Название статьи (авторы статьи): Deep Double Descent: Where Bigger Models and More Data Hurt (Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever)
Автор исследования: Сабина Даянова

Публикация
Данная статья была написана в декабре 2019 года и представлена на конференции ICLR2020 в формате постер. На arxiv числится одна версия статьи.
Все авторы статьи кроме последнего аффилированны с Гарвардским университетом. Это объясняется тем, что Preetum Nakkiran писал эту статью в рамках стажировки в OpenAI под руководством Ilya Sutskever.

Авторы
Посмотрим поближе на значимых коллабораторов этой статьи. Preetum Nakkiran во время написания статьи получал PhD степень в Гарварде, а также был сооснователем исследовательской группы ML Foundations Group, резидентами которой были все авторы статьи. Темами исследования этой группы являлись теоретический и фундаментальный ML/DL. Личными научными интересами Preetum были Generalization in ML, Optimization methods. Сейчас он занимается постдокторантурой в University of California, San Diego под менторством Михаила Белкина.
Что касается Gal Kaplun и Yamini Bansal, они сейчас являются аспирантами на направлении Computer Science. Их научные интересы сильно коллинеарны с интересами Приитума с некоторыми отступлениями (Robust Optimization у Gal, Representations in ML в Yamini). Если посмотреть на совместные статьи, которые написали они и Preetum (“Distributional Generalization: A New Kind of Generalization”, “SGD on Neural Networks Learns Functions of Increasing Complexity”), видно, что все они так или иначе связаны с темой обобщаемости.

Ссылки и цитирования
Посмотрим на список источников статьи. Данная работа достаточно сильно опирается на исследования Михаила Белкина, а именно: на “Reconciling modern machine learning and the bias-variance trade-off” и на “Two models of double descent for weak features”. В первой работе Белкин исследует феномен, при котором сложные модели по типу нейронных сетей интерполируют данные при обучении и с успехом показывают хорошие результаты на тестовых данных, несмотря на то, что, грубо говоря, переобучаются. Во второй работе приводится сухой математический анализ кривизны функции ошибки на примере двух моделей - гауссианы и фурье. Есть мнение, что в первой версии статьи на работы Белкина сослались неподобающим образом. Однако, точно можно сказать, что в финальной версии в нашей работе достаточное количество мест, где указываются ссылки на его работы и благодарности за помощь.
Перейдем к цитированиям. У статьи их более 340. Большинство работ, ссылающихся на нашу, также исследует generalization, с вкраплениями bias-variance trade-off (так как наша работа тоже опирается на это). Из самых примечательных продолжений хочу отметить следующие два. Незадолго после выхода нашей работы, первый автор, Preetum Nakkiran, опубликовал индивидуальную статью - “More Data Can Hurt for Linear Regression: Sample-wise Double Descent”, где он уходит от темы глубинного обучения и рассматривает тот же эффект, но уже на одном из методов классического машинного обучения. Вторая работа, вызвавшая у меня глубокий интерес, – попытка рассмотреть феномен double-descent с точки зрения байесовских методов. В статье “Bayesian Deep Learning and a Probabilistic Perspective of Generalization” (Andrew Gordon Wilson, Pavel Izmailov) авторам удалось показать, что байесовские модели “смягчают” пик кривой ошибки, которая отвечает за критический режим модели.

Дополнительные исследования и применения
Что касается дополнительных исследований, мне было бы интересно посмотреть на такой же многогранный и обширный анализ этого явления (и model-wise DD, epoch-wise DD, sample-wise non-monotonicity), проведенный на другой группе моделей, например на генеративных моделях или на классических методов в ML.
Применение в индустрии: исследователи в компаниях довольно часто пытаются применить модели из популярных научных статей в продакшн. Однако, зачастую, примеры моделей в статьях довольно

игрушечные: у них мало параметров, и набор данных тоже небольшой. Соответственно, появляется необходимость масштабировать модели. Инженерам приходится задумываться о том, что станет с ошибкой после масштабирования модели и после увеличения количества данных. Я считаю, что это статья дает хорошее представление о природе таких вещей.