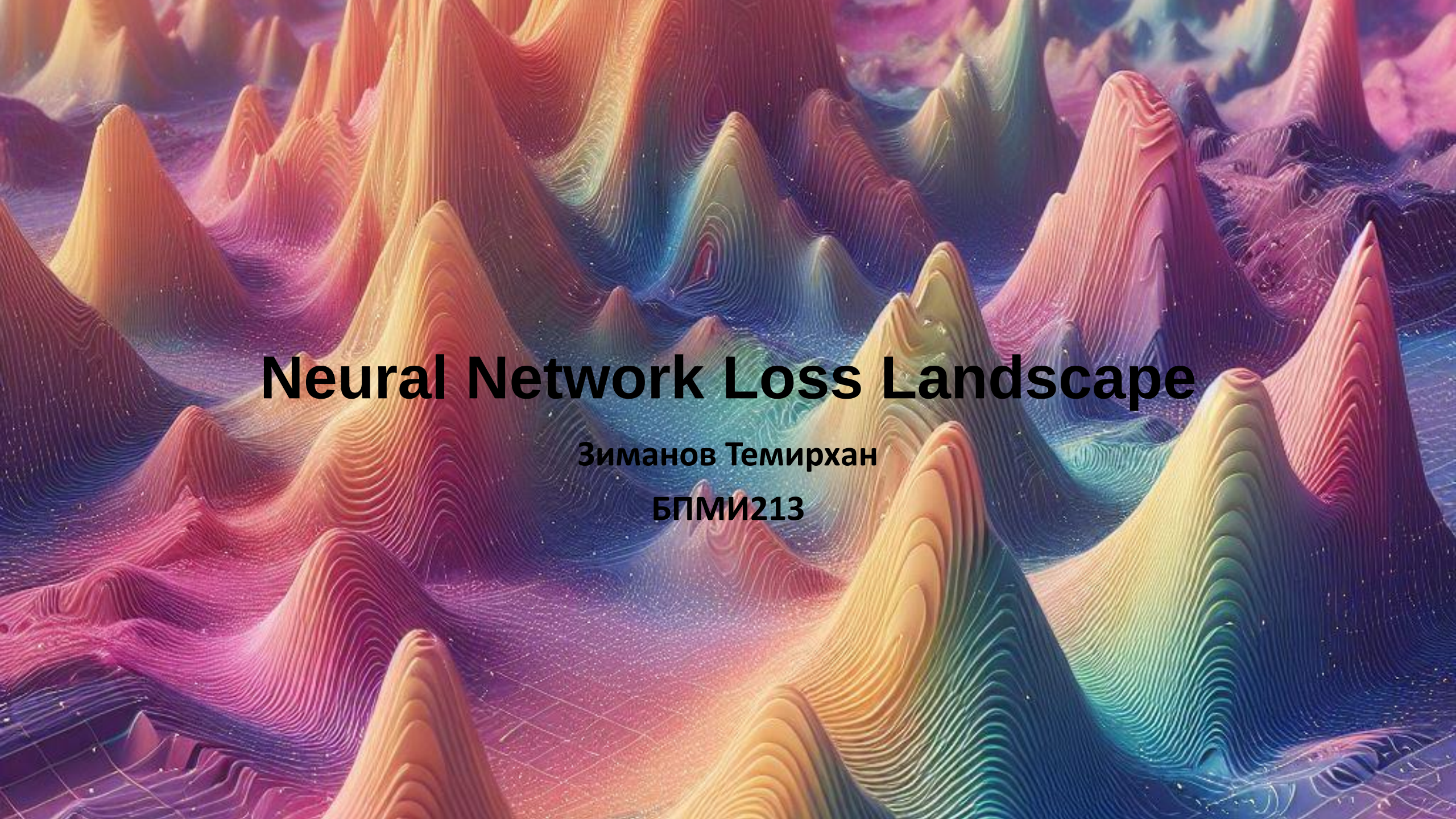


Neural Network Loss Landscape

Зиманов Темирхан

БПМИ213

1. Введение

Функция потерь нейронной сети

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \ell(x_i, y_i; \theta)$$

- Задача: найти глобальный минимум
- Дорого вычислять
- Невыпуклая

2. Мотивация

- Восхождение в плохую погоду
- Алгоритм: идти вверх, пока возможно
- Где и как скоро мы окажемся?

- Как выглядит рельеф?
- Существует ли хороший алгоритм?

- Визуализация может помочь оценить условия задачи и вероятность существования решения

3. Визуализация

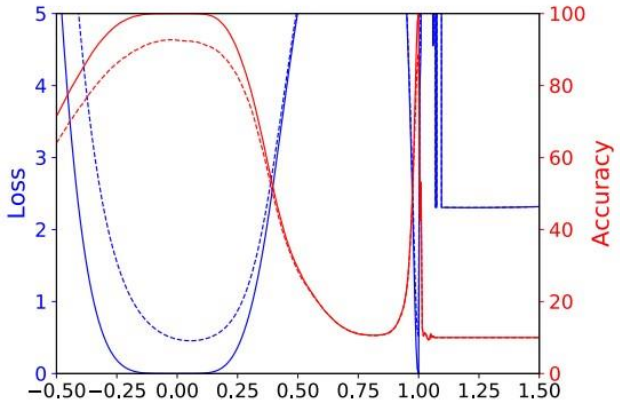
Одномерная линейная интерполяция

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

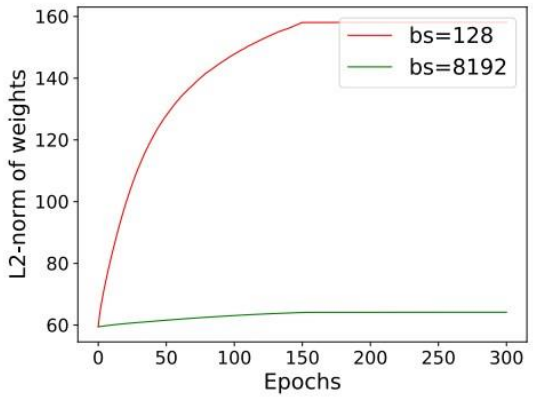
$$f(\alpha) = L(\theta(\alpha))$$

CIFAR-10, VGG9

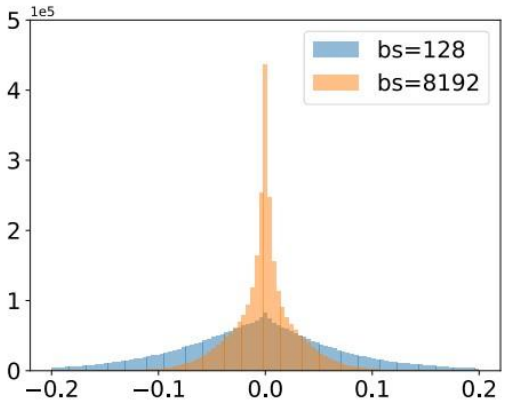
$$f(\alpha) = L(\theta^s + \alpha(\theta^l - \theta^s))$$



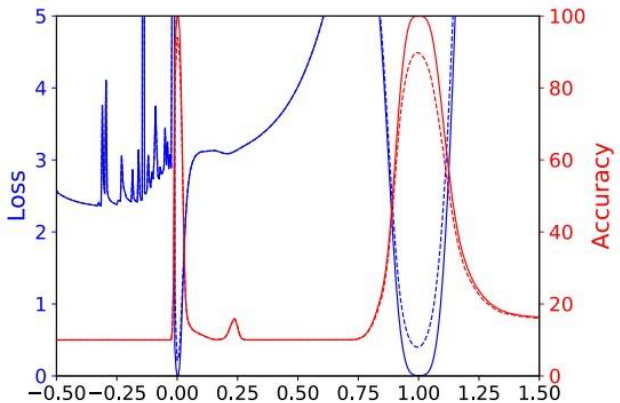
(a) 7.37% 11.07%



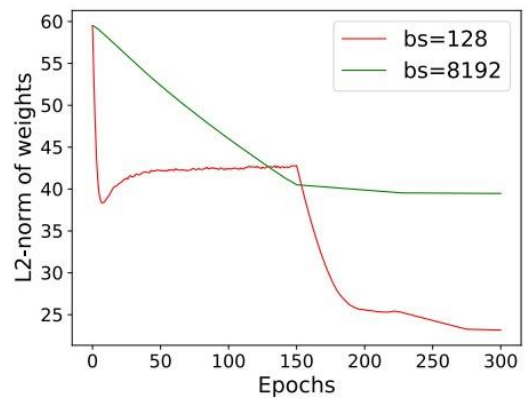
(b) $\|\theta\|_2$, WD=0



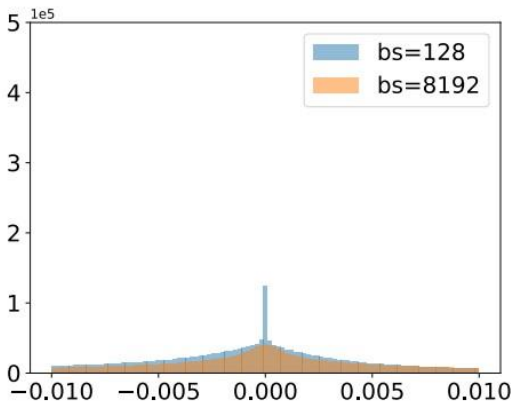
(c) WD=0



(d) 6.0% 10.19%



(e) $\|\theta\|_2$, WD=5e-4

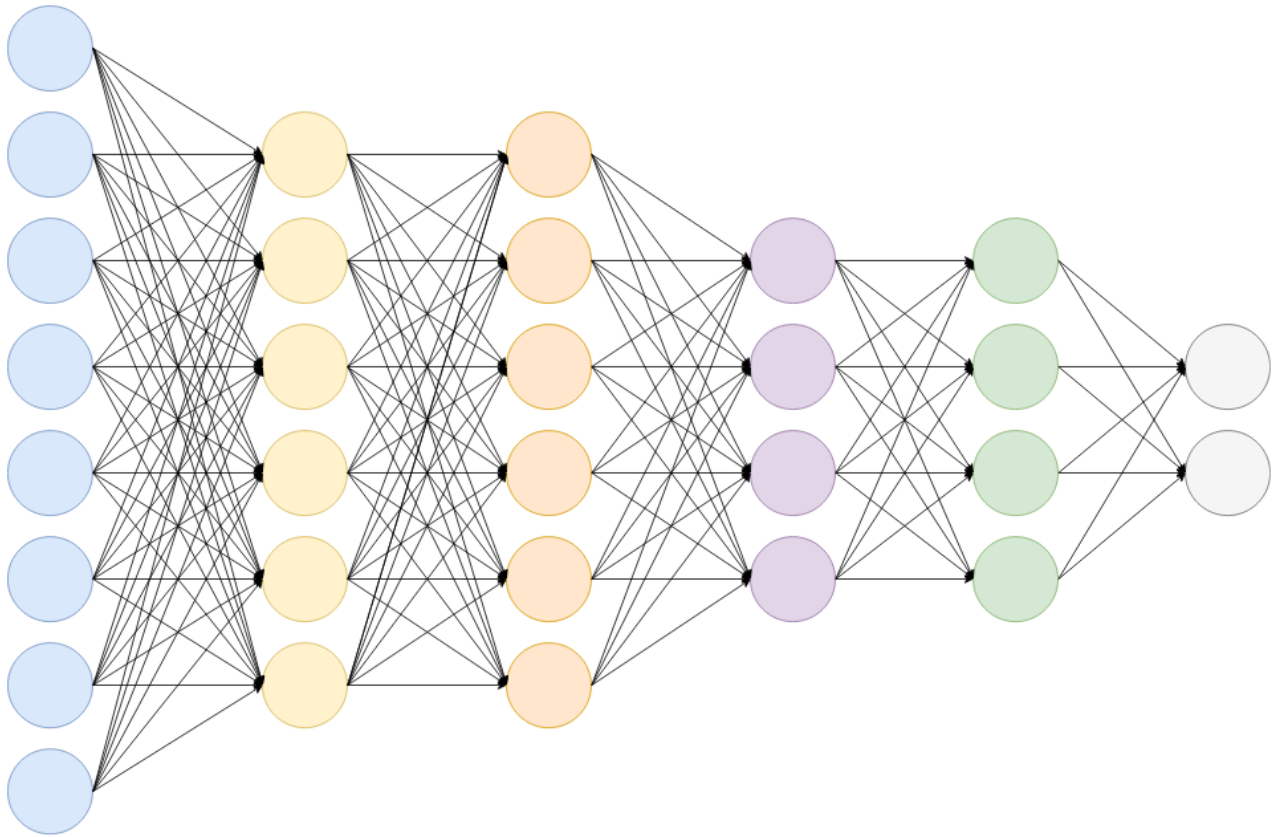


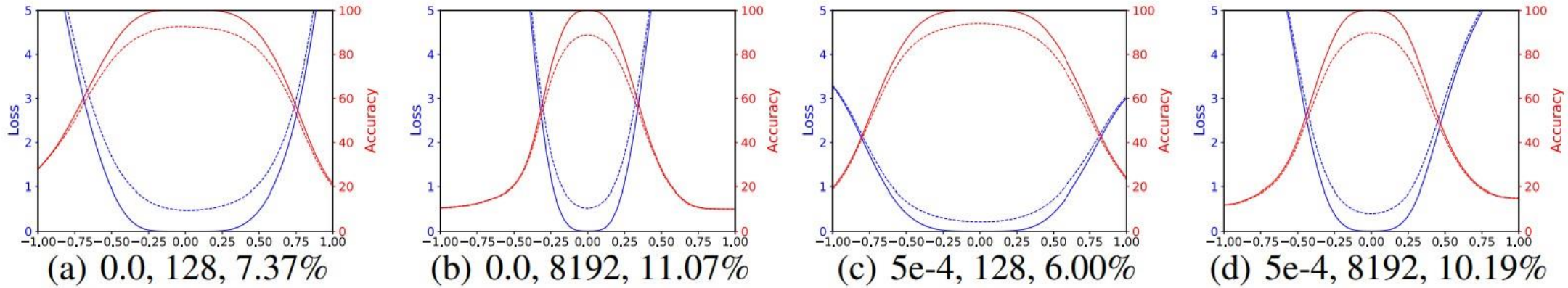
(f) WD=5e-4

Нормализация

$$d = \theta_2 - \theta_1$$

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

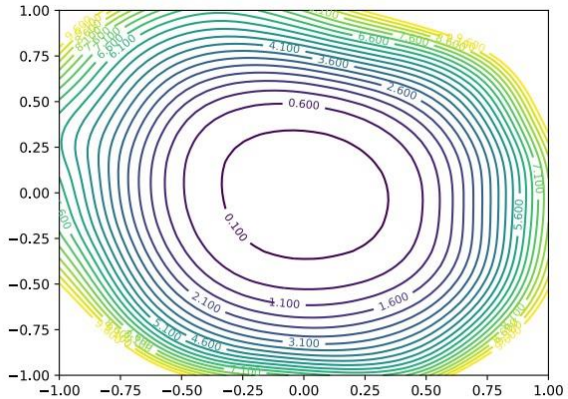




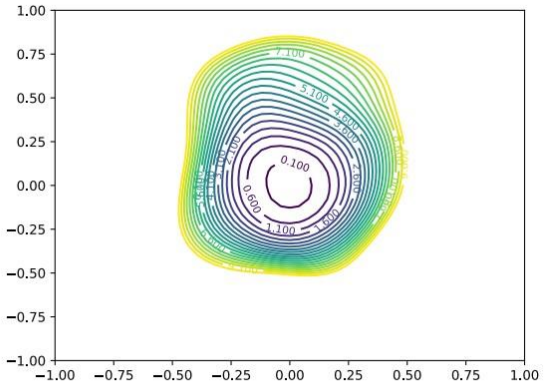
Теперь острота коррелирует с обобщающей способностью

Двухмерная интерполяция и контурные линии

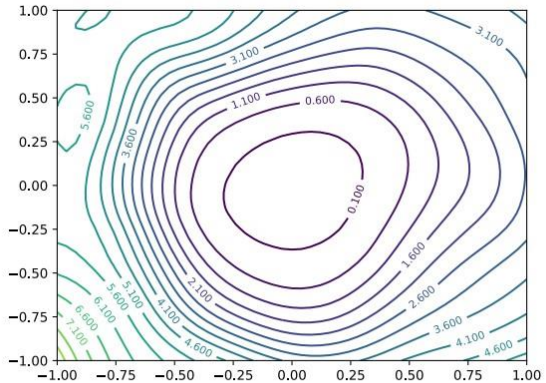
$$f(\alpha, \beta) = L(\theta_1 + \alpha d_1 + \beta d_2)$$



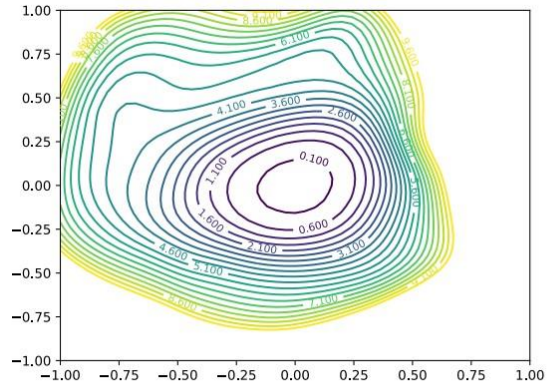
(e) 0.0, 128, 7.37%



(f) 0.0, 8192, 11.07%



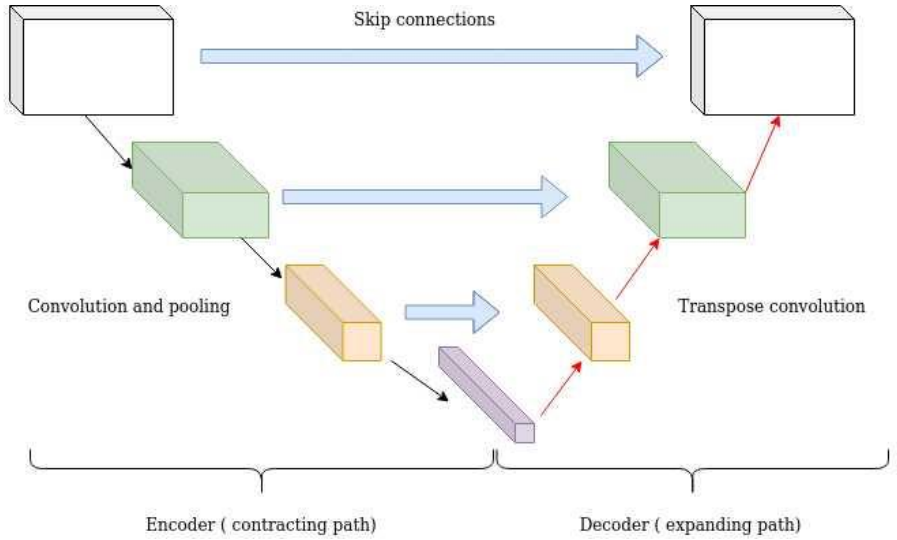
(g) 5e-4, 128, 6.00%



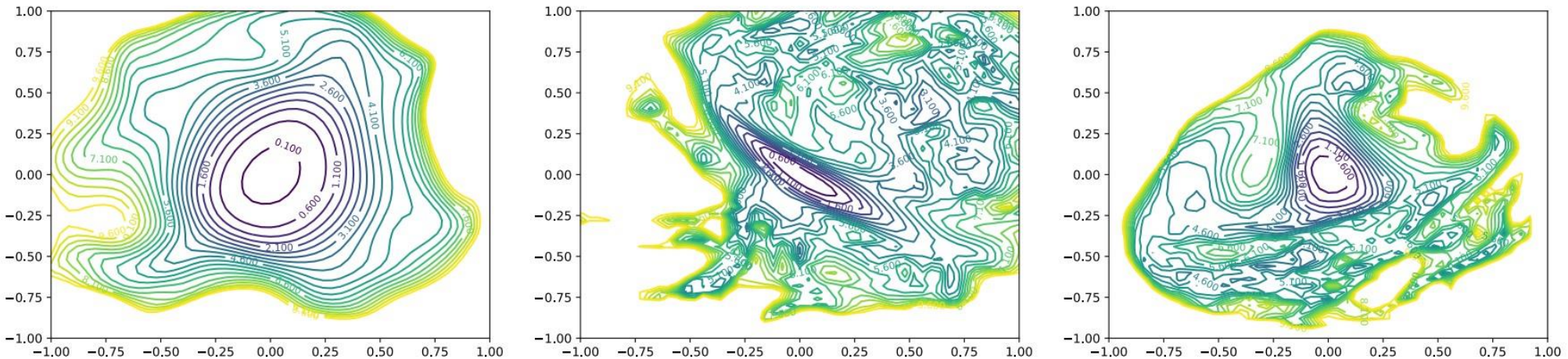
(h) 5e-4, 8192, 10.19%

4. Эксперименты

- CIFAR-10
- ResNet-20/56/110
- ResNet-20/56/110-noshort
- “Wide” ResNets

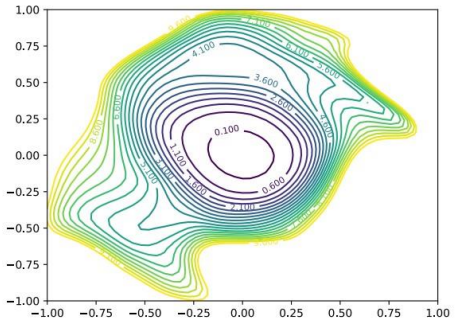


Влияние глубины

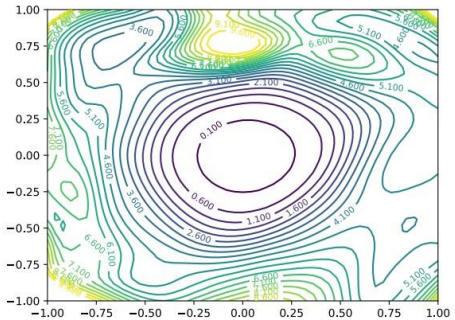


(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%

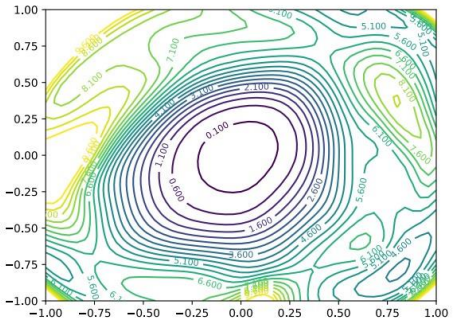
Влияние skip connections



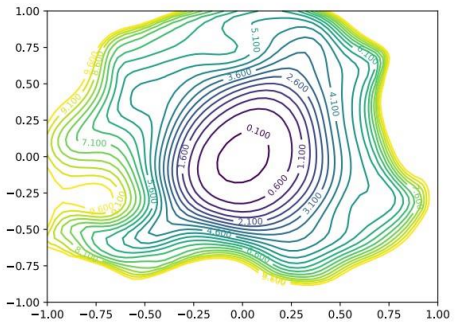
(a) ResNet-20, 7.37%



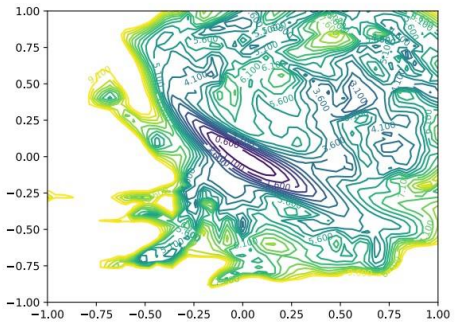
(b) ResNet-56, 5.89%



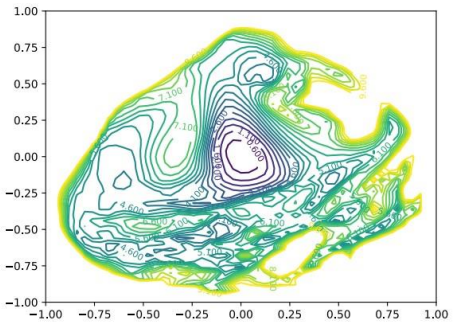
(c) ResNet-110, 5.79%



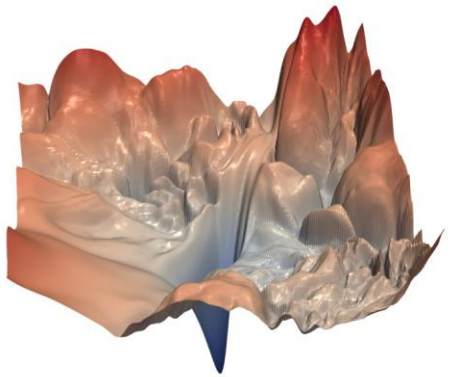
(d) ResNet-20-NS, 8.18%



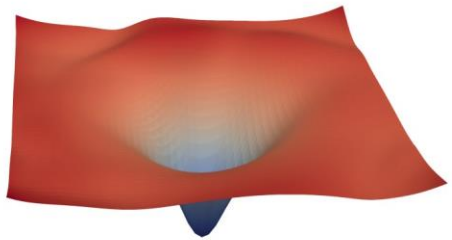
(e) ResNet-56-NS, 13.31%



(f) ResNet-110-NS, 16.44%

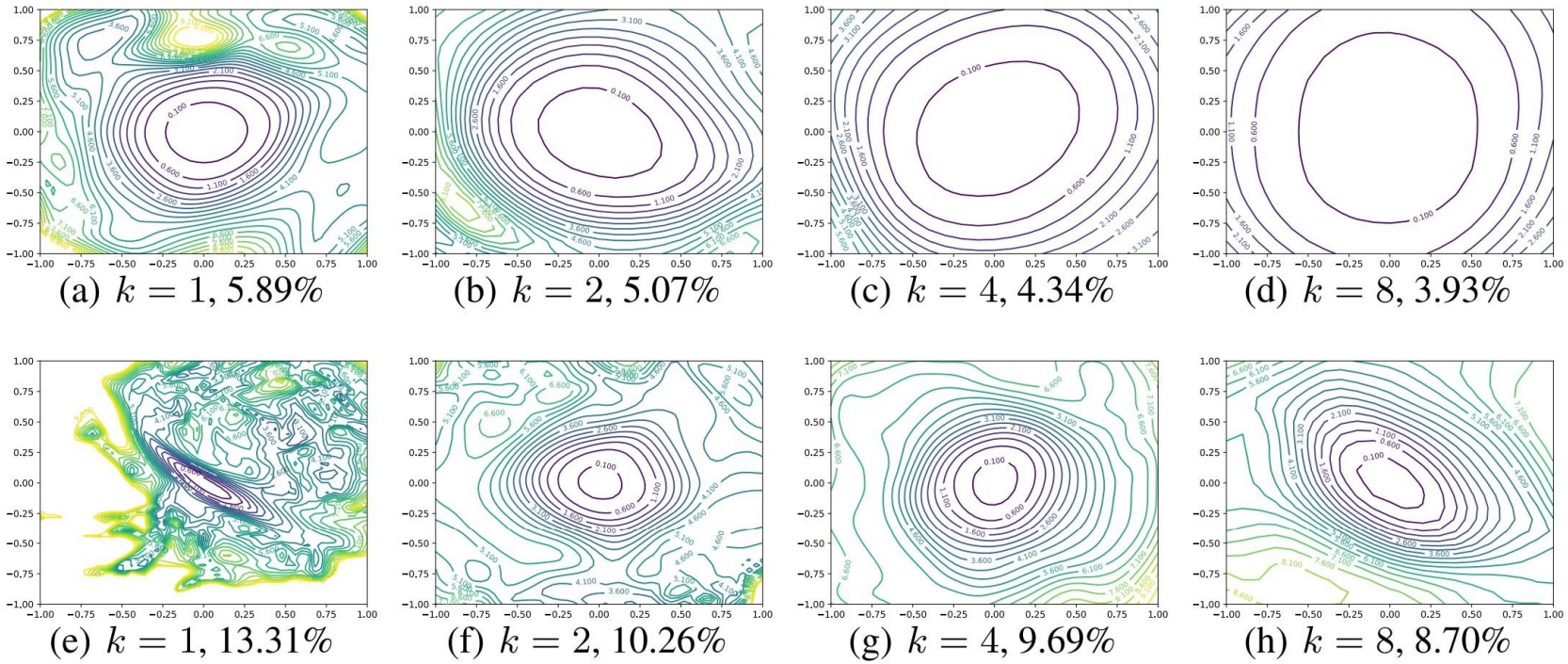


(a) without skip connections



(b) with skip connections

Влияние ширины

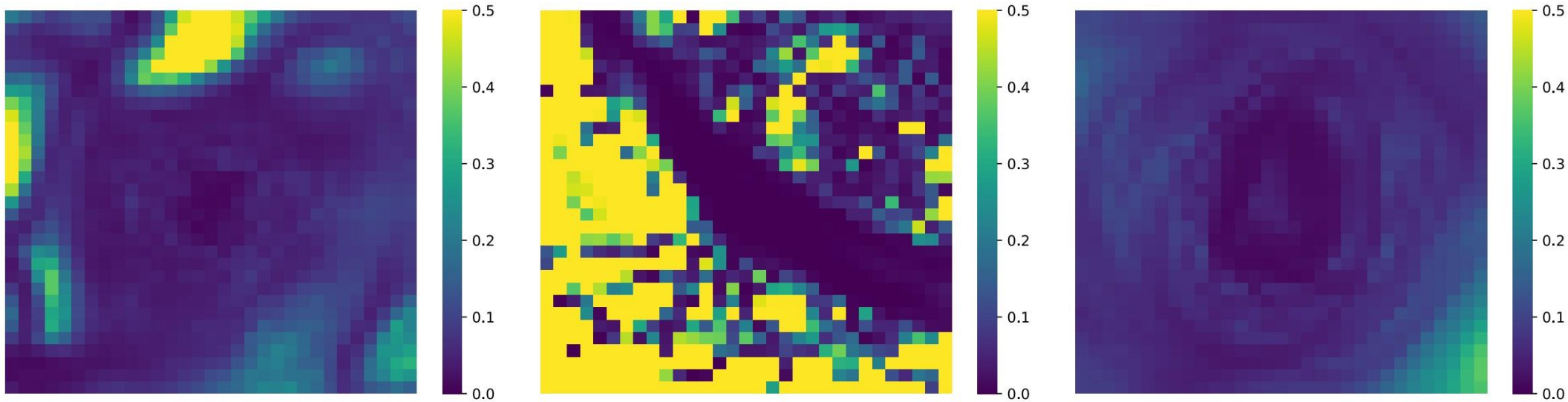


(a) $k = 1$, 5.89% (b) $k = 2$, 5.07% (c) $k = 4$, 4.34% (d) $k = 8$, 3.93%

(e) $k = 1$, 13.31% (f) $k = 2$, 10.26% (g) $k = 4$, 9.69% (h) $k = 8$, 8.70%

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.

- Но можно ли делать выводы о выпуклости функции по двухмерной интерполяции?
- В некотором смысле — да

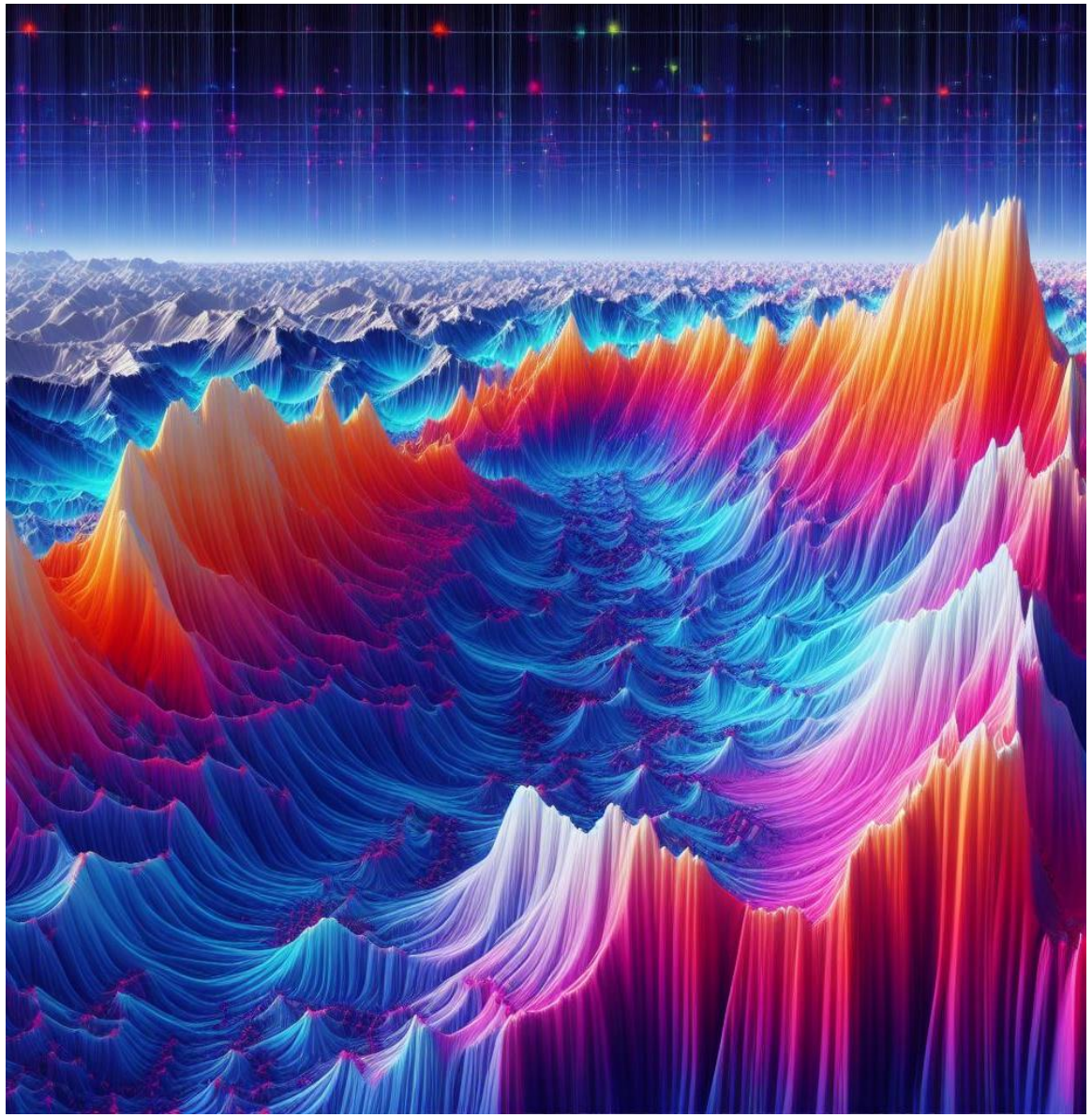


(a) Resnet-56 (b) Resnet-56-noshort (c) DenseNet-121

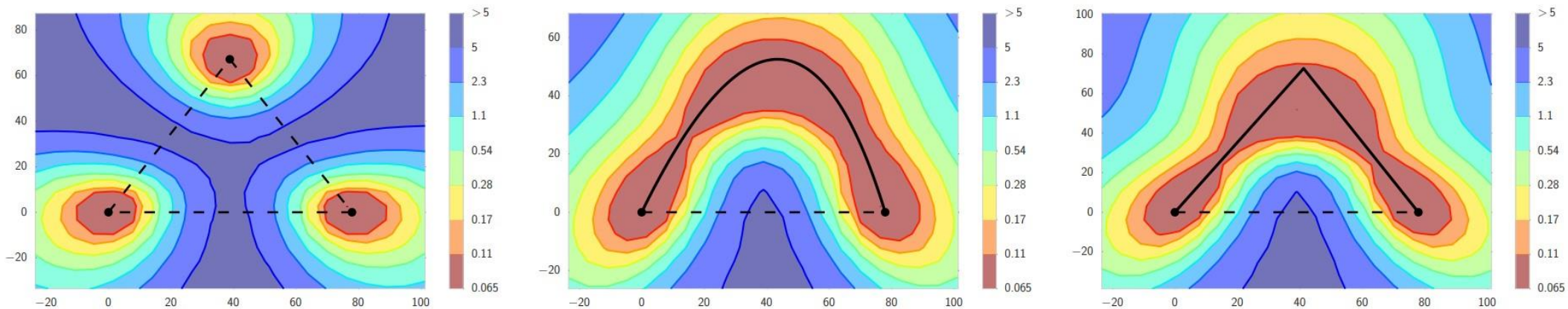
Figure 7: For each point in the filter-normalized surface plots, we calculate the maximum and minimum eigenvalue of the Hessian, and map the ratio of these two.

5. Вдохновение

- Как побыстрее пройти от одного пика к другому?
- Может существует путь, вдоль которого высота почти не меняется?



Да!


80
60
40
20
0
−20
−20
0
20
40
60
80
100
>5
5
2.3
1.1
0.54
0.28
0.17
0.11
0.065
60
40
20
0
−20
−20
0
20
40
60
80
100
>5
5
2.3
1.1
0.54
0.28
0.17
0.11
0.065
100
80
60
40
20
0
−20
−20
0
20
40
60
80
100
>5
5
2.3
1.1
0.54
0.28
0.17
0.11
0.065

6. Заключение

- Визуализации могут помочь в выборе архитектуры, оптимизатора, гиперпараметров
- Визуализации могут пролить свет на устройство локальных минимумов