# Билет 10. Обучаемые системы понимания изображений на примере системы CELL

Создание баз знаний является крайне трудозатратной операцией. Еще более сложной проблемой оказывается создание баз знаний, в которые интегрированы системы компьютерного зрения, поскольку оказывается необходимым в явном виде указывать взаимосвязи между перцептивным и семантическим (а также функциональным) знанием, чего не требовалось в классических экспертных системах. Не удивительно, что системы понимания изображений, основанные на знаниях, создаются, как правило, крайне узкоспециализированными.

Как и в случае классических экспертных систем, ключевой проблемой здесь оказывается автоматическое приобретение знаний. Таким образом, для создания систем понимания изображений общего назначения необходимо решение проблемы машинного обучения на основе перцептивного опыта. Существующие подходы в данной области заключаются в имитации процесса усвоения языка младенцем, который исходно не обладает никакими знаниями о свойствах различных объектов, однако способен научиться строить словесные описания сцен. Рассмотрим в качестве одного из типичных примеров данного подхода серию работ, выполняемых в Media Laboratory Массачусетского Технологического Института.

Развитие упомянутой серии работ началось с системы, получившей название CELL (Cross-channel Early Lexical Learning), которая впоследствии той же группой исследователей была развита в системы Describer, Newt, Ripley, Fuse. Далее мы кратко рассмотрим базовые принципы развиваемого в перечисленных выше работах подхода на примере системы CELL.

*Общая архитектура системы CELL*

Представим себе машинную систему, оснащенную сенсорами, возможно различных модальностей. В описываемом подходе предполагается, что информация, поступающая от сенсоров, разделяется на каналы, являющиеся либо лингвистическими, либо семантическими (контекстными). Введение информационных каналов в дополнение к сенсорным модальностям вызвано тем, что как лингвистическая, так и семантическая информация может поступать от сенсоров любой модальности.

К примеру, лингвистический канал может содержать извлеченную из акустического сигнала речь или извлеченные из визуального сигнала жесты, движение губ или письменную речь. Тактильный сигнал также может нести для человека лингвистическую информацию. В то же время, любая из этих сенсорных модальностей может нести и семантическую информацию. Например, семантический канал может содержать описание формы, извлеченное из визуального или тактильного сигнала, описание цвета, движения и т.д.

В системе CELL в соответствующих каналах лингвистическая или семантическая информация кодируется посредством системы признаков, процедура выделения которых считается данной априорно. Для поступающих по каждому каналу векторов признаков можно решить задачу их группирования и сформировать классы образов. Если речь идет о семантических признаках, то подобные классы соответствуют некоторым *семантическим категориям*, а если о лингвистических – *лингвистическим единицам*. В системе CELL в процессе обучения строятся классы образов, описывающиеся эталонным образом, прототипом, и величиной допустимых отклонений от прототипа, при которых образ все еще относится к данному классу.

Лексемой в данной работе называется концепт, включающий в себя некую лингвистическую единицу, ассоциированную с некоторой семантической категорией. Сложность заключается в том, что требуется не просто связать лингвистическую единицу и семантическую категорию, но одновременно с этим и построить соответствующие им классы образов, поскольку они не даются системе априорно.

Формирование классов образов для каждого из каналов в отдельности можно было бы осуществлять в стиле обучения без учителя. Однако богатство поступающей информации столь велико, что надежно выделять классы затруднительно. Таким образом, следует решать задачи построения классов и их связывание одновременно. Одна из основных идей заключается в том, что использование лингвистической информации позволит надежнее выделять релевантные семантические категории, а использование семантической информации (контекста) – надежнее вести обработку лингвистической информации (выделять и осуществлять классификацию отдельных лингвистических единиц из потока слитной речи). Посмотрим, как это было сделано в описываемой работе.

В ней для семантического и лингвистического каналов вводится понятие S-, и L-событий. Событие – это последовательность выделенных признаков в интервале времени, в котором наблюдается активность в соответствующем канале. Для лингвистического канала это соответствует цепочке некоторых лингвистических признаков, возможно разделенных паузами, соответствующих обособленному речевому высказыванию. Например, при устном вводе L-событие соответствует цепочке произнесенных вслух слов и может быть обнаружено детектором речь/молчание. S-событие соответствует заметным изменениям в сенсорном канале, например, достаточно быстрому движению в поле зрения камеры. Критерий значимости изменений в канале не вводится и предлагается определять его эвристически для каждого отдельного канала.

Итак, пусть в каждом канале выделены события как цепочки признаков. Как L-, так и S- событие может включать образы, относящиеся к различным классам. Такие протяженные во времени события необходимо разбить на некоторые более короткие временные интервалы, сегменты. Для лингвистических событий эти временные интервалы определяют гипотетические границы лингвистических единиц. Если используется несколько лингвистических каналов (например, фонемный и движение губ), то гипотезы об интервалах могут взаимно проверяться и делаться более достоверными. В семантических каналах также необходимо ввести разделение на сегменты. Границы сегментов соответствуют резкому (во времени) изменению скорости, контраста, цвета и т.д. Это означало бы изменение распознаваемых семантических категорий, если бы они были уже сформированы. В связи с этим вводится понятие L- и S-подсобытий, как подпоследовательностей в цепочках признаков, составляющих событиям.

Речь, обращенная к младенцам, обычно относится к мгновенному контексту. Значит, при овладении лексикой является принципиальным одновременное появление лингвистических единиц и соответствующих им семантических категорий, представленных в каналах L- и S-подсобытиями. В описываемой работе моделируется восприятие младенцев и используются записи игры с ними, поэтому делается предположение, что L- и S-подсобытия, относящиеся к одному и тому же объекту, присутствуют в перекрывающихся во времени L- и S-событиях. Когда такие перекрывающиеся события встречаются, то формируется единое LS-событие, которое и подлежит дальнейшему анализу. Это событие помещается в кратковременную память, являющуюся FIFO-буфером (очередью). Размер этого буфера в системе CELL был взят  LS-событий. Помимо ссылки на моделирование человеческого восприятия, другим аргументом является экономия вычислительных ресурсов, поскольку поиск в кратковременной памяти повторяющихся событий является исчерпывающим и требует экспоненциального времени.

После помещения LS-событий в буфер требуется выделить подсобытия, в частности, необходимо осуществить сегментацию слитной речи на отдельные слова при неизвестном лексиконе. При этом используется предположение, что в помещенных в буфер LS-событиях значимые подсобытия неоднократно повторяются, поэтому их можно выделить как наиболее длинные повторяющиеся (с определенной точностью) цепочки признаков. Идея поиска повторяющихся подсобытий основывается на том, что речь, обращенная к младенцам, является чрезвычайно избыточной. Если младенцу дают в руки какую-то игрушку, то ее название повторяется много раз («посмотри, какой мячик; мячик круглый, …»). Поскольку подсобытия повторяются неточно, то чтобы их искать, вводится (в качестве дополнительной априорной информации) метрика в каждом L- и S-каналах. Пары повторяющихся и совместно встречающихся L- и S-подсобытий помещаются в память следующего уровня, имеющую значительно больший объем. Если во время обнаружения следующих LS-событий появляются похожие (в смысле близости в пространстве признаков) значимые пары подсобытий, то на их основе строится лексема (концепт), состоящая из эталонной семантической категории, эталонной лингвистической единицы и радиусов в пространствах признаков, определяющих границы концепта. Это аналогично инкрементному обучению без учителя.

Принятие решения о формировании лексемы осуществляется на основе взаимной информации, содержащейся в соответствующих L- и S-подсобытиях. Пусть имеется набор пар совместно встречающихся подсобытий (*Li, Si*). Тогда для каждой из них можно оценить безусловные вероятности *P*(*L*) и *P*(*S*) их появления, а также вероятность совместного появления *P*(*L,S*). Большое значение взаимной информации, вычисляемой как , будет служить индикатором того, что появление одного события может быть предсказано при появлении другого подсобытия.

Если оцененная взаимная информация превосходит некоторый порог, то концепт формируется. В процессе функционирования системы при формировании ею концептов рассматриваются в промежуточной памяти только такие события, которые не распознаются как сформированные лексемы. Мы не будем детальнее разбирать вопросы выделения подсобытий, формирования концептов и т.д., а кратко опишем реализацию системы на основе зрительной и акустической модальностей.

*Реализация зрительной и акустической подсистем в системе CELL*

Один из возможных вариантов реализации описываемой системы (CELL) использует аудио-видео вход. В этой реализации аудиоинформация получается с микрофона, на который записывается естественная слитная речь воспитателя, обращенная к младенцу. В качестве семантических каналов используются каналы, содержащие информацию о форме и цвете объекта, присутствующего в поле зрения камеры. Ставятся задачи выявления лингвистических единиц и семантических категорий формы и цвета и их связывания в концепт таким образом, чтобы в результате обучения система могла называть предъявленные ей объекты, а также, наоборот, по названию объекта выбирать его из заданного набора.

Для выделения лингвистических и семантических признаков необходимы модально-специфичные процедуры, которые считаются заданными априорно. Посмотрим, как здесь ведется выделение семантических признаков в двух каналах: формы и цвета.

Трехмерная форма объекта в явном виде не восстанавливается. Вместо этого используются сгруппированные описания плоских форм, которые получаются в результате съемки объекта с разных ракурсов. На каждом таком снимке осуществляется разделение на объект и фон, при котором используется предположение о том, что фон является сравнительно однородным (однотонная скатерть, простыня и т.д.), на который помещен крупный предмет близко к центру кадра. Все пиксели, которые не относятся к фону, помечаются как принадлежащие к объектам. Находятся связные области из таких пикселей, и из этих областей выделяется наибольшая расположенная ближе к центру кадра. Также рассматривается случай, когда подходящий объект отсутствует (установлением порога на размер и положение) или не полностью попадает в кадр. Далее анализируется форма и цвет выделенной таким образом области, предположительно соответствующей изображению объекта.

Для представления цвета используется двухцветная гистограмма 8х8 ячеек нормализованных цветов (r, g), инвариантная к смене интенсивности освещения. Поскольку цвета каждого пикселя нормированы на его уровень яркости, то третий цвет дополнительной информации не несет и в представлении не используется.

Описание формы тоже осуществляется с помощью двухмерной гистограммы. Для ее построения находятся точки, лежащие на границе выделенной области. Для них определяются наклоны касательных, проведенных к границе. Также для каждой пары точек границы определяется нормализованное расстояние. По этим двум параметрам строится гистограмма, которая и описывает форму области. Это представление инвариантно к масштабу и повороту объекта в плоскости изображения.

Как видно, представление изображений весьма простое и опирается на сильные предположения. В нем не выделяется значимая информация, а часть ее теряется. Это связано с тем, что в описываемой работе основное внимание уделялось проблеме анализа речи (которую мы здесь не рассматриваем). Кроме того, в последующих работах анализ изображений был усложнен.

Система CELL, оснащенная акустическими и видео сенсорами, обучалась по записям игры взрослых с младенцами примерно десятимесячного возраста. Для обучения использовались записи игры разных взрослых с младенцами с привлечением различных объектов (игрушки, мячики, обувь и т.д.). Одна запись относилась к одному предмету. Результаты исследований показали относительную успешность данного подхода. Система научилась по естественным записям распознавать 70% предъявляемых объектов по их названиям, представляемым в акустической форме. Для сравнения, поиск повторяющихся фрагментов в записях речи (без использования видео) позволил выделить только 30% релевантных слов (кроме того, не связанных с какими-либо семантическими категориями).

*Дальнейшее развитие системы CELL*

Система CELL является наиболее ранней и имеет ряд расширений: это системы Describer и Newt, робот Ripley, система Fuse и ряд других.

В системах Describer и Newt семантическая основа подводится не только под отдельные слова, но и под целые выражения, описывающих пространственные отношения между объектами, такие как «большой синий квадрат справа от красного прямоугольника». Обучение системы Describer производится по примерам смоделированных сцен, сопровождающихся выполненными людьми описаниями, по которым можно выбрать нужный объект. Система Newt учится по изображениям реальных сцен, сопровождающихся устным описанием того объекта, на который система указывает лазерной указкой (на сцене может присутствовать несколько объектов, помещенных на однородный фон).

Системы Describer и Newt отличаются в том, что в первой осуществляется описание выбранного объекта на сцене, в то время как во второй – выбор объекта по его словесному описанию. Помимо проблемы связывания слов с семантическими категориями здесь возникают также две дополнительные проблемы: разделение слов на классы (например, слова «желтый», «розовый», «зеленый» должны принадлежать одному классу, а «над», «слева», «рядом» – к другому классу) и определение порядка слов в предложении (обучение синтаксису).

В системе Fuse основное внимание уделяется вопросам взаимодействия зрительной и акустической подсистем на нижнем уровне, а именно, влиянию речи на управление вниманием зрительной подсистемы и влиянию визуальной информации на улучшение сегментации и распознавания слов и понимания речи.

В роботе Ripley подход получил дальнейшее развитие по нескольким направлениям. Во-первых, расширена сенсорная система робота за счет добавления в нее проприоцепторной подсистемы, чувства гравитации и тактильной подсистемы (чувства осязания). Во-вторых, добавлена система эффекторов: у робота появилась возможность совершать движения и брать предметы манипулятором. И, в-третьих, добавлена «ментальная» модель окружающего физического мира, в которой описываются пространственные положения обнаруженных роботом объектов. Расширение системы сенсоров и эффекторов дает роботу возможность в процессе обучения устанавливать семантическую опору для таких слов, как «легкий», «тяжелый», «дотрагиваться», «поднимать», «давать» и т.д. Ментальная модель физического пространства не только позволяет осуществлять в нем навигацию, но и позволяет понимать различия в таких фразах, как «мяч слева от меня» или «мяч слева от тебя». Для этого в ментальной модели реализована возможность смены точки зрения, что осуществляется с помощью библиотеки 3D-моделирования OpenGL, на основе которой и строится воображение робота. Таким образом, слова «мой» или «твой» находят свою семантическую опору в ментальной операции переноса точки зрения.

В результате обучения упомянутые системы способны строить описания сцен и выполнять в них указанные человеком действия, что свидетельствует об определенном уровне «понимания» как изображений, так и речевых высказываний.