УДК 004.8 **DOI** 10.15622/ia.22.3.4

Н.Г. Шилов, А.В. Пономарев, А.В. Смирнов

АНАЛИЗ МЕТОДОВ ОНТОЛОГО-ОРИЕНТИРОВАННОГО НЕЙРО-СИМВОЛИЧЕСКОГО ИНТЕЛЛЕКТА ПРИ КОЛЛАБОРАТИВНОЙ ПОДДЕРЖКЕ ПРИНЯТИЯ РЕШЕНИЙ

Шилов Н.Г., Пономарев А.В., Смирнов А.В. **Анализ методов онтолого-ориентированного нейро-символического интеллекта при коллаборативной поддержке принятия решений.**

Аннотация. Нейросетевой подход к ИИ, получивший особенно широкое распространение в последнее десятилетие, обладает двумя существенными ограничениями – обучение моделей, как правило, требует очень большого количества образцов (не всегда доступных), а получающиеся модели не являются хорошо интерпретируемыми, что может снижать доверие к ним. Использование символьных знаний как основы коллаборативных процессов с одной стороны и распространение нейросетевого ИИ с другой, обусловливают необходимость синтеза нейросетевой и символьной парадигм применительно к созданию коллаборативных систем поддержки принятия решений. В статье представлены результаты аналитического обзора в области онтолого-ориентированного нейро-символического интеллекта применительно к решению задач обмена знаниями при коллаборативной поддержке принятия решений. А именно, в ходе обзора делается попытка ответить на два вопроса: 1. как символьные знания, представленные в виде онтологии, могут быть использованы для улучшения ИИ-агентов, действующих на основе нейронных сетей (передача знаний от человека к ИИ-агентам); 2. как символьные знания, представленные в виде онтологии, могут быть использованы для интерпретации решений, принимаемых ИИ-агентами, и объяснения этих решений (передача знаний от ИИ-агента к человеку). В результате проведенного обзора сформулированы рекомендации по выбору методов внедрения символьных знаний в нейросетевые модели, а также выделены перспективные направления онтолого-ориентированных методов объяснения нейронных сетей.

Ключевые слова: нейро-символический ИИ, априорные знания, машинное обучение, глубокое обучение, объяснимый ИИ, XAI, онтологии.

1. Введение. Нейросетевой подход к ИИ в последнее десятилетие получил широкое распространение, искусственные нейронные сети активно используются для решения широкого спектра задач обработки информации (особенно, слабоструктурированной – видео, аудио, тексты). Вместе с тем, одним из значительных недостатков нейросетевого подхода к ИИ при принятии решений является то, что результат работы нейронной сети не всегда является легко интерпретируемым и объяснимым. Это, в частности, ограничивает доверие экспертов к результатам работы нейронных сетей и сдерживает их применение в ответственных областях. Данная проблема осознается научным сообществом – к настоящему времени предложен широкий спектр методов, ориентированных на интерпретацию и объяснение предсказаний, получаемых с помощью нейронных сетей [1], однако значительная часть таких

методов предназначена для экспертов в области машинного обучения и искусственного интеллекта, а не для экспертов проблемной области [2].

Коллаборативные системы поддержки принятия решений, основанные на взаимодействии людей (экспертов) и агентов, действующих на основе искусственного интеллекта (ИИ-агентов), являются одной из областей применения ИИ, в которых данный недостаток является весьма существенным. В таких системах команда, состоящая из разнородных участников, в ходе работы над заданной конечным пользователем проблемной ситуацией, осуществляет сбор и обработку информации, формирует и оценивает возможные альтернативы, позволяя конечному пользователю принять взвешенное и обоснованное решение. При этом распределение задач между участниками может быть как жестким, диктуемым заранее заданным сценарием, так и более гибким, когда участники процесса непрерывно анализируют текущее состояние решения проблемы и вносят вклад в соответствии со своими возможностями [3].

Основой коллаборативных процессов является обмен знаниями и взаимное обучение (перенос знаний как от человека к ИИ-агенту, так и наоборот). При этом коммуникативные процессы в широком смысле требуют наличия некоторой символьной системы, обеспечивающей взаимодействие [4].

Использование символов как основы коллаборативных процессов с одной стороны и распространение нейросетевого ИИ с другой обусловливают необходимость синтеза нейросетевой и символьной парадигм применительно к созданию коллаборативных систем поддержки принятия решений. Подобный синтез получил название нейро-символический искусственный интеллект [5]. Под нейро-символическим искусственным интеллектом понимается очень широкий спектр методов. В данной статье рассматривается одно из направлений подобной конвергенции, особенно важное в коллаборативных системах, а именно – перенос знаний от человека к ИИ-агенту (что позволяет использовать априорные символьные знания для повышения качества работы нейросетевых ИИ-агентов) и от ИИ-агента к человеку (что позволяет объяснять результат работы ИИ-агента с помощью символов). При этом, в статье рассматривается лишь один из видов символьного представления знаний – онтологии (рисунок 1). Актуальность использования онтологий как основы коммуникативной системы при коллаборативной поддержке принятия решений обусловлена двумя факторами:

1) Онтология представляет собой формализованное представление терминологии проблемной области, а значит и непротиворечивый язык, который понятен эксперту и может обрабатываться программно.

2) К настоящему моменту разработано большое количество онтологий для многих проблемных областей. Эти знания описаны с использованием стандартизованных языков (Semantic Web) и потенциально могут быть использованы в широком спектре конкретных приложений.

Таким образом, в статье представлены результаты аналитического обзора в области онтолого-ориентированного нейро-символического интеллекта применительно к решению задач обмена знаниями в коллаборативных системах поддержки принятия решений. А именно, в ходе обзора делается попытка ответить на два вопроса:

1) Как символьные знания, представленные в виде онтологии, могут быть использованы для улучшения ИИ-агентов, действующих на основе нейронных сетей (передача знаний от человека к ИИ-агентам).

2) Как символьные знания, представленные в виде онтологиии, могут быть использованы для интерпретации решений, принимаемых ИИ-агентами, и объяснения этих решений (передача знаний от ИИ-агента к человеку).

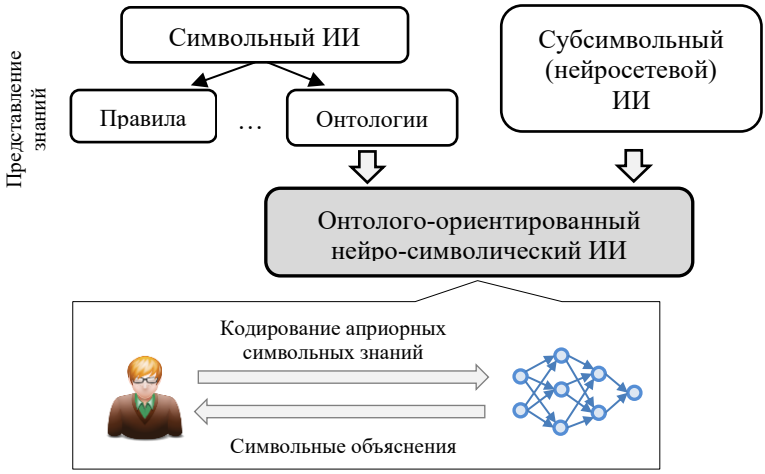

Рис. 1. Место рассматриваемых методов в общей палитре методов ИИ

2. Архитектуры интеграции символьных знаний в субсимвольные (нейросетевые) модели машинного обучения. При создании интеллектуальных систем с использованием методов машинного обучения и, в частности, нейронных сетей, комбинирование последних с символьными знаниями может осуществляться различными способами [6, 7]. В работе [8] выделены четыре различных подхода интеграции, названные «нейросетевое приближение» («neural approximative reasoning»), «нейросетевое рассуждение» («neural unification»), «интроспекция» («introspection») и «интегрированное получение знаний» («integrated knowledge acquisition»).

Под термином «нейросетевое приближение» понимаются методы, использующие нейронные сети для генерации приближенных выводов в интеллектуальных системах. Одним из наиболее очевидных путей для этого типа интеграции является реализация аппроксимирующей нейронной сети для имеющихся правил.

Подход «нейросетевое рассуждение» подразумевает повторение механизма, используемого для автоматического доказательства теорем, а именно выполнение последовательности логических утверждений, приводящих к подтверждению или опровержению исходного утверждения [9]. В нейронных сетях для этого используются методы кодирования знаний с использованием выделенных элементов сети и специальных типов связей для описания элементов утверждения. Рассуждение осуществляется на основе минимизации показателя «энергии» при обновлении состояния нейронной сети. Обучение таких сетей, как правило, выполняется с использованием «метаинтерпретатора», генерирующего обучающие шаблоны, например, посредством кодирования успешных доказательств на языке Prolog. Обученная на таких примерах доказательств нейронная сеть способна обобщать стратегию управления для выбора доводов при доказательстве утверждений.

Под «интроспекцией» понимаются методы и технологии, с помощью которых ИИ «наблюдает» за собственным поведением и улучшает свою работу. Этот подход может быть реализован с помощью нейронных сетей, которые «наблюдают» за последовательностью шагов, выполняемых ИИ при осуществлении логического вывода. Такой подход часто называют управляющим знанием (control knowledge). Когда наблюдаемое поведение ИИ закодировано соответствующим образом, нейронная сеть может научиться избегать ошибочных путей и быстрее приходить к своим выводам.

Подход «интегрированное получение знаний» основан на следующих предпосылках: а) ИИ в основном зависит от человека-эксперта, который формулирует знания в виде символьных утверждений (правил); б) эксперту практически невозможно описать свои знания в виде правил (в частности, очень трудно описать знания, приобретенные опытным путем). Поэтому, например, ИИ может оказаться не в состоянии поставить диагноз, который может поставить опытный врач [10]. Таким образом, основная проблема заключается в том, как извлечь знания из ограниченного набора примеров (малые данные) для использования ИИ. Модели машинного обучения расширяют возможности классического ИИ в области логического вывода за счет способности к обобщению и обработке неполных данных. То есть можно использовать алгоритмы обучения нейронной сети с учителем для извлечения закономерностей из примеров, а затем генератор символьных правил может преобразовать эти закономерности в правила, реализуемые, например, на языке Prolog. С другой стороны, наличие явных символьных знаний (правил) позволяет уменьшить объем обучающих данных для выявления неявных закономерностей. Сгенерированные правила и обученная нейронная сеть встраиваются в сервисы на основе ИИ в качестве базы знаний.

В работе [11] предложена классификация архитектур, объединяющих символьные знания и субсимвольные (нейросетевые) знания (называемые комбинированными нейронными системами). Данная классификация включает следующие типы архитектур (рисунок 2):

– Унифицированная архитектура: символьные знания непосредственно кодируются в нейронной сети.
  - Локальная коннекционистская архитектура (localist connectionist architecture): отдельные фрагменты нейронной сети направлены на кодирование символьных знаний.
  - Распределенная нейронная архитектура: символьные и нейросетевые знания кодируются невыделенными перекрывающимися фрагментами нейронной сети.

– Трансформационная архитектура (аналогична унифицированной архитектуре, но включает механизмы перевода (трансформации) субсимвольных представлений знаний в символьные и/или наоборот). Обычно, такая архитектура реализуется через механизмы извлечения символьных знаний (например, правил) из обученной нейронной сети.

Архитектуры интеграции символьных знаний в нейросетевые модели машинного обучения

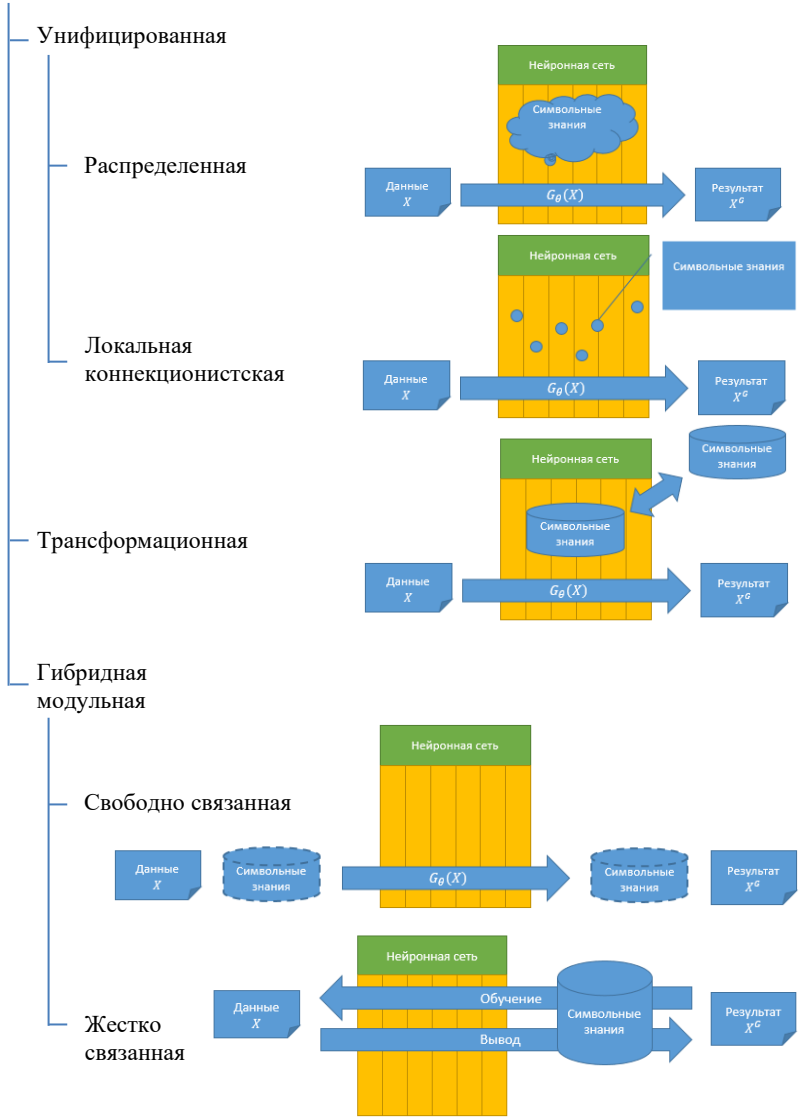

Рис. 2. Архитектуры интеграции символьных знаний в нейросетевые модели машинного обучения

– Гибридная модульная архитектура: символьные и нейросетевые знания кодируются в отдельных модулях (модуль символьных знаний и модуль нейросетевых знаний).

- Свободно связанная: информация может передаваться от одного модуля к другому только в одном направлении. Как правило, в моделях с подобной архитектурой символьные знания используются либо для предобработки и/или дополнения данных перед их передачей в нейронную сеть, либо для постобработки выходных данных нейронной сети.
- Жестко связанная: обмен информацией осуществляется через общие структуры данных в любом направлении.
- Полностью интегрированная архитектура: модули взаимосвязаны по нескольким каналам или даже на основе их перекрывающихся фрагментов.

В работе [12] рассмотрены непосредственно модели представления знаний в рамках нейронных сетей. Авторы предлагают следующую классификацию.

– Пропозиционная логика / логика высказываний (propositional logic).

- Представление на основе правил. Работы по представлению символьных знаний в коннекционистских сетях направлены на адаптацию параметров моделей для установки эквивалентности между функцией отображения, представленной средствами нейросетевой модели, и правилами логического вывода. Было показано, что ограничение значений синаптических весов нейронной сети позволяет выполнять расчеты на основе алгоритма прямого распространения, способные точно имитировать поведение правил логического вывода.
- Представление на основе формул. Одна из проблем с представлением знаний на основе правил, например, в стиле KBANN (Knowledge-Based Artificial Neural Network / основанная на знаниях искусственная нейронная сеть) или CILP (Connectionist inductive learning and logic programming / программирование коннекционсистского индуктивного обучения и логики), заключается в том, что дискриминационная структура искусственных нейронных сетей позволяет рассчитать только подмножество переменных (следствия формулы "если-то"), если только не

используются рекуррентные сети, а остальные переменные (предпосылки формулы "если-то") будут рассматриваться только как входы. Это не вполне соответствует поведению логических формул и не обеспечивает поддержку общего логического вывода, где может быть выведена любая переменная. Для решения этой проблемы можно использовать генеративные нейронные сети, поскольку они могут рассматривать все переменные как недискриминационные. В таком формульном подходе, обычно связанном с ограниченными машинами Больцмана в качестве строительного блока, основное внимание уделяется отображению логических формул на симметричные коннекционистские сети, каждая из которых характеризуется функцией энергии [13, 14].

– Логика первого порядка.

  • Пропозиционализация. Представление знаний в логике первого порядка в нейронных сетях является известной проблемой, но она может быть отчасти решена за счет изучения представления пропозиционной логики с помощью методов пропозиционализации [15]. Такие методы позволяют преобразовать базу знаний первого порядка в пропозиционную базу знаний с сохранением логических следствий. В нейро-символических вычислениях пропозиционализация конкретизированных высказываний (bottom clause prositionalisation) является популярным подходом, поскольку такие высказывания могут быть закодированы непосредственно в нейронных сетях как характеристики данных с сохранением семантики.

  • Тензоризация. Тензоризация – это класс подходов, которые ориентированы на встраивание таких символов логики первого порядка, как константы, факты и правила, в тензоры вещественных значений [16 – 18]. Обычно константы представляются в виде векторов (тензор первого порядка). Предикаты и функции представляются в виде матриц (тензор второго порядка) или тензоров более высокого порядка.

  • Темпоральная логика. Одной из самых ранних работ по темпоральной логике и нейронным сетям является подход на основе коннекционистской темпоральной логики (Connectionist Temporal Logic / CTL), в котором

используются ансамбли рекуррентных нейронных сетей для представления семантики возможного мира линейной временной логики [19]. С одним скрытым слоем и полулинейными нейронами сети могут вычислять семантику правил темпоральной логики с фиксированной точкой (fixed-point semantics). Другая работа по представлению временных знаний предложена в работе [20], представляющей последовательную коннекционистскую темпоральную логику (Sequential Connectionist Temporal Logic / SCTL), где CILP расширено для работы с нелинейной авторегрессионной моделью. В работе [21] нейро-символические когнитивные агенты представляют временные знания в рекуррентных темпоральных ограниченных машинах Больцмана. Здесь правила темпоральной логики моделируются в виде рекурсивных конъюнкций, представленных рекуррентными структурами. Встраивание темпоральных реляционных знаний было изучено и в тензорной рекуррентной нейронной сети с применением в вопросно-ответных системах [22].

Рассмотренные далее работы сгруппированы в соответствие с используемой архитектурой объединения символьных и субсимвольных знаний.

2.1. Унфицированная архитектура. В работе [23] предложено кодировать символьные правила либо путем добавления дополнительных скрытых (необучаемых) блоков (локальная коннекционистская архитектура), либо путем полного преобразования базы правил в нейронную сеть с помощью метода KBANN [24] (распределенная нейронная архитектура). Во втором случае блоки являются обучаемыми, то есть выполняется итеративное создание элементов скрытого слоя, описывающих разделяющую поверхность для конкретных примеров обучающей выборки. Представленные в статье эксперименты показывают значительное увеличение точности при интеграции правил, особенно при интеграции с помощью метода KBANN, когда правила корректируются в процессе обучения. Отличительной особенностью KBANN является его способность работать с приблизительными символьными знаниями, которые уточняются в процессе обучения.

2.1.1. Локальная коннекционистская архитектура. Отличительной особенностью некоторых подходов с унифицированной архитектурой является возможность уточнять

топологию нейронной сети и, таким образом, добавлять новые правила в (переформулированную) базу правил. Поэтому при использовании онтологий проблемной области, в которых отсутствуют правила, обобщение оказывается слабым, а обучение может испортить исходные правила – даже те, которые изначально были правильными. В статье [25] представлен алгоритм TopGen, расширяющий алгоритм KBANN, который эвристически ищет возможные расширения сети KBANN. В TopGen это достигается путем динамического добавления скрытых узлов к нейронному представлению онтологии проблемной области, что аналогично добавлению правил к базе знаний. Представленные в статье эксперименты показывают, что данный алгоритм способен эвристически находить эффективные места для добавления узлов в базы знаний. Алгоритм показал статистически значимое улучшение по сравнению с алгоритмом KBANN во всех представленных пяти областях применения.

Особенностью работы [26] является то, что в ней представлена схема, которая использует оценки функций «черного ящика» в сочетании с символьными выражениями, определяющими отношения между данными функциями. Авторы используют древовидные LSTM-сети для реализации структуры деревьев символьных выражений. Числа, присутствующие в данных оценки функций, представляются в виде десятичного представления в древовидной кодировке.

В работе [27] авторы используют символьные знания для повышения производительности графовых сверточных нейронных сетей (GCN), а также их обучения на меньших объемах обучающих данных. Авторы расширяют классические GCN посредством внедрения формул пропозиционной логики. Для создания семантически верных расширений разработаны методы распознавания неоднородности вершин и семантической регуляризации, которые включают структурные ограничения. Механизм встраивания формул проецирует логические графы, представляющие формулы, на все многообразие решений таким образом, чтобы результат логического вывода ассоциировался с расстоянием так, чтобы удовлетворяющие формуле значения находились бы ближе к встроенной формуле (рисунок 3). Такое пространство позволяет быстро проводить приблизительные проверки значений и используется для оценки функции логических потерь, которые регуляризируют нейронную сеть для решения целевой задачи.

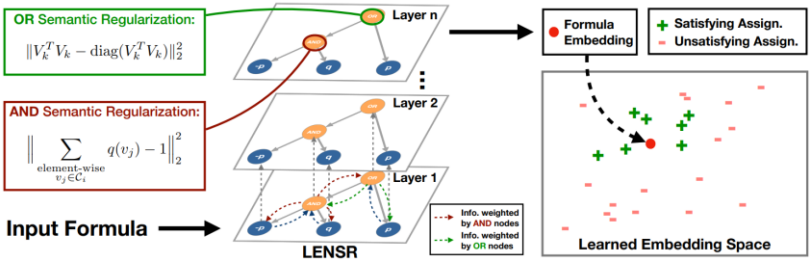

Рис. 3. Расширение классической GCN посредством внедрения формул пропозиционной логики [27]

2.1.2. Распределенная нейронная архитектура. В [28] представлена структура, которая объединяет глубокие нейронные сети с логическими правилами первого порядка, что позволило интегрировать человеческие знания и намерения в нейронные модели. В частности, предложена итеративная процедура, которая переносит структурированную информацию логических правил в веса нейронных сетей. Перенос осуществляется через обучающую сеть, построенную с использованием принципа апостериорной регуляризации. Используются две сети – учитель и ученик. Сеть-учитель сначала учится имитировать логические правила, заданные аналитически. Затем, сеть-учитель используется для обучения сети-ученика. На каждой итерации сеть-учитель обновляется путем проецирования сети-ученика в подпространство, ограниченное правилами, в результате чего она приобретает желаемые свойства. Сеть-ученик обновляется путем балансирования между аппроксимированием сети-учителя и предсказанием истинных результатов на обучающей выборке (рисунок 4).

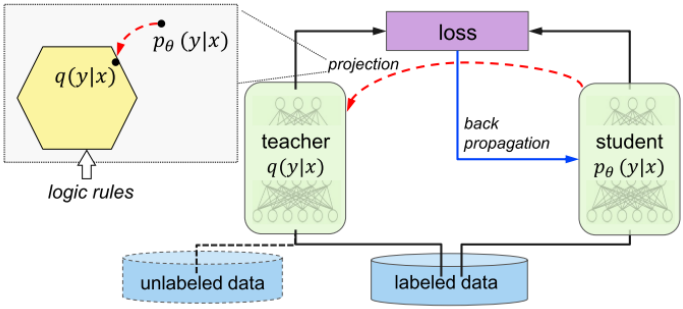

Рис. 4. Использование пары сеть-учитель – сеть-ученик для внедрения логических правил в нейронную сеть [28]

В работе [29] в качестве символьных знаний используются логические правила. Обучение выполняется в два этапа. Сначала модель машинного обучения, построенная на основе нейронной сети, обучается только на правилах, которые могут описывать требуемые закономерности довольно грубо (с вероятностью, отличной от 1). Затем, на этапе уточнения модель обучается «классическим» способом на примерах. Данный подход позволил авторам существенно ускорить сходимость модели в процессе обучения.

2.2. Трансформационная архитектура. Работа [30] представляет комплексную схему, объединяющую символьный и субсимвольный подходы, и объединяет результаты работы нескольких исследовательских групп. Эта схема (рисунок 5) рассматривает сочетание символьного и нейросетевого обучения как трехэтапный процесс: 1) введение символьной информации в нейронную сеть, тем самым (частично) определяя топологию и начальные весовые параметры сети; 2) уточнение этой сети с помощью численного метода оптимизации, такого как обратное распространение, возможно, под руководством символьных знаний; 3) извлечение символьных правил, которые точно представляют знания, содержащиеся в обученной сети. Хотя эти три компонента образуют полную картину: приблизительно правильная символьная информация на входе и более точная символьная информация на выходе, в то же время они могут изучаться независимо друг от друга.

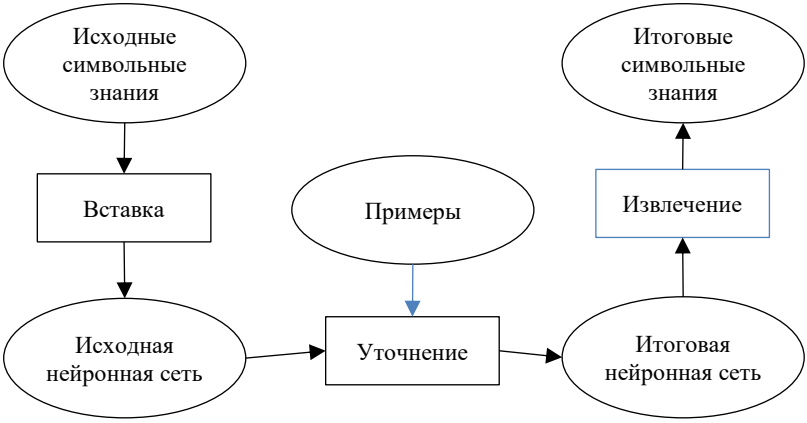

Рис. 5. Подход к объединению символьных и субсимвольных знаний

2.3. Гибридная модульная архитектура

2.3.1. Свободно связанная архитектура. В исследовании [31] предлагается совместная бустинговая система (Collaboratively Boosting Framework / CBF) для итеративного объединения модуля глубокого обучения и модуля онтологических рассуждений. Модуль глубокого обучения использует архитектуру глубокой нейронной сети для семантической сегментации (Deep Semantic Segmentation Network / DSSN) и принимает на вход интеграцию исходного изображения и логически выведенные знания (каналы логического вывода). Кроме того, модуль онтологических рассуждений состоит из внутритаксономических и внетаксономических рассуждений. Более конкретно – внутритаксономические рассуждения непосредственно исправляют неправильные классификации модуля глубокого обучения на основе знаний о проблемной области, что является ключом к улучшению эффективности классификации. Внетаксономические рассуждения направлены на создание каналов логического вывода за пределами текущей таксономии для улучшения характеристик DSSN в исходном пространстве изображений. С одной стороны, пользуясь ссылочными каналами из модуля онтологических рассуждений, модуль глубокого обучения, использующий интеграцию исходного изображения и указанных выше каналов, может достичь лучших результатов классификации, чем при использовании только исходного изображения. С другой стороны, лучшие результаты классификации, полученные с помощью модуля глубокого обучения, еще больше повышают эффективность работы модуля онтологических рассуждений.

Авторы статьи [32] рассматривают проблемы (например, проблему прогнозирования свойств химических соединений) со следующими характеристиками: 1) данные естественным образом представлены в виде графов; 2) объем доступных данных обычно невелик; и 3) имеются значительные знания о проблемной области, обычно выраженные в некоторой символьной форме (правила, таксономии, ограничения и другие). В статье рассматриваются графовые нейронные сети (GNN), к которым применяется механизм «обогащения вершин» («vertex-enrichment»), а итоговый класс нейронных сетей называется VEGNN. В отличие от классических графов, в которых отношения связывают пары вершин, механизм обогащения позволяет связывать больше, чем просто пары вершин. Например, если молекулу представить в виде графа (с атомами в качестве вершин и ребром, обозначающим связь между парой вершин), то бензольное кольцо – это связь между шестью различными

вершинами, с некоторыми специфическими ограничениями на вершины и ребра. Представленные в статье результаты подтверждают следующие выводы: а) включение знаний о проблемной области путем обогащения вершин может значительно улучшить производительность GNN (производительность VEGNN значительно выше, чем GNN); б) включение специфических для домена отношений, построенных с помощью индуктивного логического программирования (Inductive Logic Programming / ILP), улучшает производительность VEGNN. В целом, полученные результаты свидетельствуют о том, что в GNN можно включить символьные знания о проблемной области, и что ILP может играть важную роль в обеспечении высокоуровневых отношений, которые нелегко обнаружить с помощью классических GNN.

В статье [33] описан основанный на правилах способ объединения нескольких искусственных нейронных сетей с символьными рассуждениями для работы с аннотированными картами местности (рисунок 6). Система управления транспортным средством на основе аннотированных карт отслеживает положение транспортного средства на карте и обновляет данные карт. Она предоставляет модулю «арбитр» символьную информацию о направлении, в котором нужно двигаться, чтобы следовать запланированному маршруту, и о местности, с которой автомобиль сталкивается в данный момент. Нейросетевые модули в основном задействованы для оценки обстановки и вождения транспортного средства.

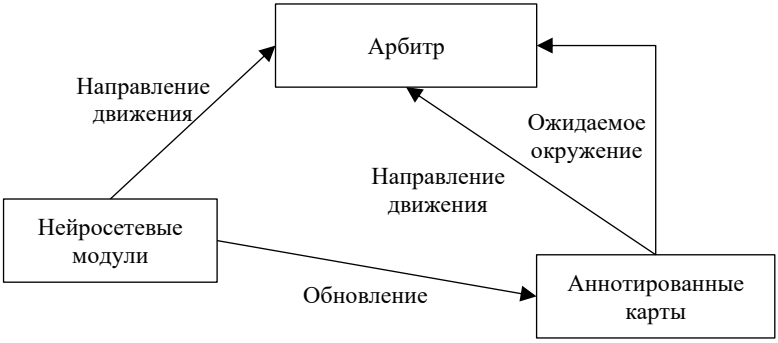

Рис. 6. Компоненты системы управления транспортным средством на основе аннотированных карт и их взаимодействие

В работе [34] предложена модель классификации функций цитирования, объединяющая онтологии со сверточными нейронными сетями. В модели онтологии используются для семантического представления характеристик автора и цитат. Это представление далее включено в нейросетевую модель для последующей классификации цитат.

В [35] описан схожий гибридный рекомендательный подход на основе онтологии и нейронной сети в области кино, который объединяет фильтрацию на основе контента и коллаборативную фильтрацию. Причем его модуль генерации рекомендаций может использовать их или по отдельности, или вместе. Показано, что подобный гибридный рекомендательный подход может решать традиционные проблемы рекомендующих систем, такие как извлечение признаков, предсказание интенсивности, разреженность матрицы и проблему холодного старта.

Еще одна похожая работа [36] ориентирована на классификацию патентов в области медицины. Исходными данными для модели машинного обучения являются ключевые слова, найденные в тексте патента, и соответствующие им фрагменты предварительно построенной онтологии проблемной области. В работе [37] выполняется кластеризация документов, основанная на том же принципе, в работе [38] – классификация воздействий лекарств, а в работе [39] – поиска и классификации отношений в текстах. Авторы работ показывают, что результаты обучения предложенных ими моделей превосходят существующие, а также требуют меньшего числа обучающих примеров.

Обратная схема используется в работе [40] – онтология используется на завершающем этапе классификации изображений. В данной работе сначала решается задача сегментации, результатом которой является набор обнаруженных на изображении объектов. Данные объекты соотносятся с концептами предварительно построенной онтологией, а затем выполняется подсчет наиболее вероятного класса, к которому может относиться рассматриваемое изображение. Подход позволил существенно уменьшить число обучающих примеров (например, для успешного распознавания баскетбольных и футбольных мячей авторы использовали набор данных всего из 15 изображений).

2.3.2. Жестко связанная архитектура. В работе [41] представлена методология использования символьных знаний в глубоком обучении, основанная на семантической функции потерь, которая устанавливает связь между векторами выходов нейронной

сети и логическими ограничениями. Эта функция потерь показывает, насколько близка нейронная сеть к удовлетворению ограничений на ее выходе. Экспериментальная оценка показывает, что она эффективно направляет обучаемую нейронную сеть на достижение (близких к лучшим) результатов в многоклассовой классификации. Более того, она значительно повышает способность нейронной сети предсказывать структурированные объекты, такие как ранжирования и пути. Такие дискретные понятия весьма сложны для обучения, и тесная интеграция глубокого обучения и методов символьных рассуждений позволяет существенно повысить его эффективность.

Авторы работы [42] предлагают архитектуру NeurASP, основанную на двух компонентах. Данный подход основан на ранее описанной идее [43], которая предлагает концепцию описания вероятностных логических моделей. Первый компонент – это нейронная сеть, которая выдает вероятности фактов, обрабатываемых вторым компонентом – машиной обработки правил / наборов ответов (rule (answer set) engine). Такая архитектура позволяет перенести нагрузку по обработке логических правил (символьных знаний) с нейронной сети на машинную обработку правил. В результате, модели машинного обучения, построенные на основе архитектуры NeurASP, демонстрируют очень быструю сходимость в процессе обучения, что позволяет обучать их на малых данных. Возможным недостатком данного подхода является то, что он ориентирован на логические знания и не может работать с алгебраическими выражениями (кроме самых простых).

В работе [44] представлены два подхода к решению проблемы абдукции (процесс поиска входных значений, которые приводят к вычислению определенного выходного значения) в нейронных сетях. В одном из них используется коннекционистская модальная логика и перевод положений Хорна в модальные положения для создания ансамбля нейронных сетей, который вычисляет абдуктивные объяснения по принципу «сверху вниз». Другой подход объединяет нейро-символьные системы и абдуктивное логическое программирование и предлагает нейронную архитектуру, которая выполняет более систематическое вычисление альтернативных абдуктивных объяснений «снизу вверх». Оба подхода используют стандартные архитектуры нейронных сетей, которые уже известны как высокоэффективные в практических приложениях для обучения. В отличие от других работ, данная работа ориентирована на усиление интеграции логического вывода и обучения так, чтобы нейронная сеть обеспечивала бы механизмы для когнитивных вычислений,

индуктивного обучения и гипотетических рассуждений, а логика обеспечивала бы строгость и возможность объяснения, облегчая взаимодействие с внешним миром. Авторы не смогли определить преимущество одного из подходов над другим.

2.3.3. Полностью интегрированная архитектура. В работе [45] представлена схема OnML, которая обучает интерпретируемую модель с помощью основанного на онтологии метода выборки, для объяснения агностических моделей предсказания (рисунок 7). В отличие от других алгоритмов, данный алгоритм учитывает контекстную корреляцию между словами, описанными в предметных онтологиях, для генерации семантических объяснений. Чтобы сузить пространство поиска объяснений, что является основной проблемой длинных и сложных текстовых данных, разработан обучаемый алгоритм формирования «якорей», позволяющий извлекать объяснения локально. Приведенные эксперименты на двух наборах данных показывают, что данный подход генерирует более точные и глубокие объяснения по сравнению с базовыми подходами.

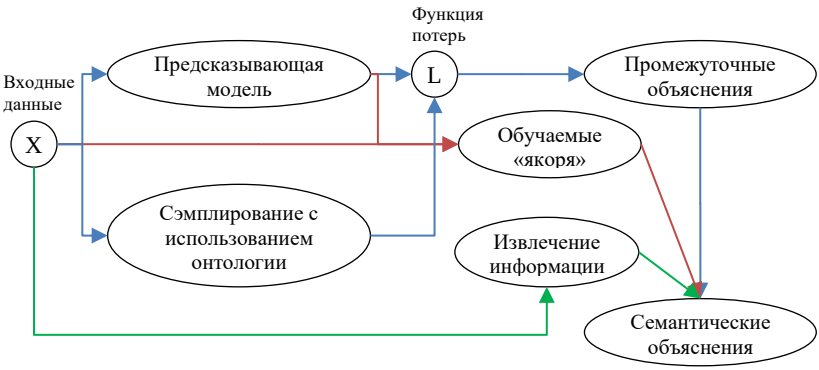

Рис. 7. Информационные потоки в схеме OnML

3. Использование онтологий для объяснения результатов нейронных сетей. Объяснимый искусственный интеллект является перспективным направлением исследований [46] и оказывается востребованным в различных прикладных областях – например, медицине [47], образовании [48]. Идет активная разработка методов объяснения нейронных сетей [1, 49] – значимым направлением при разработке методов объяснения является использование символьных структур, понятных человеку (например, правил [50], причинно-следственных связей [51]). В данной статье акцент также сделан

на методах обеспечения объяснимости за счет использования символьных структур, однако в качестве таких структур рассматриваются онтологии проблемной области.

Основными задачами обзора является идентификация существующих методов онтолого-ориентированного объяснения и выявление основных технических и методологических проблем, связанных с подобным объяснением.

В исследовании поставлены следующие вопросы:

1. Какие существуют постановки (разновидности) задачи онтолого-ориентированного объяснения?
2. Как онтологии используются для формирования объяснений?
3. Как влияет использование онтологий на «понимаемость» объяснений (способность пользователя интерпретировать предложенное объяснение и делать выводы на его основе)?
4. Как измеряется качество методов объяснения?
5. Влияет ли объяснимость на точность предсказаний нейронной сети?

Данный раздел структурирован в соответствии с ответами на поставленные вопросы.

Помимо этого, отмечается, какие именно онтологии чаще всего используются в подобных публикациях, и в каких прикладных областях делается больше всего попыток обеспечить онтолого-ориентированную объяснимость.

Что касается используемых онтологий, то в большинстве случаев это небольшие онтологии, сконструированные авторами. Иногда эти онтологии составлены на языке OWL, но в большинстве случаев описываются просто как набор высказываний дескрипционной логики или иерархии понятий (которая является лишь основой онтологии). Широко известные онтологии используются лишь в нескольких работах – это Gene Ontology – GO [52, 53], ICD-9 [54] и иерархия понятий WordNet [55].

Среди областей практических применений (вполне ожидаемо) лидирует медицина [52 – 54, 56 – 58]. На втором месте – финансы [56 – 58]. В значительной части публикаций рассматриваются достаточно условные примеры, не имеющие выраженной практической направленности.

В целом, позиционировать онтолого-ориентированные методы объяснения нейронных сетей среди всех прочих методов можно следующим образом. В подавляющем большинстве методов объяснения нейронных сетей под объяснением понимают

идентификацию входов и частей модели (например, нейронов), в наибольшей степени «ответственных» за результат. В контексте некоторых задач (например, предсказание фенотипа по экспрессии генов) подобные методы не позволяют получить понимаемое объяснение. Объяснение должно быть дополнено знаниями проблемной области [52]. В ряде публикаций отмечается, что многие методы объяснений предназначены, в первую очередь, для специалистов по ИИ, а не для конечных пользователей [2], и цель интеграции нейросетевых моделей и символьных знаний заключается именно в повышении интерпретируемости для конечного пользователя.

При этом нужно помнить, что любой подход к достижению объяснимости исходит из того, что на некотором уровне абстракции уже не требуется дальнейших объяснений [59]. В этом смысле, онтолого-ориентированные методы и «классические» (например, градиентные (GradCAM) или методы, основанные на окклюзиях), занимают разные ниши (выбирают различные уровни абстракции, при которых объяснения не нужны). В большинстве онтолого-ориентированных методов предполагается, что задача нейронной сети может быть разбита на две подзадачи – 1) переход от «сырых» наблюдений (данных) к понятиям (символам), 2) получение конечного результата посредством манипуляции с понятиями (символами). При этом объясняемой частью является именно символьная, а первая подзадача – переход к символам – считается не требующей объяснений. В то время как в классических (и более распространенных) методах наличие символьной составляющей вообще игнорируется, и объяснение строится на основе исходных признаков задачи (входов нейронной сети), что имеет два следствия. Во-первых, указанное допущение в определенной степени ограничивает область применения онтолого-ориентированных методов – для того, чтобы их применение было целесообразным должна быть возможность действительно сформировать решение задачи как символьной (в частности, должен существовать адекватный набор символов, чтобы целевая переменная могла быть с их помощью выражена). Во-вторых, онтолого-ориентированные и классические методы могут достаточно эффективно дополнять друг друга, формируя объяснения на разных уровнях абстракции.

3.1. Постановки задачи онтолого-ориентированного объяснения нейронных сетей. В области объяснимых моделей машинного обучения (не обязательно онтолого-ориентированных) принято подразделять методы по двум критериям:

1. Характер взаимосвязи процессов получения результата модели и получения объяснения.

2. Характер объяснения – связано ли оно с одним образцом или общей логикой работы модели машинного обучения.

Оба эти критерия остаются вполне валидными и применительно к онтолого-ориентированным объяснениям – практически в каждом классе методов, порождаемом подобной классификацией, к настоящему моменту предложены и способы использования онтологий. Остановимся подробнее на этих классах.

По характеру взаимосвязи процесса получения результата модели и процесса получения объяснения разделяют самообъяснимые модели и так называемые post-hoc (ретроспективные) методы. В первом случае объяснение порождается одновременно с результатом модели и, в каком-то смысле, неотъемлемо от него. Для описания этого класса моделей иногда используют метафору glass-box («стеклянный ящик»), имея в виду, что логика принятия решения видна сквозь его «стенки». Во втором случае сам способ получения результата является менее понимаемым, и для его интерпретации требуется использование какой-то дополнительной инфраструктуры (модели, алгоритма и т.п.). Для описания подобных моделей часто используется метафора black-box («черный ящик»), а ретроспективный метод объяснения дает наблюдателю (всегда косвенное) представление о том, как такая модель может работать. Существует довольно популярная точка зрения [60], что для ответственных применений плохи не только традиционные «черные ящики», но и их ретроспективные объяснения, и настоящее доверие вызывают только самообъяснимые модели. В этом смысле использование «черных ящиков» и их ретроспективных объяснений является в какой-то мере компромиссом.

По характеру объяснения методы принято делить на локальные и глобальные. Глобальные методы позволяют объяснить всю логику работу модели (например, представить ее в виде понятного человеку алгоритма), а локальные – лишь объяснить, почему был получен тот или иной результат для конкретного образца (например, какие признаки внесли в это наибольший вклад).

3.1.1. Самообъяснимые (self-explainable) модели. Основной техникой, используемой в различных самообъяснимых нейросетевых моделях, является явное определение соответствия между (каждым) концептом онтологии и каким-либо элементом (либо целым фрагментом) нейронной сети.

При этом нейроны, соответствующие концептам, могут быть как внутренними узлами сети (например, в [52] нейронная сеть имеет иерархическую структуру, определяемую структурой онтологии), так и – этот вариант встречается наиболее часто – образовывать последний слой нейронной сети [53, 55, 59, 61, 62], при этом структура сети напоминает структуру, используемую при многозначной (multi-label) классификации.

Основным механизмом установления соответствия между нейронами и концептами является функция потерь. В простейшем случае, когда все нейроны, связанные с концептами, размещаются в последнем слое, функция потерь может быть обыкновенной (например, бинарная энтропия). Здесь соответствие достигается благодаря тому, что каждый образец обучающей выборки должен быть размечен всеми концептами, к которым он относится, соответственно, предсказание каждого из нейронов-концептов непосредственно учитываются в расчете значения функции потерь. Более сложный вариант – дополнительно отразить в функции потерь какие-то из логических ограничений, описанных в онтологии. Так, в [52] функция потерь содержит штраф на силу таких связей между нейронами (коэффициентов), которые не описаны в онтологии.

После идентификации концептов сам результат модели формируется с помощью какого-либо объяснимого метода (например, логистической регрессии [53, 55, 61, 62]) или машиной логического вывода [59]. Таким образом, полученный результат оказывается трактуем уже в терминах онтологии по способу своего получения. Однако связь с исходными признаками, как правило, оказывается за пределами уровня абстракции, рассматриваемого соответствующим методом (вводное замечание про различные уровни объяснений).

Схожие идеи могут использоваться и для дополнительной верификации предсказаний «черного ящика». Так, в [63] предлагается концепция объяснимой системы распознавания, в которой используется своего рода «двойная проверка» – помимо целевой модели классификации обучаются модели для каждого из важных (по определению) свойств объекта (определенных в онтологии), и если все предсказания оказываются согласованными, то пользователь может получить и сам результат классификации и его объяснение, иначе – предупреждение о несогласованности.

Менее распространенной техникой является модификация исходного признакового пространства, попытка сделать его более нагруженным семантически и за счет этого сделать стандартные

методы объяснения более понятными неспециалисту. Подобное решение предлагается, например, в [64].

3.1.2. Ретроспективные (post-hoc) методы объяснения. Основным подходом к построению ретроспективных онтолого-ориентированных методов объяснения является построение объяснимой аппроксимации (приближения) модели «черного ящика», то есть такой модели, которая будет давать приблизительно такие же результаты, что и модель «черный ящик». При этом аппроксимация может быть как глобальной (для любой области признакового пространства), так и локальной (для «окрестности» определенного образца).

В рамках этого подхода было найдено два метода (что характерно, в обоих методах в роли аппроксимирующей модели выступает дерево решений). В первом [56 – 58] – онтология используется при построении дерева решений для повышения его интерпретируемости. При оценке потенциальных разбиений (в ходе построения дерева решений) авторы метода учитывают общность концепта по его положению в онтологии, стараясь сначала использовать более общие. Во втором [54] – онтологии используются для формирования синтетической обучающей выборки для обучения аппроксиматора в окрестности объясняемого образца. В частности, онтологии позволяют: а) отобрать семантически близкие образцы (истории болезни, в которых есть близкие коды болезней), б) сформировать синтетические истории, вычеркивая семантически близкие состояния.

Одной из серьезных проблем, связанных с ретроспективным объяснением с помощью аппроксиматоров, является то, что аппроксиматор, в сущности, лишь косвенно связан с объясняемой моделью.

Принципиально иным подходом к онтолого-ориентированному ретроспективному объяснению нейронных сетей является подход, идея которого предложена в [65] и развивается в [66, 67]. В [65, 66] показано, что внутренние представления нейронной сети, обученной решению определенной задачи, при определенных условиях могут быть сопоставлены с концептами онтологии проблемной области. Это позволяет, в частности, определить набор концептов онтологии, связанных с образцом, обрабатываемым сетью, и сформировать объяснение как онтологический вывод на базе известного набора концептов. Одной из проблем здесь является поиск множества нейронов сети, активации которых наиболее информативны для

извлечения того или иного концепта. Алгоритмы для решения этой задачи предложены в работах [65] и [67].

3.2. Использование онтологий для формирования объяснений. Выявлено несколько способов использования онтологий для формирования самих объяснений.

Возможно, наиболее очевидный способ – это использование логических машин вывода. Сначала с помощью какого-либо способа (выходящего за рамки формируемых объяснений) выясняется набор концептов, связанных с образцом. Затем, считая образец анонимным индивидом онтологии, формируется набор высказываний, связывающих его с найденными концептами. Наконец, инициируется логический вывод, позволяющий на основе полученного таким образом описания образца и определений концептов, описанных в онтологии, получить факт принадлежности концепта целевому классу (классу, на определение которого обучена объясняемая модель). Совокупность аксиом, задействованных при этом, и правил вывода образуют объяснение. Такое объяснение, опирающееся на формальную логику, предлагается в работе [65]. Для небольших онтологий такая схема может быть существенно ускорена посредством подготовки специальных тензоров и производиться на GPU [59]. Следует заметить, что выделение связанных концептов зачастую является вероятностным, что может учитываться и машиной вывода. Объяснение также может формироваться как обобщение с помощью индуктивного логического программирования [68]. В [63] также используется вариация этого способа, только производится сравнение целевого класса, полученного с помощью вывода, и класса, полученного с помощью оригинальной модели (которую нужно объяснить) – если результаты различаются, то пользователь уведомляется о низкой уверенности системы классификации.

В ряде рассмотренных методов онтологии используются косвенно. Они позволяют получить набор некоторых признаков, хорошо понятных пользователю, а значит, делают любые методы объяснения, показывающие влияние признаков на итоговый результат модели, более понятными конечному пользователю. Сюда, например, можно отнести множество самообъяснимых моделей, в которых выделяется слой концептов, над которым определяется логистическая регрессия [55, 62], и некоторые другие [52, 69]. Поскольку многие самообъяснимые модели предполагают установление однозначной связи между нейроном и концептом, то методы, показывающие вклад нейрона в предсказание (например, Layer-wise relevance propagation), становятся более содержательными [52]. Похожий принцип

используется и в работе [64], где предлагается создание специальных легко интерпретируемых семантических признаков (атрибутов) классифицируемых объектов.

Еще более опосредованно онтологии используются в методах [56 – 58]. Здесь объяснения следуют из дерева решений (определяются путем в дереве от корня до листа), но узлы этого дерева упорядочены в соответствии с общностью терминов, определяемой онтологией, что потенциально делает его более интерпретируемым.

В [54] онтологии не используются для объяснений. Они используются для формирования синтетической обучающей выборки для обучения аппроксиматора в окрестности объясняемого примера.

Основные подходы к постановке задачи онтолого-ориентированного нейро-символического интеллекта и соответствующие методы формирования объяснений сведены в таблице 1.

Таблица 1. Основные классы методов онтолого-ориентированного объяснения нейросетевых моделей

Класс	Механизм объяснения	Работы
Самообъяснимая модель	Концепты соответствуют внутренним узлам сети	[52]
	Концепты соответствуют последнему слою, логистическая регрессия	[53, 55, 61, 62]
	Концепты соответствуют последнему слою, логический вывод	[59]
	Семантические входные признаки	[64]
Ретроспективный	Аппроксимация сети деревом решений	[56 – 58]
	Сети отображения (активаций нейронной сети в концепты)	[65, 67]

3.3. Влияние онтологий на понимаемость объяснений. Несмотря на то, что «понимаемость» объяснений является одним из основополагающих вопросов, в подавляющем большинстве работ не приводится подобной оценки, авторы ограничиваются лишь общим тезисом о том, что использование понятных эксперту концептов делает объяснение более понимаемым. Поскольку никакой численной оценки обычно не проводится, то, очевидно, не сложилось и устоявшихся методик оценки подобного влияния.

Единственной работой (это цикл публикаций одних и тех же авторов, в которых излагается, в сущности, один и тот же алгоритм – TREPAN), где делается попытка экспериментально оценить

понимаемость онтолого-ориентированных объяснений, является [56 – 58]. Экспериментальное исследование основано на использовании наборов данных из области медицины и финансов. Испытуемым (не являющимся экспертами в области машинного обучения и анализа данных) после показа общего ролика о деревьях решений было предложено две задачи:

– классификация (демонстрируется объект и дерево решений, нужно осуществить классификацию, используя данное дерево);

– инспектирование (оценить истинность высказывания – например, «вы мужчина, ваш доход влияет на вероятность выдачи кредита»).

По задачам фиксировалась правильность ответа, уверенность, время, понятность дерева (по субъективной оценке испытуемого). Было показано, что построение деревьев решений с учетом онтологий действительно позволяет упростить работу с ними (и сделать их более понятными).

По всей видимости, подобный подход к оценке понимаемости, в рамках которого оценивается способность людей интерпретировать те или иные виды объяснений, является достаточно перспективным.

Достаточно опосредованной, но тем не менее потенциально полезной оценкой (применительно к некоторым видам объяснений) является структурная сложность. В частности, она оказывается хорошо применима к методам ретроспективного объяснения с помощью аппроксимации сети деревом решений [54].

3.4. Оценка качества метода объяснений. Помимо оценки понимаемости получаемых объяснений, которая, как отмечено выше, практически никем не проводится, методы формирования объяснений могут оцениваться по следующим характеристикам.

Для методов аппроксимации:

– достоверность (fidelity) – точность воспроизведения объяснимой моделью «черного ящика» на синтетических примерах, на которых обучается объясняющая модель (аппроксимация);

– попадание (hit) – совпадение предсказаний объяснимой аппроксимации и «черного ящика».

Для ретроспективных объяснений с помощью извлечения концептов важным оказывается точность выделения концептов. Соответственно, в [65] объясняющая модель оценивается по тому, использует ли объяснение только релевантные образцу концепты.

3.5. Влияние объяснимости на точность предсказаний нейронной сети. Ретроспективные методы объяснения не оказывают влияния на точность предсказаний нейронной сети (в этом

заключается один из существенных плюсов, связанных с их использованием).

Что касается самообъяснимых сетей, то ситуация не столь однозначна. В большинстве исследований привнесение в модель объяснимости приводило к небольшой деградации предсказательной силы (точности) [52, 55 – 58, 61, 63]. Насколько критична эта деградация по сравнению с приобретенным пониманием модели, очевидно, зависит от задач.

Тем не менее, в определенных случаях, онтолого-ориентированная объяснимость не вредит качеству предсказаний. Так, авторы [54] указывают, что качество предсказаний онтолого-ориентированной объяснимой модели оказалось хуже, чем неинтерпретируемой, однако лучше, чем интерпретируемой без использования онтологий. В [62] отмечается, что качество предсказаний модели такое же, как и у неинтерпретируемой модели, но обучение требует меньшего числа примеров. Наконец, в [53] качество оказалось даже лучше, чем в неинтерпретируемых моделях. Подобные внушающие оптимизм результаты связываются с тем, что в некоторых случаях объяснимость сопровождается переносом знаний, закодированных в виде онтологии, в нейронную сеть. Такой перенос способен как ускорить обучение, так и позволить добиться более качественных предсказаний.

4. Заключение. В обзоре представлены подходы к созданию интеллектуальных систем с использованием методов машинного обучения, архитектур, объединяющих символьные знания и субсимвольные (нейросетевые) знания, а также моделей представления символьных знаний в нейронных сетях с использованием онтологий.

Онтологическое моделирование применяется на разных уровнях. Так, выделяют базовые онтологии (или онтологии верхнего уровня), онтологии задач, онтологии проблемных областей. На настоящий момент, подавляющее большинство методов онтолого-ориентированного нейро-символического интеллекта ориентировано на использование онтологий проблемных областей – они применяются как в качестве априорных знаний для повышения качества чисто нейросетевых решений, так и для проблемно-ориентированных объяснений. Поскольку рассмотренные методы используют, в первую очередь, структуру онтологии, определяемую языком, на котором реализована онтология (и соответствующей дескрипционной логикой), это связано, по всей видимости, с тем, что использование концептов онтологии проблемной области является наиболее востребованным

(для эксперта), нежели с какими-то ограничениями самих методов. Так, онтология задач может оказаться востребованной при построении ИИ-агентов для классификации или формирования структур процессов.

Подходы к созданию интеллектуальных систем с использованием методов машинного обучения в большей степени ориентированы на различные макро-сценарии использования методов машинного обучения в интеллектуальных системах. Показано, что «нейросетевое приближение» в основном ориентировано на использование методов машинного обучения для аппроксимации имеющихся знаний с помощью нейронных сетей; «нейросетевое рассуждение» – на реализацию методов логического вывода и доказательства теорем; «интроспекция» – на обучение искусственных нейронных сетей в процессе «наблюдения» за другими моделями ИИ; а «интегрированное получение знаний» – на обучение искусственных нейронных сетей в процессе «наблюдения» за экспертами, что в наибольшей степени соответствует процессам, происходящим при коллаборативной поддержке принятия решений.

Архитектуры, объединяющие символьные знания и субсимвольные (нейросетевые) знания, в свою очередь, ориентированы именно на объединение указанных знаний в единые комбинированные модели не зависимо от сценариев их использования. На настоящий момент предложено достаточно много архитектур представления символьных знаний в нейронных сетях. Рассмотренные работы показывают более высокую эффективность моделей машинного обучения, комбинирующих символьные и субсимвольные (нейросетевые) знания, по сравнению с «классическими» нейросетевыми моделями по показателям точности. Кроме того, как правило, обучение таких комбинированных моделей требует существенно меньших объемов обучающих данных, что позволяет говорить о возможности их обучения на малых данных. Большинство из рассмотренных архитектур комбинированных моделей и моделей представления символьных знаний в них равно эффективны и их выбор в первую очередь определяется решаемой конкретной прикладной задачей (таблица 2).

Так, в случае наличия отдельных онтологий, описывающих знания проблемной области, которые являются динамичными (развивающимися), целесообразно применение гибридной модульной архитектуры, разделяющей символьные и нейросетевые знания, поскольку в этом случае символьные знания могут корректироваться в некоторой степени, не затрагивая нейросетевые. В противном случае

(работа со статичными знаниями), унифицированная и трансформационная архитектуры выглядят более привлекательно в силу гибкости и многообразия способов внедрения символьных знаний в нейросетевые модели.

В системах коллаборативной поддержки принятия решений, как правило, целесообразно использование нескольких онтологий [70], охватывающих различные аспекты проблемы, что обуславливает предпочтительность применения именно гибридной модульной архитектуры. При этом, необходимость двустороннего обмена знаниями между человеком и ИИ подразумевает применение ее полностью интегрированного варианта. Вопрос о выборе модели представления знаний в данном случае становится менее актуальным, поскольку знания не встраиваются напрямую в нейросетевые модели.

Таблица 2. Преимущества и недостатки архитектур, объединяющие символьные знания и нейросетевые знания

Архитектуры, объединяющие символьные знания и нейросетевые знания	Преимущества	Недостатки
Унифицированная	– Многообразие способов внедрения символьных знаний в нейросетевые модели.	– Существенная зависимость символьных и субсимвольных знаний друг от друга.
Трансформационная	– Многообразие способов внедрения символьных знаний в нейросетевые модели. – Способность извлечения символьных знаний из субсимвольных.	– Ограниченная область использования.
Гибридная модульная	– Символьные и субсимвольные знания в значительной степени независимы друг от друга.	– Ограниченные механизмы взаимодействия символьных и субсимвольных знаний.

Основные рекомендации по выбору методов объяснения, сформулированные в результате обзора, сведены в таблице 3. Преимуществом самообъяснимых сетей, как правило, является то, что

в них процесс формирования объяснения напрямую связан с получением результата, а значит, объяснение в большей степени описывает логику принятия решения сетью. Однако в таких моделях зачастую накладываются определенные ограничения на структуру сети, что может снижать ее качество (точность получаемых результатов). Преимуществом же ретроспективных методов является то, что они потенциально могут применяться к любым сетям (в том числе и к тем, процесс обучения которых не контролируется), однако между реальной логикой работы сети и ее онтолого-ориентированным объяснением возможно расхождение.

Одной из наиболее серьезных исследовательских проблем, выявленных в ходе обзора, является отсутствие принятой и широко используемой методики для оценки понимаемости объяснений и, в частности, влияния использования онтологий на понимаемость. В связи с этим целесообразным представляется проведение исследований, направленных на разработку такой методики, что потенциально может оказать интеграционный эффект на рассматриваемую область.

Таблица 3. Преимущества и недостатки подходов к онтолого-ориентированному объяснению нейронных сетей

Подход к объяснению	Преимущества	Недостатки
Самообъяснимая модель	– Процесс формирования объяснения напрямую связан с получением результата, а значит, объяснение в большей степени описывает логику принятия решения сетью.	– Накладываются определенные ограничения на структуру сети, что может снижать ее качество (точность получаемых результатов).
Ретроспективный	– Потенциально могут применяться к любым сетям (в том числе и к тем, процесс обучения которых не контролируется).	– Возможно расхождение между реальной логикой работы сети и ее объяснением.

Что касается конкретных методов онтолого-ориентированного объяснения нейронных сетей, то наиболее перспективными являются следующие направления:

1. Ретроспективные объяснения, построенные не на аппроксимации, а на исследовании внутренних представлений,

порождаемых глубокими нейронными сетями. Потенциально они применимы к любым сетям, гарантированно не ведут к деградации предсказаний и могут помочь избежать «разрыва» между предсказаниями «черного ящика» и аппроксиматора, характерного для объяснений через аппроксимацию.

2. Самообъяснимые модели, использующие знания проблемной области, закодированные в форме онтологий. Подобные разновидности самообъяснимых моделей могут обеспечивать качество предсказаний не хуже, чем у неинтерпретируемых моделей, однако при этом обеспечивают объяснимость в «наилучшем» варианте – когда сам процесс формирования решения и его объяснение совпадают. Подобные эффекты характерны, например, для вариантов гибридной модульной архитектуры использования априорных знаний (с учетом того, что объяснимой является только та часть результата, которая связана с работой машин вывода).

В целом, методы онтолого-ориентированного нейро-символического интеллекта ориентированы на решение преимущественно задач классификации, в которых используемые онтологии задают адекватную понятийную базу (если набор концептов, предусматриваемых онтологией, оказывается недостаточным для выражения правила классификации, то и ценность онтолого-ориентированного решения, скорее всего, окажется невысокой). Таким образом, ограничением для подобных методов них является наличие качественных, проработанных онтологий, а особенно перспективным представляется применение подобных методов в областях, где развитию и использованию онтологий уделяется большое внимание (например, биология и медицина).

Среди наиболее перспективных направлений развития методов онтолого-ориентированного нейро-символического интеллекта можно выделить следующие. Во-первых, преодоление основных ограничений (связанных с наличием качественных онтологий) посредством создания комплексных автоматизированных методов, обеспечивающих как извлечение концептов (и наполнение онтологий), так и их использование – одним из путей здесь может быть конвергенция с мощными языковыми моделями и исследование внутренних представлений нейросетевых классификаторов. Во-вторых, расширение области применения на более сложные задачи (за пределами классификации) – например, создание методов формирования объектов на основе генеративно-состязательных сетей с ограничениями, задаваемыми с помощью онтологии.

Литература

1. Burkart N., Huber M.F. A survey on the explainability of supervised machine learning // J. Artif. Intell. Res. 2021. vol. 70. pp. 245–317.
2. Futia G., Vetrò A. On the integration of knowledge graphs into deep learning models for a more comprehensible AI-Three challenges for future research // Inf. 2020. vol. 11. no. 2. DOI: 10.3390/info11020122.
3. Smirnov A., Ponomarev A. Human-Machine Collective Intelligence Environment for Decision Support: Conceptual and Technological Design // 27th Conference of Open Innovation Association, FRUCT. 2020. pp. 330–336. DOI: 10.23919/FRUCT49677.2020.9211077.
4. Fernback J. Symbolic Interactionism in Communication // Communication. Oxford University Press, 2019.
5. Garcez A. d'Avila, Lamb L.C. Neurosymbolic AI: The 3rd Wave. 2020. 37 p. DOI: 10.48550/arXiv.2012.05876.
6. Радюш Д.В. Применение нейро-символьных моделей в разработке вопросно-ответных систем // XI конгресс молодых учёных. Санкт-Петербург, 2022. Т. 2. С. 122–126.
7. Каширин И.Ю. Нейронные сети, использующие модели знаний // Современные технологии в науке и образовании – СТНО-2021. 2021. С. 9–13.
8. Ultsch A. The Integration of Neural Networks with Symbolic Knowledge Processing // New Approaches in Classification and Data Analysis. 1994. pp. 445–454.
9. Picco G. et al. Neural Unification for Logic Reasoning over Natural Language. Findings of the Association for Computational Linguistics: EMNLP. 2021. pp. 3939–3950. DOI: 10.48550/arXiv.2109.08460.
10. Грибова В.В., Гельцер Б.И., Шахгельдян К.И., Петряева М.В., Шалфеева Е.А., Костерин В.В. Гибридная технология оценки рисков и прогнозирования в кардиологии // Врач и информационные технологии. 2022. № 3. С. 24–35. DOI: 10.25881/18110193_2022_3_24.
11. Wermter S., Sun R. An Overview of Hybrid Neural Systems // Lect. Notes Artif. Intell. Subseries Lect. Notes Comput. Sci. Springer. 2000. vol. 1778. pp. 1–13.
12. Garcez A. d'Avila et al. Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. 2019. vol. 6 no. 4. pp. 611–632. DOI: 10.48550/arXiv.1905.06088.
13. Tran S.N., d'Avila Garcez A.S. Deep Logic Networks: Inserting and Extracting Knowledge From Deep Belief Networks // IEEE Trans. Neural Networks Learn. Syst. 2018. vol. 29. no. 2. pp. 246–258.
14. Poon H., Domingos P. Sum-Product Networks: A New Deep Architecture. IEEE International Conference on Computer Vision Workshops (ICCV Workshops). 2012. DOI: 10.48550/arXiv.1202.3732.
15. Muggleton S. Inverse entailment and progol // New Gener. Comput. 1995. vol. 13. no. 3–4. pp. 245–286.
16. Dong H., Mao J., Lin T., Wang C., Li L., Zhou D. et al. Neural Logic Machines. International Conference on Learning Representations. 2019. DOI: 10.48550/arXiv.1904.11694.
17. Evans R., Grefenstette E. Learning Explanatory Rules from Noisy Data. Journal of Artificial Intelligence Research. 2017. vol. 61. pp. 1–64. DOI: 10.48550/arXiv.1711.04574.
18. Gori M. Machine Learning: A Constraint-Based Approach. Morgan Kaufmann, 2017. 580 p.
19. Garcez A.S. d'Avila, Lamb L.C. Reasoning about time and knowledge in neural-symbolic learning systems // NIPS'03: Proceedings of the 16th International Conference on Neural Information Processing Systems. 2003. pp. 921–928.

20. Borges R.V., d'Avila Garcez A., Lamb L.C. Learning and Representing Temporal Knowledge in Recurrent Networks // IEEE Trans. Neural Networks. 2011. vol. 22. no. 12. pp. 2409–2421.
21. Penning L. de et al. A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning // Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). 2011. pp. 1653–1658. DOI: 10.5591/978-1-57735-516-8/IJCAI11-278.
22. Palangi H. et al. Question-Answering with Grammatically-Interpretable Representations. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). 2017. pp. 5350–5357.
23. Fletcher J., Obradovi Z. Combining Prior Symbolic Knowledge and Constructive Neural Network Learning // Conn. Sci. 1993. vol. 5. no. 3–4. pp. 365–375.
24. Towel G.G., Shavlik J.W., Noordewier M.O. Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks // Eighth National Conference on Artifiical Inrelligence (AAAI). 1990. pp. 861–866.
25. Pitz D.W., Shavlik J.W. Dynamically adding symbolically meaningful nodes to knowledge-based neural networks // Knowledge-Based Syst. 1995. vol. 8. no. 6. pp. 301–311.
26. Arabshahi F., Singh S., Anandkumar A. Combining Symbolic Expressions and Black-box Function Evaluations in Neural Programs. 6th International Conference on Learning Representations. 2018. DOI: 10.48550/arXiv.1801.04342.
27. Xie Y. et al. Embedding Symbolic Knowledge into Deep Networks // Adv. Neural Inf. Process. Syst. 2019. no. 32.
28. Hu Z. et al. Harnessing Deep Neural Networks with Logic Rules. 2016. pp. 2410–2420. DOI: 10.48550/arXiv.1603.06318.
29. Prem E. et al. Concept support as a method for programming neural networks with symbolic knowledge // GWAI-92: Advances in Artificial Intelligence. Berlin/Heidelberg: Springer-Verlag. 1992. pp. 166–175.
30. Shavlik J.W. Combining symbolic and neural learning // Mach. Learn. 1994. vol. 14, no. 3. pp. 321–331.
31. Li Y., Ouyang S., Zhang Y. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation // Knowledge-Based Syst. 2022. vol. 243. pp. 108469.
32. Dash T., Srinivasan A., Vig L. Incorporating symbolic domain knowledge into graph neural networks // Mach. Learn. 2021. vol. 110. no 7. pp. 1609–1636.
33. Pomerleau D.A., Gowdy J., Thorpe C.E. Combining artificial neural networks and symbolic processing for autonomous robot guidance // Eng. Appl. Artif. Intell. 1991. vol. 4. no. 4. pp. 279–285.
34. Bakhti K. et al. Citation Function Classification Based on Ontologies and Convolutional Neural Networks // Commun. Comput. Inf. Sci. 2018. vol. 870. pp. 105–115.
35. Deng Y. et al. A Hybrid Movie Recommender Based on Ontology and Neural Networks // 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. IEEE, 2010. pp. 846–851.
36. Trappey A.J.C. et al. Ontology-based neural network for patent knowledge management in design collaboration // Int. J. Prod. Res. 2013. vol. 51. no. 7. pp. 1992–2005.
37. Hung C., Wermter S. Neural Network Based Document Clustering Using WordNet Ontologies // Int. J. Hybrid Intell. Syst. 2005. vol. 1. no. 3–4. pp. 127–142.

38. Hinnerichs T., Hoehndorf R. DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions / ed. Wren J. // Bioinformatics. 2021. vol. 37. no. 24. pp. 4835–4843.
39. Lamurias A. et al. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies // BMC Bioinformatics. 2019. vol. 20. no. 1. pp. 10.
40. Breen C., Khan L., Ponnusamy A. Image classification using neural networks and ontologies // Proceedings. 13th International Workshop on Database and Expert Systems Applications. IEEE Comput. Soc, 2002. pp. 98–102.
41. Xu J. et al. A Semantic Loss Function for Deep Learning with Symbolic Knowledge // Proc. Mach. Learn. Res. 2018. vol. 80. pp. 5502–5511.
42. Yang Z., Ishay A., Lee J. NeurASP: Embracing Neural Networks into Answer Set Programming // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020. pp. 1755–1762.
43. Lee J., Wang Y. Weighted Rules under the Stable Model Semantics // Proceedings, Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016). 2016. pp. 145–154.
44. Garcez A.S. d'Avila et al. Abductive reasoning in neural-symbolic systems // Topoi. 2007. vol. 26. no. 1. pp. 37–49.
45. Lai P. et al. Ontology-based Interpretable Machine Learning for Textual Data // 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020. pp. 1–10.
46. Аверкин А.Н. Объяснимый искусственный интеллект: итоги и перспективы // Интегрированные модели и мягкие вычисления в искусственном интеллекте (ИММВ-2021). Сборник научных трудов X-й Международной научно-технической конференции. 2021. С. 153–174.
47. Карпов О.Э., Андриков Д.А., Максименко В.А., Храмов А.Е. Прозрачный искусственный интеллект для медицины // Врач и информационные технологии. 2022. № 2. С. 4–11. DOI: 10.25881/18110193_2022_2_4.
48. Захарова И.Г., Воробьева М.С., Боганюк Ю.В. Сопровождение индивидуальных образовательных траекторий на основе концепции объяснимого искусственного интеллекта // Образование и наука. 2022. Т. 24. № 1. pp. 163–190.
49. Шевская Н.В. Объяснимый искусственный интеллект и методы интерпретации результатов // Моделирование, оптимизация и информационные технологии. 2021. Т. 9. № 2(33). pp. 22.
50. Аверкин А.Н., Ярушев С.А. Обзор исследований в области разработки методов извлечения правил из искусственных нейронных сетей // Известия Российской академии наук. Теория и системы управления. 2021. № 6. С. 106–121.
51. Шевская Н.В., Охримук Е.С., Попов Н.В. Причинно-следственные связи в объяснимом искусственном интеллекте // Международная конференция по мягким вычислениям и измерениям. 2022. С. 170–173.
52. Bourgeais V. et al. Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data // BMC Bioinformatics. BioMed Central, 2021. vol. 22. pp. 1–24.
53. Ma T., Zhang A. Incorporating Biological Knowledge with Factor Graph Neural Network for Interpretable Deep Learning. 2019. DOI: 10.48550/arXiv.1906.00537.
54. Panigutti C., Perotti A., Pedreschi D. Doctor XAI An ontology-based approach to black-box sequential data classification explanations // FAT* 2020 – Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. pp. 629–639.
55. Daniels Z.A. et al. A framework for explainable deep neural models using external knowledge graphs / Ed. Pham T., Solomon L., Rainey K. // Proc. SPIE 11413,

Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IISPIE, 2020. pp. 73.
56. Confalonieri R. et al. An Ontology-based Approach to Explaining Artificial Neural Networks. 2019.
57. Confalonieri R. et al. Trepan reloaded: A knowledge-driven approach to explaining black-box models // Front. Artif. Intell. Appl. 2020. vol. 325. pp. 2457–2464.
58. Confalonieri R. et al. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models // Artif. Intell. Elsevier, 2021. vol. 296. pp. 103471.
59. Bourguin G. et al. Towards Ontologically Explainable Classifiers. Artificial Neural Networks and Machine Learning – ICANN 2021. 2021. pp. 472–484. DOI: ff10.1007/978-3-030-86340-1_38.
60. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nat. Mach. Intell. 2019. vol. 1. no. 5. pp. 206–215.
61. Voogd J. et al. Using Relational Concept Networks for Explainable Decision Support // 3rd IFIP Cross Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE). 2019. pp. 78–93. DOI: 10.1007/978-3-030-29726-8_6.
62. Fong A.C.M., Hong G. Ontology-Powered Hybrid Extensional-Intensional Learning // Proceedings of the 2019 International Conference on Information Technology and Computer Communications (ITCC2019). New York, USA: ACM Press, 2019. pp. 18–23.
63. Bellucci M. et al. Ontologies to build a predictive architecture to classify and explain // DeepOntoNLP Workshop @ESWC 2022. 2022.
64. Martin T. et al. Bridging the gap between an ontology and deep neural models by pattern mining // The Joint Ontology Workshops, JOWO. 2020. vol. 2708.
65. De Sousa Ribeiro M., Leite J. Aligning Artificial Neural Networks and Ontologies towards Explainable AI // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 6. pp. 4932–4940.
66. Agafonov A., Ponomarev A. An Experiment on Localization of Ontology Concepts in Deep Convolutional Neural Networks // The 11th International Symposium on Information and Communication Technology. NY, USA: ACM, 2022. pp. 82–87.
67. Ponomarev A., Agafonov A. Ontology Concept Extraction Algorithm for Deep Neural Networks // 2022 32nd Conference of Open Innovations Association (FRUCT). IEEE, 2022. pp. 221–226.
68. Sarker M.K. et al. Wikipedia Knowledge Graph for Explainable AI // KGSWC 2020, CCIS 1232. 2020. pp. 72–87.
69. Abbass H.A. et al. Machine Education: Designing semantically ordered and ontologically guided modular neural networks // IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2019. pp. 948–955.
70. Smirnov A. et al. Multi-aspect Ontology for Interoperability in Human-machine Collective Intelligence Systems for Decision Support // Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. SCITEPRESS – Science and Technology Publications, 2019. pp. 458–465.

Шилов Николай Германович — канд. техн. наук, доцент, старший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: искусственный интеллект, управление знаниями, управление онтологиями, моделирование и конфигурирование сложных систем, машинное обучение. Число научных публикаций — 300. nick@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Пономарев Андрей Васильевич — канд. техн. наук, доцент, старший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: коллективный интеллект, крауд-вычисления, рекомендательные системы, машинное обучение. Число научных публикаций — 70. ponomarev@iias.spb.su; 14 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Смирнов Александр Викторович — д-р техн. наук, профессор, главный научный сотрудник, руководитель лаборатории, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: системы поддержки принятия решений, интеллектуальные системы, интеллектуальное управление конфигурациями виртуальных и сетевых организаций, логистика знаний. Число научных публикаций — 400. smir@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 22-11-00214).

DOI 10.15622/ia.22.3.4

N. SHILOV, A. PONOMAREV, A. SMIRNOV

THE ANALYSIS OF ONTOLOGY-BASED NEURO-SYMBOLIC INTELLIGENCE METHODS FOR COLLABORATIVE DECISION SUPPORT

Shilov N., Ponomarev A., Smirnov A. **The Analysis of Ontology-Based Neuro-Symbolic Intelligence Methods for Collaborative Decision Support.**

Abstract. The neural network approach to AI, which has become especially widespread in the last decade, has two significant limitations – training of a neural network, as a rule, requires a very large number of samples (not always available), and the resulting models often are not well interpretable, which can reduce their credibility. The use of symbols as the basis of collaborative processes, on the one hand, and the proliferation of neural network AI, on the other hand, necessitate the synthesis of neural network and symbolic paradigms in relation to the creation of collaborative decision support systems. The article presents the results of an analytical review in the field of ontology-oriented neuro-symbolic artificial intelligence with an emphasis on solving problems of knowledge exchange during collaborative decision support. Specifically, the review attempts to answer two questions: 1. how symbolic knowledge, represented as an ontology, can be used to improve AI agents operating on the basis of neural networks (knowledge transfer from a person to AI agents); 2. how symbolic knowledge, represented as an ontology, can be used to interpret decisions made by AI agents and explain these decisions (transfer of knowledge from an AI agent to a person). As a result of the review, recommendations were formulated on the choice of methods for introducing symbolic knowledge into neural network models, and promising areas of ontology-oriented methods for explaining neural networks were identified.

Keywords: neuro-symbolic AI, domain knowledge, machine learning, deep learning, explainable AI, XAI, ontology.

References

1. Burkart N., Huber M.F. A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. 2021. vol. 70. pp. 245–317.
2. Futia G., Vetrò A. On the integration of knowledge graphs into deep learning models for a more comprehensible AI-Three challenges for future research. Inf. 2020. vol. 11. no. 2. DOI: 10.3390/info11020122.
3. Smirnov A., Ponomarev A. Human-Machine Collective Intelligence Environment for Decision Support: Conceptual and Technological Design. Conference of Open Innovation Association, FRUCT. 2020. pp. 330–336. DOI: 10.23919/FRUCT49677.2020.9211077.
4. Fernback J. Symbolic Interactionism in Communication. Communication. Oxford University Press, 2019.
5. Garcez A. d'Avila, Lamb L.C. Neurosymbolic AI: The 3rd Wave. 2020. 37 p. DOI: 10.48550/arXiv.2012.05876.
6. Radyush D.V. [Application of neuro-symbolic models in question-answer systems] XI kongress molodyh uchyonyh [The 11th congress of young scientists]. Sankt-Peterburg, 2022. vol. 2. pp. 122–126. (in Russ.).
7. Kashirin I.Yu. [Neural networks using knowledge models] Sovremennye tekhnologii v nauke i obrazovanii – STNO-2021 [Modern technologies in science and education]. 2021. pp. 9–13. (in Russ.).

8. Ultsch A. The Integration of Neural Networks with Symbolic Knowledge Processing. New Approaches in Classification and Data Analysis. 1994. pp. 445–454.
9. Picco G. et al. Neural Unification for Logic Reasoning over Natural Language. Findings of the Association for Computational Linguistics: EMNLP. 2021. pp. 3939–3950. DOI: 10.48550/arXiv.2109.08460.
10. Gribova V.V. et al. [Hybrid technology of risk assessment and forecasting in cardiology]. Vrach i informacionnye tekhnologii – Doctor and information technologies]. 2022. no. 3. pp. 24–35. (in Russ.).
11. Wermter S., Sun R. An Overview of Hybrid Neural Systems. Lect. Notes Artif. Intell. Subseries Lect. Notes Comput. Sci. Springer. 2000. vol. 1778. pp. 1–13.
12. Garcez A. d'Avila et al. Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. 2019. vol. 6 no. 4. pp. 611–632. DOI: 10.48550/arXiv.1905.06088.
13. Tran S.N., d'Avila Garcez A.S. Deep Logic Networks: Inserting and Extracting Knowledge From Deep Belief Networks. IEEE Trans. Neural Networks Learn. Syst. 2018. vol. 29. no. 2. pp. 246–258.
14. Poon H., Domingos P. Sum-Product Networks: A New Deep Architecture. IEEE International Conference on Computer Vision Workshops (ICCV Workshops). 2012. DOI: 10.48550/arXiv.1202.3732.
15. Muggleton S. Inverse entailment and progol. New Gener. Comput. 1995. vol. 13. no. 3–4. pp. 245–286.
16. Dong H., Mao J., Lin T., Wang C., Li L., Zhou D. et al. Neural Logic Machines. International Conference on Learning Representations. 2019. DOI: 10.48550/arXiv.1904.11694.
17. Evans R., Grefenstette E. Learning Explanatory Rules from Noisy Data. Journal of Artificial Intelligence Research. 2017. vol. 61. pp. 1–64. DOI: 10.48550/arXiv.1711.04574.
18. Gori M. Machine Learning: A Constraint-Based Approach. Morgan Kaufmann, 2017. 580 p.
19. Garcez A.S. d'Avila, Lamb L.C. Reasoning about time and knowledge in neural-symbolic learning systems. NIPS'03: Proceedings of the 16th International Conference on Neural Information Processing Systems. 2003. pp. 921–928.
20. Borges R.V., d'Avila Garcez A., Lamb L.C. Learning and Representing Temporal Knowledge in Recurrent Networks. IEEE Trans. Neural Networks. 2011. vol. 22. no. 12. pp. 2409–2421.
21. Penning L. de et al. A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). 2011. pp. 1653–1658. DOI: 10.5591/978-1-57735-516-8/IJCAI11-278.
22. Palangi H. et al. Question-Answering with Grammatically-Interpretable Representations. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). 2017. pp. 5350–5357.
23. Fletcher J., Obradovi Z. Combining Prior Symbolic Knowledge and Constructive Neural Network Learning. Conn. Sci. 1993. vol. 5. no. 3–4. pp. 365–375.
24. Towel G.G., Shavlik J.W., Noordewier M.O. Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks. Eighth National Conference on Artifiical Inrelligence (AAAI). 1990. pp. 861–866.
25. Pitz D.W., Shavlik J.W. Dynamically adding symbolically meaningful nodes to knowledge-based neural networks. Knowledge-Based Syst. 1995. vol. 8. no. 6. pp. 301–311.

26. Arabshahi F., Singh S., Anandkumar A. Combining Symbolic Expressions and Black-box Function Evaluations in Neural Programs. 6th International Conference on Learning Representations. 2018. DOI: 10.48550/arXiv.1801.04342.
27. Xie Y. et al. Embedding Symbolic Knowledge into Deep Networks. Adv. Neural Inf. Process. Syst. 2019. no. 32.
28. Hu Z. et al. Harnessing Deep Neural Networks with Logic Rules. 2016. pp. 2410–2420. DOI: 10.48550/arXiv.1603.06318.
29. Prem E. et al. Concept support as a method for programming neural networks with symbolic knowledge. GWAI-92: Advances in Artificial Intelligence. Berlin/Heidelberg: Springer-Verlag. 1992. pp. 166–175.
30. Shavlik J.W. Combining symbolic and neural learning. Mach. Learn. 1994. vol. 14. no. 3. pp. 321–331.
31. Li Y., Ouyang S., Zhang Y. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. Knowledge-Based Syst. 2022. vol. 243. pp. 108469.
32. Dash T., Srinivasan A., Vig L. Incorporating symbolic domain knowledge into graph neural networks. Mach. Learn. 2021. vol. 110. no 7. pp. 1609–1636.
33. Pomerleau D.A., Gowdy J., Thorpe C.E. Combining artificial neural networks and symbolic processing for autonomous robot guidance. Eng. Appl. Artif. Intell. 1991. vol. 4. no. 4. pp. 279–285.
34. Bakhti K. et al. Citation Function Classification Based on Ontologies and Convolutional Neural Networks. Commun. Comput. Inf. Sci. 2018. vol. 870. pp. 105–115.
35. Deng Y. et al. A Hybrid Movie Recommender Based on Ontology and Neural Networks. 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. IEEE, 2010. pp. 846–851.
36. Trappey A.J.C. et al. Ontology-based neural network for patent knowledge management in design collaboration. Int. J. Prod. Res. 2013. vol. 51. no. 7. pp. 1992–2005.
37. Hung C., Wermter S. Neural Network Based Document Clustering Using WordNet Ontologies. Int. J. Hybrid Intell. Syst. 2005. vol. 1. no. 3–4. pp. 127–142.
38. Hinnerichs T., Hoehndorf R. DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug–target interactions. Bioinformatics. 2021. vol. 37. no. 24. pp. 4835–4843.
39. Lamurias A. et al. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics. 2019. vol. 20. no. 1. pp. 10.
40. Breen C., Khan L., Ponnusamy A. Image classification using neural networks and ontologies. Proceedings. 13th International Workshop on Database and Expert Systems Applications. IEEE Comput. Soc, 2002. pp. 98–102.
41. Xu J. et al. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. Proc. Mach. Learn. Res. 2018. vol. 80. pp. 5502–5511.
42. Yang Z., Ishay A., Lee J. NeurASP: Embracing Neural Networks into Answer Set Programming. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020. pp. 1755–1762.
43. Lee J., Wang Y. Weighted Rules under the Stable Model Semantics. Proceedings, Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016). 2016. pp. 145–154.
44. Garcez A.S. d'Avila et al. Abductive reasoning in neural-symbolic systems. Topoi. 2007. vol. 26. no. 1. pp. 37–49.

45. Lai P. et al. Ontology-based Interpretable Machine Learning for Textual Data. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020. pp. 1–10.
46. Averkin A.N. [Explainable AI: current results and future perspectives] Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte (IMMV-2021): Sbornik nauchnyh trudov X-j Mezhdunarodnoj nauchno-tekhnicheskoj konferencii [Integrated systems and soft computing in AI: Conference proceedings]. 2021. pp. 153–174. (in Russ.).
47. Karpov O.E. et al. [Transparent AI for medicine]. Vrach i informacionnye tekhnologii – Doctor and intofmation technologies. 2022. no 2. pp. 4–11. (in Russ.).
48. Zaharova I.G., Vorob'eva M.S., Boganyuk Yu.V. [Individual educational trajectories based on AI]. Obrazovanie i nauka – Education and science. 2022. vol. 24. no. 1. pp. 163–190. (in Russ.).
49. Shevskaya N.V. [Explainable AI and results intepretation]. Modelirovanie, optimizaciya i informacionnye tekhnologii – Modelling, optimization and information technologies. 2021. vol. 9. no. 2(33). pp. 22. (in Russ.).
50. Averkin A.N., Yarushev S.A. [Review of methods for rule extraction from neural networks]. Izvestiya Rossijskoj akademii nauk. Teoriya i sistemy upravleniya – Proceedings of the Russian Academy of Sciences. Control theory and systems. 2021. no. 6. pp. 106–121. (in Russ.).
51. Shevskaya N.V., Ohrimuk E.S., Popov N.V. [Causual relationships in explainable AI] Mezhdunarodnaya konferenciya po myagkim vychisleniyam i izmereniyam [International conference on soft computing]. 2022. pp. 170–173. (in Russ.).
52. Bourgeais V. et al. Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data. BMC Bioinformatics. BioMed Central, 2021. vol. 22. pp. 1–24.
53. Ma T., Zhang A. Incorporating Biological Knowledge with Factor Graph Neural Network for Interpretable Deep Learning. 2019. DOI: 10.48550/arXiv.1906.00537.
54. Panigutti C., Perotti A., Pedreschi D. Doctor XAI An ontology-based approach to black-box sequential data classification explanations. FAT* 2020 – Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. pp. 629–639.
55. Daniels Z.A. et al. A framework for explainable deep neural models using external knowledge graphs. Proc. SPIE 11413, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IISPIE, 2020. pp. 73.
56. Confalonieri R. et al. An Ontology-based Approach to Explaining Artificial Neural Networks. 2019.
57. Confalonieri R. et al. Trepan reloaded: A knowledge-driven approach to explaining black-box models. Front. Artif. Intell. Appl. 2020. vol. 325. pp. 2457–2464.
58. Confalonieri R. et al. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. Artif. Intell. Elsevier, 2021. vol. 296. pp. 103471.
59. Bourguin G. et al. Towards Ontologically Explainable Classifiers. Artificial Neural Networks and Machine Learning – ICANN. 2021. pp. 472–484. DOI: 10.1007/978-3-030-86340-1_38.
60. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 2019. vol. 1. no. 5. pp. 206–215.
61. Voogd J. et al. Using Relational Concept Networks for Explainable Decision Support. 3rd IFIP Cross Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE). 2019. pp. 78–93. DOI: 10.1007/978-3-030-29726-8_6.
62. Fong A.C.M., Hong G. Ontology-Powered Hybrid Extensional-Intensional Learning. Proceedings of the 2019 International Conference on Information Technology and

Computer Communications (ITCC2019). New York, USA: ACM Press, 2019. pp. 18–23.
63. Bellucci M. et al. Ontologies to build a predictive architecture to classify and explain. DeepOntoNLP Workshop @ESWC 2022. 2022.
64. Martin T. et al. Bridging the gap between an ontology and deep neural models by pattern mining. The Joint Ontology Workshops, JOWO. 2020. vol. 2708.
65. De Sousa Ribeiro M., Leite J. Aligning Artificial Neural Networks and Ontologies towards Explainable AI. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35. no. 6. pp. 4932–4940.
66. Agafonov A., Ponomarev A. An Experiment on Localization of Ontology Concepts in Deep Convolutional Neural Networks. The 11th International Symposium on Information and Communication Technology. NY, USA: ACM, 2022. pp. 82–87.
67. Ponomarev A., Agafonov A. Ontology Concept Extraction Algorithm for Deep Neural Networks. 2022 32nd Conference of Open Innovations Association (FRUCT). IEEE, 2022. pp. 221–226.
68. Sarker M.K. et al. Wikipedia Knowledge Graph for Explainable AI. KGSWC 2020, CCIS 1232. 2020. pp. 72–87.
69. Abbass H.A. et al. Machine Education: Designing semantically ordered and ontologically guided modular neural networks. IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2019. pp. 948–955.
70. Smirnov A. et al. Multi-aspect Ontology for Interoperability in Human-machine Collective Intelligence Systems for Decision Support. Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. SCITEPRESS – Science and Technology Publications, 2019. pp. 458–465.

Shilov Nikolay — Ph.D., Associate Professor, Senior researcher, Laboratory of computer-aided integrated systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: artificial intelligence, knowledge management, ontology management, complex system modelling and configuration, machine learning. The number of publications — 300. nick@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Ponomarev Andrew — Ph.D., Associate Professor, Senior researcher, Laboratory of computer-aided integrated systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: collective intelligence, crowd computing, recommender systems, applied machine learning. The number of publications — 70. ponomarev@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Smirnov Alexander — Ph.D., Dr.Sci., Professor, Chief researcher, Head of laboratory, Laboratory of computer-aided integrated systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: decision support systems, intelligent systems, intelligent configuration management in virtual and network organizations, knowledge logistics. The number of publications — 400. smir@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Acknowledgements. This research is funded by the Russian Science Foundation (grant 22-11-00214).