %) по трем разным подходам трех архитектур на двух датасетах.

Методы: ***Ind*** – независимо обученные модели, ***SSE*** – SnapShot Ensembling(также основан на циклическом изменении шага с использованием косинуса), ***FGE*** – Fast Geometric Ensambling.

Заключение.

- Между оптимумами существуют кривые, вдоль которых лосс практически постоянно низкий.
- Такие кривые имеют простой вид (ломаная, кривая Безье).
- На этом наблюдении основан алгоритм построения ансамблей FGE.
- Цикл FGE ставится обычно 2-4 эпохи (например, для Snapshot Ensabmling - 20-40 эпох).
- FGE способен улучшать SOTA-архитектуры.

Спасибо за внимание!

Вопросы?

Вопросы!

- В чем заключается идея поиска кривых между оптимумами функции потерь? Какие кривые предлагают строить между оптимумами потерь авторы статьи?

- Опишите алгоритм Fast Geometric Ensembling.