**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

***Садртдинов И.А. Исследование применения постулированной (posit) системы счисления и системы чисел с фиксированной точкой (fp) для представления весов глубоких сверточных нейронных сетей // Академия педагогических идей «Новация». Серия: Студенческий научный вестник. – 2018. – №6 (июнь). – АРТ 390-эл. – 0,4 п.л. - URL: http: //akademnova.ru/page/875550***

***РУБРИКА: ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ***

**УДК 004.032.26**

**Садртдинов Илья Айдарович**
магистрант 2 курса, факультет техника и технологии
*Научный руководитель:* Маков С.В., к.т.н., доцент
ФГБОУ ВПО «Институт сферы обслуживания и предпринимательства»
г. Шахты, Российская Федерация
e-mail: mail@sssu.ru

# ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ ПОСТУЛИРОВАННОЙ (POSIT) СИСТЕМЫ СЧИСЛЕНИЯ И СИСТЕМЫ ЧИСЕЛ С ФИКСИРОВАННОЙ ТОЧКОЙ (FP) ДЛЯ ПРЕДСТАВЛЕНИЯ ВЕСОВ ГЛУБОКИХ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

*Аннотация:* В этой статье представлено исследование использования постулированной системы счисления для представления весов глубоких сверточных нейронных сетей. В работе не применялись какие-либо методы квантования, и поэтому веса сети не требовали переобучения. Результаты исследования показывают, что использование постулированной системы счисления превосходит систему чисел с фиксированной точкой с точки зрения точности и использования памяти.

*Ключевые слова:* сверточные нейронные сети, глубокое обучение, метод представления весов сети.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

**Sadrtdinov Ilya Aydarovich**
2nd year master, Faculty of Engineering and Technology
Supervisor: S.V. Makov, Ph.D., Associate Professor
FGBOU HPE "Institute of Service and Entrepreneurship"
Shakhty, Russian Federation

# INVESTIGATION OF APPLICATION OF POSIT NUMBER SYSTEM AND FIXED POINT NUMBER SYSTEM FOR REPRESENTATION OF WEIGHTS OF DEEP CONVOLUTIONAL NEURAL NETWORK

*Abstract:* This article presents an investigation of the use of the posit number system for representing the weights of deep convolutional neural networks. The work did not use any quantization methods, and therefore the network weights did not require retraining. The results of the study show that the use of the posit number system outperforms the system of fixed-point numbers in terms of accuracy and memory usage.

Keywords: neural networks, deep learning, method of representation of network weights.

## 1. Введение

Глубокое обучение, как особая форма иерархического репрезентативного обучения [1], зарекомендовала себя в таких областях, как компьютерное зрение [2], обработка естественного языка [3], распознавание речи [4], робототехника [5] и медицина [6]. Успех глубокого обучения связан со способностью учиться на необработанных и неструктурированных данных [1]. В глубоком обучении обычно используются глубокие сверточные нейронные сети (DCNN), в которых в качестве механизма обучения применятся стохастический градиентный спуск [7].

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

Хотя DCNN достигают самой современной точности, по сравнению с другими подходами к компьютерному обучению они имеют такие недостатки, как неэффективная мощность и длительный период обучения. Например, для обучения ResNet-50 (50 слоев) [8] на наборе данных ImageNet [9] требуется 256 графических процессоров (GPU) [10]. Другой пример: нейронная сеть AlphaGo проходила обучение в течение нескольких месяцев на оборудовании с 1202 процессорами и 176 GPU, чтобы обыграть Ли Седола, 18-кратного чемпиона мира, в стратегической настольной игре «Go» [11]. Как можно видеть по приведенным примерам, обучение глубоких нейронных сетей даже с ресурсами в центрах обработки данных имеет множество ограничений. Глубокое изучение менее сложное, чем глубокое обучение. Кроме того, ограничения для внедрения глубокого изучения на обычные аппаратные средства, такие как процессоры и графические процессоры, были устранены цифровыми нейроморфными чипами, такими как TPU [12]. Но этот чип был разработан для центров обработки данных. Поэтому, в настоящее время, исследуются конструкции цифровых нейроморфных чипов для внедрения DCNN, с эффективностью в реальном времени, в встраиваемые платформы малой мощности, мобильные устройства и устройства IoT. Арифметика с низкой точностью - это общий подход к снижению энергопотребления и повышению производительности приложений глубокого обучения, в режиме реального времени, на встраиваемых устройствах.

Среди различных систем счисления, используемых для выполнения глубокого изучения, с применением арифметики с низкой точностью, система с фиксированной точкой показывает наиболее перспективный компромисс между точностью и вычислительной сложностью [15-18]. Однако, равномерно распределенные действительные числа с

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

фиксированной точкой не подходят для приложений глубокого обучения, поскольку веса и данные имеют неравномерное распределение [16]. В исследовании [19] в качестве альтернативы системе счисления с плавающей запятой была предложена постулированная система чисел (posit number system). Эта система счисления имеет уникальную нелинейную характеристику численного представления для всех чисел в динамическом диапазоне, которая отличает ее от других числовых систем, таких как с фиксированной и плавающей точкой. В этой работе представлено исследование применения постулированной системы счисления в DCNN для задач распознавания и классификации изображений.

В работе сравнивается постулированная система счисления и система чисел с фиксированной точкой для представления весов трех DCNN, с 4, 5 и 8 слоями, обученных на наборах данных MNIST [20], Cifar-10 [21] и ImageNet [9] соответственно. Постулированная система счисления превзошла систему с фиксированной точкой с точки зрения точности и использования памяти, при условии, что обе числовые системы, в процессе сравнения, имели одинаковый динамический диапазон ([-1,1]).

## 2. Обзор существующих решений

Judd и др. в своем исследовании DCNN сосредоточили внимание на улучшении вычислительной эффективности, с использованием ограниченной точности весов. Они представили веса в динамической системе с фиксированной точкой, и выполнили вычисления с использованием системы с плавающей точкой [22]. При таком подходе потребление энергии для операций доступа к памяти, полученное при обучении различных глубоких нейронных сетей на разных наборах данных, уменьшается в среднем на 15% [22]. После этого исследования, 8-разрядная система чисел с плавающей запятой применялась для представления весов

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

AlexNet и VGG-16 [23] и была оценена в наборе данных ImageNet [24]. Исследования показали, что можно представить 20% весов в 8-битовом представлении с плавающей запятой с менее чем 1% – ным снижением точности. Gysel и др. успешно провели глубокое обучение, используя архитектуру AlexNet на наборе данных ImageNet, с 8-битными динамическими весами с фиксированной точкой и 8-битными динамическими данными с фиксированной точкой, при снижении точности менее 1% [16]. Однако, чтобы достичь такого уровня точности, необходимо произвести переобучение нейронной сети. После успешного выполнения глубокого обучения, с использованием 8-битного представления точности весов и данных, исследователям удалость добиться представления менее чем 8 бит, в частности, 1-битное (двоичное) [17], [25] и 2-битное (тернаризованное) представление. Несмотря на то, что, используя эти представления, операции умножения в глубокой нейронной сети удаляются или преобразуются в операции обнаружения знака, значительное падение в точности подавляет все вычислительное преимущество. Поэтому, оценка модели глубокого обучения с 8 слоями или более (например, AlexNet, GoogLeNet) на больших наборах данных (например, ImageNet), с представлением каждого из значений веса и данных менее чем 8 битами, без существенного падения точности и/или переподготовки, остается открытым вопросом.

Постулированная система счисления является системой с конической точностью т.е. числа с малыми показателями более точны, чем числа с большими показателями. Формат постулированной системы чисел определяется как $P_{(n,\ es)}$, где n относится к общему количеству бит в этой системе, а es указывает количество бит степени (*exponent*). Каждое число в этой системе, как показано в уравнении 1, обозначается следующим

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

образом: *sign* (0 для положительных чисел, 1 для отрицательных чисел), *useed*, *exponent*, $r_{value}$ (отвечает за поиск бит режима числа) и *fraction* (для указания точности). [19]

$$X = (-1)^{sign} \cdot (useed)^{r_{value}} \cdot 2^{exponent} \cdot (1 + fraction) \quad (1)$$

Например, число 2.56 в этой системе счисления, в формате $P_{(16,1)}$, будет представленно 4 как *useed*, 1 как *exponent*, 0 как $r_{value}$ и 0.280 как *fraction* (рисунок 1). Более подробно преобразование будет описано далее.

**3. DCNNs с постулированной системой чисел**

В этой главе представлены исследования эффекта использования постулированной системы чисел для представления весов, точности и использования памяти DCNN в процессе обучения. Для достижения поставленной задачи, весы преобразуются из первоначальной единой системы с плавающей запятой в новую постулированную систему чисел, во время операций чтения и записи в памяти. Далее, постулированная система чисел преобразуется обратно в единую систему счисления с плавающей запятой по мере необходимости, так как этого требует стандартная вычислительная архитектура. Предлагаемая архитектура DCNN показана на рисунке 2. Она похожа на архитектуру DCNN, которая предложена в [22], за исключением того, что используется постулированная система счисления, которая имеет преимущества для представления весов неравномерной DCNN. Эта архитектура может быть фрагментирована на три подмодуля, которые поясняются в последующих подразделах.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

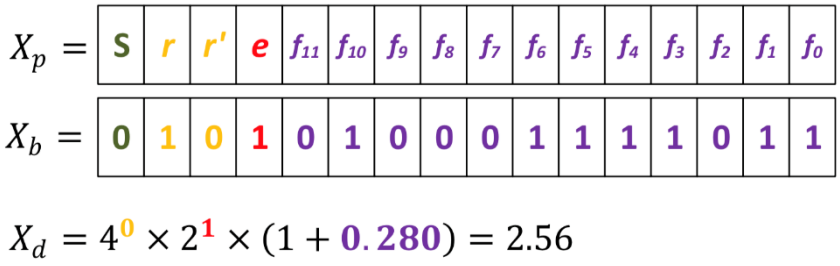



Рисунок 1. Представление числа в системе **posit** чисел в формате $P_{(16,1)}$ [19]

Первым шагом является преобразование постулированных чисел в десятичное число с плавающей точкой, а затем в двоичное число с плавающей точкой. Преобразование из постулированного числа в десятичное число с плавающей точкой делится на четыре этапа [19]: извлечение бит знака (*bit sign*), извлечение бит режима (*bit regime*), извлечение бит степени (*exponent*), извлечение бит дробной части (*fraction*). Самым старшим битом в постулированной системе является бит знака. Он определяет, будет ли число положительным или отрицательным. Бит режима представлен унарной арифметикой [19]. Поэтому, при извлечении значения бит режима ($r_{value}$), алгоритм начинает подсчитывать количество последовательных единиц или нулей после знакового бита, пока не достигнет противоположного значения (ноль или единица соответственно). Затем результат отрицается, если подсчитанные биты являются нулями или уменьшается на 1, если подсчитанные биты были единицами. Биты степени представлены целым числом без знака, и поэтому легко извлекаются из битовой строки постулированной системы. Остальные биты в битовой строке являются битами дробной части. Десятичное число с плавающей точкой преобразуется в двоичное число с плавающей точкой путем деления

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

или умножения на 2 до тех пор, пока число не будет находиться в диапазоне [1, 2) [19].

В DCNNs признаки извлекаются с использованием сверточных слоев. Признаки представлены вектором $F_n = (f_1, f_2, \dots f_m)$ где $n$ это количество изображений в наборе данных, а $m$ показывает размерность вектора признаков. Затем признаки классифицируются сетью полносвязных слоев. В последнем слое слой softmax используется в качестве классификатора для минимизации $y - f^*(x, w)$ где $y$ определяет метку (*label*), $x$ обозначает вход слоя *softmax*, $w$ обозначает веса в слое *softmax*, а $f$ - наилучшая аппроксимационная функция [7].

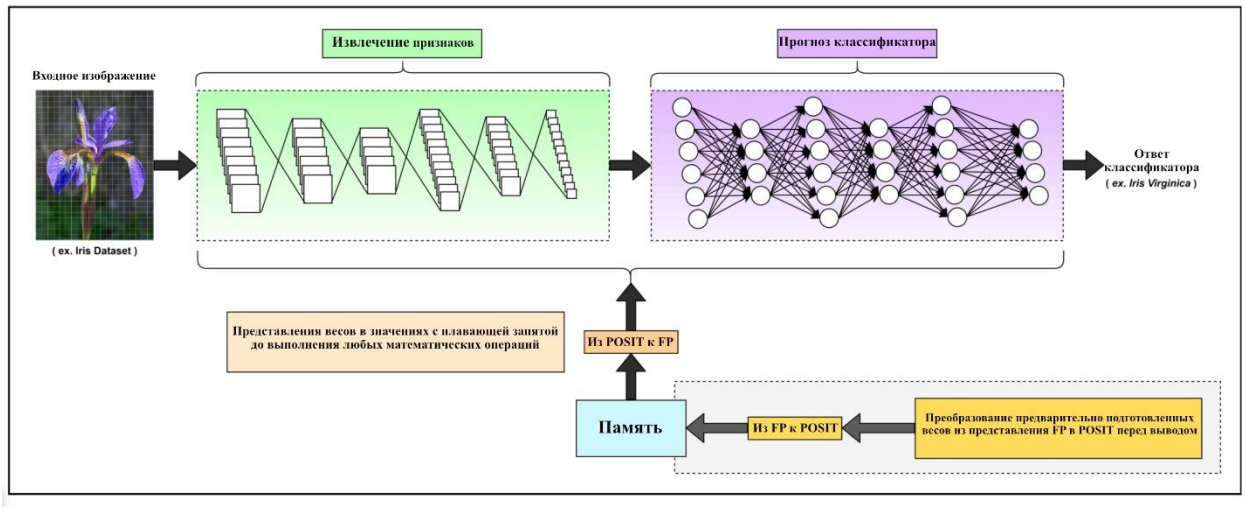



Рисунок 2 – Схема DCNN с использованием постулированной системы чисел для представления весов

Преобразование из системы с плавающей точкой в постулированную систему счисления состоит из двух этапов [19]: преобразование двоичной системы чисел с плавающей точкой в десятичную систему чисел с плавающей точкой; преобразование системы двоичных чисел с плавающей запятой в постулированную систему чисел. Первое преобразование

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

выполняется путем умножения дробной части (*fraction*) на два, в степени экспоненты (*exponent*). Затем, десятичное число с плавающей точкой, полученное на первом шаге, преобразуется в постулированное число путем деления или умножения на два, пока оно не окажется в диапазоне [1, *useed*], чтобы найти бит режима. Затем, этот процесс продолжается до тех пор, пока число не окажется в диапазоне [1, 2), чтобы найти показатель степени. Остальные биты - это дробная часть (*fraction*) [19].

**4. Оценка**

Разработанный подход оценивается по трем наборам данных: набор данных MNIST; набор данных CIFAR-10; и подмножество набора данных ImageNet. Набор данных MNIST, содержащий рукописные цифры, и другие наборы данных собираются для оценки эффективности новых методов распознавания образов. Для каждого набора данных используются разные DCNN, а также контрольная реализация DCNN с применением только системы счисления с плавающей точкой. Реализации используют Keras API [28]. Результаты точности приведены в таблице 1.

Таблица 1 – Точность трех различных нейронных сетей

| Задача | Набор данных | Обучающий набор | Сеть | Слои | Точность |
|---|---|---|---|---|---|
| Классификация цифр | MNIST | 10000 | LeNet | 2 Conv и 2 FC | 99.03% |
| Классификация изображений | CIFAR-10 | 10000 | Convnet | 3 Conv и 2 FC | 68.45% |
| Классификация изображений | ImageNet | 10000 | AlexNet | 5 Conv и 3 FC | 55.45% |

В этой работе веса сети представлены системой с фиксированной точкой переменной длины (с максимальной длиной бита 16 бит) и 8-битной постулированной системы счисления. Для представления весов в

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

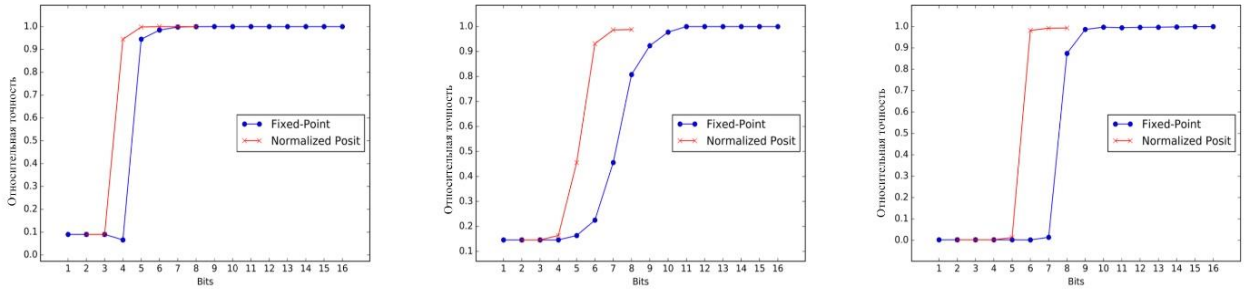
переменной системе с фиксированной запятой учитывается только один бит для целочисленной части, а дробная часть (*fractional*) изменяется в диапазоне [0,15] бит, так как большая часть весов в DCNNs находится в интервале [ -1, 1]. Для представления весов в постулированной системе счисления был выбран формат $P_{(i, 0)}$, где i изменяется в пределах диапазона [2,8]. Обратите внимание, что степень (*exponent*) равна нулю. Причина этого выбора заключается в том, что динамический диапазон постулированной числовой системы с нулевым показателем является ближайшим приближением к динамическому диапазону весов нейросети по сравнению с другими возможными вариантами значения степени (*exponent*). Среди этих форматов $P_{(2,0)}$ имеет наименьший динамический диапазон ([-1,1]), в то время как другие форматы имеют больший динамический диапазон. Однако, системы с фиксированной точкой и переменной длиной имеют тот же динамический диапазон. Поэтому, мы нормируем все форматы в постулированной системе чисел и получаем нормированную систему постулированных чисел. В этой версии постулированной системы чисел все форматы имеют одинаковый динамический диапазон [-1, 1]. Результаты относительной точности для разных задач показаны на рисунке 3.

Нормированная постулированная система чисел превосходит систему с фиксированной точкой с точки зрения точности при меньшем количестве бит. Результаты показывают, что при постулированной системе чисел LeNet, ConvNet и AlexNet можно использовать в 5, 7 и 7 битовом представлении соответственно, с ухудшением точности менее 1% по сравнению с 7, 11 и 9 бит соответственно при использовании системы с фиксированной точкой переменной длиной. Это уменьшает использование памяти на 28,6%, 36,4% и 23% [22], [29], а также может значительно

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

уменьшить количество обращений к памяти. Результаты были достигнуты без использования квантования или переобучения DCNN.




а) б) в)

Рисунок 3 – Относительная точность DCNN на различных наборах данных при представлении весов с использованием фиксированной точки переменной длины (Fixed-Point) и нормализованной постулированной системы (Normalized Posit). (а) Относительная точность для LeNet на наборе данных MNIST, (б) относительная точность ConvNet на наборе данных Cifar-10, (в) относительная точность AlexNet на наборе данных ImageNet.

**5. Заключение**

В этой работе было проведено исследование применения постулированной системы чисел для представления весов трех DCNNs, обученных на наборах данных MNIST, Cifar-10 и ImageNet. Нормализованная постулированная система чисел превосходит систему чисел с фиксированной точкой с точки зрения точности, при меньшем количестве бит, используемых для представления весов. В результате сокращается количество обращений к памяти, необходимое для передачи весов, а также общее потребление энергии для одной и той же задачи. В будущем будет исследовано представление данных с низкой точностью с использованием постулированной системы чисел и реализована DCNNs с использованием системы чисел, как для хранения, так и для вычисления.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

## Список использованной литературы:

1. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, C. 436–444, 2015.
2. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 1, no. 2, C. 3, 2017.
3. M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vie´gas, M. Wattenberg, G. Corrado et al., “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” Transactions of the Association for Computational Linguistics, vol. 5, C. 339–351, 2017.
4. D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen et al., “Deep speech 2: End-to-end speech recognition in english and mandarin,” in International Conference on Machine Learning, C. 173–182, 2016.
5. S. Gu, E. Holly, T. Lillicrap, and S. Levine, “Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates,” in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, C. 3389–3396, 2017.
6. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, C. 115–118, 2017.
7. Deep Learning. MIT Press [Электронный ресурс]. – Режим доступа: http://www.deeplearningbook.org, свободный. Дота обращения: 21.01.2018
8. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, C. 770–778. 2016.
9. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis., vol. 115, no. 3, C. 211–252, 2015.
10. P.Goyal,P.Dolla´r,R.Girshick,P.Noordhuis,L.Wesolowski,A.Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” arXiv preprint arXiv:1706.02677, 2017.
11. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., “Mastering the game of go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, C. 484–489, 2016.
12. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-L. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. Mackean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, “In Datacenter Performance Analysis of a Tensor Processing Unit TM,” C. 1–17, 2017.
13. D. Li, X. Wang, and D. Kong, “Deeprebirth: Accelerating deep neural network execution on mobile devices,” arXiv preprint arXiv:1708.04728, 2017.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

14. S. Kodali, P. Hansen, N. Mulholland, P. Whatmough, D. Brooks, and G. Y. Wei, "Applications of deep neural networks for ultra low power iot," in 2017 IEEE 35th International Conference on Computer Design (ICCD). IEEE, С. 589–592, 2017.
15. P. Judd, J. Albericio, T. Hetherington, T. M. Aamodt, N. E. Jerger, and A. Moshovos, "Proteus: Exploiting numerical precision variability in deep neural networks," in Proceedings of the 2016 International Conference on Supercomputing. ACM, С. 23, 2016.
16. P. Gysel, M. Motamedi, and S. Ghiasi, "Hardware-oriented Approximation of Convolutional Neural Networks," Iclr, С. 8, 2016.
17. I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," arXiv:1609.07061, 2016.
18. A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," arXiv preprint arXiv:1711.05852, 2017.
19. J. L. Gustafson and I. T. Yonemoto, "Beating floating point at its own game: Posit arithmetic," Supercomputing Frontiers and Innovations, vol. 4, no. 2, С. 71–86, 2017.
20. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner,"Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, С. 2278–2324, 1998.
21. B. Graham, "Fractional max-pooling," arXiv preprint arXiv:1412.6071, 2014.
22. P. Judd, J. Albericio, T. Hetherington, T. Aamodt, N. E. Jerger, R. Urtasun, and A. Moshovos, "Reduced-precision strategies for bounded memory in deep neural nets," arXiv preprint arXiv:1511.05236, 2015.
23. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
24. Q. C. Zhaoxia Deng, Cong Xu and P. Faraboschi, "Reduced-precision memory value approximation for deep learning," HP, December 2015.
25. M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," arXiv preprint arXiv:1602.02830, 2016.
26. F. Li, B. Zhang, and B. Liu, "Ternary weight networks," arXiv preprint arXiv:1605.04711, 2016.
27. R. Morris, "Tapered floating point: A new floating-point representation," IEEE Transactions on Computers, vol. 100, no. 12, С. 1578–1579, 1971.
28. Keras: Deep Learning [Электронный ресурс]. – Режим доступа: https://github.com/keras-team/keras, свободный. Дата обращения: 04.03.2018.
29. P. Judd, J. Albericio, T. Hetherington, T. Aamodt, N. E. Jerger, R. Urtasun, and A. Moshovos, "Proteus: Exploiting precision variability in deep neural networks," Parallel Computing, 2017.

***Дата поступления в редакцию: 20.06.2018 г.***

***Опубликовано: 25.06.2018 г.***