Исследование метода дистилляции данных

Медведев Дмитрий Владимирович

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

ДНК

- Дано: алгоритм дистилляции данных.
- Найти:
  - исследовать работу алгоритма при менее экстремальном сжатии, чем в оригинальной статье (сжатие в 3 раза против сжатия в 600 раз).
  - исследовать обобщаемость данных не только на другие инициализации, но и на другие архитектуры.
- Критерий: средняя точность решения задачи на 10 перезапусках.

Алгоритм Дистилляции

$$\begin{cases} \theta_0 \sim P_0(\theta) \\ \theta_{k+1} = \theta_k - \tilde{\eta}_k \nabla_\theta l(\tilde{x}_{i(k)}, \theta_k); k = 1, ..., n; i(k) = k \mod s \\ \mathcal{L} = l(x, \theta_n) \to \min\limits_{\tilde{x}, \tilde{\eta}} \end{cases}$$

$$d\mathcal{L} = \frac{\partial \mathcal{L}}{\partial \theta_n} d\theta_n = \frac{\partial \mathcal{L}}{\partial \theta_n} d\Big(\theta_{n-1} - \underbrace{\tilde{\eta}_{n-1} \nabla_\theta l(\tilde{x}_{i(n-1)}, \theta_{n-1})}_{g(\tilde{\eta}_{n-1}, \tilde{x}_{i(n-1)}, \theta_{n-1}) = g_{n-1}}\Big) =$$

$$= \left(\frac{\partial \mathcal{L}}{\partial \theta_n} - \frac{\partial \mathcal{L}}{\partial \theta_n} \frac{\partial g_{n-1}}{\partial \theta_{n-1}}\right) d\theta_{n-1} - \left(\frac{\partial \mathcal{L}}{\partial \theta_n} \frac{\partial g_{n-1}}{\partial \tilde{\eta}_{n-1}}\right) d\tilde{\eta}_{n-1} - \left(\frac{\partial \mathcal{L}}{\partial \theta_n} \frac{\partial g_{n-1}}{\partial \tilde{x}_{i(n-1)}}\right) d\tilde{x}_{i(n-1)}$$

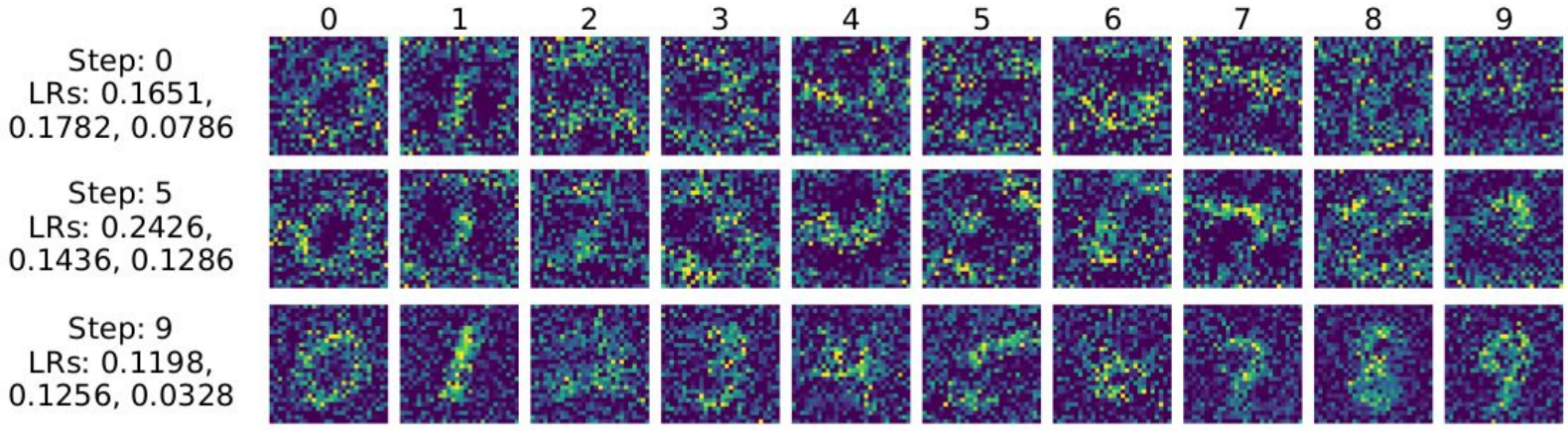
Алгоритм Дистилляции

$$\nabla_{\tilde{\eta}_k} \mathcal{L} = \boxed{\frac{\partial \mathcal{L}}{\partial \theta_{k+1}}} \cdot \boxed{\frac{\partial g_k}{\partial \tilde{\eta}_k}};$$

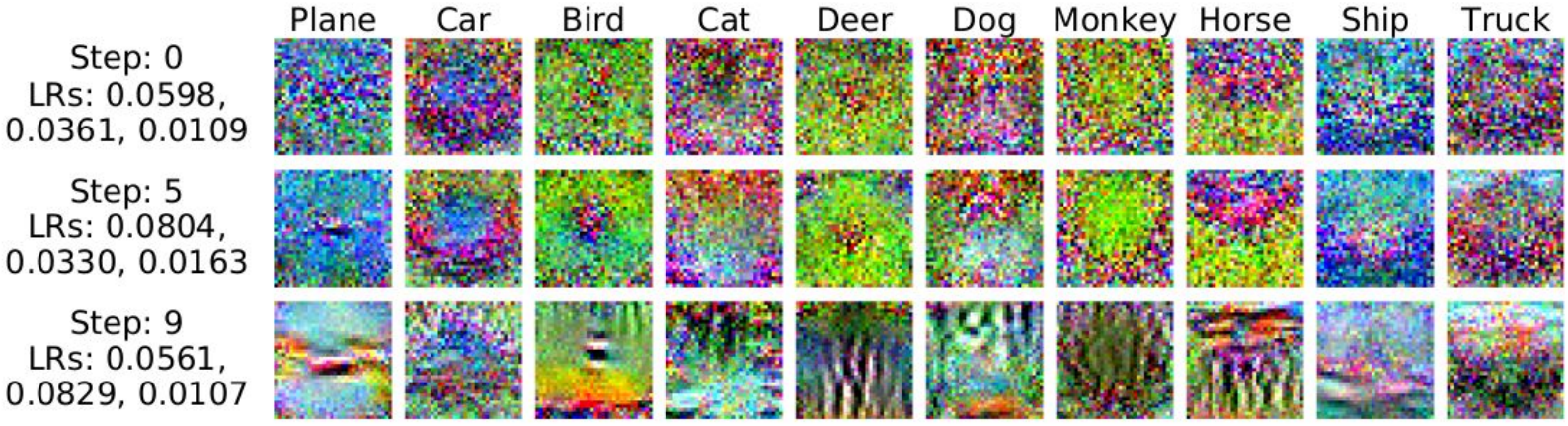
$$\nabla_{\tilde{x}_i(k)} \mathcal{L} = \sum_{j=1}^{n} I[j = i(k)] \cdot \boxed{\frac{\partial \mathcal{L}}{\partial \theta_{k+1}}} \cdot \boxed{\frac{\partial g_k}{\partial \tilde{x}_i(k)}};$$

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = \boxed{\frac{\partial \mathcal{L}}{\partial \theta_{k+1}}} \cdot \boxed{\left(1 - \frac{\partial g_k}{\partial \theta_k}\right)}$$

Проблема обобщаемости



(a) MNIST. These distilled images unknown random initializations to $79.50\% \pm 8.08\%$ test accuracy.



(b) CIFAR10. These distilled images unknown random initializations to $36.79\% \pm 1.18\%$ test accuracy.

Figure 3: Distilled images trained for *random initialization* with ten GD steps and three epochs (100 images in total). We show images from selected GD steps and the corresponding learning rates for all three epochs.

Качество на синтетических данных

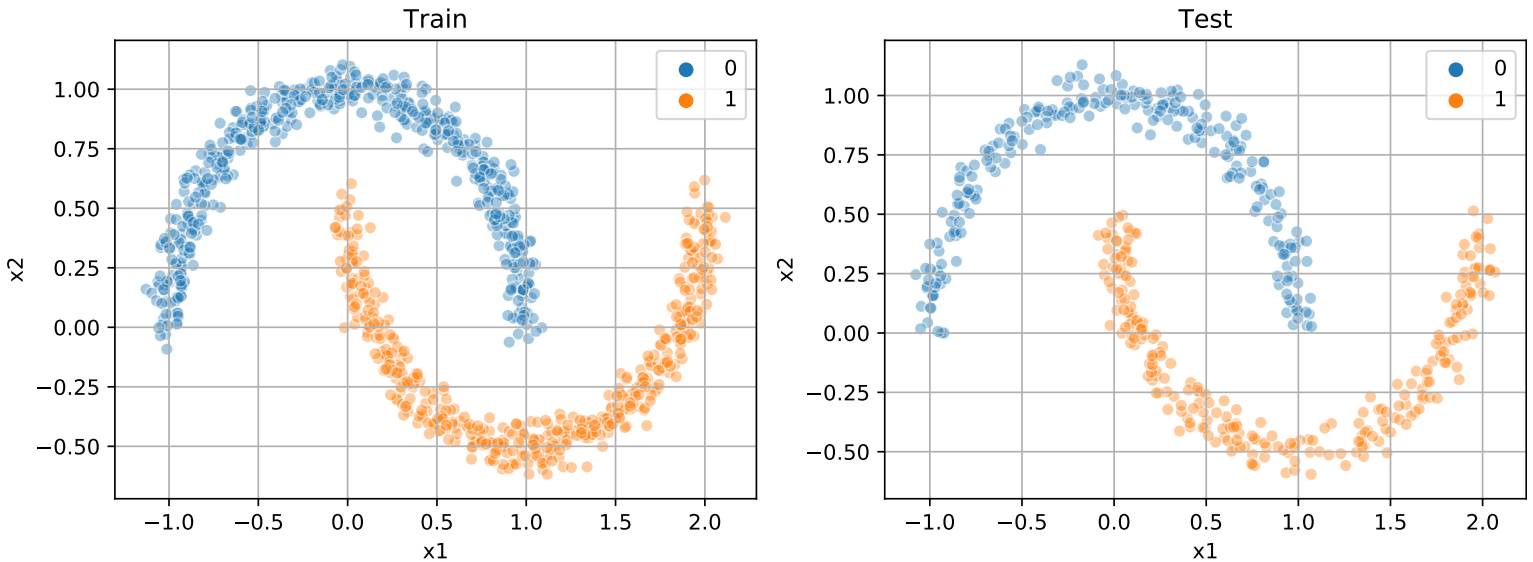
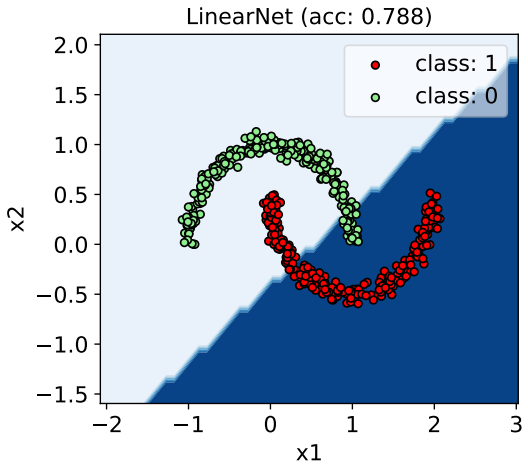


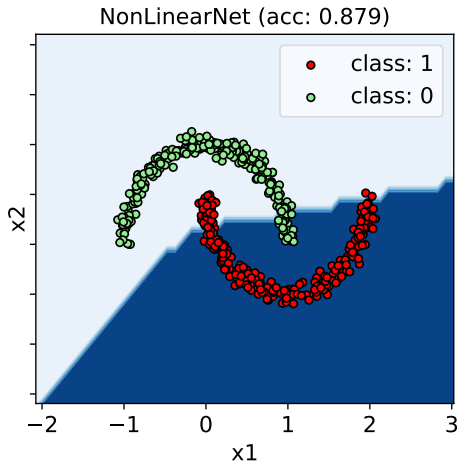
Рис.: Тренировачная и тестовая части оригинальной выборки.

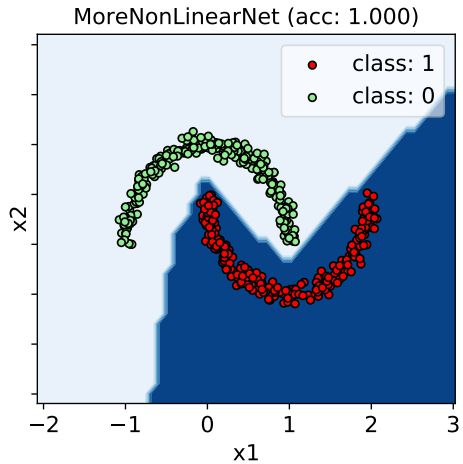
Модель	Качество ориг.	Качество дистилл.
LinearNet	0.766 ± 0.089	$\mathbf{0.871 \pm 0.003}$
NonLinearNet	0.877 ± 0.006	$\mathbf{0.941 \pm 0.043}$
MoreNonLinearNet	$\mathbf{0.995 \pm 0.015}$	0.906 ± 0.054

Таблица: Итоговая точность (и её среднеквадратичное отклонение) на тестовой выборке при обучении на оригинальных и синтетических данных.

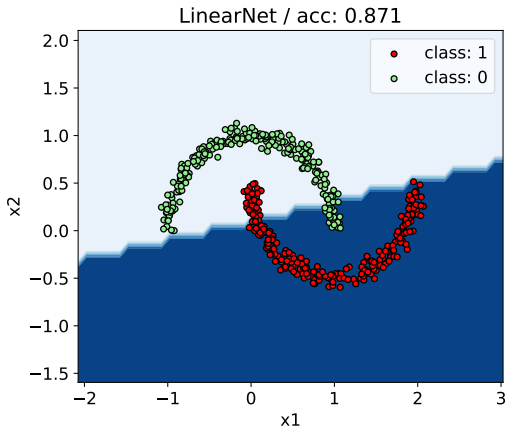
Решающее правило

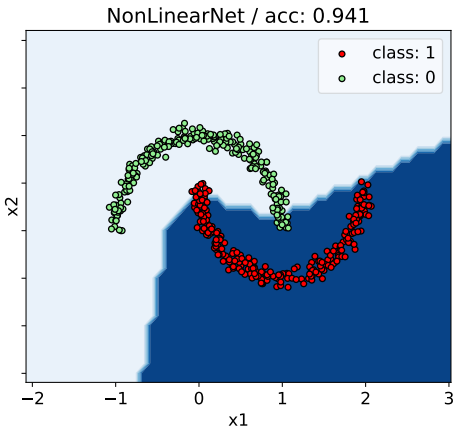


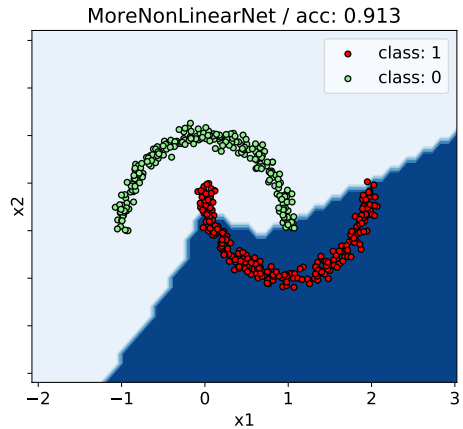




ориг. выборка

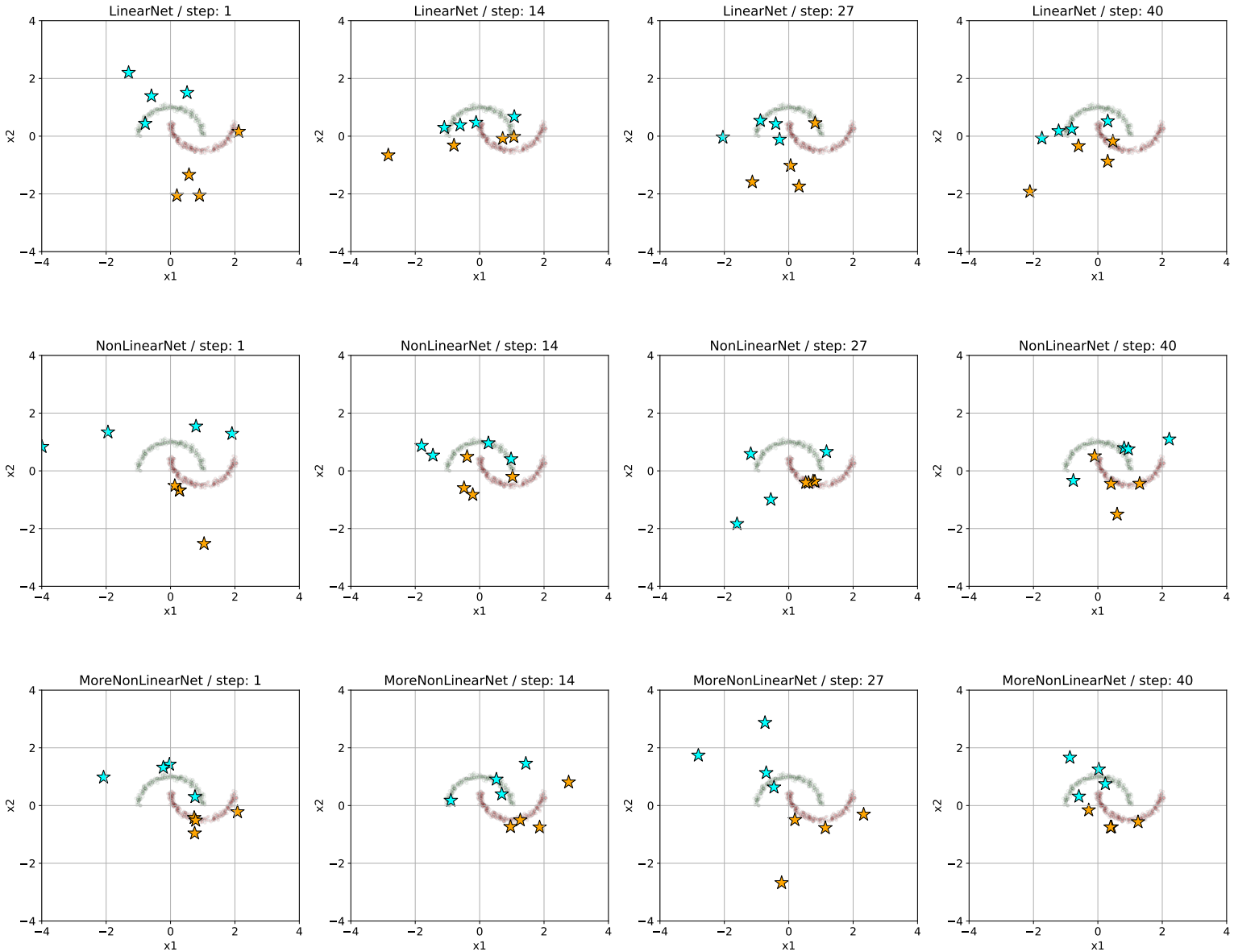






дистил. выборка

Синтетические объекты



Обучаемые длины шага

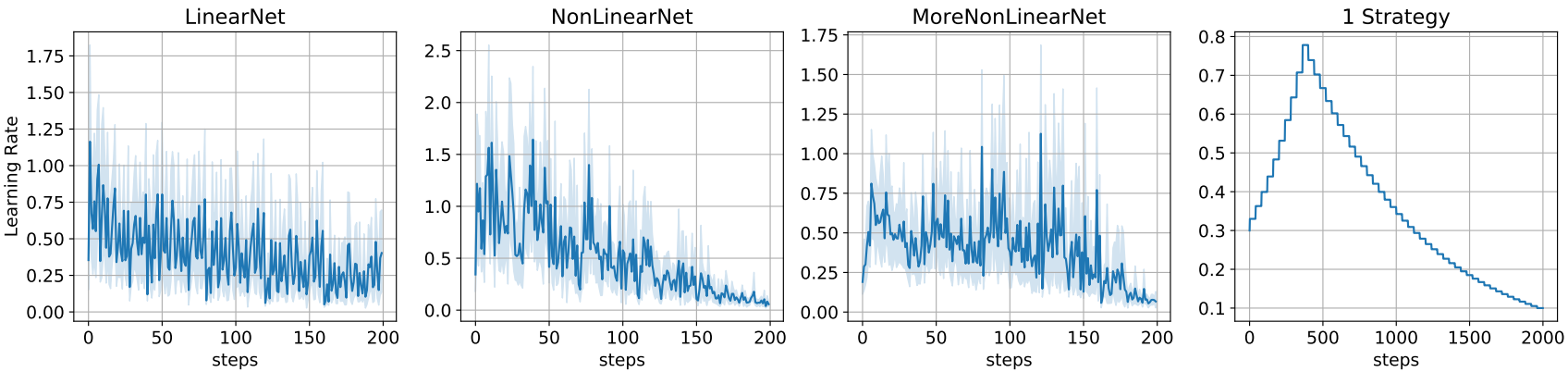


Рис.: Обученные длины шагов, усредненные для 10-и различных инициализаций.

Обобщаемость

Модели Данных	Тестовые Модели		
	LinearNet	NonLinearNet	MoreNonLinearNet
LinearNet	**0.871 ± 0.003**	0.869 ± 0.004	0.864 ± 0.006
NonLinearNet	0.808 ± 0.014	**0.941 ± 0.043**	0.691 ± 0.182
MoreNonLinearNet	0.825 ± 0.014	0.879 ± 0.013	**0.906 ± 0.054**
Strategy1 + LinearNet	**0.867 ± 0.005**	0.860 ± 0.008	0.860 ± 0.010
Strategy1 + NonLinearNet	0.808 ± 0.010	**0.937 ± 0.039**	**0.985 ± 0.015**
Strategy1 + MoreNonLinearNet	0.818 ± 0.012	0.911 ± 0.059	0.926 ± 0.055

Таблица: Итоговая точность (и её среднеквадратичное отклонение) на тестовой выборке для разных наборов синтетических данных и моделей. Жирным выделено наибольше значение в столбце.

Обучение на всех трех архитектурах

Модели Данных	Тестовые Модели		
	LinearNet	NonLinearNet	MoreNonLinearNet
raw steps	**0.859 ± 0.005**	0.881 ± 0.004	0.867 ± 0.122
strategy 1	0.851 ± 0.007	**0.970 ± 0.028**	**0.986 ± 0.014**

Таблица: Итоговая точность (и её среднеквадратичное отклонение) на тестовой выборке для разных моделей, для данных обученных на всех трёх архитектурах. Жирным выделено наибольше значение в столбце.

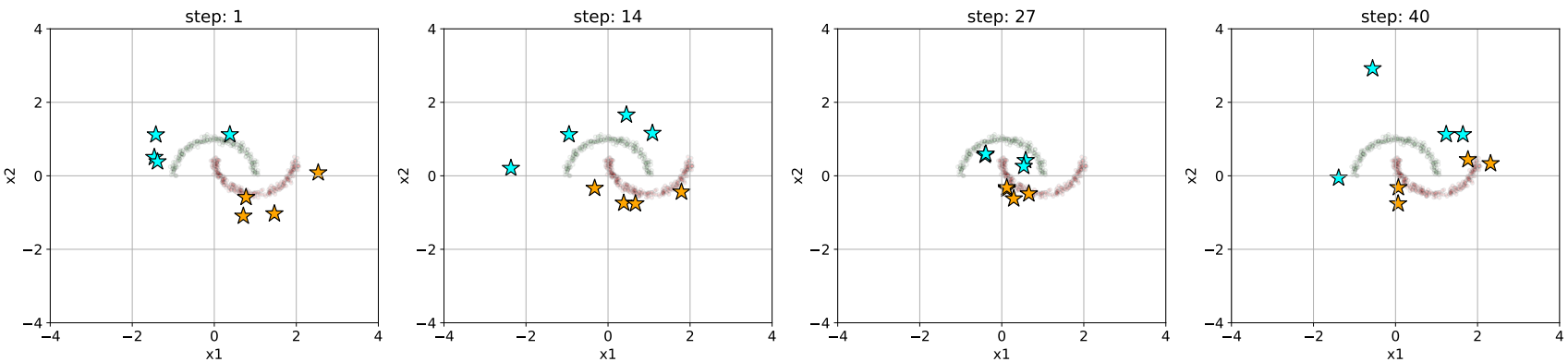


Рис.: Синтетические объекты, при обучении на всех трёх архитектурах.

Список литературы

1. Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. "Dataset Distillation arXiv preprint, 2018.
2. D. Maclaurin, D. Duvenaud, and R. P. Adams. Gradientbased hyperparameter optimization through reversible learning. In ICML, 2015.