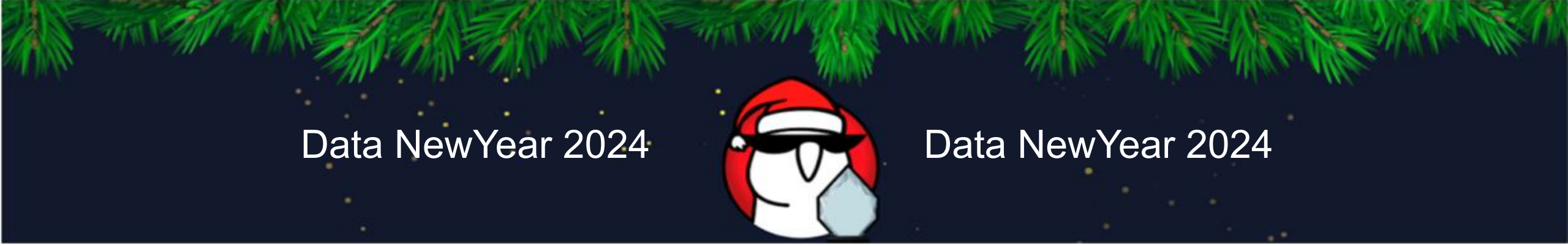


Окончательное решение проблемы затухания градиента через инструменты наблюдения за обучающейся сетью, без скипконнекшенов и дополнительных модулей

Влад kraidiky Голощапов

независимый исследователь

Канал для обсуждений: @GradientWitnesses

kraidiky@gmail.com, t.me/kraidiky

Полезные ссылки:

- Предыдущий доклад: https://www.youtube.com/watch?v=Npm-awHtfeM – «Моментум истины: Не всем известные свойства оптимизаторов с импульсом - SGD,Adam и т.д. »
- Статья про визуализации траектории: https://habr.com/ru/articles/221049/ - «Подглядываем за метаниями нейросети»
- Ссылка на коллаб, в котором видна ошибка в BatchNorm: https://colab.research.google.com/drive/1FVslR7zYxpybXQt_UN6rtG4GZQP9MQsN из-за этого некоторые метрики сетей его использующих могут быть не совсем точны и не воспроизводимы.

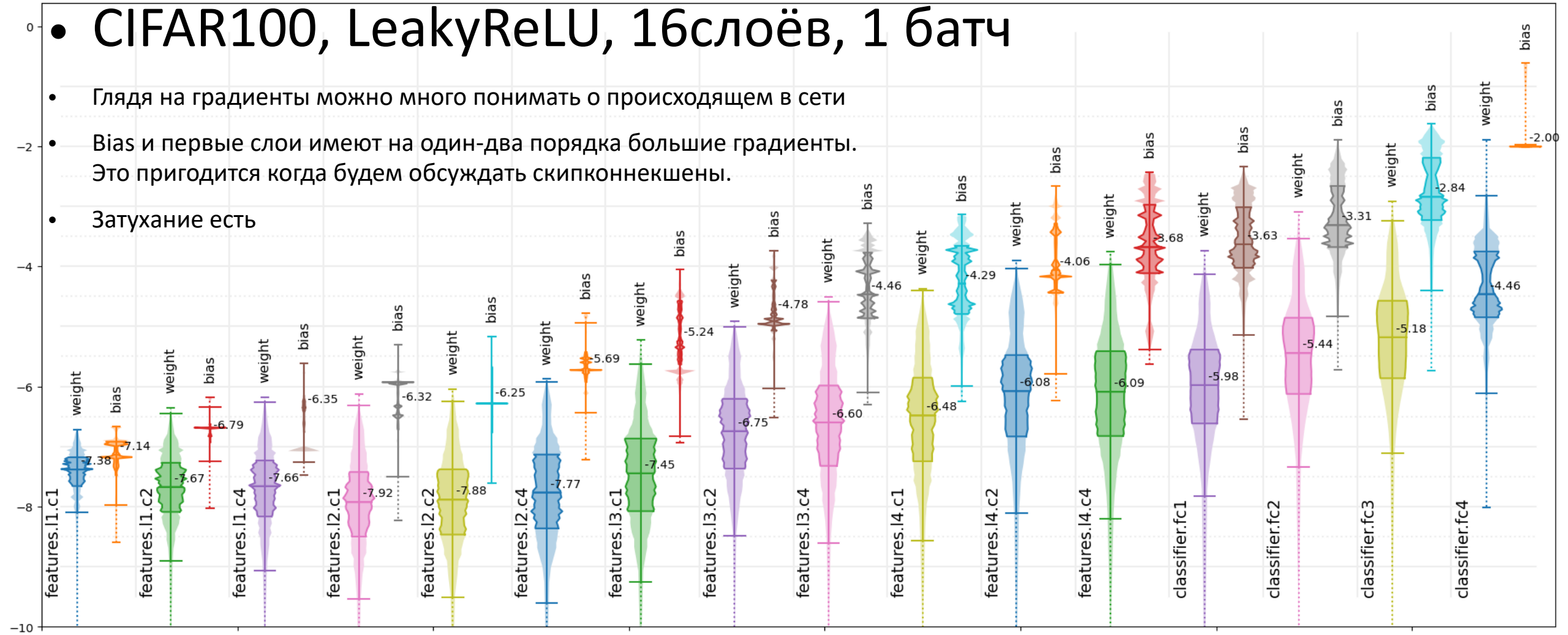
- Почти все сталкивались с затуханием градиентов
- Затухание градиентов может доставлять неудобства.
- Почти всем слышали при обучении почему затухание происходит и слышали несколько приёмов как с ним бороться, но ни понимание ни способы борьбы не точны и не окончательны.
- Затухания градиентов мало кто видел! Трудно формировать интуицию относительно явления, которого не видел.

Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # CIFAR100, LeakyReLU, 16слоёв, 1 батч

- Глядя на градиенты можно много понимать о происходящем в сети
- Bias и первые слои имеют на один-два порядка большие градиенты.
  Это пригодится когда будем обсуждать скипконнекшены.
- Затухание есть

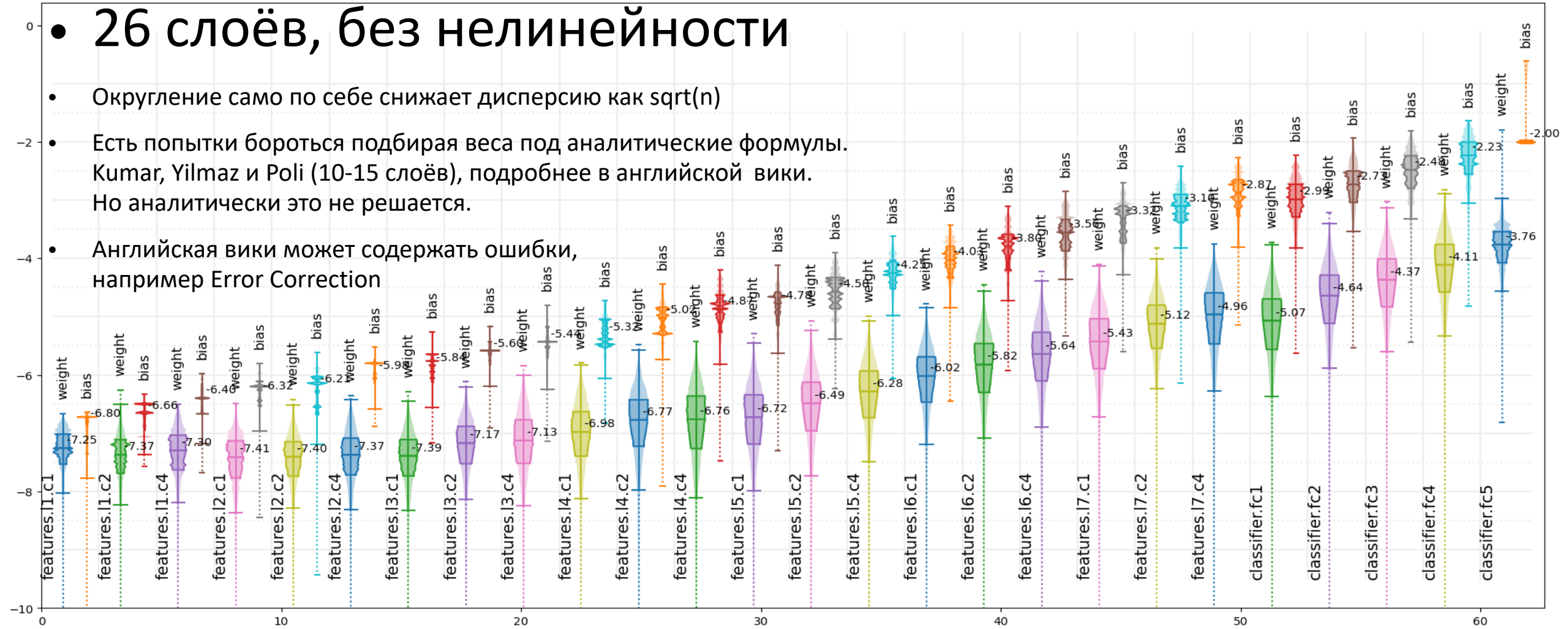


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

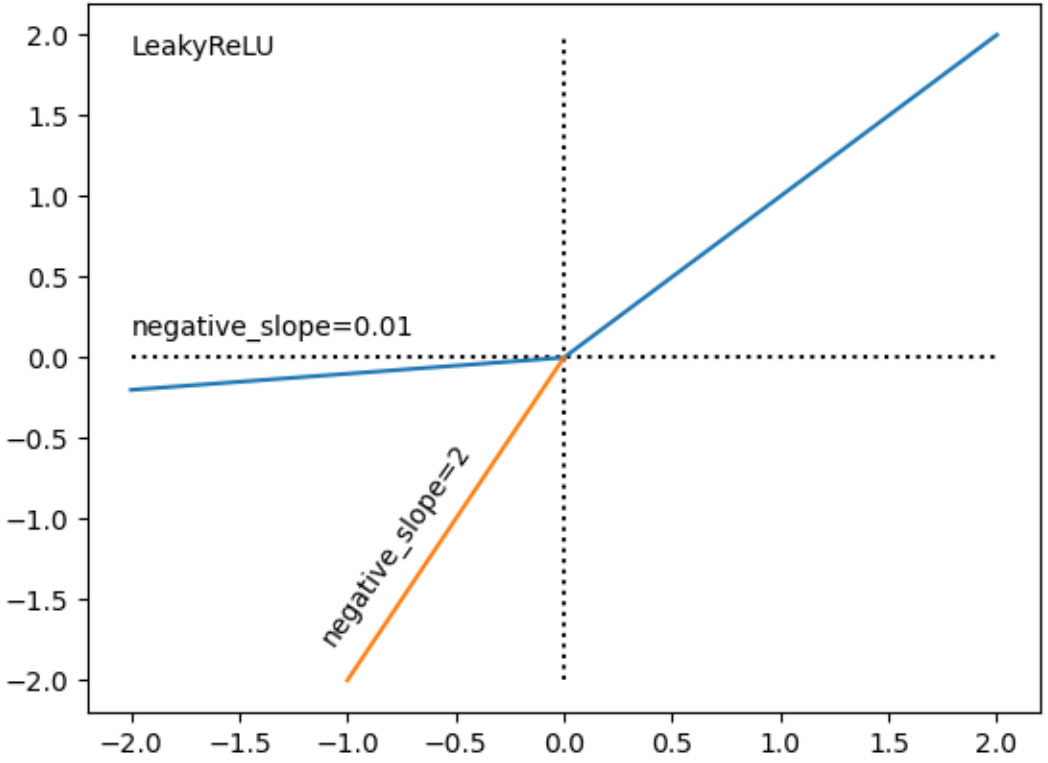
• 26 слоёв, без нелинейности

- Округление само по себе снижает дисперсию как sqrt(n)
- Есть попытки бороться подбирая веса под аналитические формулы. Kumar, Yilmaz и Poli (10-15 слоёв), подробнее в английской вики. Но аналитически это не решается.
- Английская вики может содержать ошибки, например Error Correction



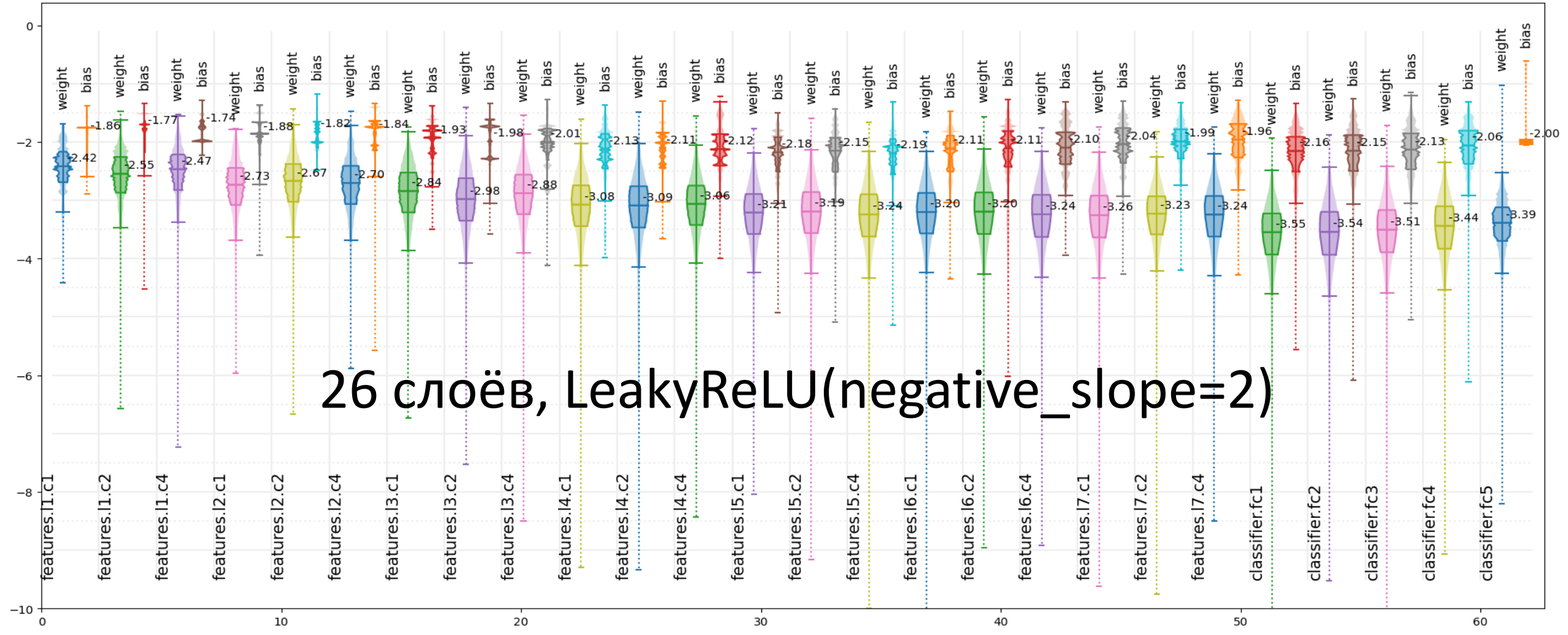
Заменим функцию активации на такую, которая усиливает градиенты. Из коробки:

torch.nn.LeakyReLU(negative_slope=2)



Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого



Data NewYear 2024
Голощапов Влад

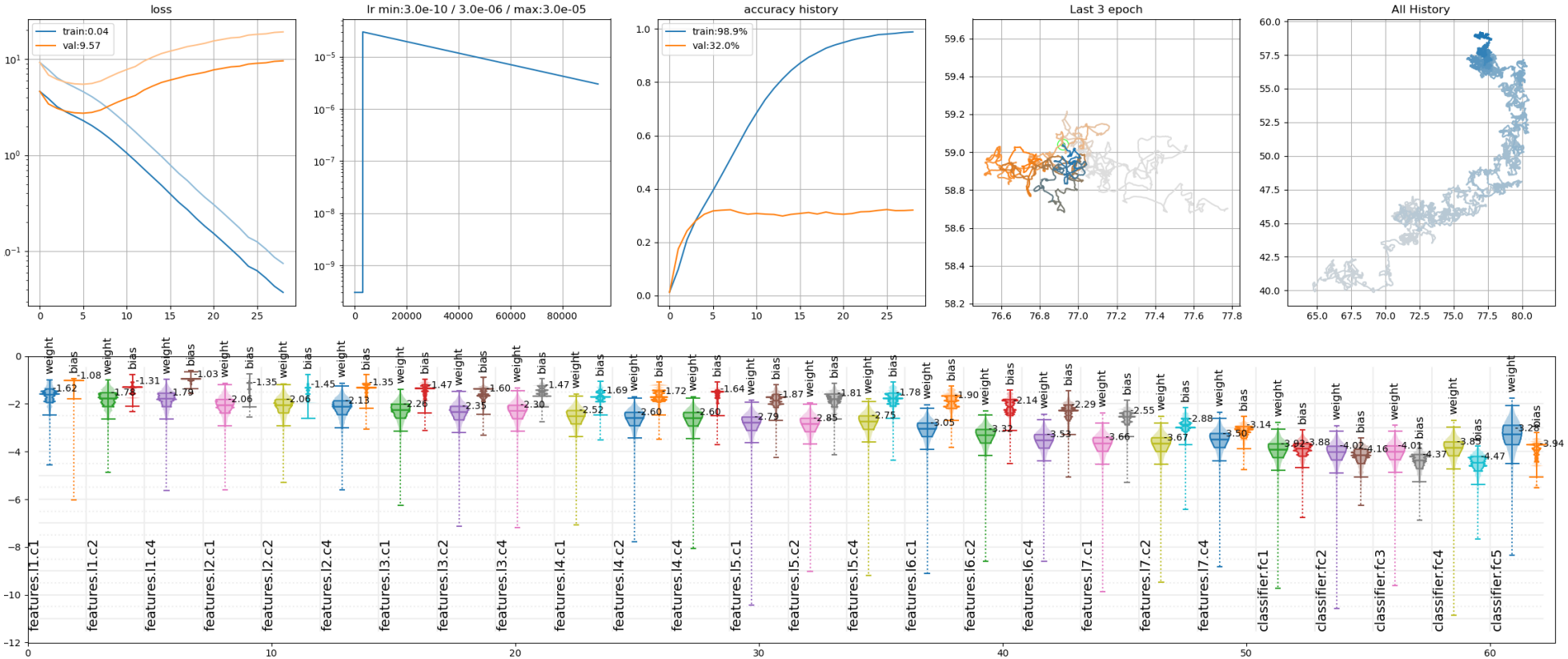
Про не затухание градиентов и видение невидимого

Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель способна учиться с такой нетрадиционной нелинейностью, получается нормально.

- Эта архитектура явно избыточна для этой задаче.
- Обратите внимание, что loss может расти когда accuracy уже не портится.
- Но 32 слоя это совсем не много, мы должны пойти глубже.

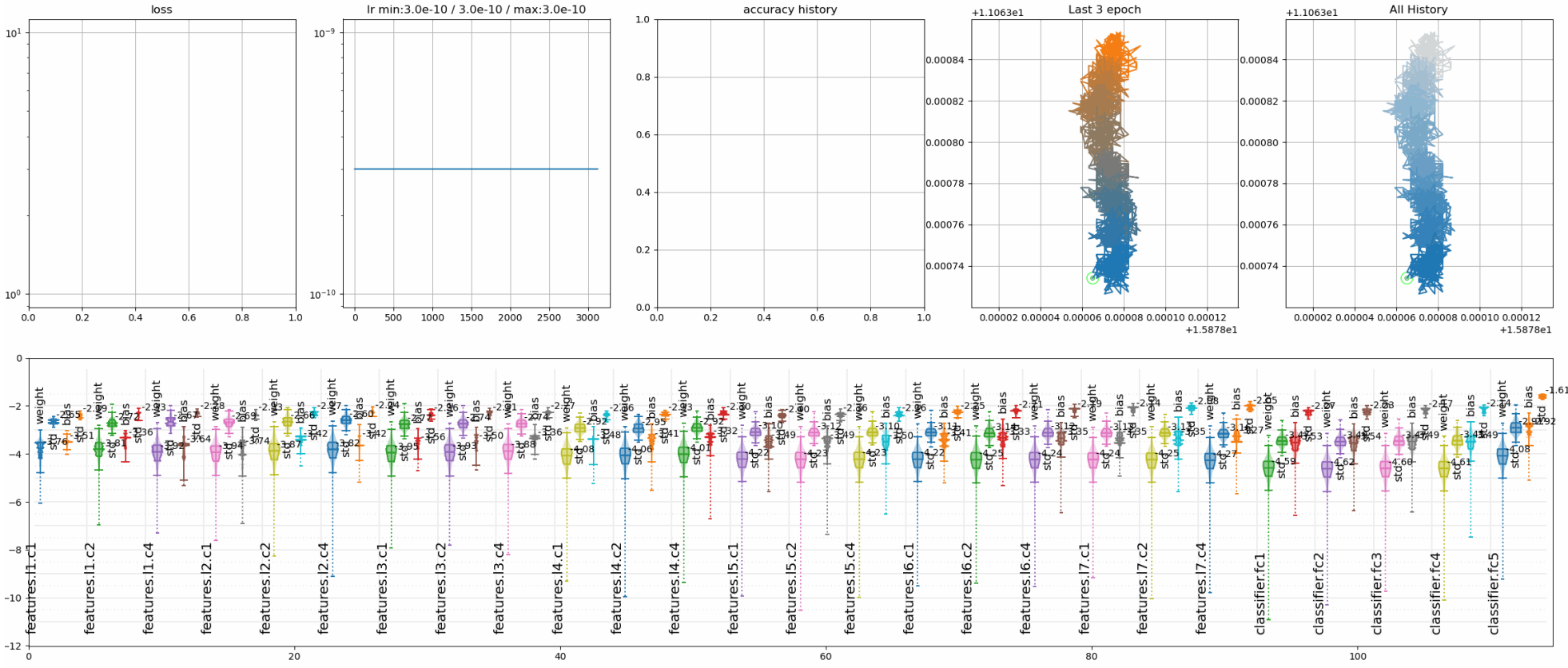


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель способна учиться с такой нетрадиционной нелинейностью, получается нормально.

- Эта архитектура явно избыточна для этой задаче.
- Обратите внимание, что loss может расти когда accuracy уже не портится.
- Но 32 слоя это совсем не много, мы должны пойти глубже.

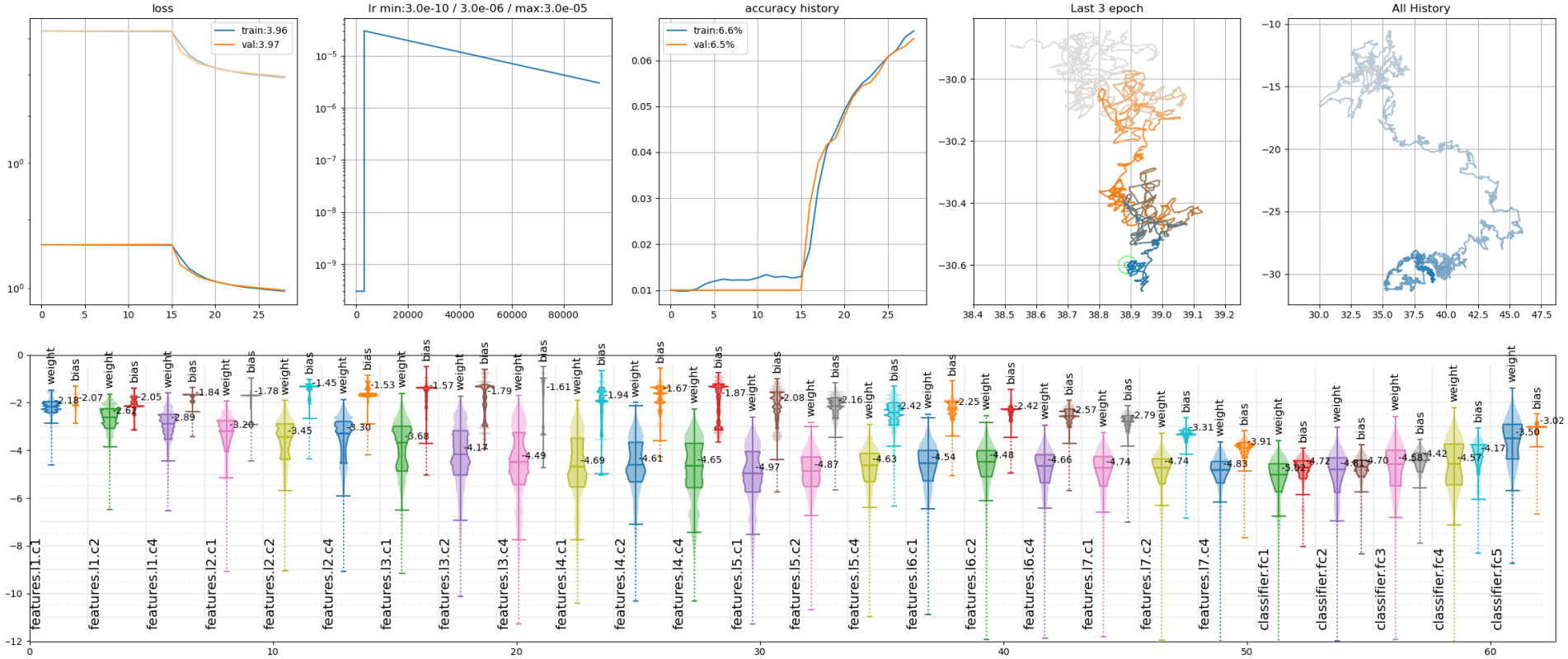


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием, хотя это ей трудно дается. lr: (3e-5,3e-6)

- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Это становится возможным благодаря алгоритму Adam
- Это требует умного подбора гиперпараметров.
- Обратите внимание на градиенты
- И на траекторию

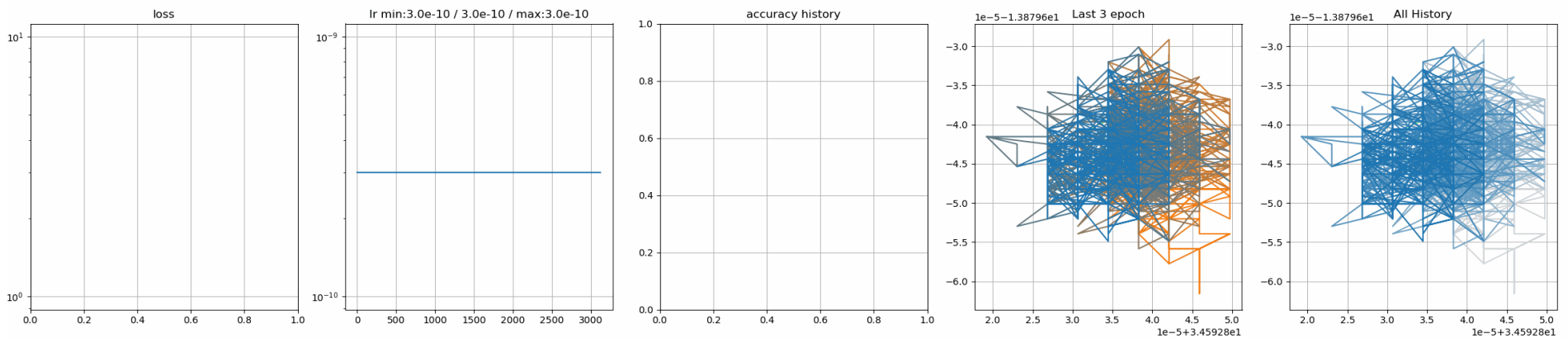


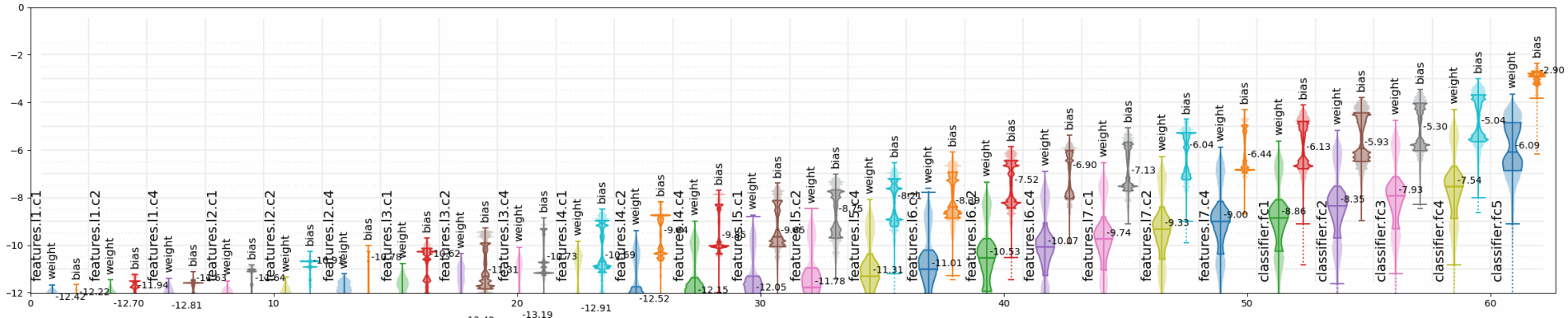
Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием, хотя это ей трудно дается. lr: (3e-5,3e-6)

- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Это становится возможным благодаря алгоритму Adam
- Это требует умного подбора гиперпараметров.
- Обратите внимание на градиенты
- И на траекторию



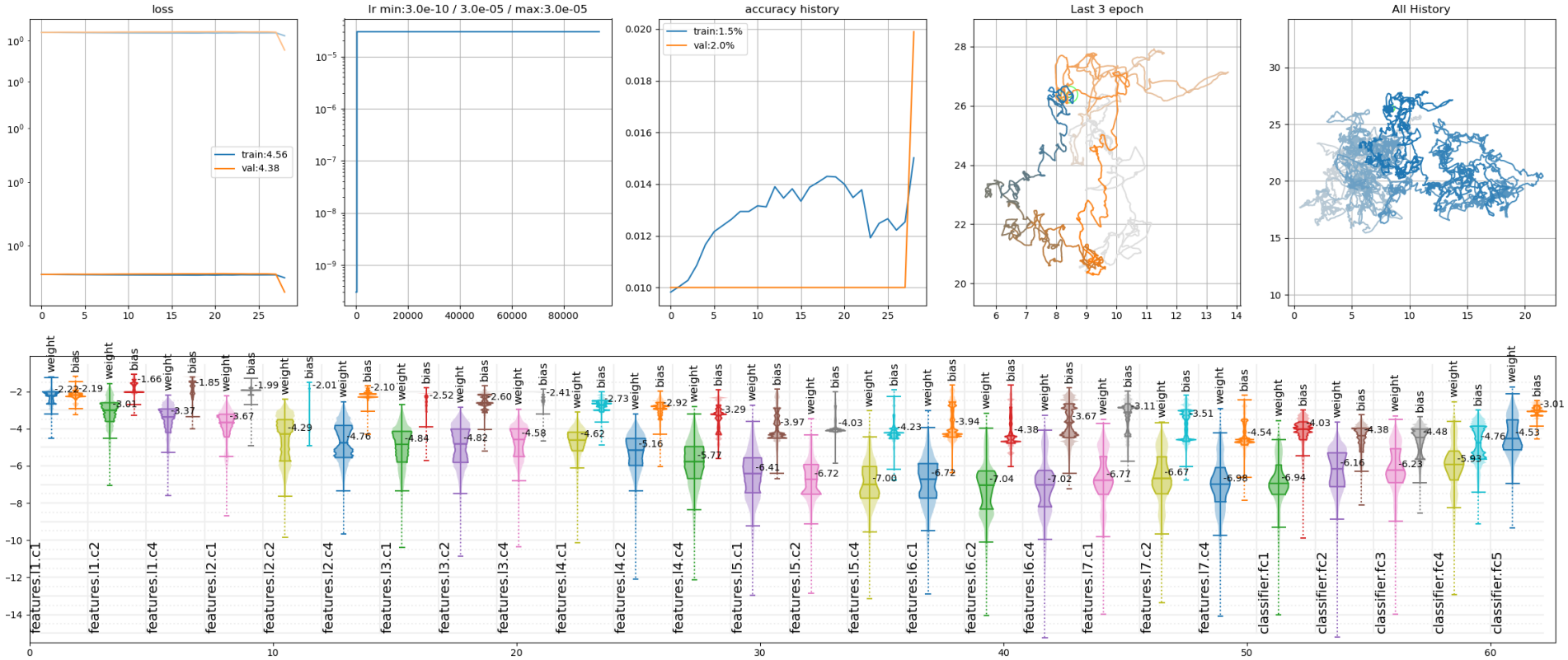


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием. lr: 3e-5

- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Это требует умного подбора гиперпараметров.
- Не только в скорости дело.

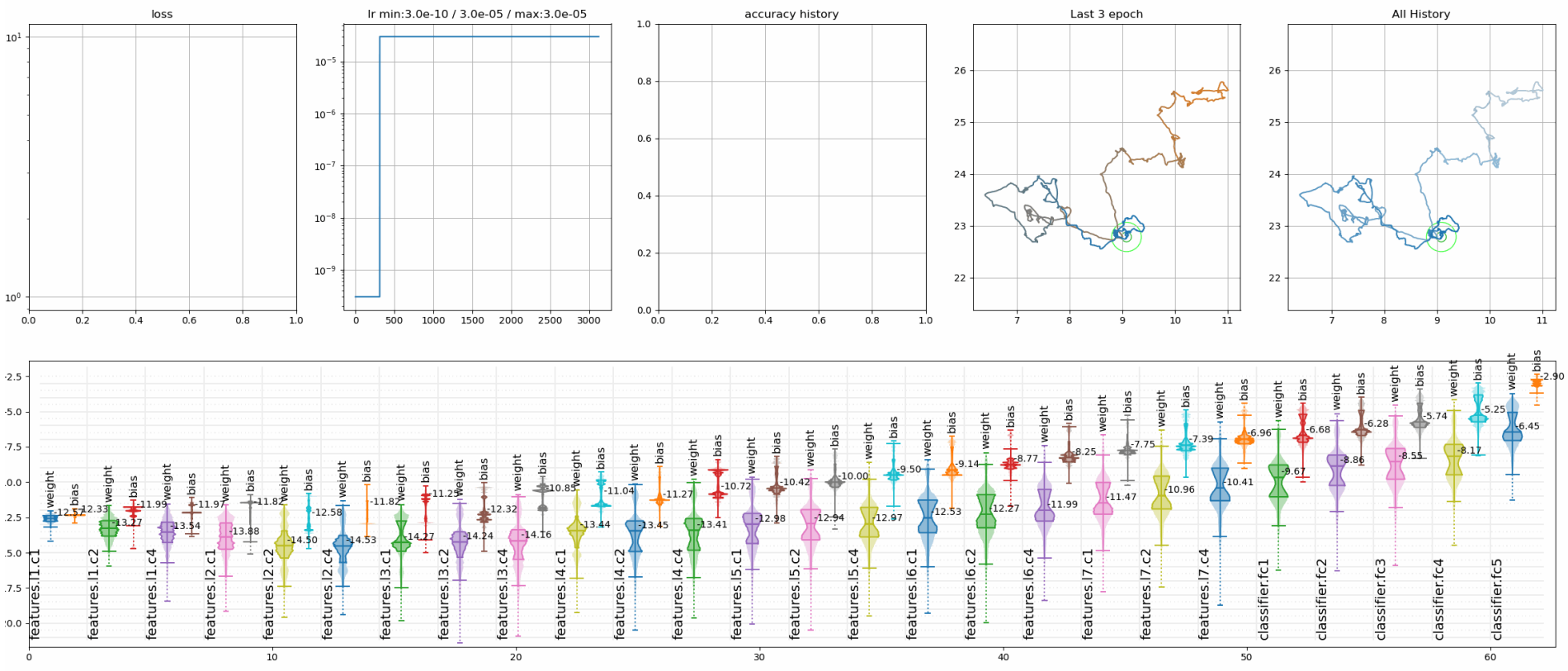


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием. lr: 3e-5

- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Это требует умного подбора гиперпараметров.
- Не только в скорости дело.

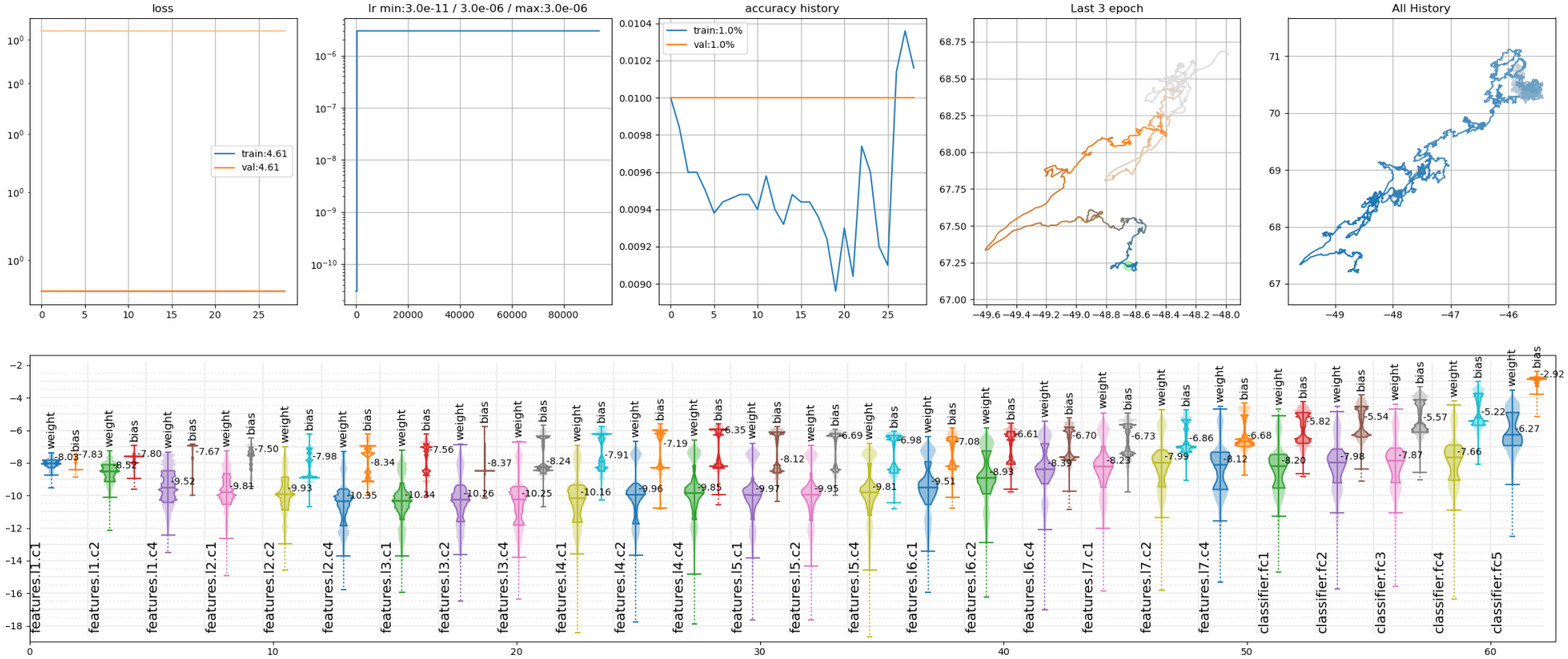


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием. lr: 3e-6

- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Это требует умного подбора гиперпараметров.
- Не только в скорости дело.

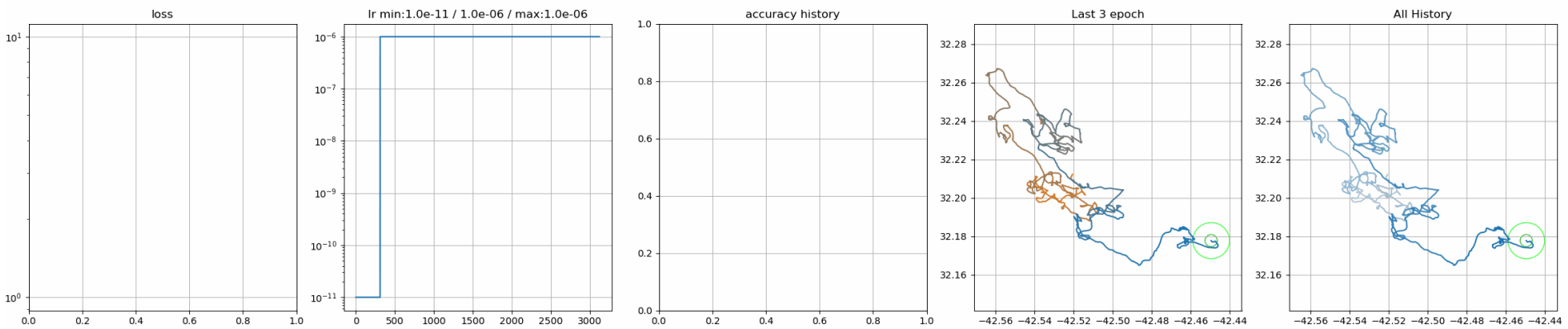


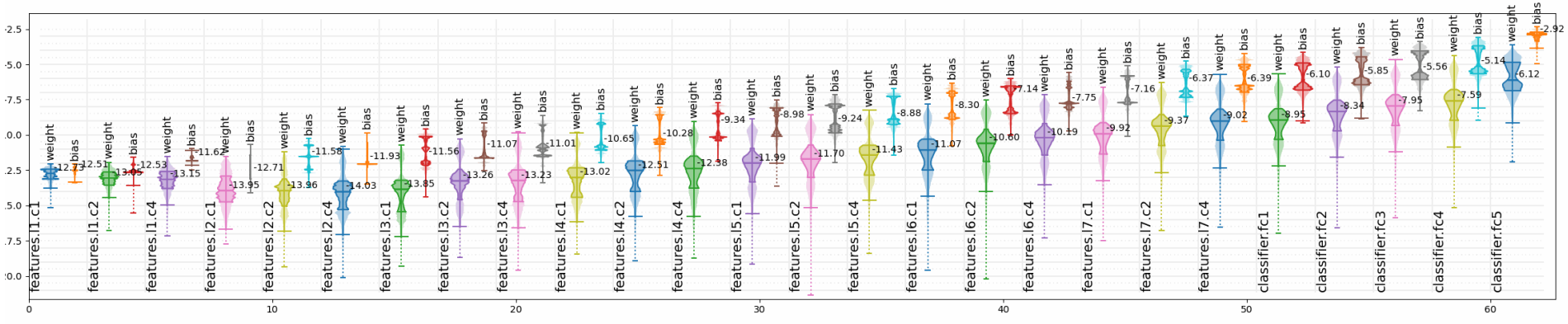
Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 26 слоёв, способна научится и сама без борьбы с затуханием. lr: 3e-6

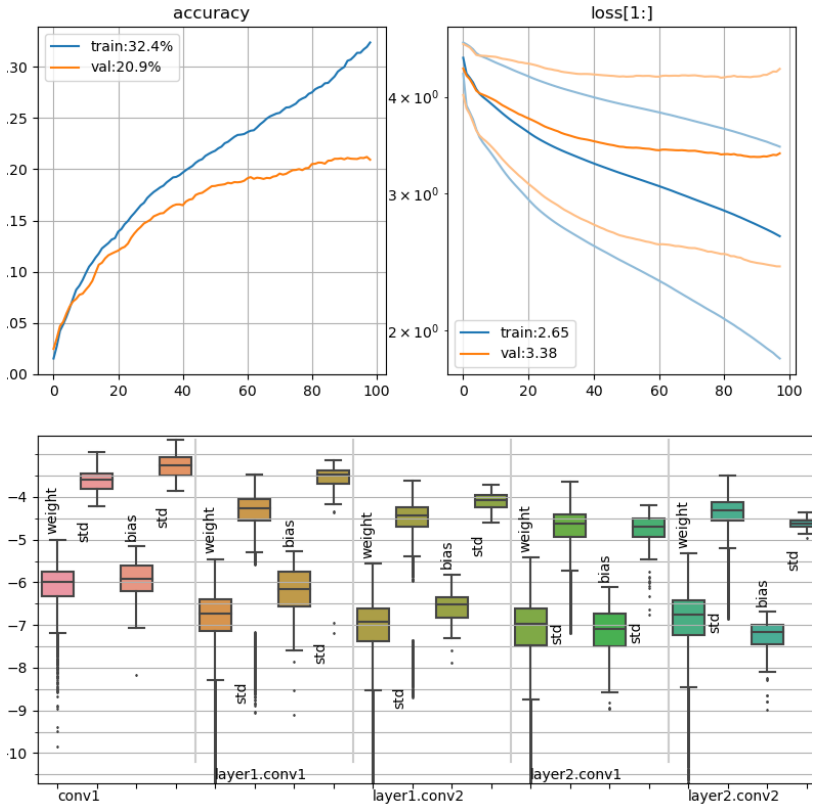
- Помехой являются, скорее всего, особенности рельефа адаптивного пространства (ландшафта функции потерь).
- Не только в скорости дело.
- Диапазон подходящих параметров страшно узкий.



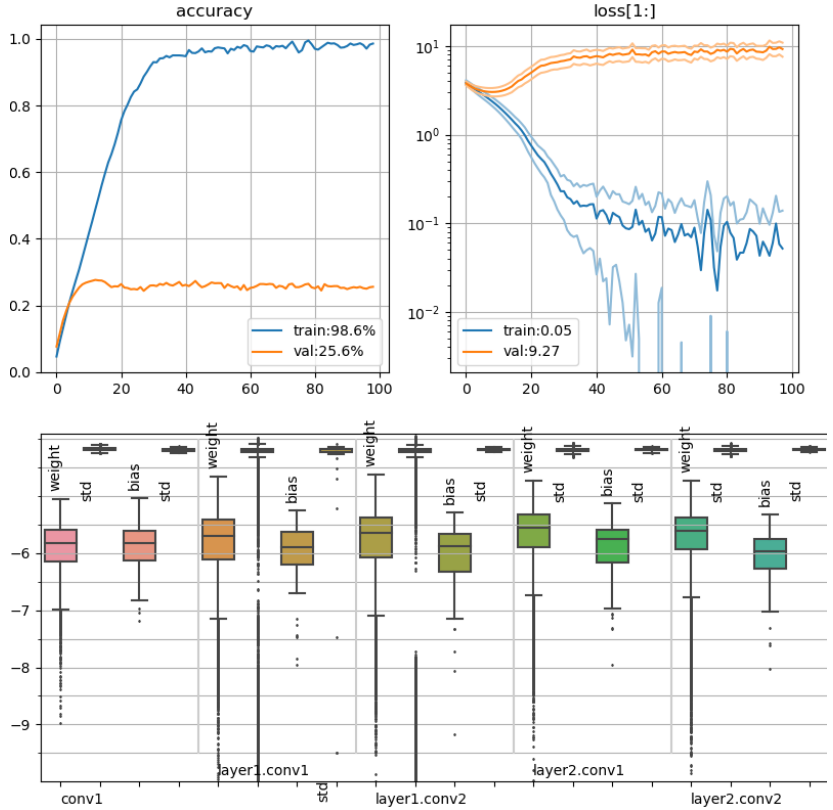


- Секрет алгоритма Adam в раздельном манипулировании скоростями и дисперсией градиентов.

SGD

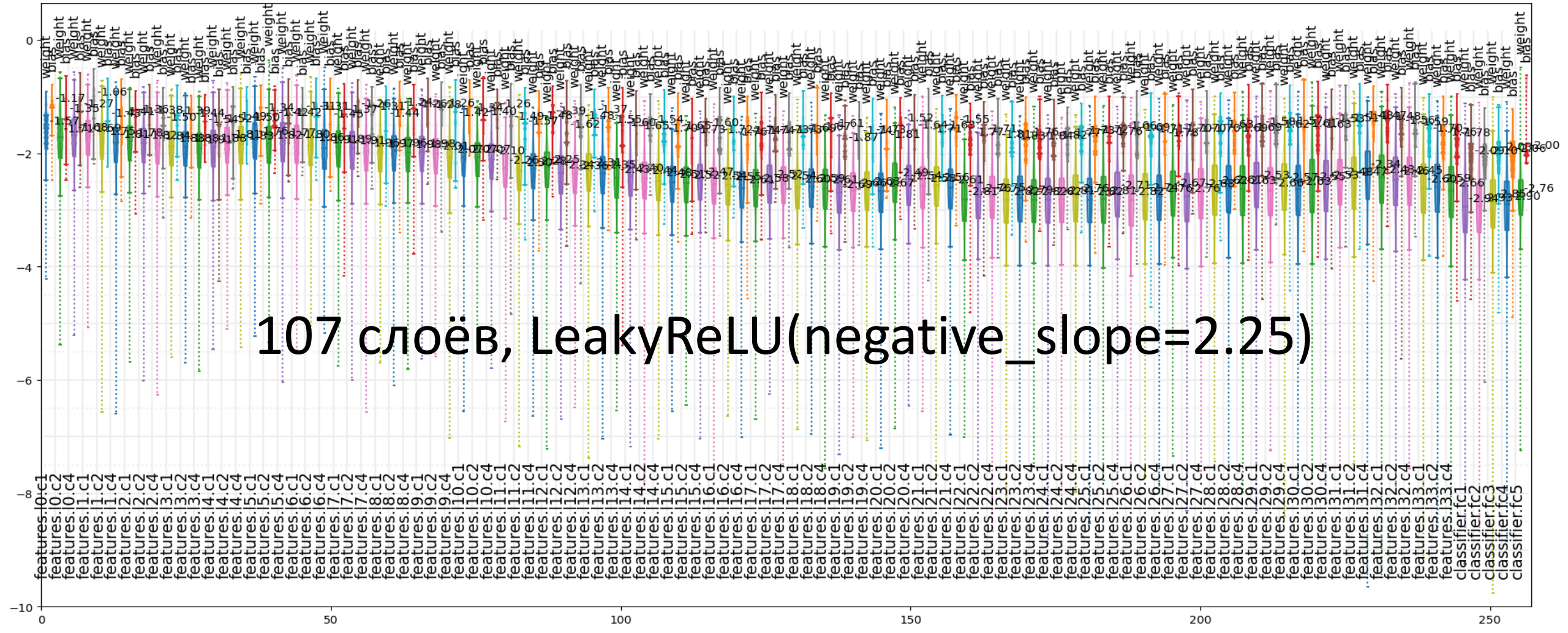


Adam



Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

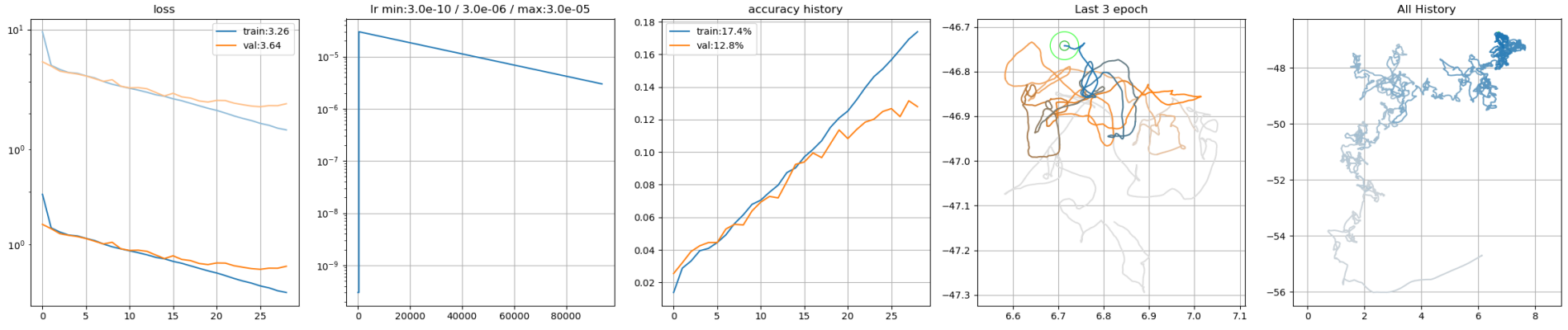


Data NewYear 2024
Голощапов Влад

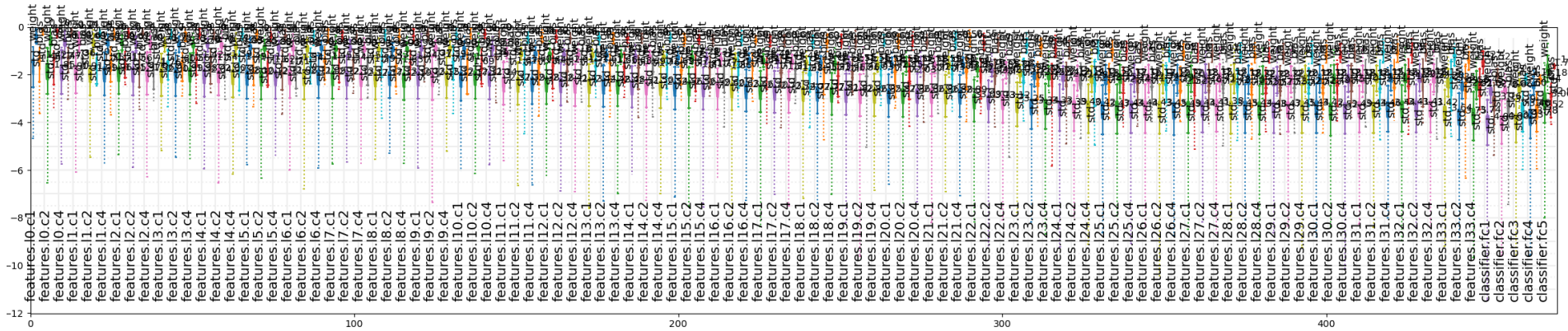
Про не затухание градиентов и видение невидимого

- # Модель 107 слоёв, LeakyReLU(negative_slope=2.25), lr: (3e-5, 3e-6),

- Заводится сходу без сложных поисков параметров. Подготовительный этап хорошо виден на траектории и укладывается в первую эпоху.



- Ума не приложу зачем вам может понадобиться такое чудовище, но теперь вы по крайней мере можете себе это позволить.

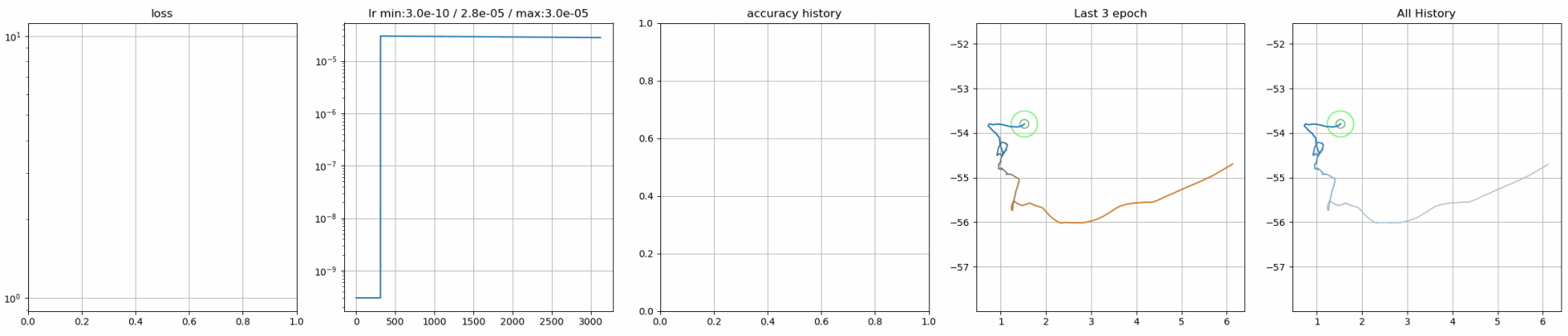


Data NewYear 2024
Голощапов Влад

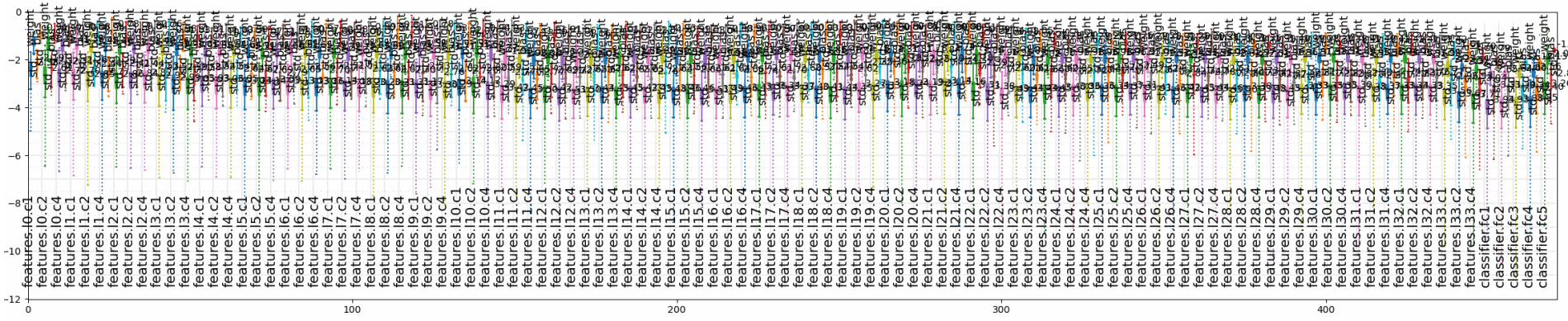
Про не затухание градиентов и видение невидимого

- # Модель 107 слоёв, LeakyReLU(negative_slope=2.25), lr: (3e-5, 3e-6),

- Заводится сходу без сложных поисков параметров. Подготовительный этап хорошо виден на траектории и укладывается в первую эпоху.



- Ума не приложу зачем вам может понадобиться такое чудовище, но теперь вы по крайней мере можете себе это позволить.



- Второй способ модифицировать не функцию активации, а следующие за ней веса, в формулу градиентов они входят как множитель.
- Это примерно то что предлагали делать Yilmaz и Poli в прошлом 2022-ом году (см. вики), только мы не будем пытаться аналитически вычислить множитель, а посмотрим на диаграмму градиентов и подберём его. Отрицательными веса делать тоже не обязательно.
- А ещё у них получилось только 10-15 слоёв, а у нас сотни, я думаю, потому что они не поняли про скорость.

Data NewYear 2024
Голощапов Влад

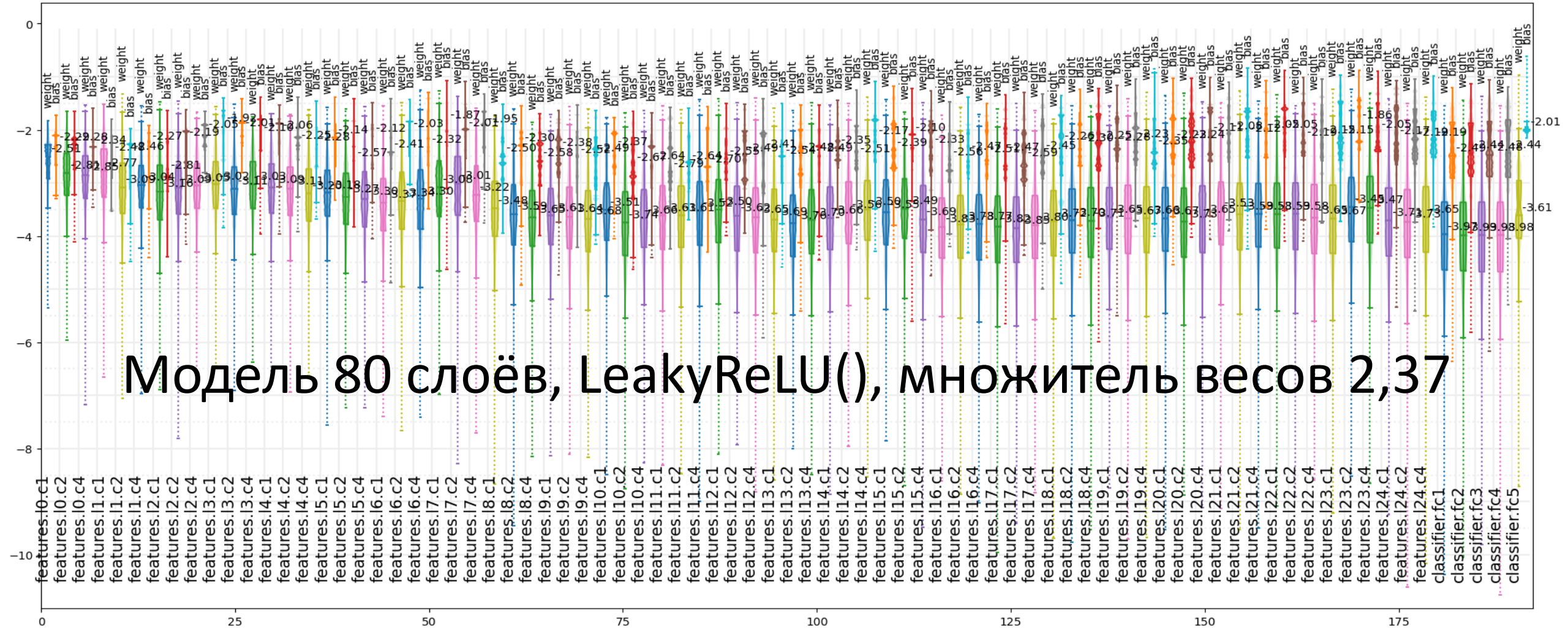
Про не затухание градиентов и видение невидимого

• Модель 80 слоёв, LeakyReLU()

- 107 слоёв, 30 порядков. 10^(30/80) = 2,37
- Для других активаций, не линейных везде кроме 0 будет на так легко сделать «в лоб», но поскольку один тест занимает пару секунд множитель можно просто подобрать под любую функцию активации и любую архитектуру.

Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

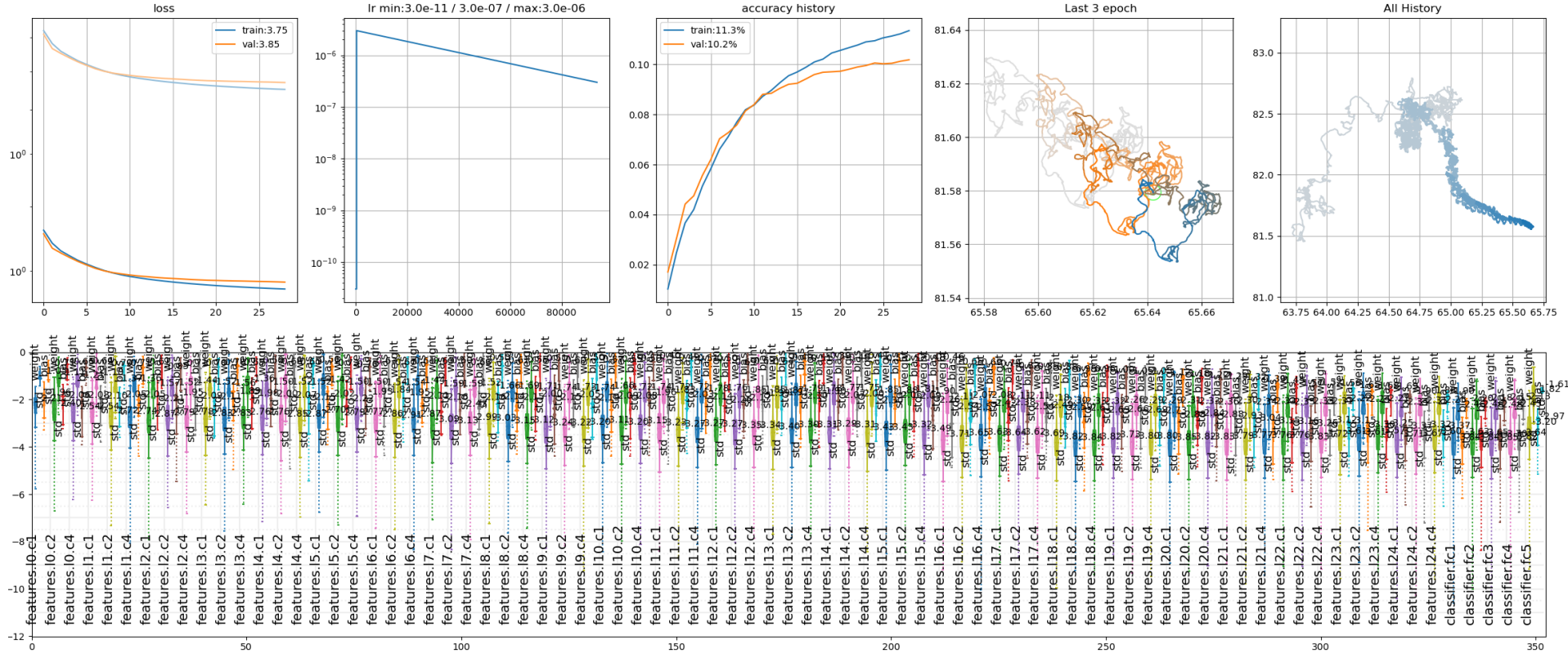


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(negative_slope=0,01), множитель: 2.37, lr: (3e-5, 3e-6),

- Заводится на тех же параметрах, свезло. Начальный этап хорошо виден на траектории в первую эпоху.
- Ума не приложу зачем вам может понадобиться такое чудовище, но теперь вы по крайней мере можете себе это позволить.
- Обучению может мешать всякое, но не градиенты.

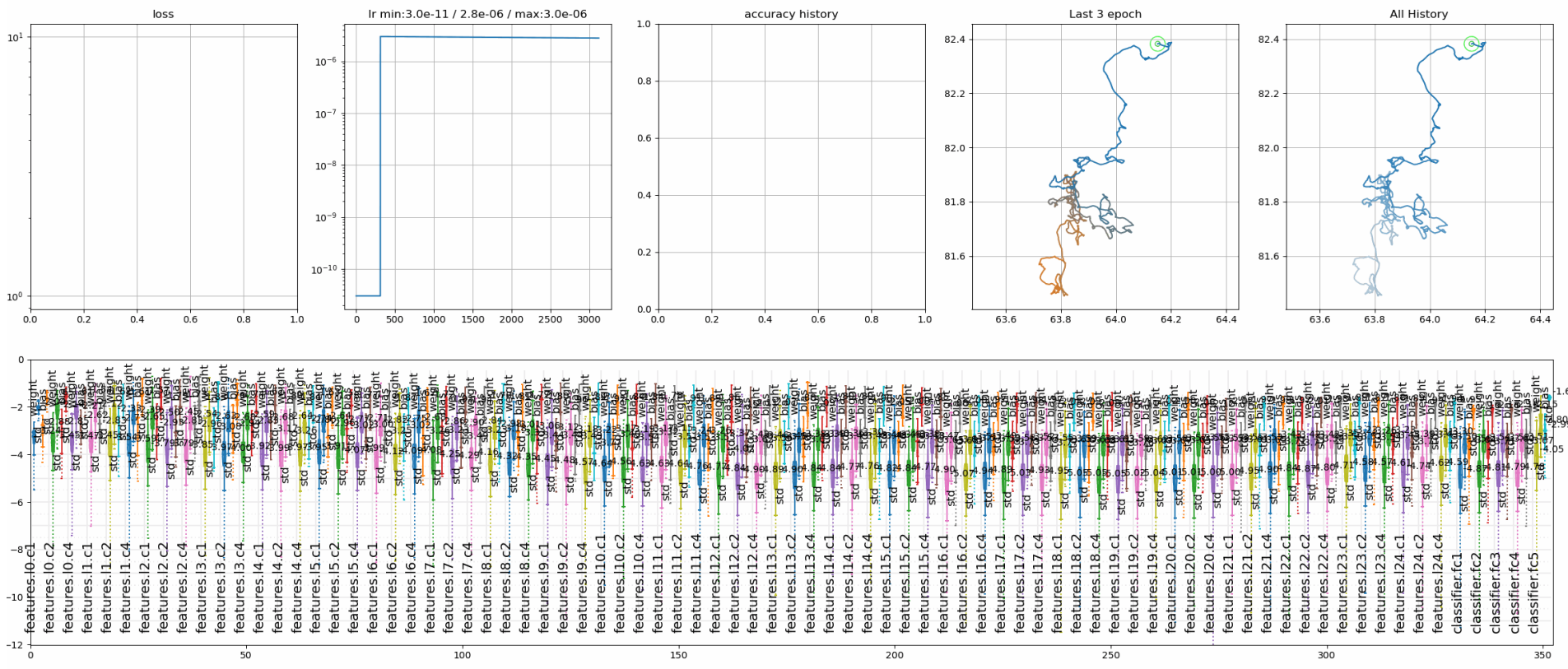


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(negative_slope=0,01), множитель: 2.37, lr: (3e-5, 3e-6),

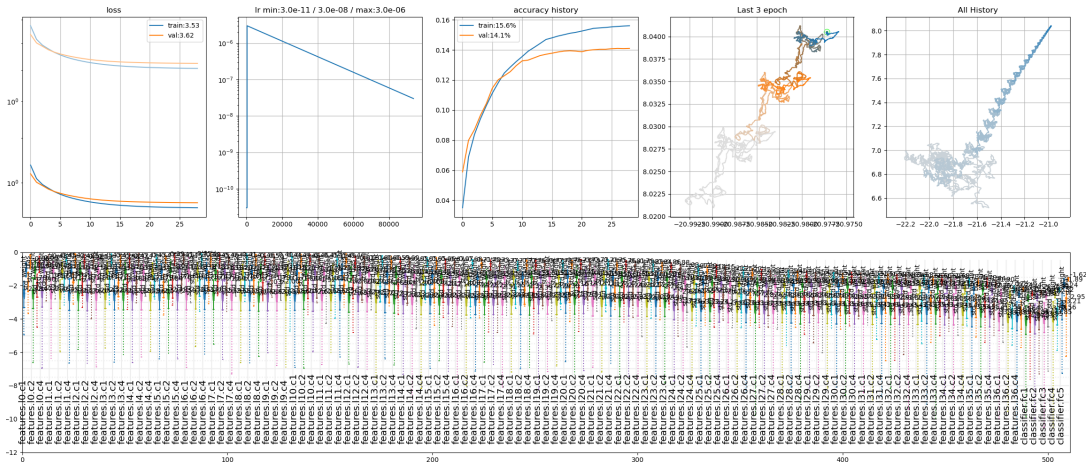
- Заводится на тех же параметрах, свезло. Начальный этап хорошо виден на траектории в первую эпоху.
- Ума не приложу зачем вам может понадобиться такое чудовище, но теперь вы по крайней мере можете себе это позволить.
- Обучению может мешать всякое, но не градиенты.



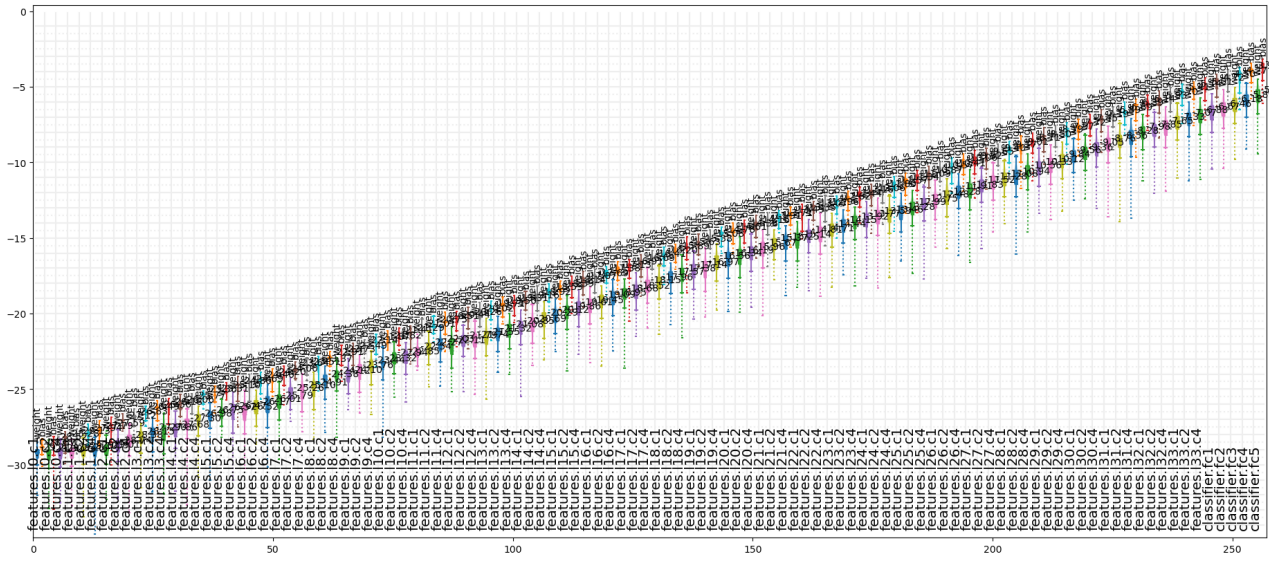
Data NewYear 2024
Голощапов Влад

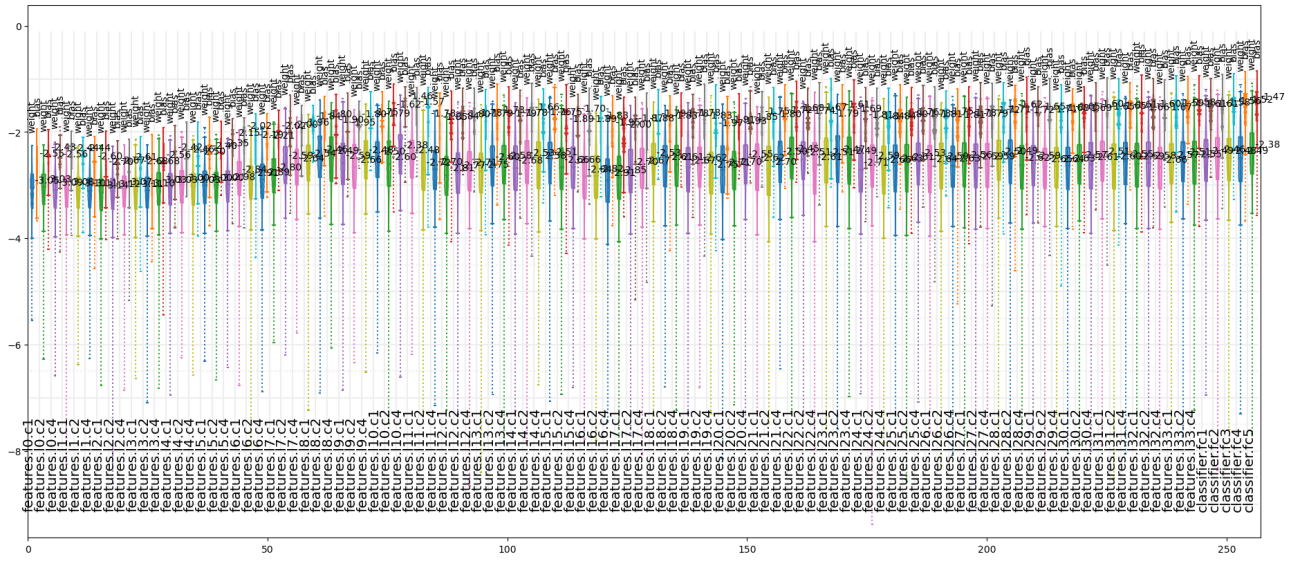
Про не затухание градиентов и видение невидимого

- Для ограниченной активации типа TanH() 116 слоёв тоже работает. Множитель 1.8



- На слайде мы тоже видим первоначальный этап, не смотря на то, что градиенту протекать как будто бы и не нужно.
- Протекание градиента само по себе ничего не гарантирует, потому что есть нюанс…



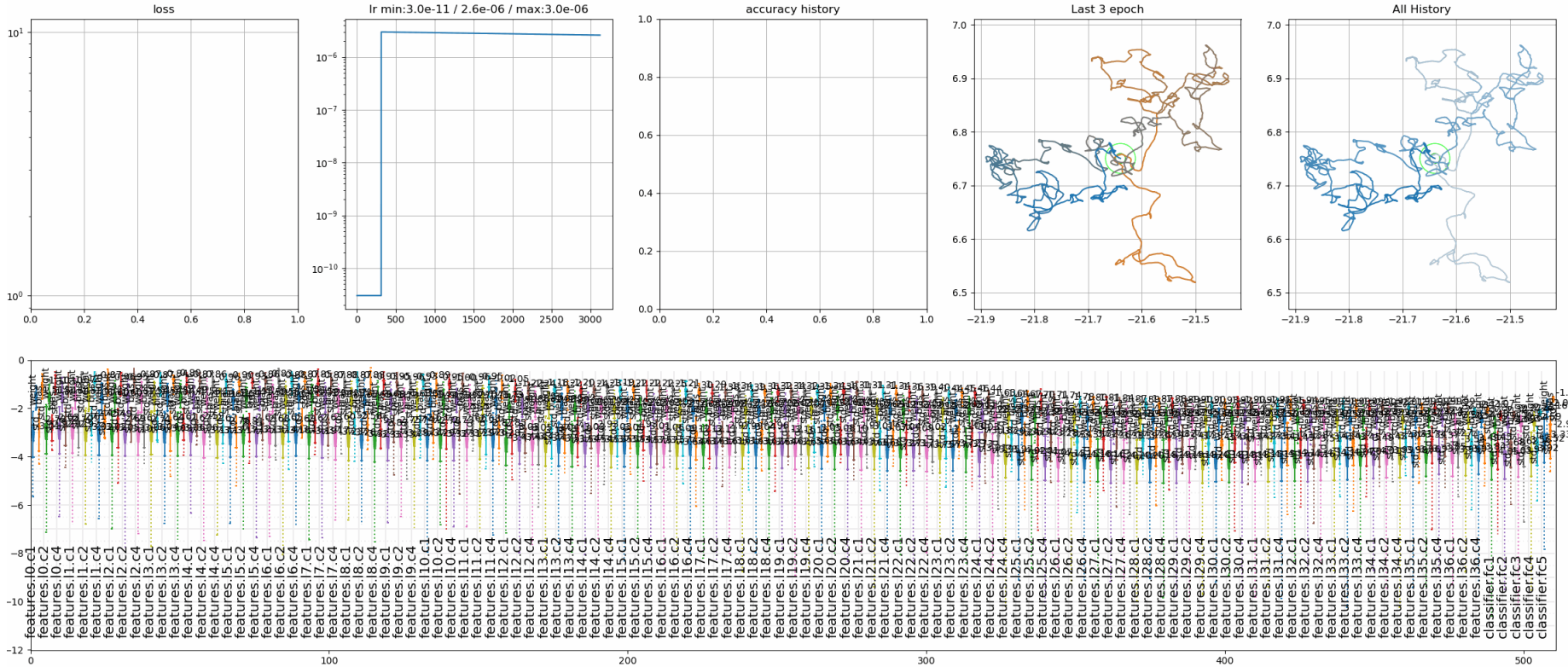


Data NewYear 2024
Голощапов Влад

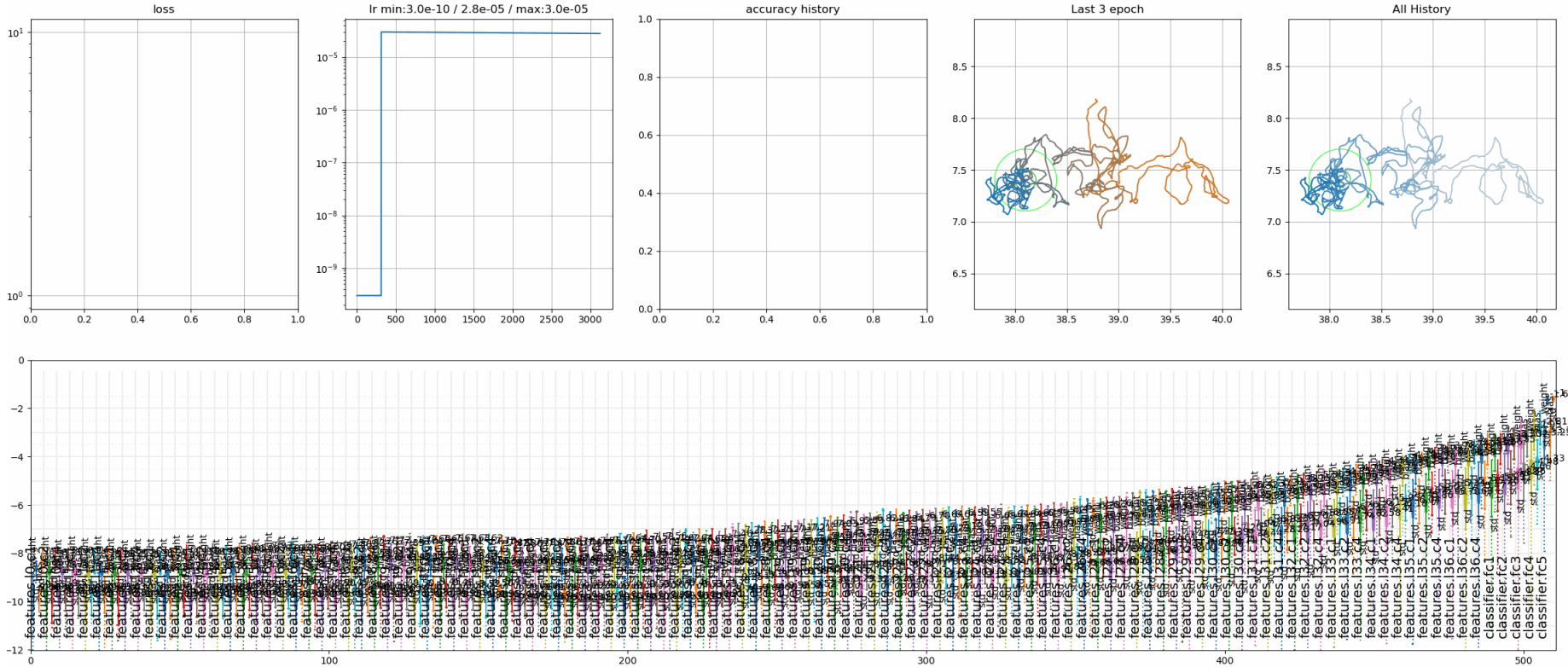
Про не затухание градиентов и видение невидимого

- 116 слоёв TanH() Множитель 1.8,
- Успешные гиперпараметры, lr: (3e-6, 3e-8)

- Правильная скорость непривычно низкая, обычно в таких маленьких числах даже не ищут.
- Интуитивно предположу, что чем больше толщина тем больше пространство изрезано разделяющими плоскостями и, соответственно, характерный размер рельефа мельче.

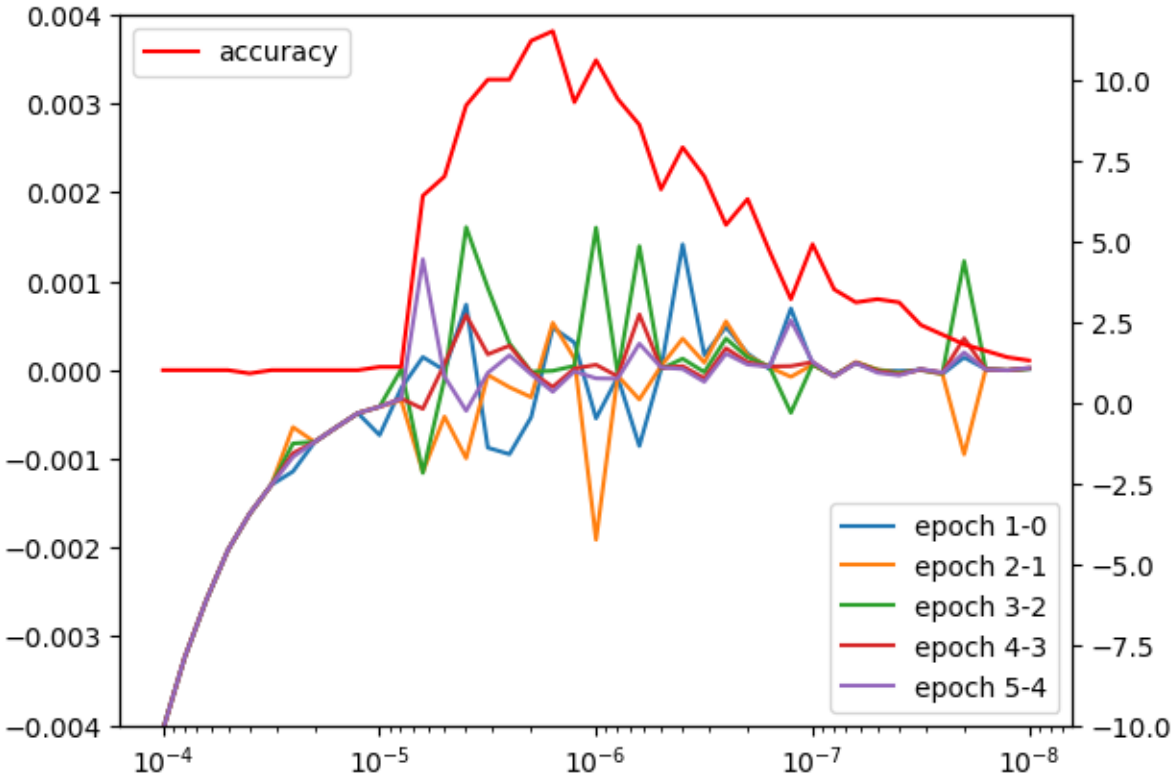


- 116 слоёв TanH() Множитель 1.8,
- Неудачные гиперпараметры, lr: 3e-5
  - На неправильной скорости, слишком большой, происходит быстрое обрушение первоначально протащенного градиента. Это наблюдение нам может пригодиться в дальнейшем.



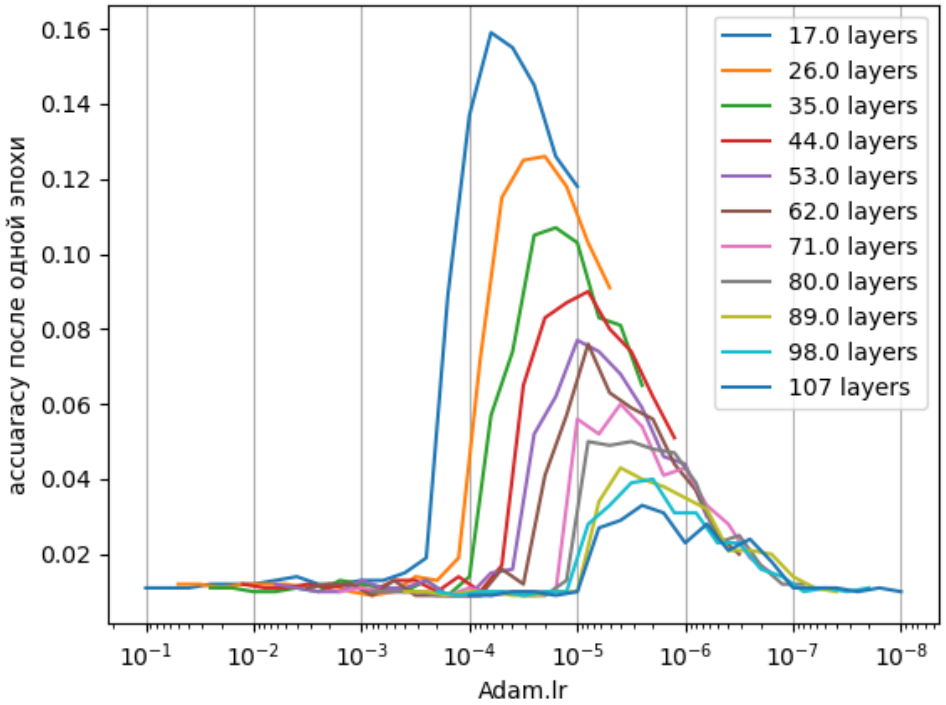
- # Сделаем перебор параметров, чтобы попытаться нащупать закономерности.

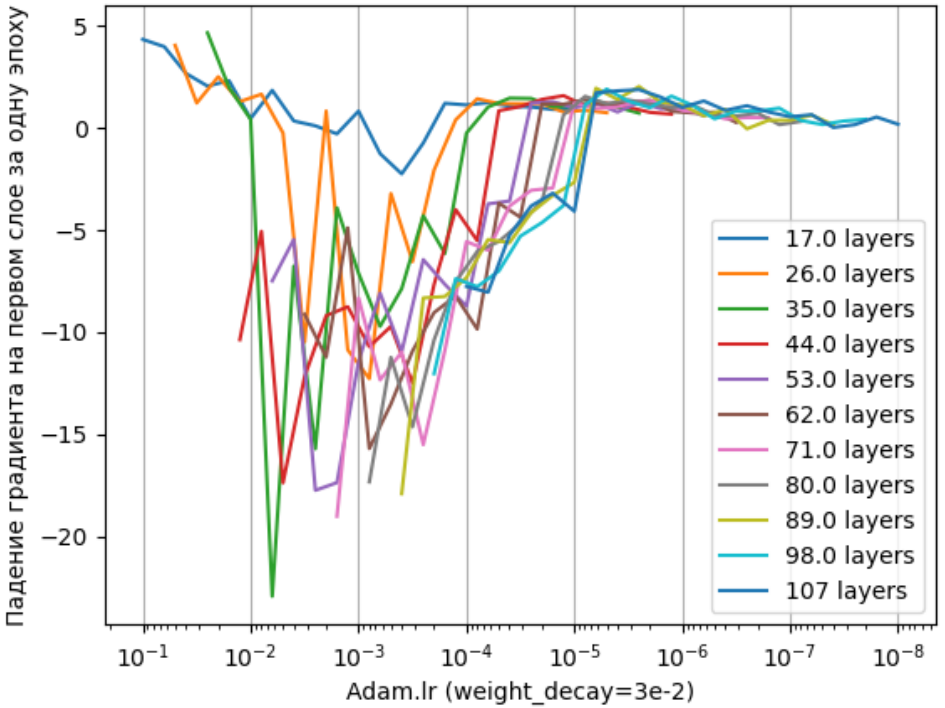
- При скоростях выше 6.3e-6 обучение за разумное время не начинается, не взирая на градиенты.
- Медианный градиент на первом слое может колбаситься при приемлемых параметрах, при неправильных он недвусмысленно обваливается.
- Итоговая метрика в общем случае – выпуклая функция, так что если вам приспичило вы можете решить задачу делением пополам.
- Это только один пример. Другие архитектуры и другие глубины могут раскрыть вам ещё какую-нибудь хтонь, экспериментируйте и визуализируйте результаты экспериментов, потому что если это все сложить циферками в табличку не понятно будет вообще ничто и никак.



- # Сделаем перебор параметров, для закономерностей.

  - Чем толще сеть, тем ниже допустимые скорости, и тем вреднее для сети превышение скорости.
  - У не слишком толстых сетей, 17-26 слоёв слишком высокая скорость не обязательно ведёт к разрушению.





- Существует много разных способов протаскивать градиенты. Об этом написаны тонны литературы. Все они либо хуже либо чего-то не учитывают либо плодят лишний код, а главное, вы не видите происходящего во внутренностях сети – вам остаётся только применить метод и молиться.
- Страничку русской и английской википедии можно смело переписывать, всё что там написано – устарело.
- За 20 минут вы своими глазами увидели своими глазам и о протекании градиента больше чем всё человечество за несколько лет.

- Когда библиотечка будет хороша я выложу её в общий доступ, но всё то же самое можно узнать и без красивых библиотек:

```
[(n,p.grad.view(-1)[p .grad.view(-1)!=0].abs().log10().median().item()) for n, p in model.named_parameters()]
```

- Того, что можно заметить, если смотреть не только на loss может хватить на десять докладов. Прозрейте же и станьте Свидетелями Градиента! @GradientWitnesses

Заметим на визуализации что-то новое,

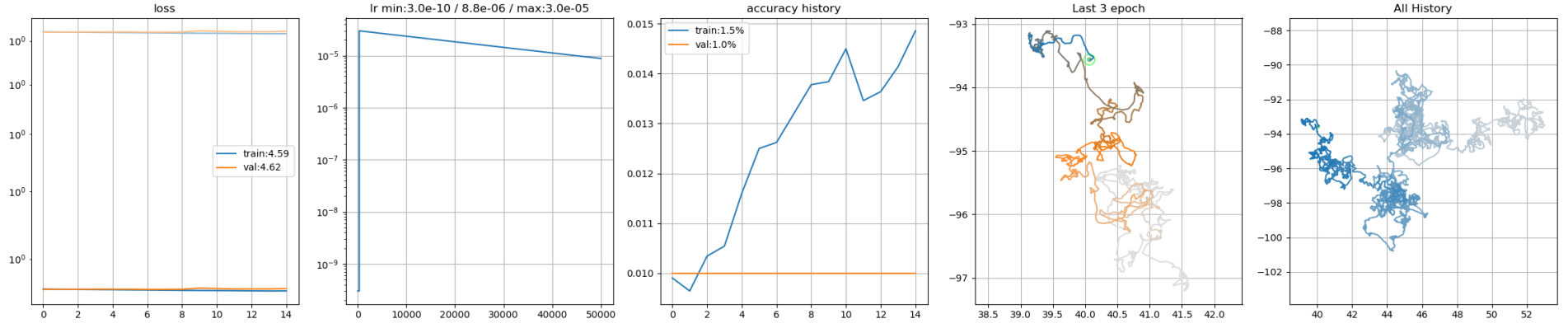
если осталось время, что вряд ли.

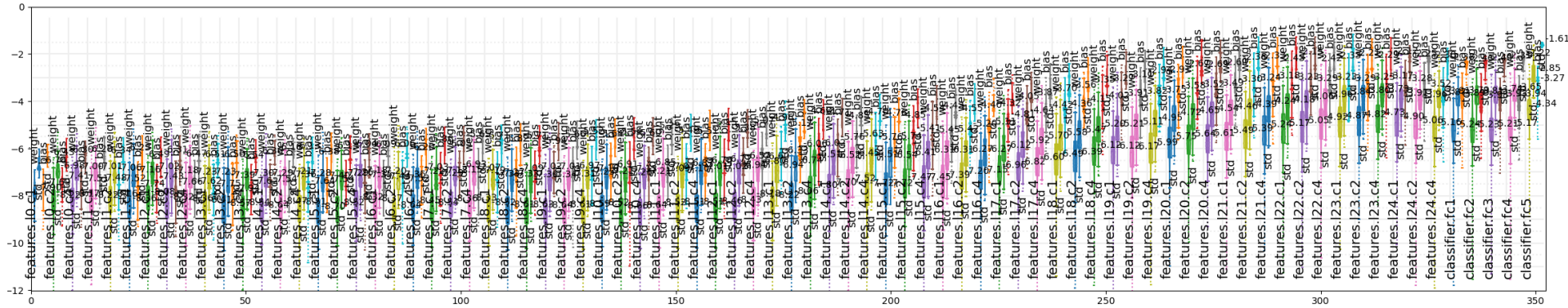
Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(), множитель 2.37 lr: (3e-5, 3e-6), эпоха 16

- Смотрим на поведение градиентов и вид траектории обучения. Прорыв сети в понимании данных хорошо виден по поведению градиентов, но не сводится к нему.



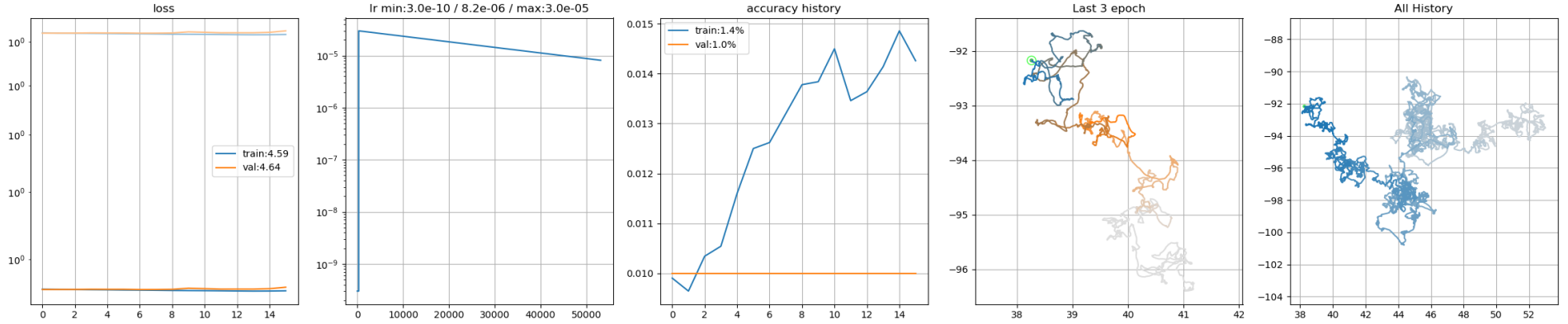


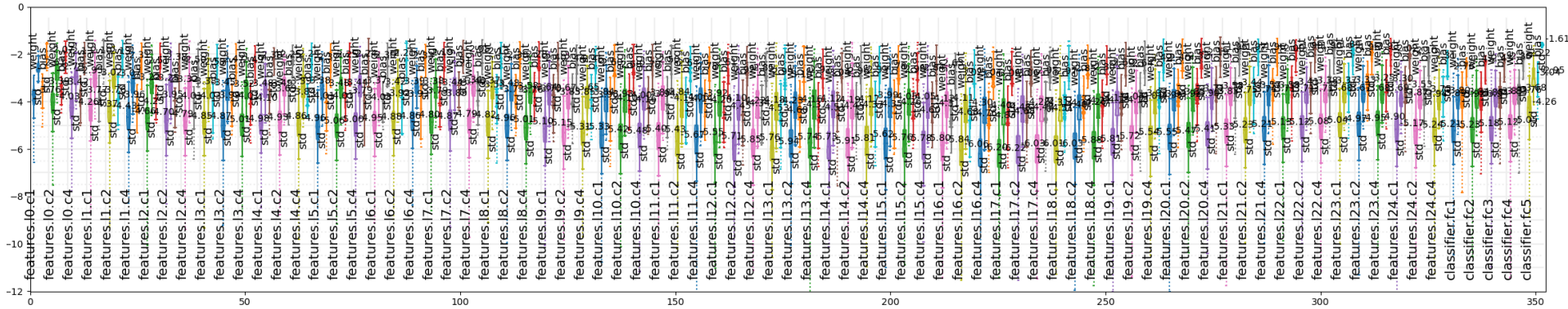
Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(), множитель 2.37 lr: (3e-5, 3e-6), эпоха 17

- Смотрим на поведение градиентов и вид траектории обучения. Прорыв сети в понимании данных хорошо виден по поведению градиентов, но не сводится к нему.



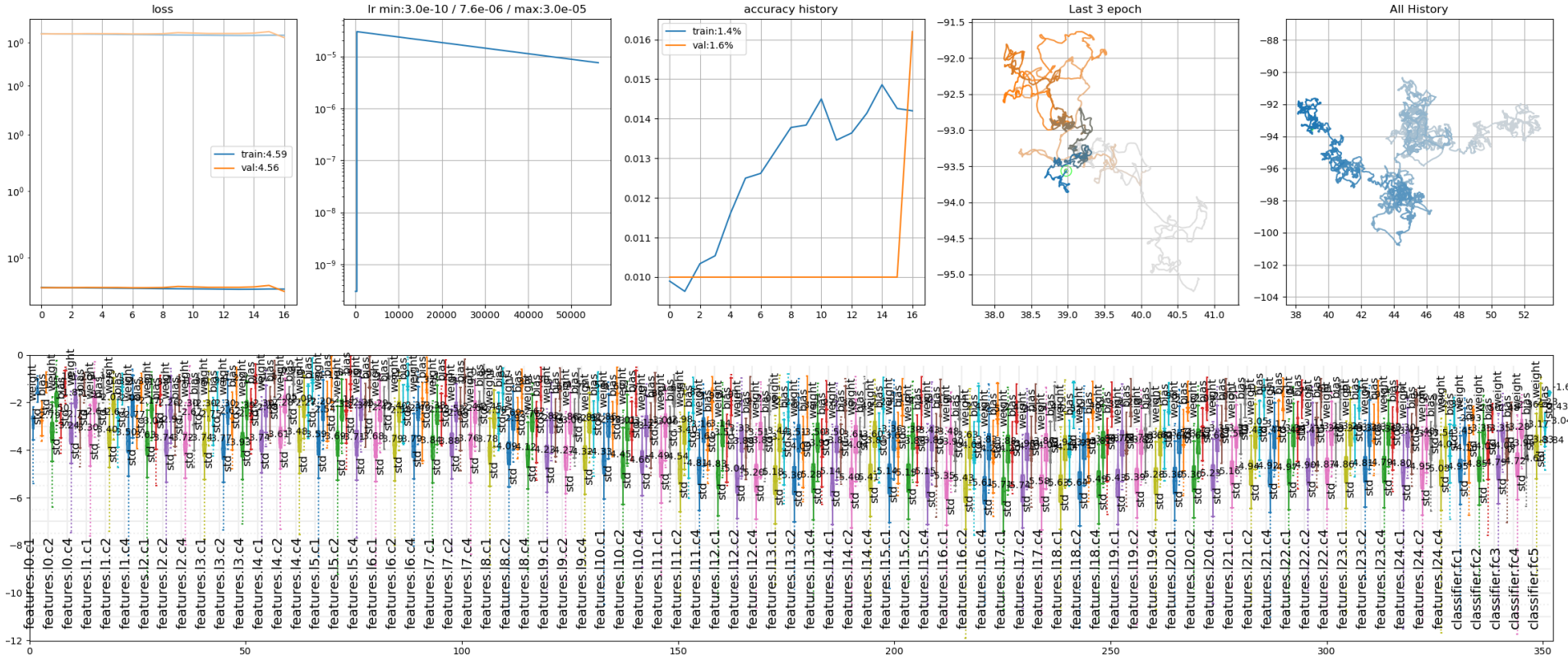


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(), множитель 2.37 lr: (3e-5, 3e-6), эпоха 18

- Смотрим на поведение градиентов и вид траектории обучения. Прорыв сети в понимании данных хорошо виден по поведению градиентов, но не сводится к нему.
- Он происходит до, собственно роста loss-а. Интересно что там происходит, я не знаю.

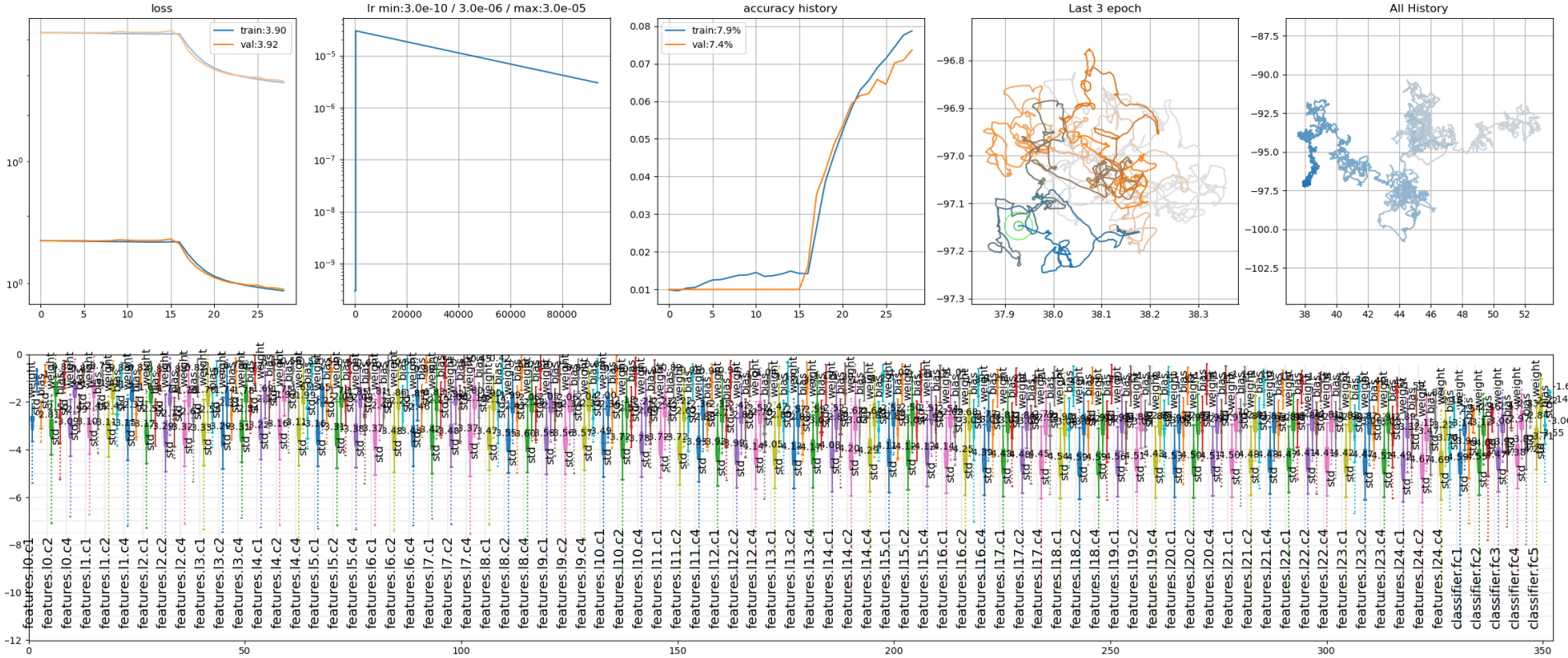


Data NewYear 2024
Голощапов Влад

Про не затухание градиентов и видение невидимого

- # Модель 80 слоёв, LeakyReLU(), множитель 2.37 lr: (3e-5, 3e-6), эпоха 30

- Смотрим на поведение градиентов и вид траектории обучения. Прорыв сети в понимании данных хорошо виден по поведению градиентов, но не сводится к нему.
- Он происходит до, собственно роста loss-а. Интересно что там происходит, я не знаю.



А про скип, денс и резидуал коннекшены мы поговорим как-нибудь в следующий раз. Вступайте в Свидетели Градиента, чтобы не пропустить анонсы.

https://t.me/GradientWitnesses