Экстраполяция и интерполяция нейронных сетей.

Искусственные нейронные сети — это мощные инструменты логического вывода. Однако это не означает, что они могут изучать правила так, как это делают люди. Например, мы можем применять арифметику к произвольно большим числам. Если мы изучили правило сложения однозначных чисел, то при сложении двузначных чисел необходимо учесть составляющую десятков и единиц. Когда число единиц получилось больше десяти, необходимо добавить единицу в столбец десятков. Это ключевое понимание, на котором основана арифметика. Может ли нейронная сеть делать обобщения на случаи, которые далеки от тренировочных данных?

Введем определение тренировочного пространства и тренировочного множества [10]. Каждый вход в нейронную сеть состоит из множества n признаков, принимающих бинарные значения. Это множество может быть использовано для определения n-мерного пространства, которое назовем входящим пространством. Любой входящий элемент соответствует точке во входящем пространстве. Тренировочное множество – это множество точек во входящем пространстве, на которых модель была натренированна. Входящие данные, которые не лежат в тренировочном множестве, но которые полностью состоят из значений признаков, присутствующих в тренировочном множестве, лежат в **тренировочном пространстве**. Любые входящие данные, содержащие значение признака, отсутствующее в тренировочном множестве, лежат за пределами тренировочного пространства. Таким образом, тренировочное пространство это подпространство во входящем пространстве, ограниченное значениями признаков, присутствующих в тренировочном множестве. Например, пусть тренировочное множество состоит из двух векторов: (101) и (011). Вектор (111) не лежит в тренировочном множестве, но состоит из значений признаков из него. Поэтому данный вектор лежит в тренировочном пространстве.

Различают два типа входящих данных:

1. данные, лежащие в пределах тренировочного пространства (то есть состоящие исключительно из значений признаков, на которых была обучена модель);

2. данные, лежащие вне тренировочного пространства (то есть включающие значения признаков, на которых модель не обучалась).

Обобщение нейронной сети в тренировочном пространстве называется **интерполяцией**, а обобщение вне тренировочного пространства – **экстраполяцией**.

Историю об экстраполяции нейронных сетей можно проследить, начиная с работ [3], [10], где исследовались нейронная сеть прямого распространения и простая рекуррентная нейросеть. Рассмотрим функцию идентичности, которую будем тренировать на следующих данных [10]:

вход	выход
0 0 0 1 0	0 0 0 1 0
0 0 1 0 0	0 0 1 0 0
0 0 1 1 0	0 0 1 1 0
0 1 0 0 0	0 1 0 0 0
0 1 0 1 0	0 1 0 1 0
0 1 1 0 0	0 1 1 0 0
0 1 1 1 0	0 1 1 1 0
1 0 0 0 0	1 0 0 0 0
1 0 0 1 0	1 0 0 1 0
1 0 1 0 0	1 0 1 0 0
1 0 1 1 0	1 0 1 1 0
1 1 0 0 0	1 1 0 0 0
1 1 0 1 0	1 1 0 1 0
1 1 1 0 0	1 1 1 0 0
1 1 1 1 0	1 1 1 1 0

Каков будет ответ на вход (11111) ? Для человека нетрудно предсказать, что ответ будет равен входу (11111), не смотря на то, что вход (11111) лежит за пределами тренировочного пространства (значение признака 1 в крайнем правом положении не появлялось в тренировочном множестве). Стандартная нейрон-

ная сеть прямого распространения дает ответ (11110), даже если менять скорость обучения, число нейронов скрытого слоя, число скрытых слоев и последовательность тренировочных примеров. Важно понимать, что ответ нейронной сети — это вполне обоснованное обобщение, математически согласующееся с входными данными. Заметим, что в обучающем наборе условная вероятность того, что самая правая цифра на выходе будет «1», равна нулю. Таким образом, нейронная сеть не является неправильной в абсолютном смысле, скорее обобщение, которое она делает отличается от того, которое делают люди.

Рассмотрим простую рекуррентную нейронную сеть или нейронную сеть Элмана, состоящую из входного слоя, скрытого слоя, слоя контекстных элементов и выходного слоя. На каждом временном шаге на вход подается заданное слово из предложения. Задача нейросети – предсказать следующее слово в этом предложении. Например, нейросеть обучается на серии предложений типа «X – это X» для заданного набора слов. Затем берется новое слово, которое находится за пределами тренировочного пространства. В то время, как для человека не составит труда завершить предложение типа «X – это X» для нового слова, простая рекуррентная нейронная сеть либо активирует какое-то слово из тренировочного пространства, либо никакое слово не будет активировано. Какое слово будет активировано, зависит от набора случайных весов, которые изначально присвоены словам.

То, что нейросети прямого распространения и простые рекуррентные нейросети не могут обобщать за пределы тренировочного пространства не должно вызывать удивление. Данные нейронные сети обучаются с помощью алгоритма обратного распространения ошибки, согласно которому нейроны выходного слоя обучаются независимо друг от друга, также как и нейроны входного слоя. Это следует непосредственно из уравнений корректировки весов.

Методы глубокого обучения способствовали достижению значительных результатов в таких областях, как распознавание образов, компьютерное зрение, машинный перевод и обучение с подкреплением. Однако, они имеют ограниченные возможности обобщения за пределами тренировочных данных [5], [7], [8], [11], [14]. Поэтому они далеки от достижения надежности и гибкости, которую демонстрируют люди.

В работе [2] авторы поставили под сомнение идею о том, что для высоко размерных наборов данных (число признаков >100) нейронные сети интерполируют. Авторы используют следующее определение интерполяции и экстраполяции: интерполяция нейронной сети происходит всякий раз, когда тестовые данные лежат в выпуклой оболочке тренировочных данных, в противном случае происходит экстраполяция. Основной аргумент их точки зрения основан на том, что в пространстве высокой размерности расстояние между точками становится большим и, следовательно, любой новый сэмпл почти наверняка будет лежать вне выпуклой оболочки тренировочных данных. Учитывая многомерность данных, с которыми мы обычно имеем дело, авторы приходят к выводу, что нейронные сети не могут выполнять интерполяцию.

В нейронауках в последние годы наблюдается быстрое развитие и все более широкое использование технологий для записи большого количества биологических нейронов, а также методов для анализа нейронной активности. Анализ нейронной динамики различных отделов головного мозга у некоторых животных показал, что паттерны активности нейронов лежат на многообразии низкой размерности, вложенном в пространство, которое определяется всей популяцией нейронов (обзор работ можно посмотреть в [6]). Это многообразие определяется небольшим числом независимых паттернов коррелирующей активности (нейронными модами). Каждая нейронная мода включает огромное число нейронов всей популяции, при этом каждый нейрон может участвовать сразу в нескольких нейронных модах.

Рис. 1: A) В каждом нейроне синяя и зеленая области показывают величину вклада активации каждой моды в активность этого нейрона. B) Траектория динамики популяционной активности в пространстве трёх нейронов n_1, n_2, n_3. Траектория сосредоточенна на двумерном многообразии, определяемом модами u_1, u_2.

В искусственных нейронных сетях существует понятие внутренней размерности данных, то есть количество переменных, необходимых для их описания без потери существенной информации. Входящие данные, такие как изображения, часто являются многомерными лишь на поверхностном уровне, а на самом деле они живут в пространстве более низкой размерности, называемом внутренним многообразием. В работе [13] показано, что для некоторых баз изображений, таких как MNIST, CIFAR-10, ImageNet, SVHN, CelebA, CIFAR-100, MS-COCO, внутренняя размерность данных существенно меньше размерности пиксельного представления. Например, для базы ImageNet, несмотря на то, что каждое изображение содержит $224 \times 224 \times 3 = 150528$ пикселей, оценка внутренней размерности лежит между 26 и 43. Кроме того, в этой работе исследовалось влияние внутренней размерности данных на производительность глубокой генеративно-состязательной нейросети. В [1], [4], [9] исследовалось влияние внутренней размерности представлений данных в скрытых слоях глубоких нейросетей на производительность (в [4] – для многослойного перцептрона, в [1], [4], [9] – для свёрточных нейронных сетей). Во всех указанных работах установлено, что чем меньше внутренняя размерность данных или представлений данных в скрытых слоях, тем выше производительность нейронной сети.

Далее мы приводим исследования из работы [4], где рассматривается простая нейронная сеть в виде многослойного перцептрона. В данной работе исследуется не только внутренняя размерность представления данных в скрытом слое, но и на сколько подходящим является определение интерполяции через выпуклую оболочку. Многослойный перцептрон обучается распознаванию рукописных цифр из базы MNIST и классификации естественных изображений из базы CIFAR-10. Первый слой соответствует входящему пространству. Например, для базы MNIST входящим пространством будет пиксельное пространство размерностью $28 \times 28 = 786$ пикселей, а для CIFAR-10 размерностью $32 \times 32 \times 3 = 3072$. В обоих случаях это пространство имеет высокую размерность. Далее за входящим слоем следует один или несколько скрытых слоев. Исследуется последний слой нейросети перед категориальным решением, где категории организованны в линейно разделимые кластеры. Назовем пространство, которое соответствует этому слою, нейронным пространством (рис. 2A). Во всех экспериментах размерность этого пространства 128. Для исследования нейронного пространства используется автоэнкодер, который сжимает данные из нейронного пространства для представления их в скрытом внутреннем пространстве меньшей размерности, а затем реконструирует из этих представлений данные, наиболее близкие к данным из нейронного пространства (рис. 2B). Это делается с помощью обучения автоэнкодера, используя тренировочный набор. Важно отметить, что исследование нейронного пространства проводится после того, как нейронная сеть уже обучилась.

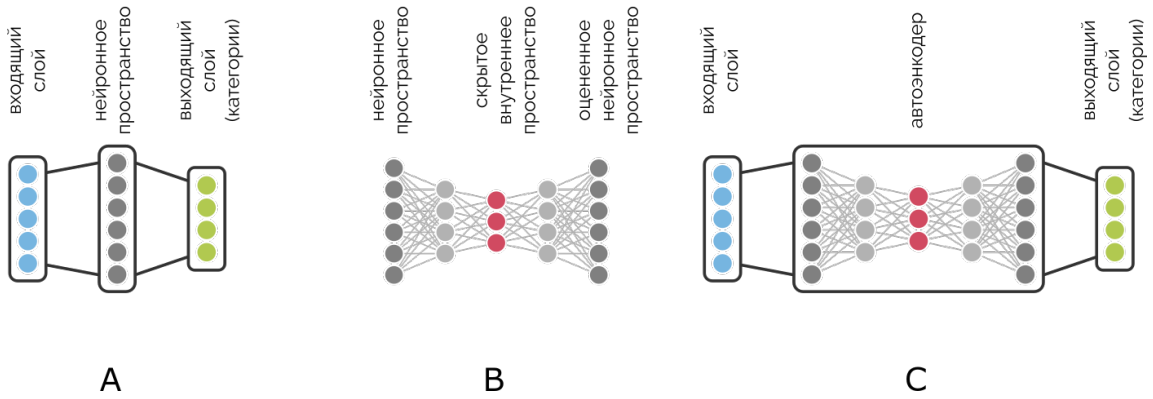

Рис. 2: (A) Нейронная сеть в виде перцептрона; (B) автоэнкодер; (C) гибридная нейронная сеть.

После этого нейронное пространство заменяется на автоэнкодер, и вычисляется точность классификации этой гибридной нейросети (рис. 2C). Эта процедура повторяется для разных размерностей скрытого внутреннего пространства (число нейронов в самом маленьком слое автоэнкодера). Предполагается, что если для заданной размерности скрытого внутреннего пространства автоэнкодер реконструирует данные из нейронного пространства достаточно хорошо, то производительность этой гибридной нейросети должна быть близка к производительности оригинальной нейросети на тестовом наборе. Если эта размерность слишком мала, то точность классификации будет ниже, чем у оригинальной нейронной сети. Начиная от истинной размерности скрытого пространства, производительность гибридной нейросети должна быть равна проиводительности оригинальной и не меняться для больших значений внутренней размерности. Еще раз отметим, что во время этого исследования все части исходной нейросети остаются нетронутыми, и что автоэнкодер обучается не задаче классификации, а только восстановлению данных из нейронного пространства.

На следующем рисунке представлена архитектура многослойного перцептрона. Чтобы исследовать разные варианты одной и той же архитектуры, число нейронов в первом скрытом слое варьируется, таким образом влияя на точность классификации. Нейронное пространство, которое определяется скрытым слоем перед выходящим слоем, выделено темно-серым цветом.

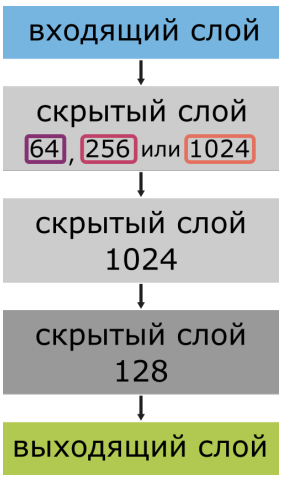

Рис. 3: Архитектура многослойного перцептрона.

Автоэнкодер также является многослойным перцептроном с 256 нейронами в первом и последнем слое. Самый маленький слой автоэнкодера (скрытое внутреннее пространство) рассматривается состоящим из 2, 4, 8 или 16 нейронов. Количество нейронов в этом слое и будет внутренней размерностью нейронного пространства. Все нейроны автоэнкодера используют функцию активации ReLU, за исключением самого маленького слоя. В этом слое используется линейная функция активации. При оценки нейронного пространства используется среднеквадратическая ошибка в качестве функции потерь. Все результаты усредняются по 10 испытаниям с различной случайной инициализацией.

На рисунке 4 представлены результаты исследований для скрытого внутреннего пространства. Рисунок A соответствует базе данных MNIST, рисунок B – базе CIFAR-10. Цвета графиков указывают на количество нейронов в первом скрытом слое нейронной сети: фиолетовый цвет– для нейронной сети с 64 нейронами в первом скрытом слое и наихудшей производительностью, жёлтый – для нейронной сети с 1024 нейронами в первом скрытом слое и наилучшей производительностью. Серая закрашенная область представляет собой отклонение 1% от средней точности модели. Ось X имеет масштаб $log2$.

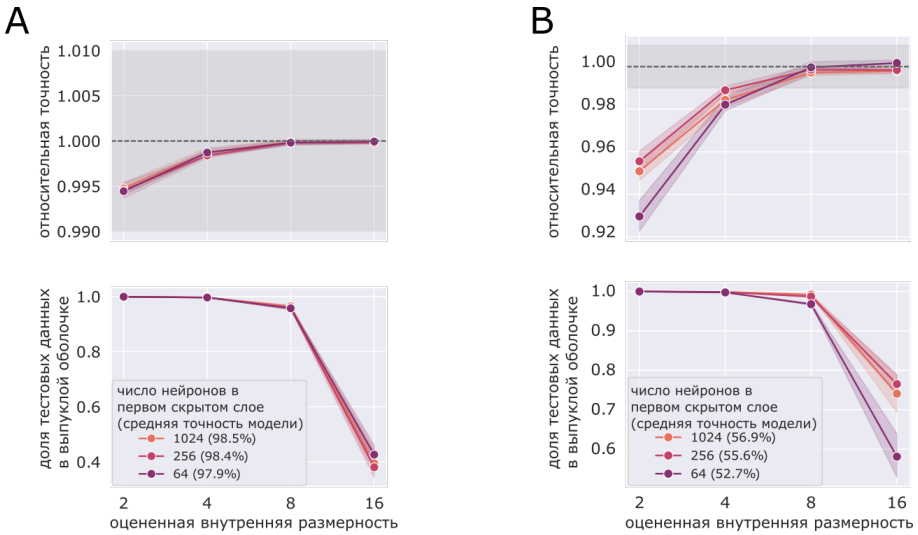

Рис. 4: Относительная точность модели и доля тестовых данных, лежащих в выпуклой оболочке тренировочных данных в зависимости от оцененной внутренней размерности: A – база MNIST, B – база CIFAR-10.

Если сравнивать результаты, полученные по этим двум базам, то в обоих случаях внутренняя размерность нейронного пространства, равная 8 и даже 4, будет достаточной, чтобы получить высокую точность классификации. Доля тестовых данных, лежащих в выпуклой оболочке тренировочных данных, значительно уменьшается с увеличением количества измерений, как и ожидалось. Однако, для заданной низкой внутренней размерности нейронного пространства 2, 4 и 8 подавляющее большинство тестовых данных лежит в выпуклой оболочке тренировочных данных, и адаптируя определение интерполяции из [2], модель работает в интерполяционном режиме. Это противоречит выводам, сделанным в этой работе, где авторы показали, что большинство тестовых данных как во входящем слое, так и в нейронном пространстве не могут лежать в выпуклой оболочке тренировочных данных, из-за высокой размерности. Возникает вопрос, на сколько подходящим является определение интерполяции через выпуклую оболочку.

Существует понятие линейной интерполяции, которое используется во многих областях. По определению линейная интерполяция генерирует промежуточный объект, который лежит строго на линии, соединяющей два заданных объекта. В глубоком обучении для многомерных данных часто используется интерполяция на многообразии, которая генерирует промежуточный объект, находящиеся на кривой во вложенном пространстве (многообразии) [12]. Эта кривая соединяет низкоразмерные представления входящих объектов в высокоразмерном пространстве. Интерполяция на многообразии в глубоком обучении используется для выявления более сложных и нелинейных связей между входящими объектами, в то время как линейная интерполяция может привести к неточностям и некорректным результатам. В работе [2] авторы используют интерполяцию в смысле линейной интерполяции, где новые сэмплы могут лежать только на прямой в Евклидовом пространстве.

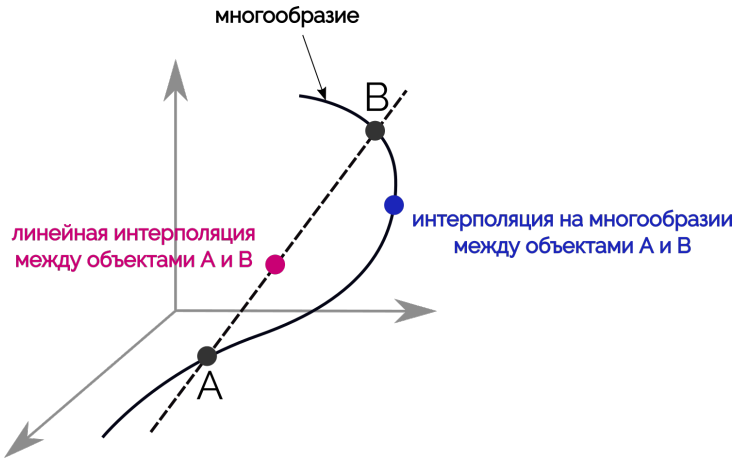

Рис. 5: Интерполяция на многообразии и линейная интерполяция.

Для иллюстрации разницы этих определений рассмотрим картинки двух рукописных цифр 2 и 0 из базы MNIST. Интерполяция их представлений на многообразии это цифра 6, так как каждая точка на многообразии – это представление цифры [5]. Возможно появление несколько неоднозначных изображений, которые близки к границам двух классов, но даже они будут очень похожи на цифры. Интерполяция между цифрами 2 и 0 в исходном пространстве не является настоящей цифрой, так как это среднее значение между пикселями.

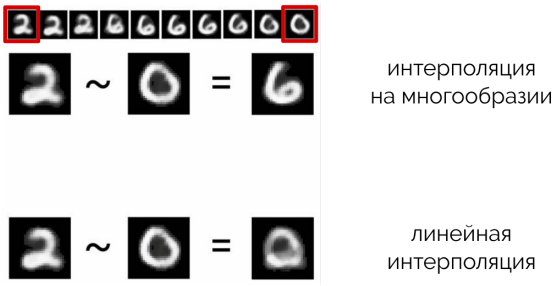

Рис. 6: Интерполяция двух цифр из базы MNIST на многообразии и линейная интерполяция.

Вернемся к графикам на Рисунке 4. Для заданной оцененной внутренней размерности не выявляется очевидной связи между относительной точностью и долей тестовых данных, лежащих в выпуклой оболочке тренировочных данных. Например, для 16-мерного скрытого пространства не более половины тестовых данных попадают в выпуклую оболочку, при этом точность классификации не уменьшается. Изучим подробнее связь между фактом попадания в выпуклую оболочку и вероятностью правильной классификации в 16-мерном скрытом пространстве. Эксперимент проводится для базы MNIST с использованием евклидово расстояния в качестве меры близости тестового сэмпла до ближайшей точки тренировочных данных. Нейронная сеть рассматривается с 1024 нейронами в первом скрытом слое.

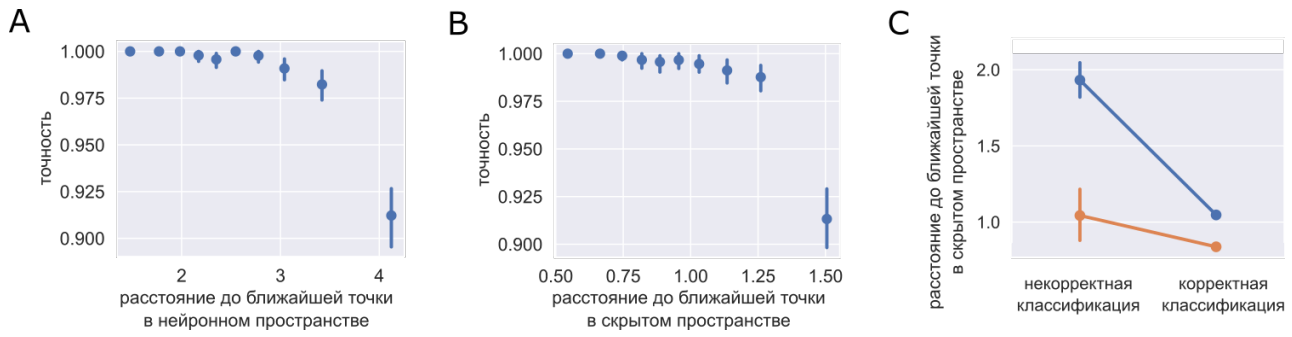

Рис. 7: Результаты для данных из базы MNIST: A – точность классификации, как функция расстояния до тренировочных данных в нейронном пространстве при первом тестировании; B – тоже самое для скрытого пространства; C – среднее расстояние до тренировочных данных, как функции от сэмпла, который корректно классифицирован или нет. Оранжевый цвет обозначает сэмплы, лежащие в выпуклой оболочке тренировочных данных, синий – вне выпуклой оболочки. Каждый кружок представляет 10% тестовых данных, а каждый отрезок – это планка погрешностей, представляющая 95%-ый доверительный интервал для 10 испытаний.

Из результатов, полученных для базы MNIST, можно сделать вывод (рис. 7A и 7B): чем ближе новый сэмпл до тренировочных данных, тем выше шанс корректной классификации. Точки, лежащие в выпуклой оболочке, ближе до тренировочных данных, чем точки, лежащие вне выпуклой оболочки. Но в обоих случаях корректно классифицированные сэмплы ближе до тренировочных данных, чем некорректно классифицированные (рисунок 7C). Также в работе [4] с помощью логистической регрессии предсказывается вероятность корректной классификации нового сэмпла, учитывая как факт нахождения его в выпуклой оболочке, так и факт расстояния от него до тренировочного множества. Оказалось, что расстояния является ключевым фактом. Даже если сэмпл находится внутри выпуклой оболочки, но далеко от тренировочных данных, вероятность некорректной классификации увеличивается с увеличением этого расстояния. И, наоборот, если сэмпл лежит вне выпуклой оболочки, но находится достаточно близко до тренировочных данных, то вероятней всего он будет корректно классифицирован. Как только новый сэмпл оказывает далеко от тренировочного набора, вне зависимости от того лежит он в выпуклой оболочке или нет, производительность падает. Такая картина согласуется с "локальным обобщением" для глубоких нейронных сетей, описанным в работе [5].

Несмотря на высокую размерность данных, понятие расстояния до тренировочных данных остается важным аспектом, определяющим эффективность обобщения искусственной нейронной сети, по крайней мере, для многослойных перцептронов и свёрточных нейронных сетей, для которых получены аналогичные результаты [4]. Эти нейронные сети сталкиваются с трудностями при экстраполяции за пределами локальной области, близкой к обучающему набору. Человек в отличии от нейронных сетей постоянно сталкивается с новыми ситуациями. Используя очень мало данных или не используя их вообще, он с помощью рассуждения, логики, абстракции, символического представления мира способен к экстраполяции. Поэтому глубокие нейронные сети пока не достигли интеллекта уровня человека.

Литература

[1] A. Ansuini, A. Laio, J.H. Macke and D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, In Advances in Neural Information Processing Systems, pp. 6111–6122, 2019.

[2] R. Balestriero, J. Pesenti and Y. LeCun, Learning in High Dimension Always Amounts to Extrapolation, *arXiv preprint arXiv:2110.09485*, 2021.

[3] E. Barnard and L. Wessels, Extrapolation and interpolation in neural network classifiers. IEEE Control Systems Magazine, 12(5), 1992.

[4] L. Bonnasse-Gahot, Interpolation, extrapolation, and local generalization in common neural networks, *arXiv preprint arXiv:2207.08648*, 2022.

[5] F. Chollet, Deep Learning with Python, Second Edition. Manning. 2021.

[6] J.A. Gallego, M. G. Perich, L. E. Miller and S.A. Solla, Neural manifolds for the control of movement. Neuron, 94(5), 2017.

[7] D. Hupkes, V. Dankers, M. Mul, and E. Bruni, Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research, 67, 2020.

[8] B. Lake and M. Baroni, Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In International conference on machine learning, 2018.

[9] X. Ma, Y. Wang, M. E Houle et al., Dimensionality-driven learning with noisy labels, Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[10] G. Marcus, Rethinking eliminative connectionism. Cognitive psychology, 37(3), 1998.

[11] G. Marcus, Deep learning: A critical appraisal, *arXiv preprint arXiv:1801.00631*, 2018.

[12] I. Mayrhofer-Hufnagl and B. Ennemoser, From Linear to Manifold Interpolation, 41st international eCAADe Conference, Austria, Graz University of Technology, v.2, 2023.

[13] P. Pope, C. Zhu, A. Abdelkader et al., The intrinsic dimension of images and its impact on learning, *arXiv preprint arXiv:2104.08894*, 2021.

[14] D. Saxton, E. Grefenstette, F. Hill and P. Kohli, Analysing mathematical reasoning abilities of neural models, International Conference on Learning Representations, 2019.