



Факультет компьютерных наук
Образовательная программа
“Прикладная математика и информатика”

1

Neural network loss landscape

Выполнил:
Разин Арслан Дмитриевич, БПМИ202

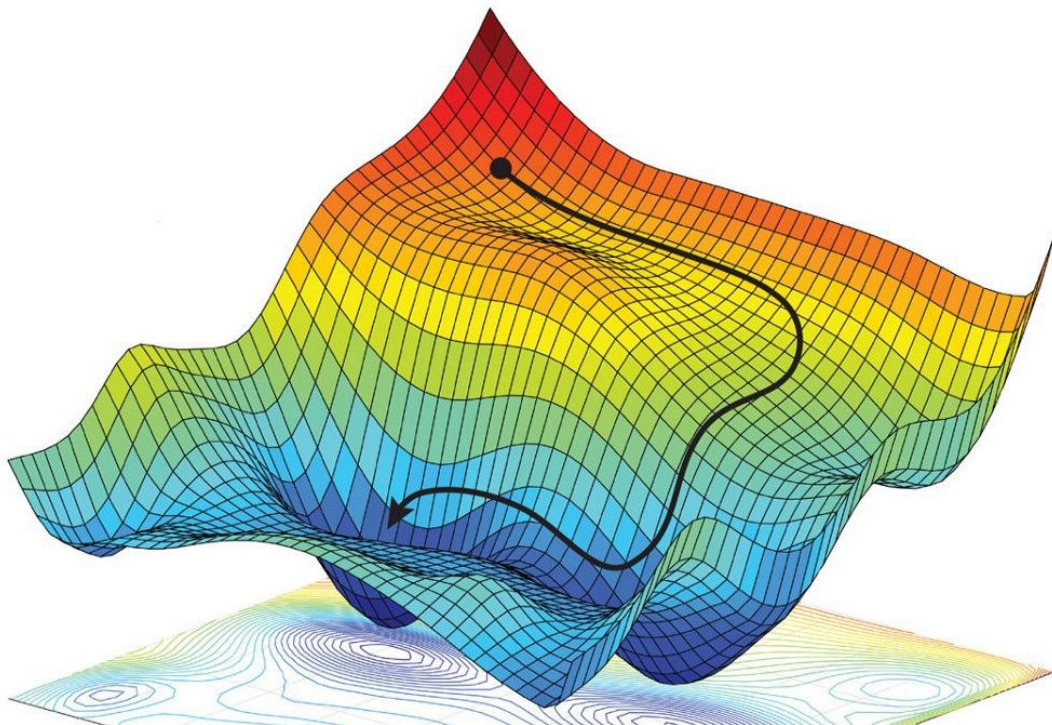
Москва, 2022

1. Визуализация функции потерь нейронных сетей
2. Применение визуализации в методе Snapshot Ensembles
3. Применение визуализации в методе FGE
4. Выводы
5. Красивые картинки
6. Ответы на вопросы

Зачем визуализировать loss?

3





Почему вообще можно минимизировать сильно невыпуклые функции потерь?



Зачем визуализировать loss?

4

Почему результирующие минимумы обобщаются?

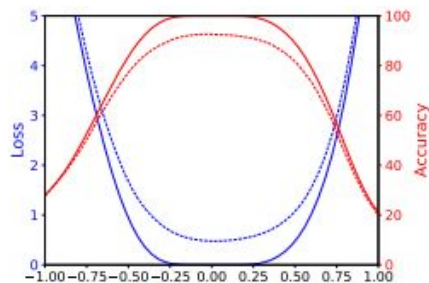
<i>Input</i>	<i>Label</i>	<i>Prediction</i>
	CAT	
	NOT CAT	 ?
	CAT	

Немного математики

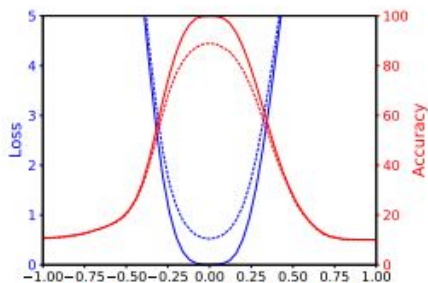
Одномерная линейная интерполяция

$$\theta(\alpha) = (1 - \alpha)\theta_1 + \alpha\theta_2$$

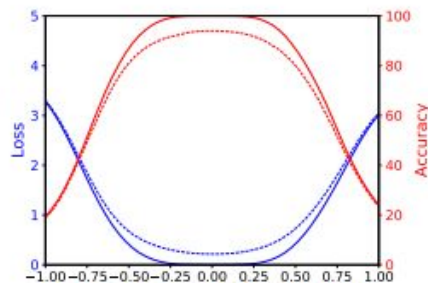
$$f(\alpha) = L(\theta(\alpha))$$



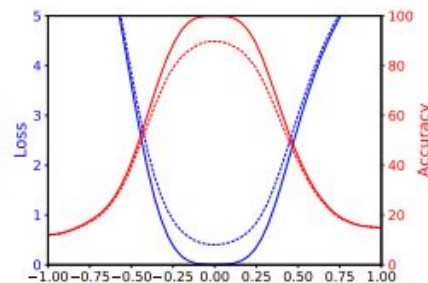
(a) 0.0, 128, 7.37%



(b) 0.0, 8192, 11.07%



(c) 5e-4, 128, 6.00%

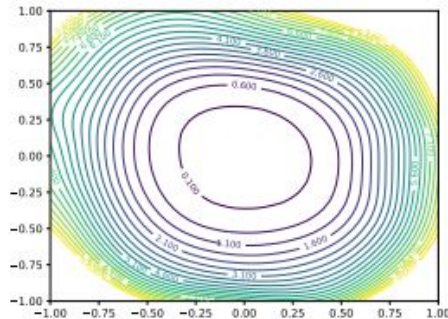


(d) 5e-4, 8192, 10.19%

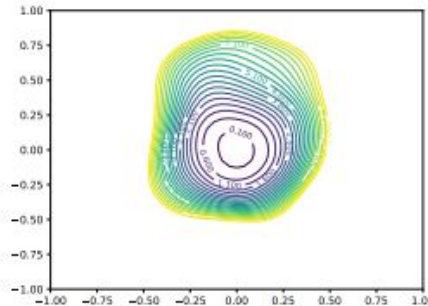
Примеры одномерной линейной интерполяции
(здесь обучали SGD с разными параметрами)

Контурные графики по случайным направлениям

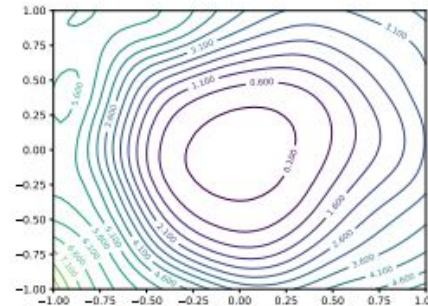
$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta)$$



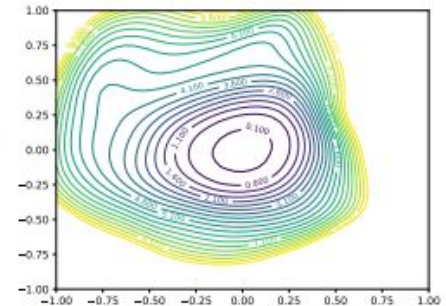
(e) 0.0, 128, 7.37%



(f) 0.0, 8192, 11.07%



(g) 5e-4, 128, 6.00%



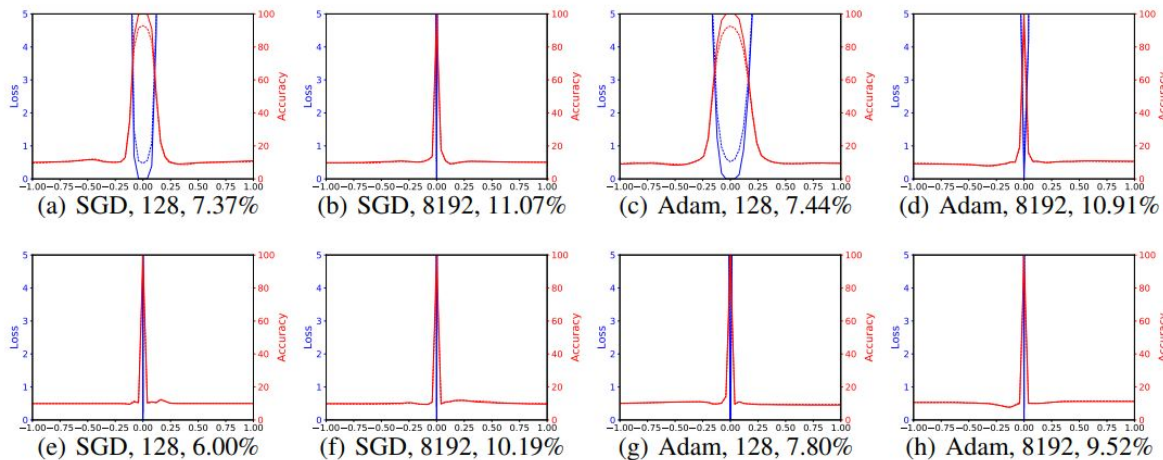
(h) 5e-4, 8192, 10.19%

Примеры контурных графиков по случайным направлениям
(здесь обучали SGD с разными параметрами)

Не забываем про нормализацию

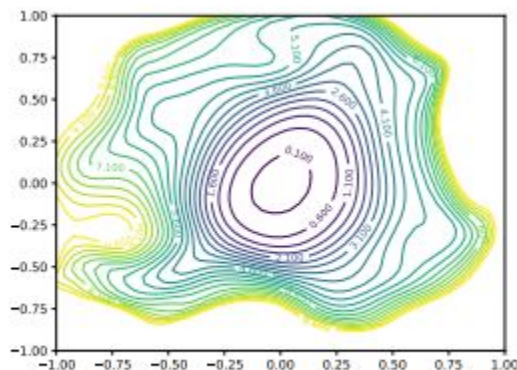
7

$$d_{i,j} \leftarrow \frac{d_{i,j}}{\|d_{i,j}\|} \|\theta_{i,j}\|$$

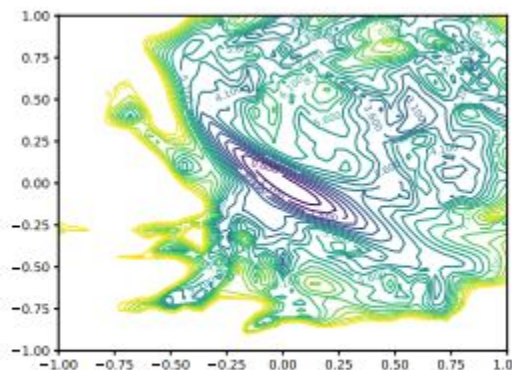


Графики без
нормализации

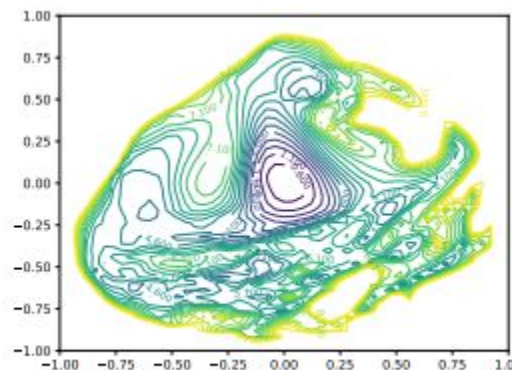
Figure 11: 1D loss plots for VGG-9 without normalization. The first row has no weight decay and the second row uses weight decay 0.0005.



(d) ResNet-20-NS, 8.18%



(e) ResNet-56-NS, 13.31%



(f) ResNet-110-NS, 16.44%

Влияние глубины нейросети

Использование skip-connection

9

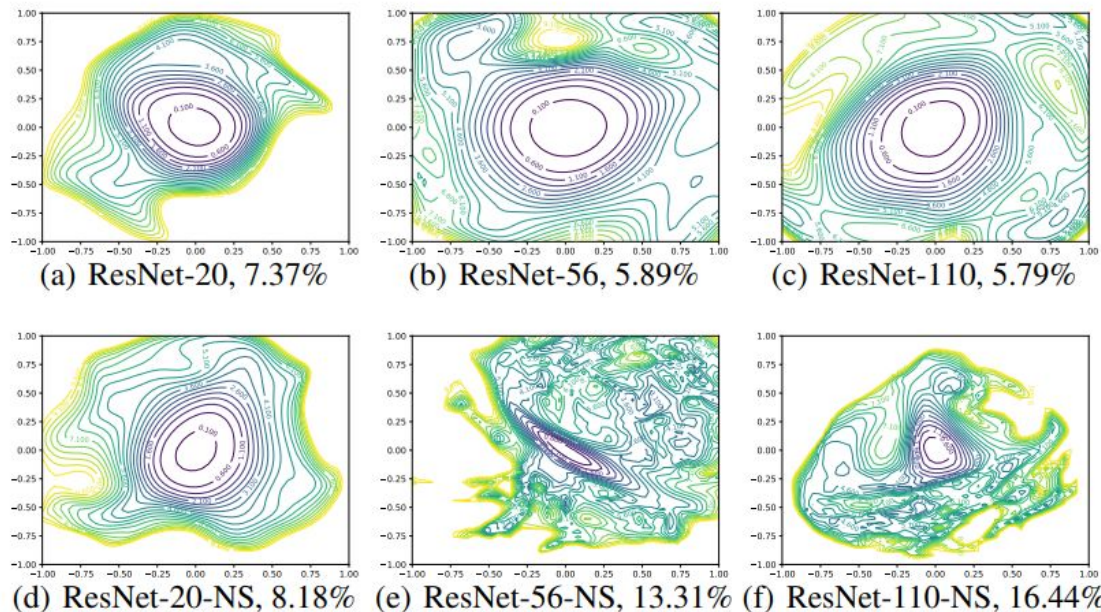


Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.

Зависимость от наличия skip-connections

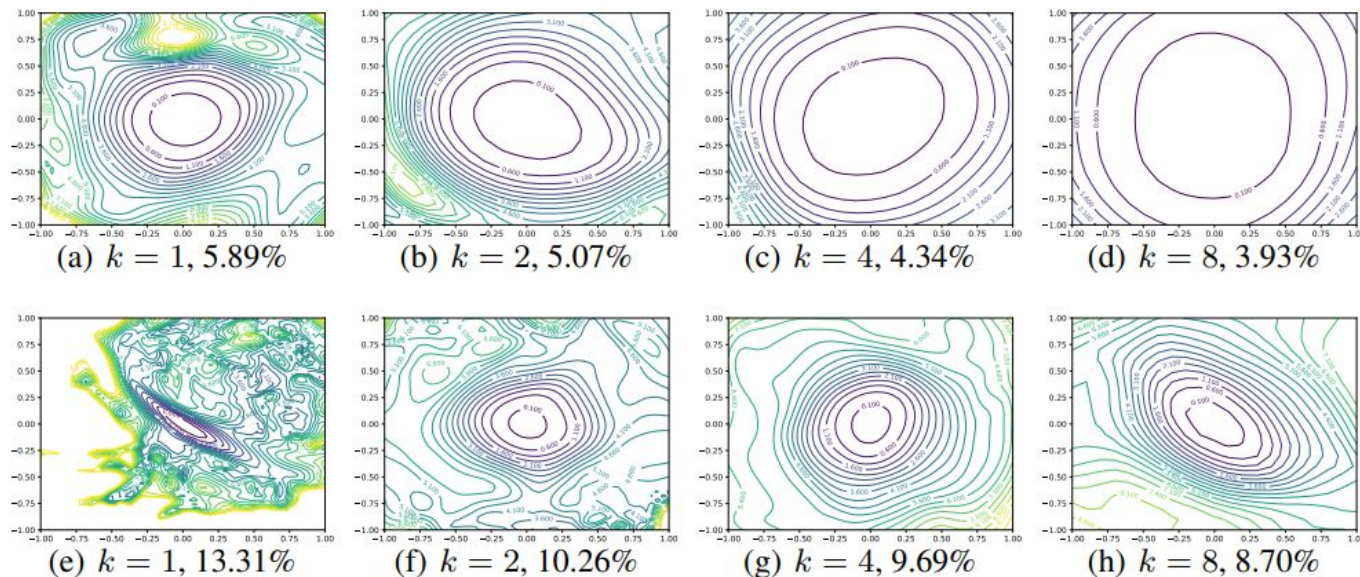
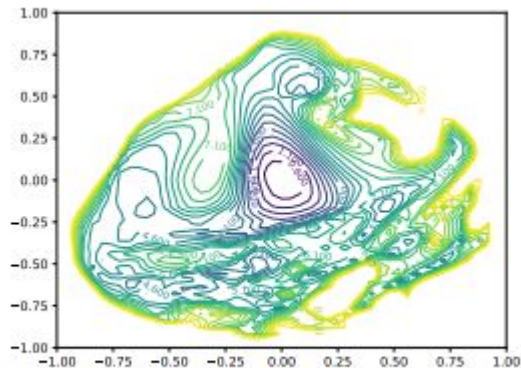


Figure 6: Wide-ResNet-56 on CIFAR-10 both with shortcut connections (top) and without (bottom). The label $k = 2$ means twice as many filters per layer. Test error is reported below each figure.

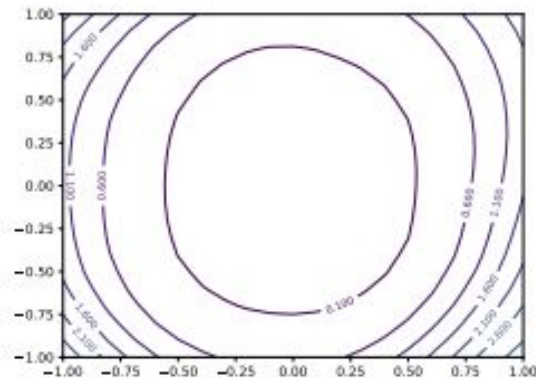
А что там с обобщаемостью?

11



(f) ResNet-110-NS, 16.44%

Пример наихудшей
обобщающей
способности



(d) $k = 8$, 3.93%

Пример наилучшей
обобщающей
способности

А наш метод точно работает?

12

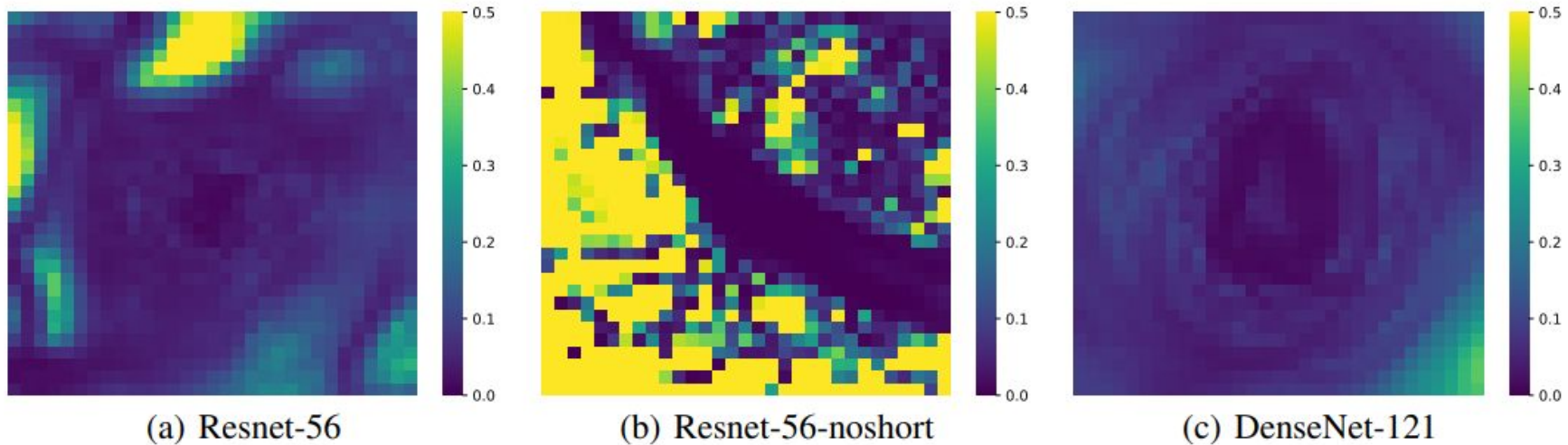
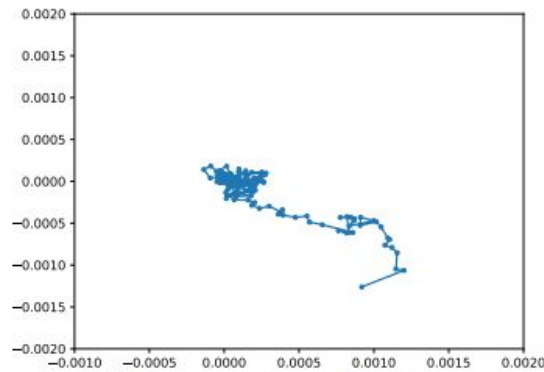


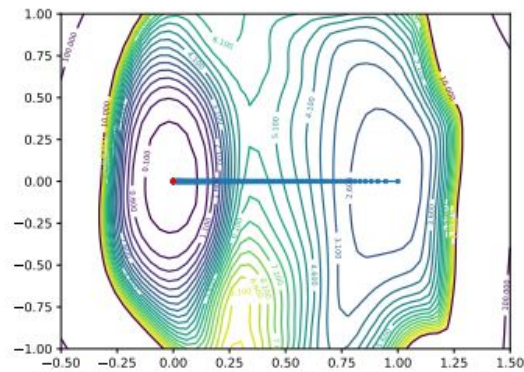
Figure 7: For each point in the filter-normalized surface plots, we calculate the maximum and minimum eigenvalue of the Hessian, and map the ratio of these two.

Проверка надежности описанного метода визуализации

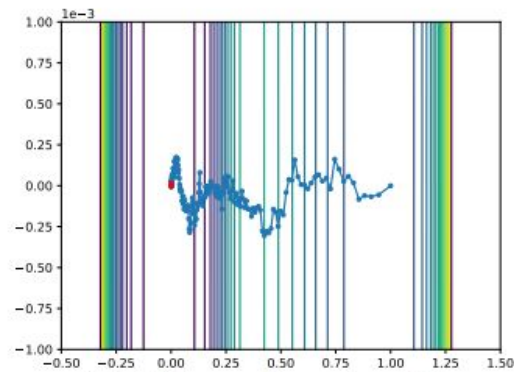
Проблема:



(a) Two random directions



(b) Random direction for y-axis



(c) Enlarged version (b)

Figure 8: Ineffective visualizations of optimizer trajectories. These visualizations suffer from the orthogonality of random directions in high dimensions.

Проблема малоразмерных пространств оптимизации

Решение: PCA (principal component analysis) - метод главных компонент

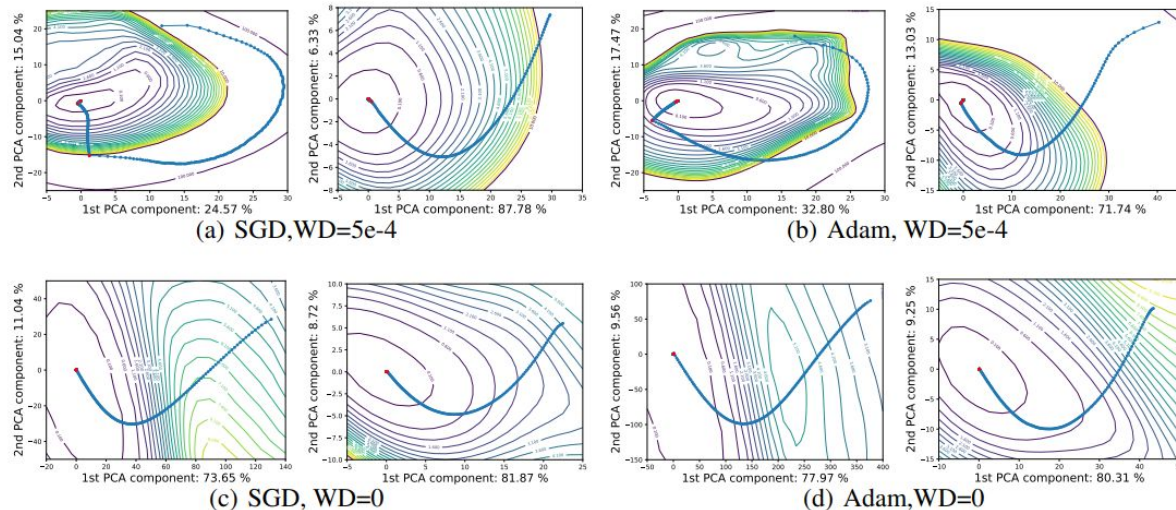
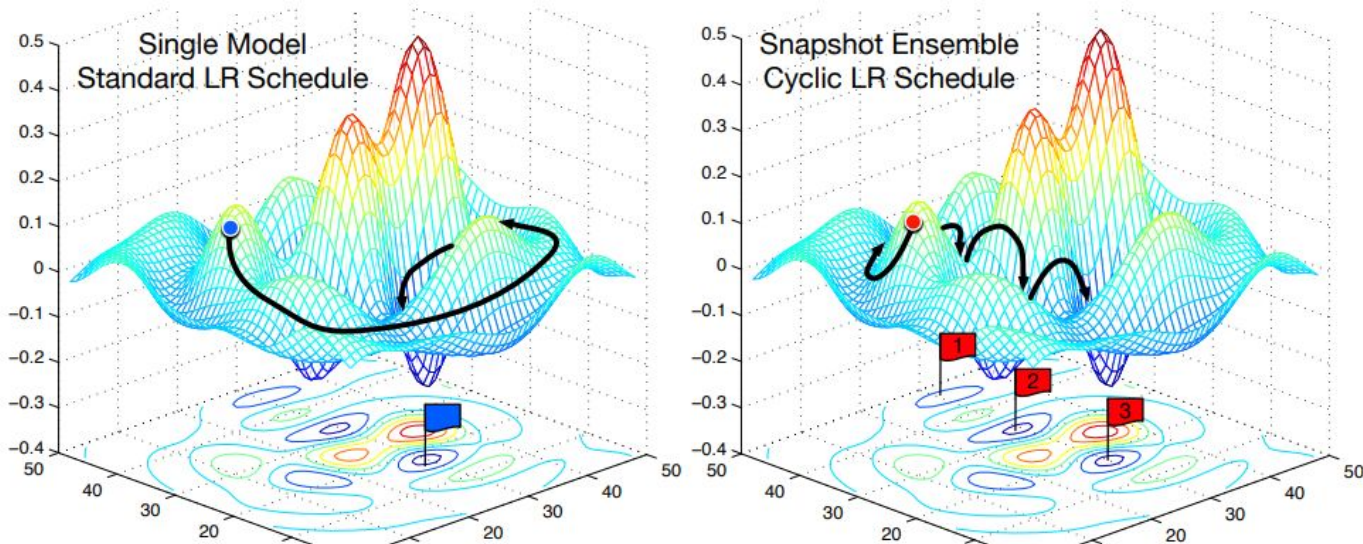


Figure 9: Projected learning trajectories use normalized PCA directions for VGG-9. The left plot in each subfigure uses batch size 128, and the right one uses batch size 8192.

Пример того, почему оптимизация работает

Ссылки на статьи:

- 1) Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs:
<https://arxiv.org/pdf/1802.10026.pdf>
- 2) SNAPSHOT ENSEMBLES: TRAIN 1, GET M FOR FREE: <https://arxiv.org/pdf/1704.00109.pdf>



Траектории градиентного спуска для Snapshot Ensembling

Snapshot Ensembling

17

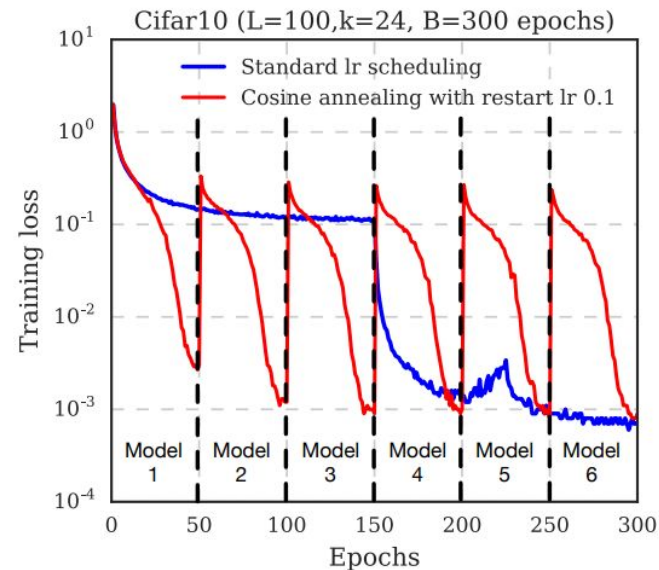
$$\alpha(t) = \frac{\alpha_0}{2} \left(\cos \left(\frac{\pi \text{mod}(t - 1, \lceil T/M \rceil)}{\lceil T/M \rceil} \right) + 1 \right)$$

t – номер итерации

T – количество итераций

M – количество циклов

α_0 - начальный lr

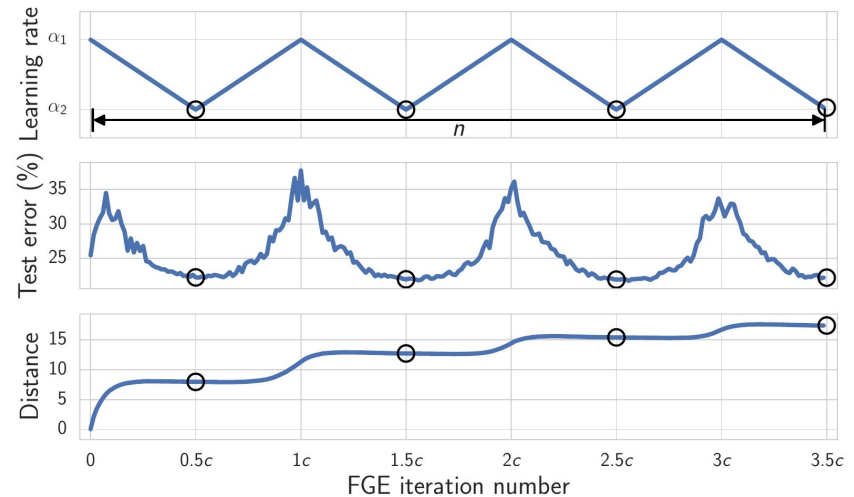


FGE – Fast Geometric Ensembling

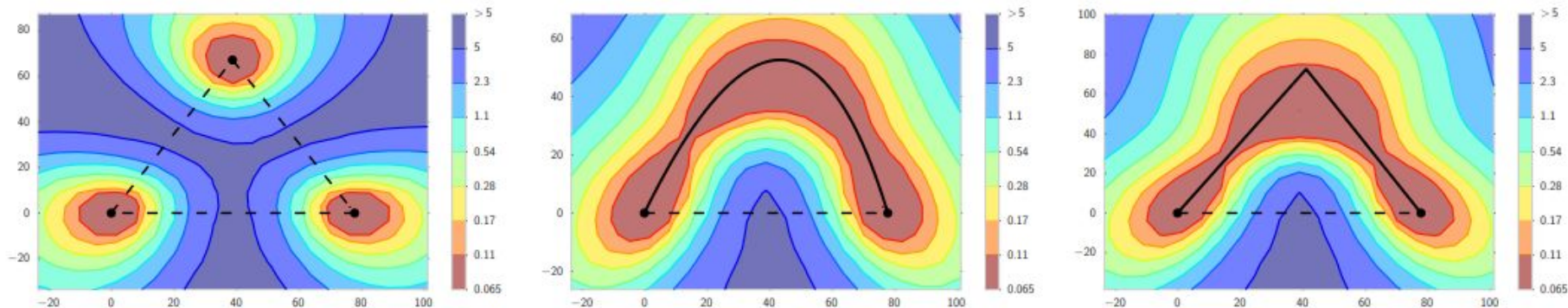
Данный метод является улучшением Snapshot Ensembling. В нём визуализация позволяет показать, почему в пространстве весов модели есть области, соединяющие локальные минимумы так, что функция потерь не сильно отличается от локальных минимумов.

$$\alpha(i) = \begin{cases} (1 - 2t(i))\alpha_1 + 2t(i)\alpha_2 & 0 < t(i) \leq \frac{1}{2} \\ (2 - 2t(i))\alpha_2 + (2t(i) - 1)\alpha_1 & \frac{1}{2} < t(i) \leq 1 \end{cases}$$

$$t(i) = \frac{1}{c}(\text{mod}(i - 1, c) + 1)$$



Пример визуализации:



Наглядный пример того, что FGE работает (линии соединяют локальные минимумы)

Table 1: Error rates (%) on CIFAR-100 and CIFAR-10 datasets for different ensembling techniques and training budgets. The best results for each dataset, architecture, and budget are **bolded**.

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

In this section we compare the proposed Fast Geometric Ensembling (**FGE**) technique against ensembles of independently trained networks (**Ind**), and SnapShot Ensembles (**SSE**) [11], a recent state-of-the-art fast ensembling approach.

Визуализировать функцию потерь нужно, чтобы специалист по нейронным сетям мог правильно выбрать:

- архитектуру нейросети
- оптимизатор
- размер батча

А также в таких методах, как FGE и Snapshot Ensembling визуализация помогает лучше понять суть метода.

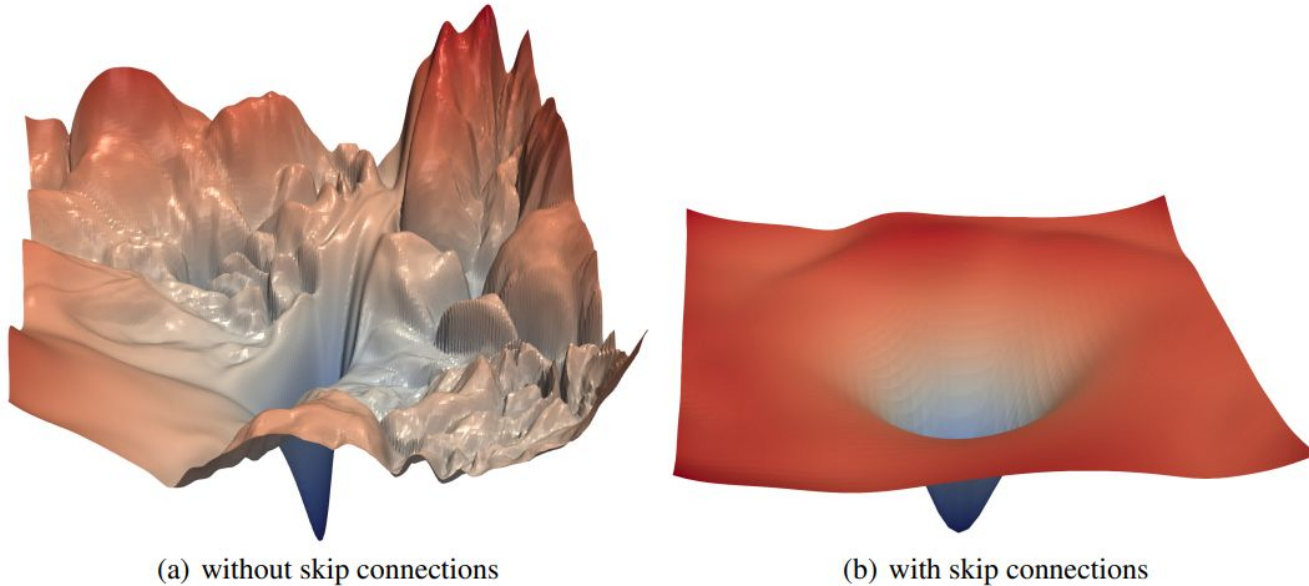
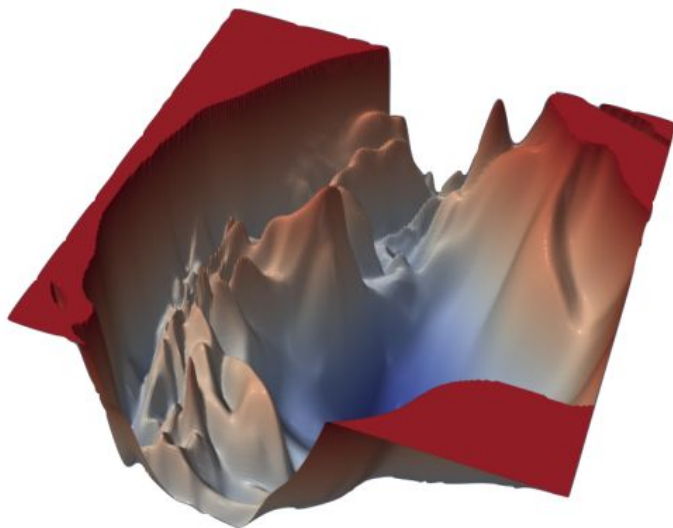
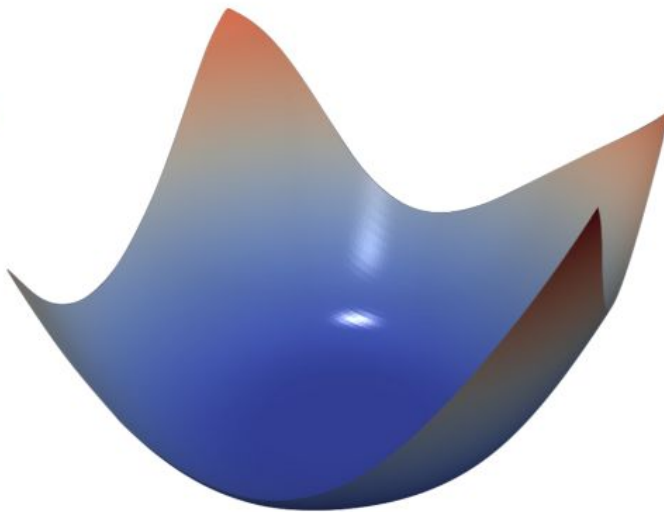


Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.



(a) ResNet-110, no skip connections



(b) DenseNet, 121 layers

Figure 4: The loss surfaces of ResNet-110-noshort and DenseNet for CIFAR-10.

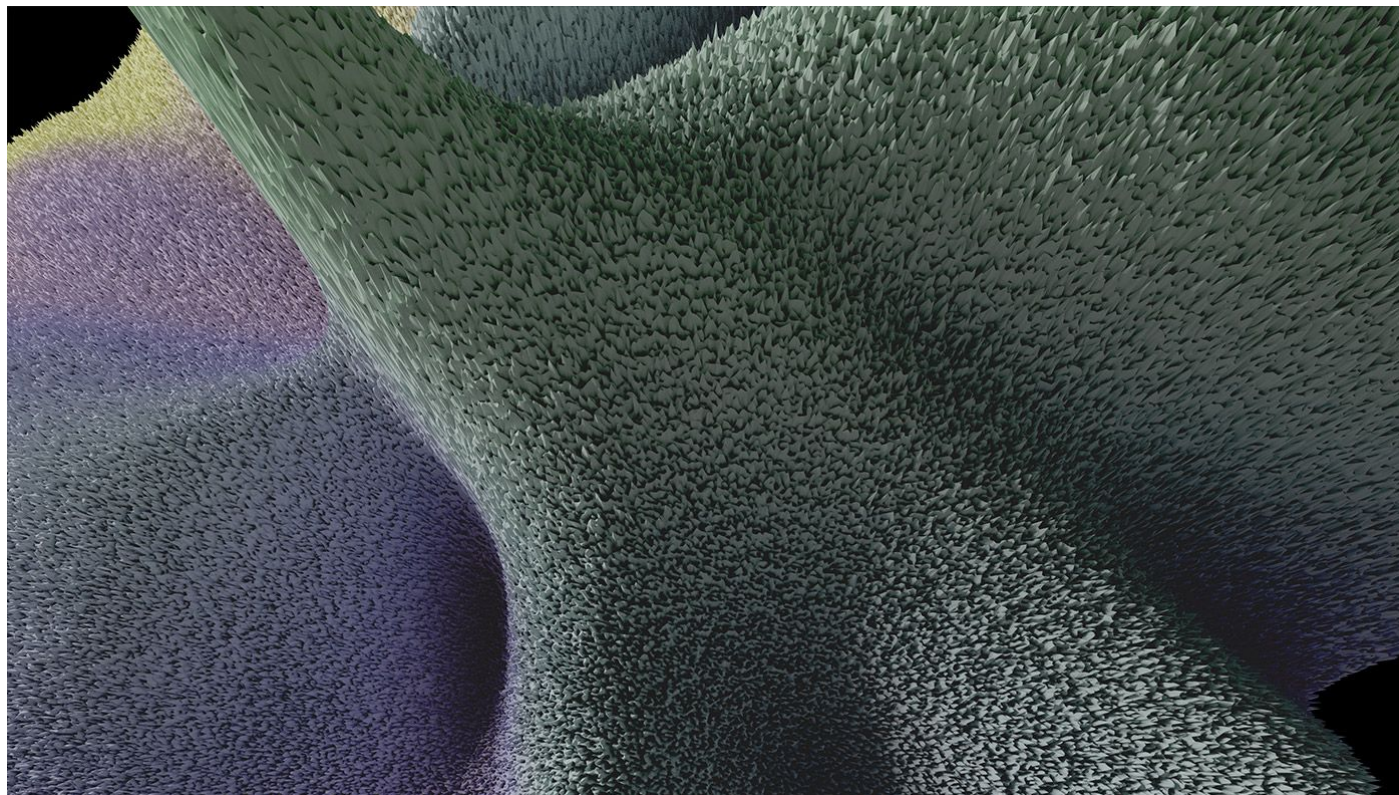
Очень красивые картинки

25



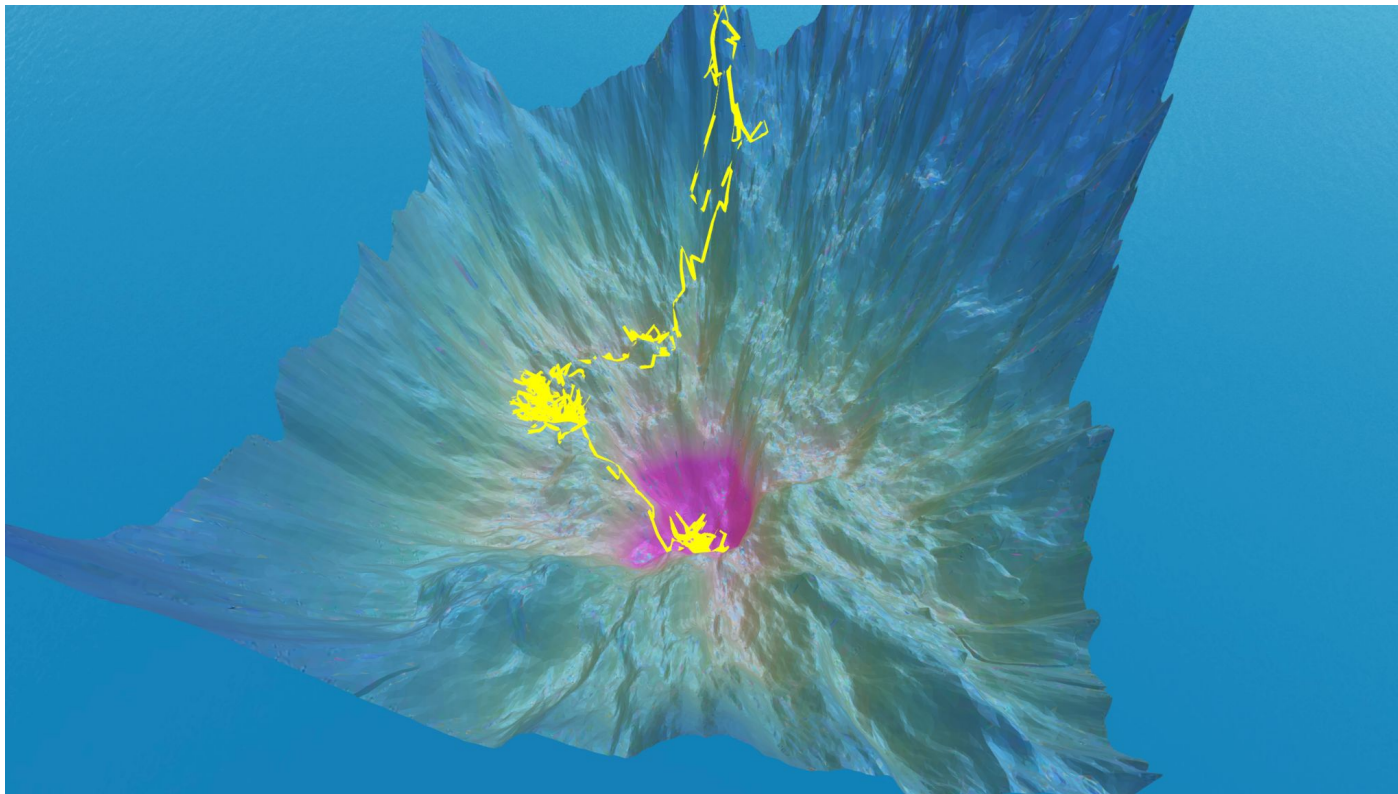
Dropout visualizes

Очень красивые картинки

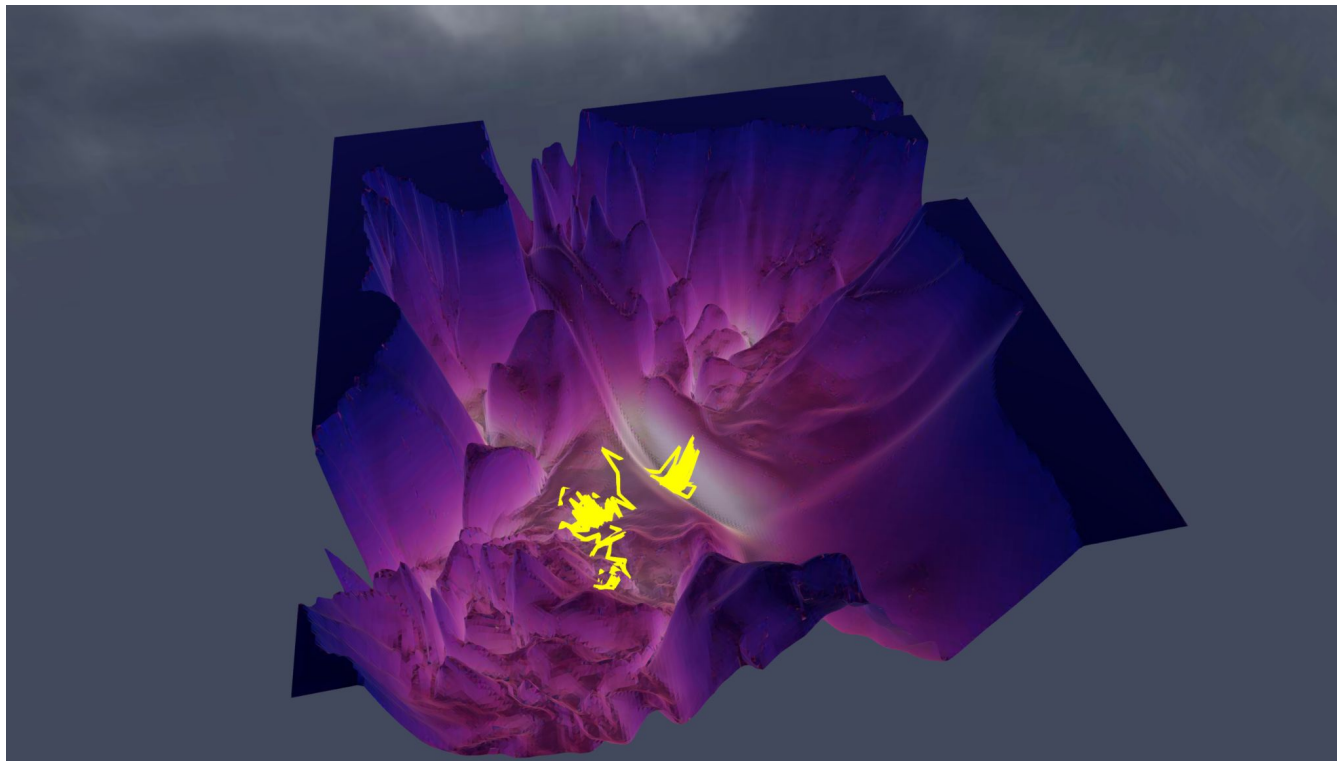


LATENT visualizes the initial stages of the training of a Wasserstein GP GAN network.

Очень красивые картинки



Визуализация градиентного спуска



Визуализация градиентного спуска

Последние две картинки – результат моих экспериментов на сайте <https://losslandscape.com/explorer>. На нём много красоты, на которую можно смотреть часами.



А это красивое представление контурных графиков по случайным направлениям

У вас есть вопросы?