Слюсар В.И., д.т.н., проф.
Центральный научно-исследовательский институт вооружения и военной техники Вооруженных Сил Украины, Киев

ТЕНЗОРНО-МАТРИЧНАЯ МОДЕЛЬ НЕЙРОСЕТЕЙ

Для уменьшения времени реакции нейронных сетей может быть предложена реализация развитой автором тензорно-матричной теории на основе проникающего торцевого произведения матриц [1]. Целью доклада является описание соответствующей такому подходу модели типовой нейросети.

Согласно определению, предложенному автором в 1998 г. [1], проникающим торцевым произведением p×g-матрицы **A** и n-мерного тензора **B**, развёрнутого в блочную матрицу строк или столбцов $\mathbf{B}=[B_n]$ ($n > 1$), содержащих p×g-блоки, является матрица вида:

$$\mathbf{A} \boxdot \mathbf{B}=[\mathbf{A}\circ B_r], \qquad (1)$$

где $\mathbf{A}\circ B_r$ представляет собой произведение Адамара.

Если тензор **B** записан как блок-строка, получим:

$$\mathbf{A} \boxdot \mathbf{B}=[\mathbf{A}\circ B_r]=[\mathbf{A}\circ B_1 \vdots \mathbf{A}\circ B_2 \vdots \ldots \vdots \mathbf{A}\circ B_r \vdots \ldots].$$

Например:

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}, \ \mathbf{B}=\left[\begin{array}{cc:cc:cc} b_{111} & b_{121} & b_{112} & b_{122} & b_{113} & b_{123} \\ b_{211} & b_{221} & b_{212} & b_{222} & b_{213} & b_{223} \\ b_{311} & b_{321} & b_{312} & b_{322} & b_{313} & b_{323} \end{array}\right],$$

$$\mathbf{A} \boxdot \mathbf{B} = \left[\begin{array}{cc:cc:cc} a_{11}\cdot b_{111} & a_{12}\cdot b_{121} & a_{11}\cdot b_{112} & a_{12}\cdot b_{122} & a_{11}\cdot b_{113} & a_{12}\cdot b_{123} \\ a_{21}\cdot b_{211} & a_{22}\cdot b_{221} & a_{21}\cdot b_{212} & a_{22}\cdot b_{222} & a_{21}\cdot b_{213} & a_{22}\cdot b_{223} \\ a_{31}\cdot b_{311} & a_{32}\cdot b_{321} & a_{31}\cdot b_{312} & a_{32}\cdot b_{322} & a_{31}\cdot b_{313} & a_{32}\cdot b_{323} \end{array}\right].$$

В данном случае матрицу **A** можно рассматривать как исходную матрицу пикселей изображения на входе нейросети. При этом каждый блок матрицы **B** будет соответствовать блоку весовых коэффициентов нескольких нейронов в одном слое нейронной сети. Дальнейшие шаги обработки данных в такой модели нейросети зависят от структуры и типа слоев. В случае свёрточной нейросети блок-строку $\mathbf{A} \boxdot \mathbf{B}$ следует умножить на единичный вектор. При этом можно получить: вектор-строку $\mathbf{1}^T(\mathbf{A} \boxdot \mathbf{B})$; вектор $(\mathbf{A} \boxdot \mathbf{B})\times\mathbf{1}$, где × символ обычного матричного умножения, **1** – вектор единиц; матрицу $(\mathbf{A} \boxdot \mathbf{B})\,[\times]\mathbf{1}$, где [×] – символ блочного обычного произведения матриц, **1** – блок-вектор единиц; скаляр $\mathbf{1}^T(\mathbf{A} \boxdot \mathbf{B})\mathbf{1}$. Результаты такого умножения необходимо использовать в качестве аргумента функции активации нейрона, например: $\tanh[(\mathbf{A} \text{◙} \mathbf{B})\times\mathbf{1}+\mathbf{d}]$ или $\mathrm{SReLU}[\mathbf{1}^{\mathrm{T}}(\mathbf{A} \text{◙} \mathbf{B})\mathbf{1}+\mathbf{d}]$, где $\mathbf{d}$ является вектором (скаляром).

В числе свойств проникающего произведения заслуживает внимания его связь с торцевым произведением матриц:

$$\mathbf{A} \square \mathbf{A} = \mathrm{A} \text{◙} (\mathbf{A} \otimes \mathbf{1}^{\mathrm{T}}),\ \mathbf{c} \square \mathbf{A} = \mathbf{A} \square \mathbf{c} = \mathbf{c} \text{◙} \mathbf{A} = \mathbf{A} \text{◙} \mathbf{c},$$

где $\square$ – символ торцевого произведения [2], $\otimes$ - произведение Кронекера, $\mathbf{1}^{\mathrm{T}}$ – вектор-строка единиц, $\mathbf{c}$ – вектор. Эти свойства позволяют применить для выполнения проникающего торцевого произведения оператор «tf.multiply», встроенный в библиотеку машинного обучения TensorFlow, поскольку данный оператор обеспечивает торцевое умножение вектора и матрицы. Однако такой подход не работает в отношении произведения матриц и требует предварительной векторизации матрицы меньшей размерности **A** в сочетании с векторизацией блоков [3] согласованной с ней блочной матрицы **B**. Необходимая для решения этой задачи процедура блочной векторизации также реализуется в рамках TensorFlow.

В интересах обработки данных в многоуровневых иерархиях кластеров нейронных сетей предлагается использовать обобщённое проникающее произведение или его транспонированную версию [3, 4]. В тех случаях, когда одно и то же изображение или видеопоток анализируются параллельно несколькими нейросетями, следует применить *проникающее прямое (кронекеровское) произведение*. Суть его сводится к обобщению операции (1) на случай, когда тензоры **A** и **B** развёрнуты в блочные матрицы с блоками одинаковой размерности:

$$\mathbf{A}\ [\text{◙}]\ \mathbf{B} = [\mathrm{A}_{ij} \text{◙} \mathbf{B}] = [\mathrm{A}_{ij} \circ \mathrm{B}_{mr}].$$

При этом каждый блок матрицы пикселей **A** поэлементно умножается на все блоки матрицы коэффициентов нейросети **B**.

Данный подход позволяет обеспечить обработку данных в периферийных устройствах в реальном масштабе времени.

Список использованных источников

1. Слюсар В.И. Семейство торцевых произведений матриц и его свойства// Кибернетика и системный анализ. – 1999.- Том 35; № 3.- С. 379-384.- DOI: 10.1007/BF02733426

2. Слюсар В.И. Торцевые произведения матриц в радиолокационных приложениях// Известия высших учебных заведений. Радиоэлектроника.- 1998. - Том 41, № 3.- С. 71 - 75.

3. Основы военно-технических исследований. Теория и приложения. Том. 2. Синтез средств информационного обеспечения вооружения и военной техники. / А.И. Миночкин, В.И. Рудаков, В.И. Слюсар. – Киев:«Гранмна, 2012. – С. 7 – 98; 354 – 521.

4. Слюсар В.И. Обобщенные торцевые произведения матриц в моделях цифровых антенных решеток с неидентичными каналами.//Известия высших учебных заведений. Радиоэлектроника.- 2003. - Том 46, № 10. - С. 15 - 26.