МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра биомедицинской информатики

МОДЕЛИРОВАНИЕ СТРУКТУРЫ БЕЛКА

Курсовой проект

Рака Алексей Степановича
студента 3 курса,
специальность
«информатика»

Научный руководитель:
А.Ю. Хадарович

Минск, 2017

Оглавление

Введение....3
Нейронные сети....4
Функции активации....5
Персептроны и полносвязная нейронная сеть....6
Свёрточная нейронная сеть....11
Слой свёртки....13
Пулинг....14
Полносвязная сеть в свёрточной сети....15
Обобщение свёрточных нейронных сетей....15
Атомарная свёрточная нейронная сеть....16
Конструкция матрицы расстояний и списка соседей....17
Атомная свёртка....17
Слой радиального пулинга....17
Полносвязный слой в атомарной конволюционной нейронной сети…18
Обучение....19
Данные....20
Разделение выборки....20
Результаты....20
Вывод....21
Список использованной литературы....23

Введение.

Вероятность успеха на начальной фазе исследования лекарственных средств зависит от предсказаний, или измерений, близости кандидата(лиганда) и цели(белка). Пространство синтетически доступных лигандов очень велико, поэтому на данный момент невозможно исследовать всё это пространство. Так что сейчас задача состоит в тестировании как можно большего числа маленьких молекул, пока не будет достигнут достаточный уровень точности. В настоящее время существует значительный компромисс, как в экспериментальных, так и в вычислительных методах скрининга лекарств, между скоростью, стоимостью и точностью.

Внедрение машинного обучения в обнаружение лекарственных средств улучшило виртуальный скрининг лекарств, а также другие физически обоснованные оценки малых молекул. Модели, такие как случайный лес, логистическая регрессия и метод опорных векторов, широко использовались в виртуальном скрининге и биоинформатике в последние десятилетия. Однако такие модели имеют фундаментальный недостаток: молекулы нужно представлять как вектор признаков фиксированных размеров.

Достаточно недавно нейронные сети показали свой потенциал в исправлении этого недостатка. Гибкость глубоких нейронных сетей позволяет моделям использовать признаки более высоких порядков из достаточно простого представления данных. В компьютерном зрении, например, свёрточные нейронные сети применённые к изображениям лиц на первых уровнях учатся различать различные углы, чуть дальше различают глаза, нос и т.д., а на последних уровнях обучаются различать сами лица. В то время, как такие нейронные сети привели к огромным прорывам в области компьютерного зрения и обработки языка, они совсем недавно стали проникать в другие области.

Все возможности нейронных сетей лишь недавно начали реализовываться в областях химии и физики. Учёные продемонстрировали использование нейронных сетей для молекулярных силовых полей, предсказания электронных свойств малых молекул, белок-лигандных соединений, и многого другого. Предыдущие работы, которые рассматривали белок-лигандные соединения использовали лишь полносвязные нейронные сети, основанные на ручном выборе признаков или прямом применении свёрточных нейронных сетей в задаче двух классовой классификации задач разделения малых молекул на связующие и не связующие.

Атомная свёрточная архитектура впервые описанная в статье «Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity» Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, and Vijay S. Pande была вдохновлена атомной нейронной сетью первоночально предложенной Behler и Parrinell для предсказания потенциальной поверхностной энергии с несколькими ключевыми отличиями. Вместо выбора фиксированного параметра для образования признаков в атомарных векторах, сейчас эти параметры оптимизируются одновременно с нейронной сетью, что позволяет модели принимать самостоятельные решения о признаках важных для предсказания связывания лиганда. Кроме того, в атомарной свёрточной сети удаляется часть соседей, что уменьшает время вычислений и позволяет моделям хорошо масштабироваться.

Нейронные сети.

Нейронная сеть — это последовательность нейронов, соединённых между собой синапсами. Структура нейронной сети пришла в мир программирования прямиком из биологии. Благодаря такой структуре, машина обретает способность анализировать и даже запоминать различную информацию.

Нейронные сети используются для решения сложных задач, которые требуют аналитических вычислений подобных тем, что делает человеческий мозг. Самыми распространёнными применениями нейронных сетей является:
1) *Классификация* — распределение данных по параметрам. Например, на вход даётся набор людей и нужно решить, кому из них давать кредит, а кому нет. Эту работу может сделать нейронная сеть, анализируя такую информацию как: возраст, платёжеспособность, кредитная история и так далее.
2) *Предсказание* — возможность предсказывать следующий шаг. Например, рост или падение акций, основываясь на ситуации на фондовом рынке.
3) *Распознавание* — в настоящее время, самое широкое применение нейронных сетей. Используется в Google, когда вы ищете фото или в камерах телефонов, когда оно определяет положение вашего лица и выделяет его и многое другое.

Для того, чтобы понять как работают нейронные сеть сразу разберёмся с базовой составляющей всех нейронных сетей – с нейронами:

Нейрон — это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передаёт её дальше. Они делятся на три

основных типа: входной, скрытый и выходной. В том случае, когда нейросеть состоит из большого количества нейронов, вводят термин слоя. Соответственно, есть входной слой, который получает информацию, n скрытых слоёв, которые её обрабатывают и выходной слой, который выводит результат. У каждого из нейронов есть 2 основных параметра: входные данные (input data) и выходные данные (output data). В случае входного нейрона: input=output. В остальных, в поле input попадает суммарная информация всех нейронов с предыдущего слоя, после чего, она нормализуется, с помощью функции активации f(x) и попадает в поле output.

Также может вводится нейрон смещения, значение которого всегда равно некоторой константе (обычно -1). У такого нейрона нет входных синапсов, однаео есть выходные синапсы. Он используется для того, чтобы сместить на некоторое значение суммарную информацию полученную с предыдущего слоя.

Синапс это связь между двумя нейронами. У синапсов есть 1 параметр — вес. Благодаря нему, входная информация изменяется, когда передаётся от одного нейрона к другому. Допустим, есть 3 нейрона, которые передают информацию следующему. Тогда у нас есть 3 веса, соответствующие каждому из этих нейронов. У того нейрона, у которого вес будет больше, та информация и будет доминирующей в следующем нейроне (пример — смешение цветов). На самом деле, совокупность весов нейронной сети или матрица весов — это своеобразный мозг всей системы. Именно благодаря этим весам, входная информация обрабатывается и превращается в результат.

Функции активации

Как отмечалось ранее у нейронов есть функция активации – нелинейное преобразование к информации полученной нейроном. Нелинейность нужна для формирования более высокоуровневых признаков. Линейные функции обычно не используются, так как суперпозиция линейных функций является линейной функций, то есть не имеет смысла вводить больше одного слоя в нейронной сети, ибо признаки полученные на более глубоких уровнях можно было бы также получить и на первом уровне просто преобразовав веса.

Самыми популярными на сегодняшний день являются следующие функции активации:

1) Сигмоида:

$$f(x)=\frac{1}{1+\exp(-x)}$$

2) ReLu:

$$f(x)=max(0,x)$$

Персептроны и полносвязная нейронная сеть

Объединение нейронов в общую сеть осуществляется различными способами. В зависимости от выбранной топологии выделяют следующие виды нейронных сетей:

1) многослойные нейронные сети (персептроны) – нейроны объединяются в слои, содержащие совокупность нейронов с едиными входными сигналами. Могут содержать входной, выходной и N промежуточных слоев.

2) полносвязные нейронные сети – структуры, в которых каждый нейрон сети имеет прямую связь с другими нейронами.

Перцептрон, или *персептрон* (англ. Perceptron от лат. Perceptio – восприятие) – математическая и компьютерная модель восприятия информации мозгом (кибернетическая модель мозга), предложенная Фрэнком Розенблаттом в 1957 году.

Согласно современной терминологии, персептроны могут быть классифицированы как искусственные нейронные сети:

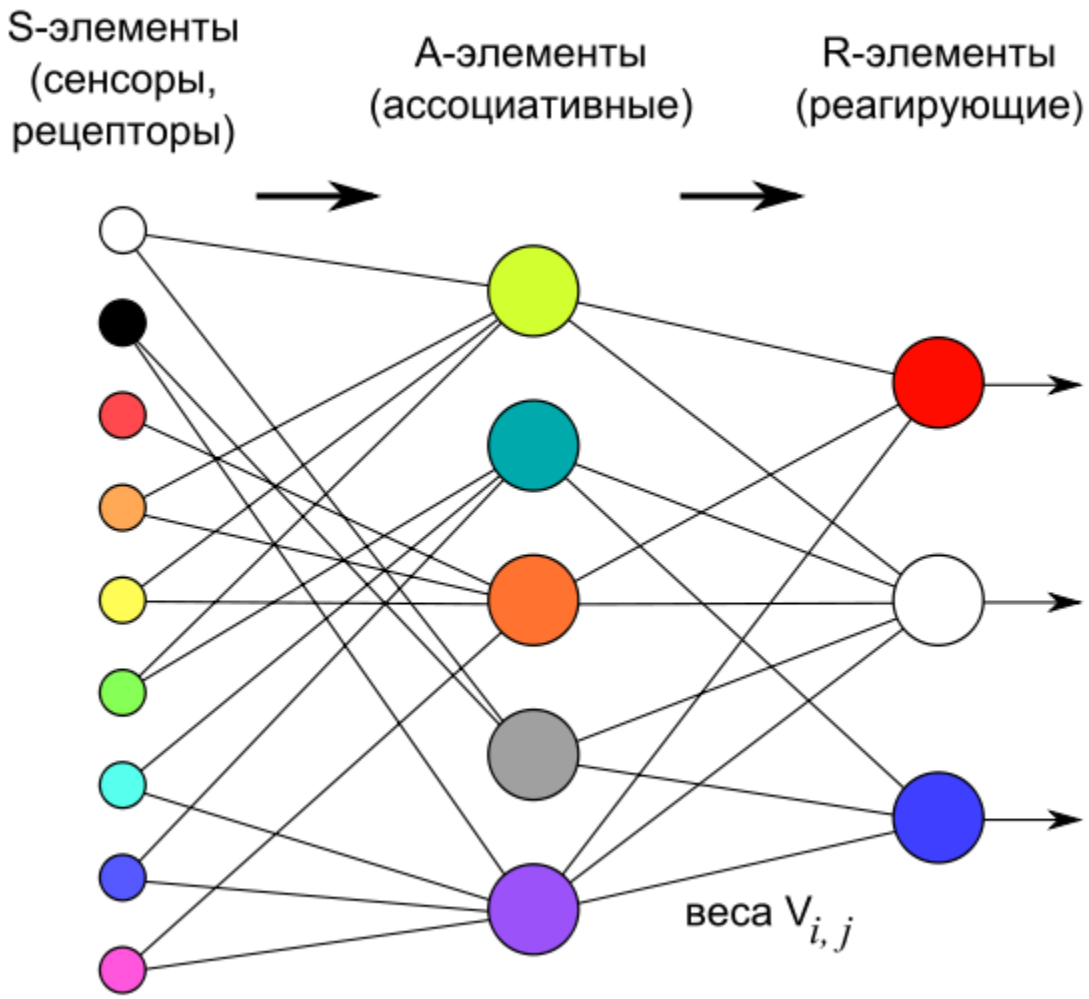
– с одним скрытым слоем;

– с пороговой передаточной функцией;

– с прямым распространением сигнала.

Элементарный персептрон состоит из элементов трёх типов: S-элементов, A-элементов и одного R-элемента.

S-элементы – это слой сенсоров, или рецепторов. Каждый рецептор может находится в одном из двух состояний – покоя или возбуждения, и только в последнем случае он передаёт единичный сигнал в следующий слой, ассоциативным элемента.



A-элементы называются ассоциативными, потому что каждому такому элементу, как правило, соответствует целый набор (ассоциация) S-элементу. A-элемент активизируется, как только количество сигналов от S-элементов на его входе превысило некоторую величину Θ.

Сигналы от возбудившихся A-элементов, в свою очередь передаются в сумматор R, причём сигнал от i-го ассоциативного элемента предаётся с коэффициентом V_i . Этот коэффициент называется весом A—R связи. Так же как и A-элементы, R-элемент подсчитывает сумму значений входных сигналов,

помноженных на веса. R-элемент, а вместе с ним и элементарный перцептрон, выдаёт 1, если линейная форма превышает порог Θ, иначе на выходе будет -1.

Обучение элементарного перцептрона состоит в изменении весовых коэффициентов V_i связей A—R. Веса связей S—A (которые могут принимать значения {-1, 0, 1}) и значения порогов A-элементов выбираются случайным образом в самом начале и затем не изменяются.

После обучений персептрон готов работать в режиме распознавания или обобщении. В этом режиме персептрон должен установить, к какому классу они принадлежат.

Можно выделить 4 довольно обособленных класса персептронов:

1) Персептрон с одним скрытым слоем – классический персептрон, который имеет по одному слою S-, A- и R-элементов;

2) Однослойный персептрон – модель, в которой входные элементы напрямую соединены с выходными с помощью системы весов. Является частным случаем классического персептрона, в котором каждый S-элемент однозначно соответствует одному A- элементу, S—A связи имеют вес +1 и все A-элементы имеют порог Θ = 1;

3) Многослойный персептрон (по Розенблатту) – персептрон, в котором присутствуют дополнительные слои A-элементов.

4) Многослойный персептрон (по Румельхарту) – персептрон, в котором присутствуют дополнительные слои A-элементов, причём, обучение такой сети проводится по методу обратного распространения ошибок, и обучаемыми являются все слои персептрона (в том числе S—A). Является частным случаем многослойного персептрона Розенблатта.

В настоящее время в литературе под термином «персептрон» понимается чаще всего однослойный персептрон. В противоположность однослойному ставят «многослойный персептрон», чаще всего подразумевая многослойный персептрон Румельхарта.

Среди отличий многослойного персептрона Румельхарта от персептрона Розенблатта можно выделить следующие:

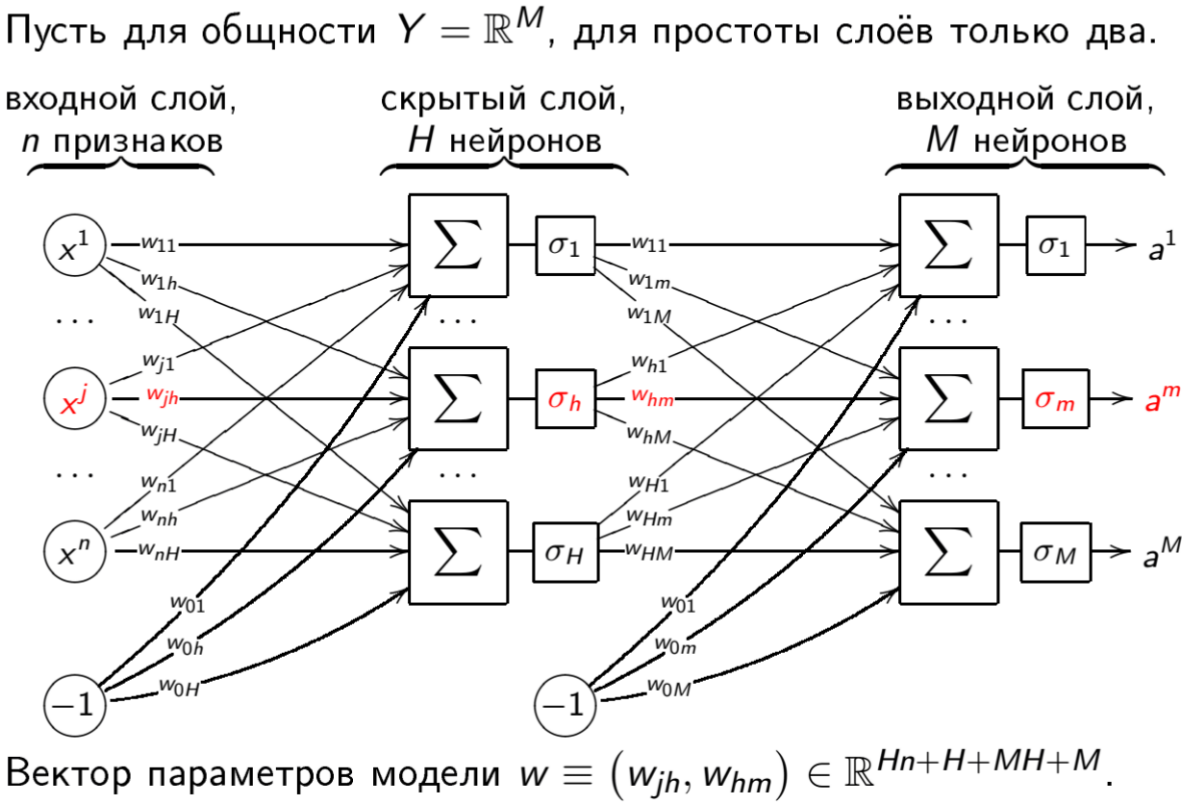
– использование нелинейной функции активации, как правило сигмоидальной;

– число обучаемых слоёв больше одного. Чаще всего в приложения используется не больше трёх;

– сигналы, поступающие на вход, и получаемые с выхода не бинарные, а могут кодироваться десятичными числами, которой нужно нормализовать, так чтобы значения на отрезке от 0 до 1 (нормализация необходима как минимум для выходных данных, в соответствии с функцией активации – сигмоидальной)

– допускается произвольная архитектура связей (в том числе, и полносвязные сети)

– обучение проводится не до отсутствия ошибок после обучения, а до стабилизации весовых коэффициентов при обучении или прерывается ранее, чтобы избежать переобучения.



Теперь подробнее разберём *полносвязную* нейронную сеть. Здесь каждый нейрон передаёт свой выходной сигнал остальным нейронам, включая самого себя. Выходными сигналами сети могут быть все или некоторые выходные

сигналы нейронов после нескольких тактов функционирования сети. Все входные сигналы подаются всем нейронам. Элементы слоистых и полносвязных сетей могут выбираться по-разному. Существует, впрочем, стандартный выбор: нейрон с адаптивным неоднородным линейным сумматором на входе. Для полносвязной сети входной сумматор нейрона фактически распадается на два: первый вычисляет линейную функцию от входных сигналов сети, второй линейную функцию от выходных сигналов других нейронов, полученных на предыдущем шаге. Функция активации нейронов (характеристическая функция) это нелинейный преобразователь выходного сигнала сумматора. Если функция одна для всех нейронов сети, то сеть называют однородной (гомогенной). Если же характеристическая функция зависит еще от одного или нескольких параметров, значения которых меняются от нейрона к нейрону, то сеть называют неоднородной (гетерогенной).

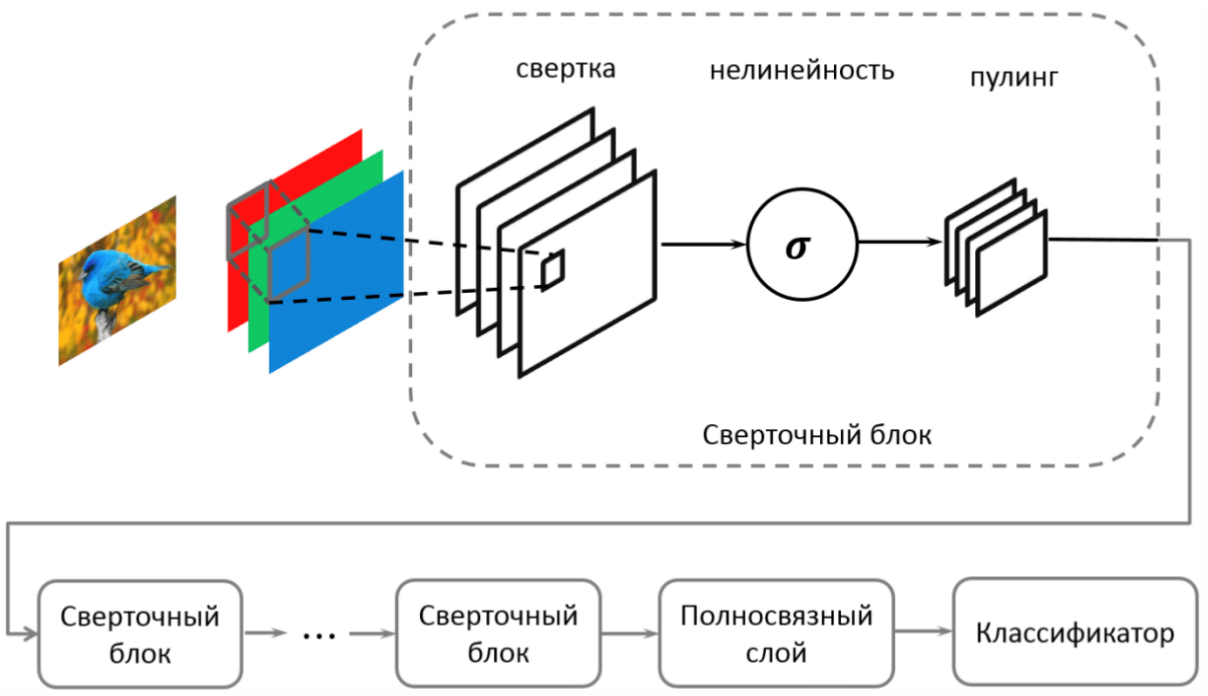
Составлять сеть из нейронов стандартного вида не обязательно. Слоистая или полносвязная архитектуры не налагают существенных ограничений на участвующие в них элементы. Единственное жёсткое требование, предъявляемое архитектурой к элементам сети, это соответствие размерности вектора входных сигналов элемента (она определяется архитектурой) числу его входов. Если полносвязная сеть функционирует до получения ответа заданное число тактов k, то её можно представить как частный случай k-слойной сети, все слои которой одинаковы и каждый из них соответствует такту функционирования полносвязной сети.

Существенное различие между полносвязной и слоистой сетями становится очевидным, когда число тактов функционирования заранее не ограничено слоистая сеть так работать не может.

Доказаны теоремы о полноте: для любой непрерывной функции нескольких переменных можно построить нейронную сеть, которая вычисляет эту функцию с любой заданной точностью. Так что нейронные сети в каком-то смысле могут все.

Однако серьёзным недостатком такой архитектуры сетей является огромное количество обучаемых параметров, из-за этого очень редко можно встретить полносвязную нейронную сеть с глубиной больше 3. Огромное число параметров возникает из-за того, что между двумя слоями размера n и m в такой сети будет nm обучаемых синапсов.

Свёрточная нейронная сеть



Работа свёрточной нейронной сети обычно интерпретируется как переход от конкретных особенностей изображения к более абстрактным деталям, и далее к ещё более абстрактным деталям вплоть до выделения понятий высокого уровня. При этом сеть самонастраивается и вырабатывает сама необходимую иерархию абстрактных признаков (последовательности карт признаков), фильтруя маловажные детали и выделяя существенное.

Подобная интерпретация носит скорее метафорический или иллюстративный характер. Фактически «признаки», вырабатываемые сложной сетью, малопонятны и трудны для интерпретации настолько, что в практических системах не особенно рекомендуется пытаться понять содержания этих признаков или пытаться их «подправить», вместо этого рекомендуется усовершенствовать саму структуру и архитектуру сети чтобы получить лучшие результаты. Так, игнорирование системой каких-то существенных явлений может говорить о том, что либо не хватает данных для обучения, либо структура сети обладает недостатками и система не может выработать эффективных признаков для данных явлений.

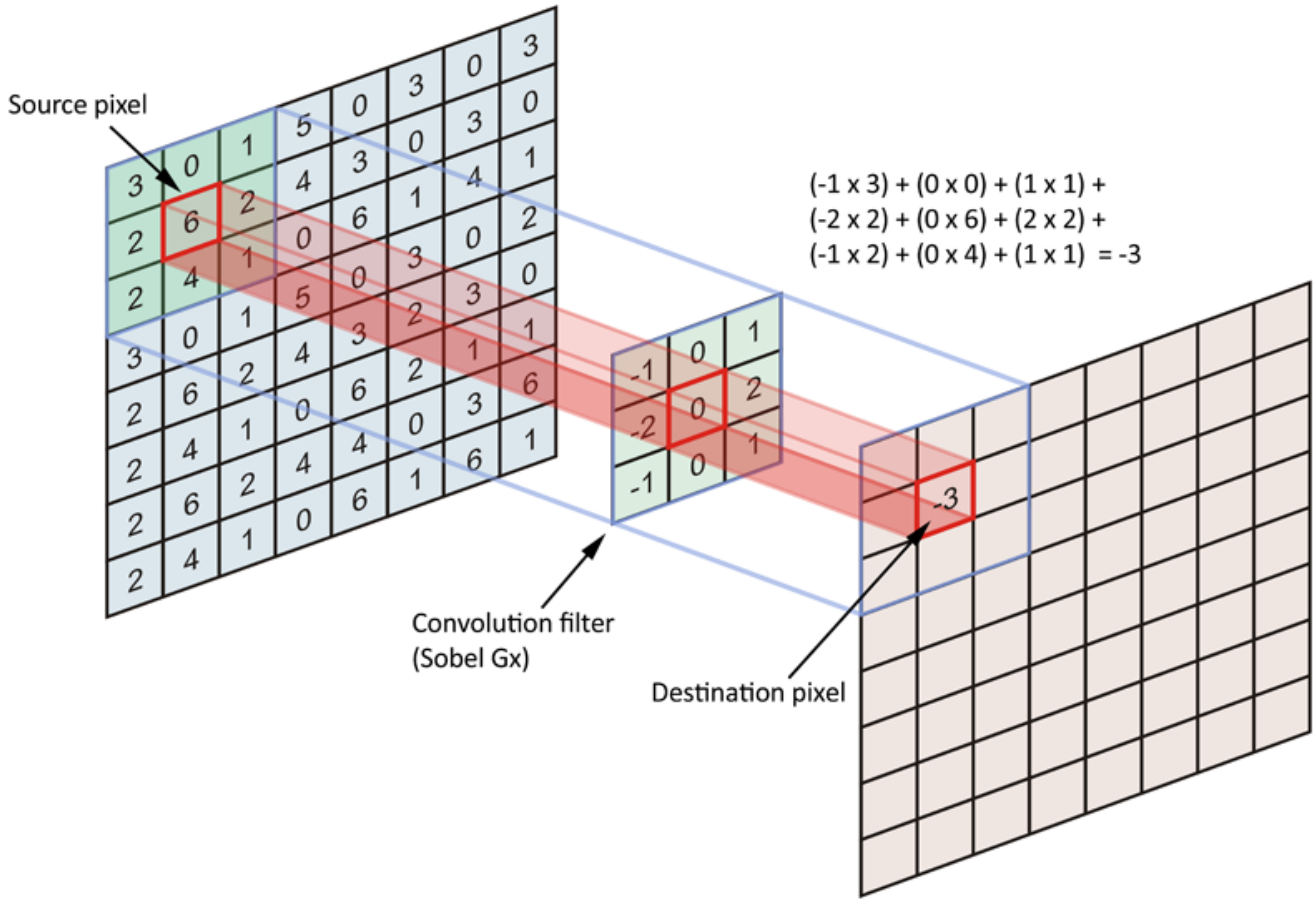
В обычном перцептроне, который представляет собой полносвязную нейронную сеть, каждый нейрон связан со всеми нейронами предыдущего слоя, причём каждая связь имеет свой персональный весовой коэффициент. В свёрточной нейронной сети в операции свёртки используется лишь ограниченная матрица весов небольшого размера, которую «двигают» по всему обрабатываемому слою (в самом начале – непосредственно по входному изображению), формируя после каждого сдвига сигнал активации для нейрона следующего слоя с аналогичной позицией. То есть для различных нейронов выходного слоя используются одна и та же матрица весов, которую также называют ядром свёртки. Её интерпретируют как графическое кодирование какого-либо признака, например, наличие наклонной линии под определенным углом. Тогда следующий слой, получившийся в результате операции свёртки такой матрицей весов, показывает наличие данного признака в обрабатываемом слое и её координаты, формируя так называемую карту признаков. Естественно, в свёрточной нейронной сети набор весов не один, а целая гамма, кодирующая элементы изображения (например линии и дуги под разными углами). При этом такие ядра свёртки не закладываются исследователем заранее, а формируются самостоятельно путём обучения сети классическим методом обратного распространения ошибки. Проход каждым набором весов формирует свой собственный экземпляр карты признаков, делая нейронную сеть многоканальной (много независимых карт признаков на одном слое). Также следует отметить, что при переборе слоя матрицей весов её передвигают обычно не на полный шаг (размер этой матрицы), а на небольшое расстояние. Так, например, при размерности матрицы весов 5×5 её сдвигают на один или два нейрона (пикселя) вместо пяти, чтобы не «перешагнуть» искомый признак.

Операция пулинга, выполняет уменьшение размерности сформированных карт признаков. В данной архитектуре сети считается, что информация о факте наличия искомого признака важнее точного знания его координат, поэтому из нескольких соседних нейронов карты признаков выбирается максимальный и принимается за один нейрон уплотнённой карты признаков меньшей размерности. За счёт данной операции, помимо ускорения дальнейших вычислений, сеть становится более инвариантной к масштабу входного изображения.

Рассмотрим типовую структуру свёрточной нейронной сети более подробно. Сеть состоит из большого количества слоёв. После начального слоя (входного изображения) сигнал проходит серию свёрточных слоёв, в которых чередуется собственно свёртка и субдискретизация (пулинг). Чередование слоёв позволяет

составлять «карты признаков» из карт признаков, на каждом следующем слое карта уменьшается в размере, но увеличивается количество каналов. На практике это означает способность распознавания сложных иерархий признаков. Обычно после прохождения нескольких слоев карта признаков вырождается в вектор или даже скаляр, но таких карт признаков становится сотни. На выходе свёрточных слоёв сети дополнительно устанавливают несколько слоев полносвязной нейронной сети (перцептрон), на вход которому подаются оконечные карты признаков.

Слой свёртки



Слой свёртки -- это основной блок свёрточной нейронной сети. Слой свёртки включает в себя для каждого канала свой фильтр, *ядро свёртки* которого

обрабатывает предыдущий слой по фрагментам (суммируя результаты матричного произведения для каждого фрагмента). Весовые коэффициенты ядра свёртки (небольшой матрицы) неизвестны и устанавливаются в процессе обучения.

Особенностью свёрточного слоя является сравнительно небольшое количество параметров, устанавливаемое при обучении. Так например, если исходное изображение имеет размерность 100×100 пикселей по трём каналам (это значит 30000 входных нейронов), а свёрточный слой использует фильтры с ядром 3x3 пикселя с выходом на 6 каналов, тогда в процессе обучения определяется только 9 весов ядра, однако по всем сочетаниям каналов, то есть 3×3×3×6 =162, в таком случае данный слой требует нахождения только 162 параметров, что существенно меньше количества искомых параметров полносвязной нейронной сети.

Пулинг

Слой пулинга (иначе подвыборки, субдискретизации) представляет собой нелинейное уплотнение карты признаков, при этом группа пикселей (обычно размера 2×2) уплотняется до одного пикселя, проходя нелинейное преобразование. Наиболее употребительна при этом функция максимума. Преобразования затрагивают непересекающиеся прямоугольники или квадраты, каждый из которых ужимается в один пиксель, при этом выбирается пиксель, имеющий максимальное значение. Операция пулинга позволяет существенно уменьшить пространственный объём изображения. Пулинг интерпретируется так. Если на предыдущей операции свёртки уже были выявлены некоторые признаки, то для дальнейшей обработки настолько подробное изображение уже не нужно, и оно уплотняется до менее подробного. К тому же фильтрация уже ненужных деталей помогает не переобучаться. Слой пулинга, как правило, вставляется после слоя свёртки перед слоем следующей свёртки.

Кроме пулинга с функцией максимума можно использовать и другие функции – например, среднего значения или *L2*-нормирования. Однако практика показала преимущества именно пулинга с функцией максимума, который включается в типовые системы.

Помимо задач уплотнения изображения (что полезно против переобучения), находят распространение также идеи использования малых фильтров or discarding the pooling layer altogether.

Полносвязная сеть в свёрточной сети.

После нескольких прохождений свёртки изображения и уплотнения с помощью пулинга система перестраивается от конкретной сетки пикселей с высоким разрешением к более абстрактным картам признаков, как правило на каждом следующем слое увеличивается число каналов и уменьшается размерность изображения в каждом канале. В конце концов остаётся большой набор каналов, хранящих небольшое число данных (даже один параметр), которые интерпретируются как самые абстрактные понятия, выявленные из исходного изображения.

Эти данные объединяются и передаются на обычную полносвязную нейронную сеть, которая тоже может состоять из нескольких слоёв. При этом полносвязные слои уже утрачивают пространственную структуру пикселей и обладают сравнительно небольшой размерностью (по отношению к количеству пикселей исходного изображения).

Обобщение свёрточных нейронных сетей

Свёрточные нейронные сети – это один из главных алгоритмов, который сейчас используется в задачах машинного обучения. Они добивается больших успехов в различных сферах, таких как распознавание изображений, распознавание речи, компьютерное зрение, машинный перевод и даже игра в Go.

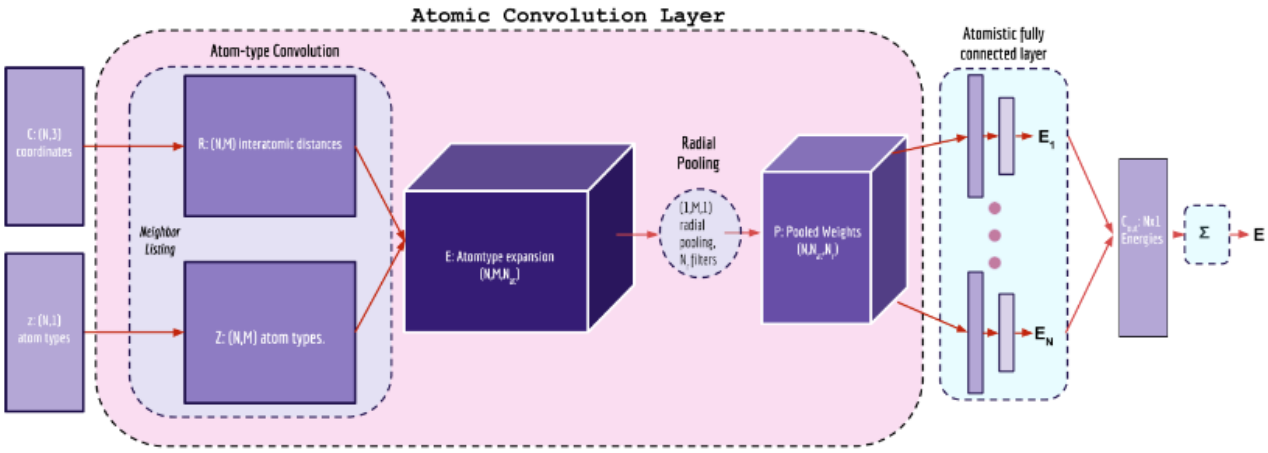
Главный успех свёрточных сетей справедливо приписывается свёртке. Но любое успешное применение свёрточных нейронных сетей неявно использует базовые атрибуты входных данных. В частности, стандартный слой свёртки может быть применён только к структурированному сеткой входу, так как он изучает локализованные прямоугольные фильтры, неоднократно свёртывая их через несколько патчей входного сигнала. Кроме того, для того чтобы конволюция была эффективной, вклад должен быть локально связным, что означает, что сигнал

должен быть сильно коррелирован в местных регионах в основном несвзан в глобальных регионах.

Это также требует, чтобы входные данные были неподвижными, чтобы сделать фильтры свёртки инвариантными к сдвигам, так как они могу выбирать локальные объекты независимо от пространственного расположения.

Поэтому свёрточные нейронные сети изначально ограничены некоторым поднабором возможных данных. Теме не менее, впечатляющие успехи свёрточных сетей побуждают к обобщению их на неструктурированные данные, которые имеют локальные связи и стационарность свойств. Так изменяя представление данных и свёрточные/пулинговые операции для некоторой отдельной задачи можно перейти к использованию свёрточных сетей.

Атомарная свёрточная нейронная сеть



Архитектура атомарной свёрточной нейронной сети показан на рисунке 1. В ней используются 2 свёрточные операций: атомная свёртка и радиальный пулинг. Атомная свёртка использует матрицу расстояний до M ближайших соседей для извлечений признаков некоторой окрестности атома из данного представления молекулы.

Конструкция матрицы расстояний и списка соседей.

Матрица расстояний R и матрица атомных номеров Z строится из атомных координат. Я использую лишь M ближайших соседей для того, что бы сложность этого этапа была O(NM), а не O(N^2). Я использовал значение M равное 12. Т.е. мы превращаем начальную матрицу размерна (N, 3) в матрицу расстояний размера (N, M). Также будем использовать типы ближайших атомов. Дли них заводим матрицу Z размера (N, M).

Атомная свёртка.

Выход атомной свёртки конструируется из матрицы расстояний R и матрицу типов атомов Z. К матрицу R применяется фильтр размера (1 x 1) с шагом 1 и глубиной N_{at}, где N_{at} – это число уникальных атомных типов присутствующих в молекулярной системе. Ядром атомной свёрточной операции является фунция:

$$(K * R)^a_{i,j} = R_{i,j} K^a_{i,j}$$

где

$$K^a_{i,j} = \begin{cases} 1, Z_{i,j} = N_a \\ 0, иначе \end{cases}$$

где N_a – это атомный номер из a = {1, …., N_{at} }. Атомарный конволюционный слой применяется к матрице расстояний до соседей и создаёт матрицу E размера (N, M, N_{at}). Об атомарной свёртке также можно думать как о one-hot encoding.

Слой радиального пулинга

Радиальный пулинг – это процесс уменьшающий размер матрицы признаков, который принимает на вход тензор размера (N, M, N_{at}). Уменьшение размерности используется для предотвращения переобучения. В дополнение к этим свойствам после применения радиального пулинга мы на выходе получаем тензор, который инвариантен к перестановкам соседей у атомов.

Как говорилось выше на этом слое мы принимаем тензор размера (N, M, N_{at}). Затем выбираем число N_r – число желаемых радиальных фильтров. Каждый радиальный фильтр объединяет часть тензора (1 x M x 1). Т.е. после применения всех радиальных фильтров мы получаем тензор размера (N, N_{at} , N_r). Каждый радиальный фильтр применяет следующие операции:

$$f_s(r_{i,j}) = \exp\frac{-(r_{i,j} - r_s)^2}{\sigma_s^2} f_c(r_{i,j})$$

$$f_c(r_{i,j}) = \begin{cases} \frac{1}{2}\cos\left(\frac{\pi r_{i,j}}{R_c}\right), 0 < r_{i,j} < R_c \\ 0, r_{i,j} \geq R_c \end{cases}$$

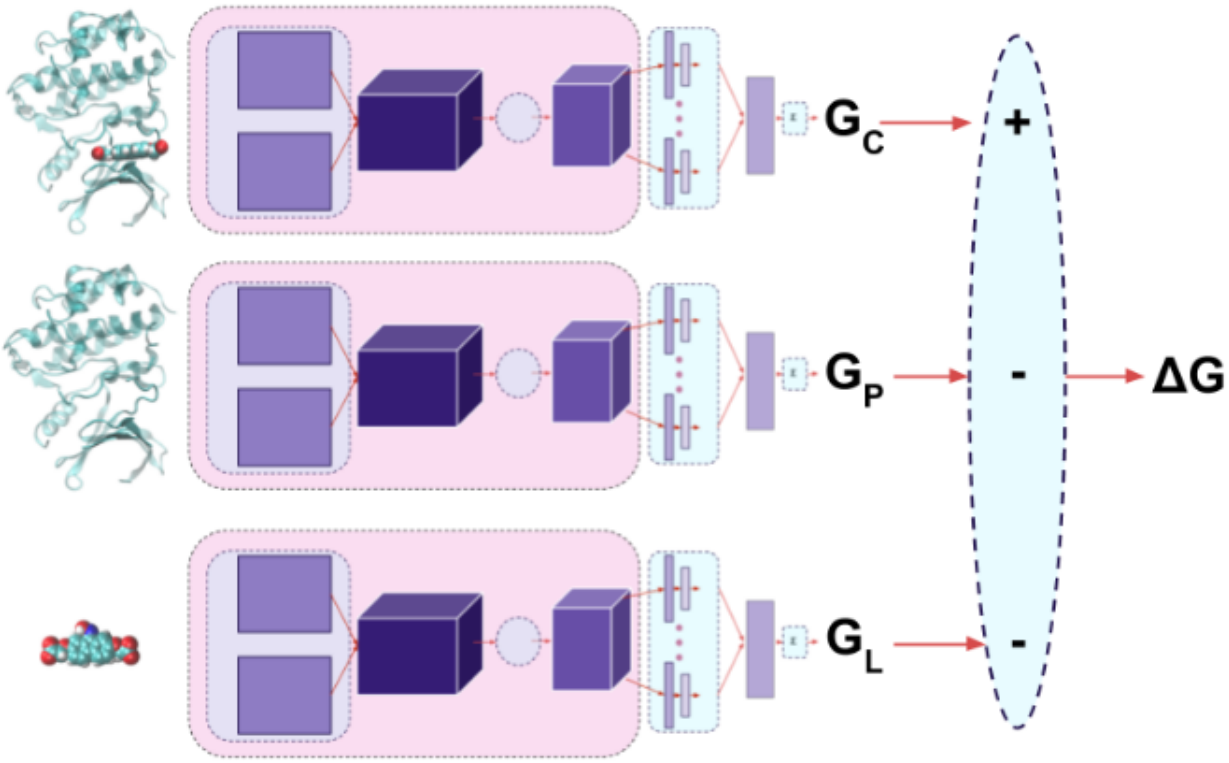
$$P_{i,n_a,n_r} = \beta_{n_r} \sum_{j=1}^{M} f_{n_r}(E_{i,j,n_a}) + b_{n_r}$$

где r_s и σ_s – это обучаемые параметры, а β и b – это константы.

Полносвязный слой в атомарной конволюционной нейронной сети

Получив из пулингового слоя тензор P размера (N, N_{at} , N_r) выровняем его координаты для каждого атома переведя его в матрицу размера (N, $N_{at} * N_r$). Далее будем брать по очереди каждую строку этой матрицы и передавать её в полносвязный слой, который работает также как и в обычных нейронных сетях, а затем, когда получим N различных выходов просуммируем их, это и будет нашим ответом.

Обучение



Структура атомарной свёрточной сети получилась полностью дифференцируемой по отношению к положению атомов. Для каждой системы нам нужно посчитать освободившуюся энергию по следующей формуле:

$$\Delta G_{complex} = G_{complex} - G_{protein} - G_{ligand}$$

Имея такой ответ от свёрточной сети осталось записать лишь функцию ошибки, которая будет равна:

$$L = (\Delta G_{complex} - \gamma_{complex})^2$$

Где $\gamma_{complex}$ это логарифм от Ki (энзимная ингибиторная константа)

Данные

Для обучения нейронной сети я использовал данные BindingDB, в которых есть более миллиона записей о различных белоково-лигандных комплексах. Однако лишь о 10000 записей из этой базы данных известы Ki (Enzyme Inhibition Constant) – энзимная ингибиторная константа, а также структуры белка, лиганда и белково-лигандного комплекса. При этом структура лиганда дана сразу как трёхмерная структрура, а о структурах белка и комплекса известны их структуры в id Protein Data Bank. Тренировал эту модель с использование стохастического градиентного спуска с размером батча 24 и ADAM оптимизатором на 100 эпох.

Разделение выборки

Я разделил полученную выборку на тренировочную и тестовую в соотношении 70/30. Предварительно все данные были случайно перемешаны, затем первую часть (примерно 7000) комплексов отправили в тренировочную выборку, оставшиеся данные были тестовым.

Результаты

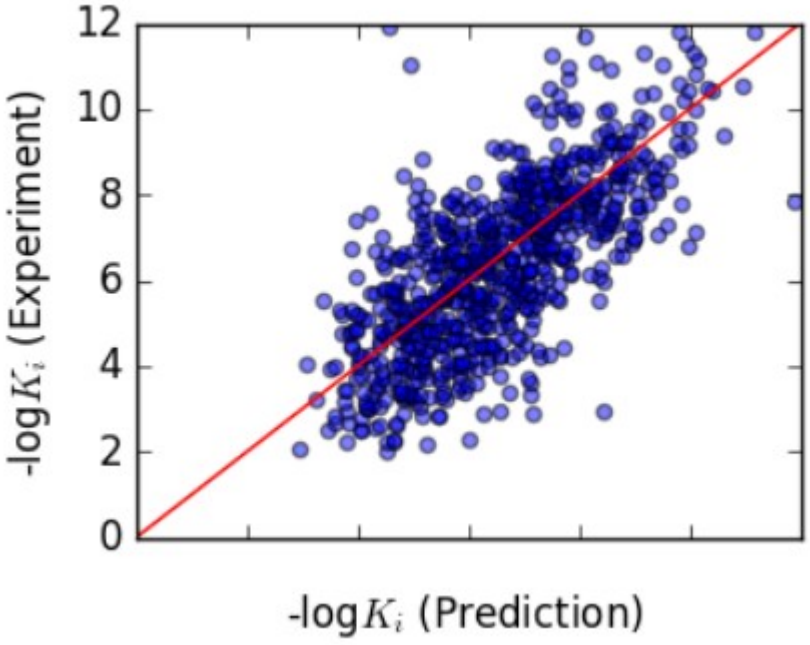
Результаты работы я решил оценить используя коэффициент детерминации Пирсона:

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{SS_{res}/n}{SS_{tot}/n} = 1 - \frac{SS_{res}}{SS_{tot}}$$

где $SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – сумма квадратов остатков регрессии, y_i , $\hat{y}_i$ – фактические и расчётные значения объясняемой переменной соответственно.

$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$ – общая сумма квадратов, $\bar{y}$ – среднее значение фактической переменной.

По данной формуле я получил, что $R^2 \approx 0.4$, что чуть меньше чем в оригинальной статье.



Вывод

Результаты эмпирических функций базирующиеся на межмолекулярных силах или на химических признаках широко используются в соединении с молекулярным докингом во время ранних стадий исследований лекарств для предсказания потенциальных возможностей и силы связывания лекарство-подобных молекул и данной цели. Эти модели требуют некоторых знаний в области физической химии и биологии, чтобы закодировать параметры или признаки лучше чем это позволяют сделать базовые модели машинного обучения. В данной работе я использую общую конволюционную операцию в трёхмерном пространстве для изучения химических взаимодействии на атомном уровне используя атомные координаты в трёхмерном пространстве и демонстрирую применение в структурном биоактивном предсказании. Атомная конволюционная нейронная сеть (АКНН) обучается предсказывать энергию связывания белково-лигандных комплексов через прямой подсчёт энергии ассоциированной с комплесом, белком и лигандом. Нековалентные взаимодействия, присутствующих в комплексе и отсутствующие в белке или лиганде, выявляются и модель изучает силы взаимодействия, связанные с этими особенностями.

Я утверждаю, что атомарная свёрточная нейронная сеть представляет собой фундаментальный прогресс в прогнозировании сродства белка-лиганда в связи с тем, что она используют полностью дифференцируемое, от начала и до конца обучаемое представление.

Также, как в машинном переводе, Богданов с соавторами ввёл первую от начала и до конца обучаемую нейросетевую систему машинного перевода. Изначально такая архитектура работала лишь наравне с более старыми алгоритмами машинного обучения. Однако спустя два года, последний нейросетевой алгоритм показал резкое улучшение на 60% по сравнению со старой система. Я считаю, что такой же потенциал есть и у атомарной свёрточной нейронной сети.

Однако существенным недостатком такой сети является то, что для предсказания энергии связывания требуется структура не только лиганда и белка, но также и их комплекса, а это сильно сказывается на применимости такой архитектуры, ибо зачастую сразу известны только структуры белка и лигандов, а получение достаточно хороших предсказаний структуры их комплекса является достаточно сложной задачей.

Наконец отметим, что атомарные свёрточные нейронные сети также могут быть применимы к ряду приложений, непосредственно не связанных с обнаружением лекарственных средств. Например, атомарные свёрточные сети смогут улучшить существующие нейронные сети на основе потенциалов для автоматизированного подбора потенциальной энергии поверхности (силовые поля), которые имеют широкое применение в теоретической химии для ускорения начальной симуляции молекулярной динамики многочастичных систем. Кроме того, атомарные свёрточные нейронные сети являются общей техникой выбора признаков, которая должна хорошо работать на традиционных задачах виртуального скрининга, где требуется трёхмерная структурная информация, таких как предсказание электронных свойств молекул и виртуального скрининга новых материалов, например, органических фотовольтаических и светоизлучающих диодов.

Список использованной литературы

1) «Neural Machine Translation by Jointly Learning to Align and Translate». Dzmitry Bahdanau, Kyunghyun Cho, Y. Bengio.

2) «Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity». Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, and Vijay S. Pande

3) «ImageNet Classification with Deep Convolutional Neural Network». Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton.

4) «A Generalization of Convolutional Neural Networks to Graph-Structured Data». Yotam Hechtlinger, Purvasha Chakravarti, Jining Qin