Лекция 9

Ассоциативные запоминающие нейронные сети

Еще один важный класс составляют сети с наличием обратных связей между различными слоями нейронов. Такие НС называются рекуррентными. Наличие обратных связей добавляет в процесс функционирование динамические зависимости между слоями. При этом различные начальные наборы параметров сети могут приводить как к стабилизации нейросети на некотором этапе функционирования (то есть начиная с определённой итерации отсутствуют значительные изменения параметров сети), так и к дестабилизации и скачковым изменениям параметров.

Сети с обратными связями

В общем случае может быть рассмотрена нейронная сеть, содержащая произвольные обратные связи, по которым переданное возбуждение возвращается к данному нейрону, и он повторно выполняет свою функцию. Наблюдения за биологическими локальными нейросетями указывают на наличие множественных обратных связей. Нейродинамика в таких системах становится итерационной. Это свойство существенно расширяет множество типов нейросетевых архитектур, но одновременно приводит к появлению новых проблем.

Безитерационная динамика состояний нейронов является, очевидно, всегда устойчивой. Обратные связи могут приводить к возникновению *неустойчивостей,* подобно тем, которые возникают в усилительных радиотехнических системах при положительной обратной связи. В нейронных сетях неустойчивость проявляется в блуждающей смене состояний нейронов, не приводящей к возникновению стационарных состояний. В общем случае ответ на вопрос об устойчивости динамики произвольной системы с обратными связями крайне сложен и до настоящего времени является открытым.. Далее мы рассмотрим некоторые классы сетей, функционирующих в качестве ассоциативных запоминающих устройств. Ассоциативная память играет роль системы, определяющей взаимную зависимость векторов. Главная задача ассоциативной памяти сводится к запоминанию обучающей выборки таким образом, чтобы при предъявлении новой выборки система смогла сгенерировать ответ – какой из запомненных векторов выборки наиболее близок к поданному на вход.

Модель Хопфилда

Модель Хопфилда (J.J.Hopfield, 1982) занимает особое место в ряду нейросетевых моделей. В ней впервые удалось установить связь между нелинейными динамическими системами и нейронными сетями. Образы памяти сети соответствуют устойчивым предельным точкам (аттракторам) динамической системы. Особенно важной оказалась возможность переноса математического аппарата теории нелинейных динамических систем (и статистической физики вообще) на нейронные сети. При этом появилась возможность теоретически оценить объём памяти сети Хопфилда, определить область параметров сети, в которой достигается наилучшее функционирование. Сеть Хопфилда является одной из первых архитектур реализующих автоассоциативную память.

Нейродинамика в модели Хопфилда

Рассмотрим сеть из N формальных нейронов, в которой степень возбуждения каждого из нейронов S_i, i=1..N, может принимать только два значения {-1, +1}. Любой нейрон имеет связь со всеми остальными нейронами S_j, которые в свою очередь связаны с ним. Силу связи от i-го к j-му нейрону обозначим как w_{ij}.

В модели Хопфилда предполагается условие *симметричности* связей $w_{ij} = w_{ji}$, с нулевыми диагональными элементами $w_{ii} = 0$. К сожалению, это условие имеет весьма отдаленное отношение к известным свойствам биологических сетей, в которых, наоборот, если один нейрон передает возбуждение другому, то тот, в большинстве случаев, непосредственно не связан с первым. Однако именно симметричность связей, как будет ясно из дальнейшего, существенно влияет на устойчивость динамики.

Можно выделить два режима функционирования сети: режим обучения и режим классификации.

Изменение состояния каждого нейрона s_i в модели Хопфилда в режиме классификации происходит по известному правилу для формальных нейронов МакКаллока и Питтса. Поступающие на его входы сигналы s_i в момент t взвешиваются с весами матрицы связей w_{ij} и суммируются, определяя полный уровень силы входного сигнала:

$$h_j = \sum_{i \neq j} w_{ij} s_i$$

Далее в момент $t+1$ нейрон изменяет состояние своего возбуждения в зависимости от уровня сигнала h и индивидуального порога каждого нейрона:

$$s_j(t+1) = \begin{cases} -1, & h_j(t) < T_j, \\ +1, & h_j(t) > T_j, \\ s_j(t), & h_j(t) = T_j. \end{cases}$$

Изменение состояний возбуждения всех нейронов может происходить одновременно, в этом случае говорят о *параллельной* динамике. Рассматривается также и *последовательная* нейродинамика, при которой в данный момент времени происходит изменение состояния только одного нейрона. Многочисленные исследования показали, что свойства памяти нейронной сети практически не зависят от типа динамики. При моделировании нейросети на обычном компьютере удобнее последовательная смена состояний нейронов. В аппаратных реализациях нейросетей Хопфилда применятся параллельная динамика.

Алгоритм. Стандартное (синхронное) функционирование сети Хопфилда.

1. Формируется сеть, имеющая некоторое число стандартных режимов.
2. На вход сети подаётся вектор $\bar{s}(0)$.
3. Сеть функционирует в соответствии с формулами:

$$\bar{h}(t) = \bar{s}(t-1) \cdot W,$$
$$\bar{s}_i(t) = sign(\bar{h}_i(t)).$$

4. Шаг 3 повторяется пока сеть не придёт (с достаточной точностью) в некоторой стандартный режим.
5. Этот стандартный режим соответствует тому образу, который вспоминает сеть.

Совокупность значений возбуждения всех нейронов s_i в некоторый момент времени образует *вектор состояния* $\bar{s}$ сети. Нейродинамика приводит к изменению вектора состояния $\bar{s}(t)$. Вектор состояния описывает траекторию в *пространстве состояний* нейросети. Это пространство для сети с двумя уровнями возбуждения каждого нейрона, очевидно, представляет собой множество вершин гиперкуба размерности,

равной числу нейронов N. Возможные наборы значений координат вершин гиперкуба (см. Рис.8.2) и определяют возможные значения вектора состояния.

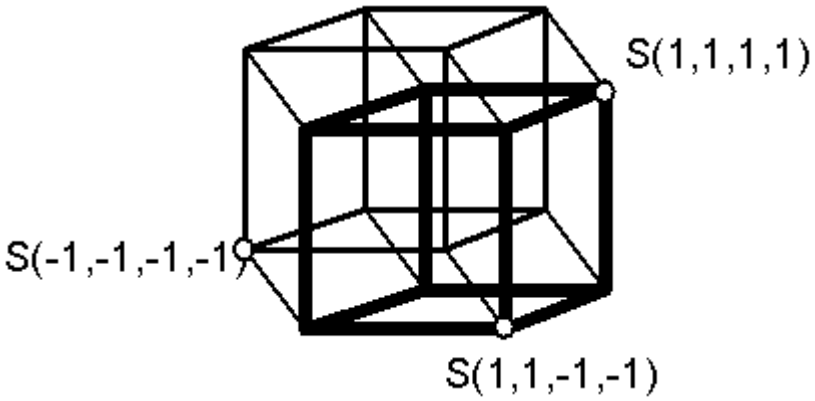

Рис. 8.1. Проекция 4-х мерного гиперкуба на плоскость. Указанные на рисунке три точки служат примерами возможных состояний нейронной сети из 4-х нейронов.

Рассмотрим теперь проблему устойчивости динамики изменения состояний. Поскольку на каждом временном шаге некоторый нейрон i изменяет свое состояние в соответствии со знаком величины $h_i(t) - T_i$, то приведенное ниже соотношение всегда неположительно:

$$\Delta E_i = -(s_i(t+1) - s_i(t)) \cdot (h_i(t) - T_i) \le 0$$

где

$$E_i = s_i(t) \cdot (h_i(t) - T_i)$$

Таким образом, соответствующая величина E, являющаяся суммой отдельных значений E_i, может только убывать, либо сохранять свое значение в процессе нейродинамики. Определим «энергию сети» E следующим образом:

$$E(t) = -\frac{1}{2}\sum_i \sum_{j \ne i} w_{ij} s_i(t) s_j(t) + \sum_i s_i(t) T_i$$

или, в матричной форме

$$E(t) = -\frac{1}{2}\bar{s}(t) \cdot W \cdot \bar{s}^T(t) + \bar{s}(t) \cdot \bar{T}$$

Введенная таким образом величина $E(t)$ является функцией состояния $E(t) = E(\bar{s}(t))$ и называется энергетической функцией (энергией) нейронной сети

Хопфилда. Поскольку она обладает свойством невозрастания при динамике сети, то одновременно является для нее функцией Ляпунова (А.М. Ляпунов, 1892). Поведение такой динамической системы устойчиво при любом исходном векторе состояния $\bar{s}(0)$ и при любой симметричной матрице связей $W = (w_{ij})$ с нулевыми диагональными элементами. Динамика при этом заканчивается в одном из минимумов функции Ляпунова, причем активности всех нейронов будут совпадать по знаку с входными сигналами h.

Поверхность энергии $E(\bar{s})$ в пространстве состояний имеет весьма сложную форму с большим количеством локальных минимумов, образно напоминая стеганое одеяло. Стационарные состояния, отвечающие минимумам, могут интерпретироваться, как *образы* памяти нейронной сети. Эволюция к такому образу соответствует процессу извлечения из памяти. При произвольной матрице связей W образы также произвольны. Для записи в память сети какой-либо осмысленной информации требуется определенное значение весов W, которое может получаться в процессе обучения.

Правило обучения Хебба

Правило обучения для сети Хопфилда опирается на исследования Дональда Хебба (D.Hebb, 1949), который предположил, что синаптическая связь, соединяющая два нейрона, будет усиливаться, если в процессе обучения оба нейрона согласованно испытывают возбуждение либо торможение. Простой алгоритм, реализующий такой механизм обучения, получил название *правила Хебба*. Рассмотрим его подробно.

Пусть задана обучающая выборка образов $x^{(a)}$, $a = 1\mathbf{K}\,p$. Требуется построить процесс получения матрицы связей W, такой, что соответствующая нейронная сеть будет иметь в качестве стационарных состояний образы обучающей выборки (значения порогов нейронов T обычно полагаются равными нулю).

В случае одного обучающего образа правило Хебба приводит к требуемой матрице:

$$w_{ij} = x_i \cdot x_j .$$

Покажем, что состояние $\bar{s} = \bar{x}$ является стационарным для сети Хопфилда с указанной матрицей. Действительно, для любой пары нейронов i и j энергия их взаимодействия в состоянии $\bar{x}$ достигает своего минимально возможного значения

$$E_{ij} = -\frac{1}{2} x_i x_j x_i x_j = -\frac{1}{2}.$$

При этом E -полная энергия равна E = -(1/2) N ^{2}, что отвечает глобальному минимуму.

Для запоминания других образов может применяется итерационный процесс:

$$w_{ij}^{(a)} = w_{ij}^{(a-1)} + x_i^{(a)} \cdot x_j^{(a)}, \quad w_{ij}^{(0)} = 0, \quad a = 1\mathbf{K}\, p,$$

который приводит к полной матрице связей в форме Хебба:

$$w_{ij} = \sum_{a=1}^{p} x_i^{(a)} \cdot x_j^{(a)}.$$

Устойчивость совокупности образов не столь очевидна, как в случае одного образа. Ряд исследований показывает, что нейронная сеть, обученная по правилу Хебба, может в среднем, при больших размерах сети N, хранить не более чем $p = \frac{N}{4 \ln N}$ различных образов. Устойчивость может быть показана для совокупности ортогональных образов, когда

$$\frac{1}{N} \sum_{k=1}^{N} x_j^{(a)} \cdot x_j^{(b)} = d_{ab} = \begin{cases} 1, a = b, \\ 0, a \neq b. \end{cases}$$

В этом случае для каждого состояния $x^{(a)}$ произведение суммарного входа i-го нейрона h_i на величину его активности $s_i = x_i^{(a)}$ оказывается положительным, следовательно само состояние $x^{(a)}$ является состоянием притяжения (*устойчивым аттрактором*):

$$h_i \cdot x_i^{(a)} = \sum_j \left(\left(\sum_b x_i^{(b)} x_j^{(b)} \right) x_j^{(a)} \right) \cdot x_i^{(a)} = N > 0$$

Таким образом, правило Хебба обеспечивает устойчивость сети Хопфилда на заданном наборе относительно небольшого числа ортогональных образов. В следующем пункте мы остановимся на особенностях памяти полученной нейронной сети.

Ассоциативность памяти и задача распознавания образов

Динамический процесс последовательной смены состояний нейронной сети Хопфилда завершается в некотором стационарном состоянии, являющемся локальным минимумом энергетической функции $E(\bar{s})$. Невозрастание энергии в процессе динамики приводит к выбору такого локального минимума $\bar{s}$, в область притяжения которого попадает начальное состояние (исходный, предъявляемый сети образ $\bar{s}_0$. В этом случае также говорят, что состояние $\bar{s}_0$ находится в чаше минимума $\bar{s}$.

При последовательной динамике в качестве стационарного состояния будет выбран такой образ $\bar{s}$, который потребует минимального числа изменений состояний отдельных нейронов. Поскольку для двух двоичных векторов минимальное число изменений компонент, переводящее один вектор в другой, является расстоянием Хемминга $r_H(\bar{s}, \bar{s}_0)$, то можно заключить, что динамика сети заканчивается в ближайшем по Хеммингу локальном минимуме энергии.

Пусть состояние $\bar{s}$ соответствует некоторому идеальному образу памяти. Тогда эволюцию от состояния $\bar{s}_0$ к состоянию $\bar{s}$ можно сравнить с процедурой постепенного восстановления идеального образа $\bar{s}$ по его искаженной (зашумленной или неполной) копии $\bar{s}_0$. Память с такими свойствами процесса считывания информации является *ассоциативной*. При поиске искаженные части целого восстанавливаются по имеющимся неискаженным частям на основе ассоциативных связей между ними.

Ассоциативный характер памяти сети Хопфилда качественно отличает ее от обычной, адресной, компьютерной памяти. В последней извлечение необходимой информации происходит по *адресу* ее начальной точки (ячейки памяти). Потеря адреса (или даже одного бита адреса) приводит к потере доступа ко всему информационному фрагменту. При использовании ассоциативной памяти доступ к информации производится непосредственно по ее *содержанию,* т.е. по частично известным искаженным фрагментам. Потеря части информации или ее информационное зашумление не приводит к катастрофическому ограничению доступа, если оставшейся информации достаточно для извлечения идеального образа.

Поиск идеального образа по имеющейся неполной или зашумленной его версии называется задачей *распознавания образов*. В нашей лекции особенности решения этой

задачи нейронной сетью Хопфилда будут продемонстрированы на примерах, которые получены с использованием модели сети на персональной ЭВМ.

Несмотря на интересные качества, нейронная сеть в классической модели Хопфилда далека от совершенства. Она обладает относительно скромным объемом памяти, пропорциональным числу нейронов сети N, в то время как системы адресной памяти могут хранить до 2N различных образов, используя N битов. Кроме того, нейронные сети Хопфилда не могут решить задачу распознавания, если изображение смещено или повернуто относительно его исходного запомненного состояния. Эти и другие недостатки сегодня определяют общее отношение к модели Хопфилда, скорее как к теоретическому построению, удобному для исследований, чем как повседневно используемому практическому средству.