УДК 004.008 **DOI** 10.15622/ia.25.1.6

Н.П. Колмаков, А.Н. Голубинский

ОЦЕНКА ВЛИЯНИЯ БИТНОСТИ ЧИСЕЛ С ПЛАВАЮЩЕЙ ЗАПЯТОЙ НА ТОЧНОСТЬ РАСПОЗНАВАНИЯ ДИКТОРОВ

Колмаков Н.П., Голубинский А.Н. **Оценка влияния битности чисел с плавающей запятой на точность распознавания дикторов.**

Аннотация. В статье проводится анализ изменения точности распознавания личности по голосу при выделении разного количества бит на число с плавающей запятой (квантование) выходного тензора нейронной сети. Тензор характеризирует скрытое пространство нейронной сети, которое содержит скрытые признаки, используемые при решении задачи распознавания дикторов. Обычно, на каждое число выходного пространства выделяется тридцать два бита (выходной тензор, исследуемых методов содержит 512 чисел), поэтому для поддержки постоянно актуализируемой базы данных требуется большое количество памяти. Из-за этого, особый интерес представляет тип чисел с плавающей запятой – minifloat, позволяющий работать с численным представлениями, на которые выделяются восемь, шесть или четыре бита. Для обеспечения полноты результатов исследования, выбраны три нейросетевых решения, показывающие лучшие результаты распознавания на тестовой выборке: CAM++, WavLM, ReDimNet. Модели обладают уникальными архитектурными особенностями, что позволяет оценить изменение точности распознавания дикторов при уменьшении битности в зависимости от используемого типа архитектуры нейронной сети. Точность распознавания оценивается с помощью точки пересечения ошибок первого и второго рода. При проведении оценки точности распознавания используется англоязычный набор данных VoxCeleb-1, по характеристикам содержащихся аудиозаписей соответствует небольшой базе данных биометрической системы. Актуальность представленного материала обусловлена возрастающим количеством научных работ, которые предлагают использовать голос в качестве верификационного ключа. Поэтому, при работе с большим набором биометрических данных необходимо выделять большие объёмы памяти как на жёстких дисках, так и ОЗУ. Современные базы данных постоянно актуализируются и расширяются, что приводит к увеличению необходимых ресурсов на её поддержку. Одним из возможных методов решения может являться применение операции квантования к выходному тензору нейронной сети. Однако, преждевременное уменьшение количества выделяемых бит на число в выходном тензоре может привести к значительному ухудшению качества распознавания, относительно базовой версии сети. Основным направлением исследования является минимизация ресурсов для поддержки биометрической системы без дополнительного обучения нейронной сети.

Ключевые слова: распознавание дикторов, нейронные сети, числа с плавающей запятой, квантование.

1. Введение. Решением задачи распознавания дикторов исследователи занимаются не одно десятилетие. Одним из основных критериев, предъявляемых биометрической системе, является выставление оптимального значения решающего порога, который обеспечивает низкие значения ошибок первого или второго рода [1]. В последнее время для решения задачи активно используются методы, базирующиеся на технологии машинного обучения, что негласно

ставит дополнительный критерий – проводить распознавание при минимизации ресурсов на поддержку базы данных и выполнения вычислений. Одним из возможных методов решения проблемы является выделение меньшего количества бит на числа выходного тензора нейронной сети (эмбеддинг) [2]. Перед проведением математических операций с числами необходимо их преобразовать в бинарный вид. Общепринятая методика описана в стандарте IEEE 754 [3], однако, существуют такие методы [4, 5], которые отличаются от стандарта (поэтому в них могут отсутствовать бесконечности и NaN, тогда максимальным или минимальным преобразованным числом будет являться верхнее или нижнее допустимое значение доступного диапазона). При формировании чисел выделяется три типа бит: S – знак (0 – если число положительное, 1 – если число отрицательное); E – экспонента (смещённая экспонента двоичного числа); M – мантисса (остаток мантиссы двоичного нормализированного числа). Общий математический способ преобразования десятичного числа в бинарный вид [3]:

$$F = (-1)^S \times 2^{E-2^{(b-1)}+1} \times \left(1+\frac{M}{2^n}\right), \quad (1)$$

где b – количество бит, выделяемых на экспоненту; n – количество бит, выделяемых на мантиссу.

В машинном обучении используются различные типы чисел с плавающей запятой: float32, float16 и bloat16, охватывающие большой диапазон чисел. Однако, появление моделей машинного обучения количество параметров, которых насчитывает десятки и сотни миллиардов привело к использованию более мощных вычислительных ресурсов и к увеличению времени обучения, что эквивалентно удорожанию всего процесса исследования. Использование minifloat – тип числа с плавающей запятой, на который выделяется восемь, шесть или четыре бита, может быть, одним из способов удешевления процесса. Математический вид преобразования числа в minifloat представлен в (2). При наличии значащих бит в экспоненте [6]:

$$F = (-1)^S \times 2^{E-B} \times \left(1+2^{-m} \times M\right), \quad (2)$$

где B – количество бит, выделенных на экспоненту в эмбеддинге; m – количество бит, выделенных на мантиссу в эмбеддинге.

При отсутствии значащих бит в экспоненте [6]:

$$F = (-1)^S \times 2^{1-B} \times \left(0 + 2^{-m} \times M\right). \quad (3)$$

Операция квантования может применяется в процессе обучения, что позволяет быстрее обучать модель. Нередко квантование применяют в процессе инференса, что позволяет запускать модели с большим количеством параметров на менее производительных аппаратных средствах, относительно используемых на этапе обучения. В задаче распознавания дикторов используются модели с количеством обучаемых параметров исчисляемыми десятками или сотнями миллионов. Выходной последовательностью рассматриваемых нейронных сетей является одномерный тензор с длиной 512, в котором на каждое число выделяется тридцать два бита. В работах [7 – 9] исследуется возможность сохранения точности распознавания дикторов при сжатии модели путём квантования весов нейронной сети. Результаты, представленные в работах, показывают, что для сохранения исходных значений качества распознавания, квантованные сети проходят через этап дополнительного обучения. Подобный подход ограничивает количество доступных для использования методов, т.к. не все они обладают инструкцией или необходимым кодовым сопровождением для проведения обучения. Поэтому, в статье рассмотрен вариант оценки изменения значений ошибок первого и второго рода при уменьшении количества бит, выделяемых на число выходного пространства сети, описывающего внутреннее представление нейронной сети. Использование чисел с меньшим количеством бит позволит существенно сократить объём базы данных с информацией о дикторах, что в свою очередь удешевляет поддержку биометрической базы данных. Целью работы является оценка SOTA нейросетевых методов распознавания дикторов и исследование возможности минимизации ресурсов, требуемых для постоянной поддержки биометрической системы без дополнительного и трудозатратного обучения нейронной сети.

2. Используемые типы чисел с плавающей запятой. Для работы с внутренним представлением нейронной сети используются несколько типов чисел с плавающих запятой.

Float32 (одинарная точность / полная точность) – тип числа с плавающей запятой, определён в стандарте IEEE 754, который

содержит восемь бит на экспоненту и двадцать три бита на мантиссу. Диапазон чисел с плавающей запятой находится на отрезке $\pm 3,4 \times 10^{38}$.

Float16 (полуточность) – тип числа с плавающей запятой, определён стандарте IEEE 754, который содержит пять бит на экспоненту и десять бит на мантиссу. Диапазон чисел, помещающихся в этот тип, находится на отрезке $\pm 6,55 \times 10^{4}$.

BFloat16 (brain floating point) – тип числа с плавающей запятой, представлен для эффективного вычисления на TPU [10], который содержит восемь бит на экспоненту и семь бит на мантиссу. Диапазон чисел для экспоненты практически эквивалентен одинарной точности.

Float8 (E5M2 и E4M3) – тип числа с плавающей запятой, minifloat, представлен в [6, 10, 11]. Для E5M2 – диапазон чисел $\pm 5,73 \times 10^{4}$. Для E4M3 – диапазон чисел $\pm 4,48 \times 10^{2}$.

Float6 (E3M2, E2M3) – тип чисел с плавающей запятой, minifloat, описан в [10]. E3M2 – диапазон чисел ± 28. E2M3 – диапазон чисел $\pm 7,5$.

Float4 (E2M1) – тип чисел с плавающей запятой, minifloat описан в [6]. Принимает пятнадцать значений на отрезке ± 6. [-6; -4; -3; -2; -1,5; -1; -0,5; 0; 0,5; 1; 1,5; 2; 3; 4; 6].

На основе формул 1-3, представлен пример преобразования числа из полной точности в minifloat с выделением 4 бит на число (E=2). Возьмём число $4,5$ (находится в допустимом диапазоне чисел float4), в бинарной полной точности представляется следующим образом $0100,1_2$. Очевидно, что исходное число находится между двумя допустимыми значениями: 4 (B=2, m=0) и 6 (B=2, m=1). Поскольку, разность допустимых чисел и исходным составляет 0,5 и 2, соответственно, тогда исходное число примет то значение, которое обладает наименьшей разницей с ним, т.е. 4. По аналогии, число 5 будет так же преобразовано в число 4; другое число 5,1 будет преобразовано в 6. Остальные положительные числа, которые выходят за диапазон допустимого значения так же будут преобразованы в 6.

3. Тестовый набор данных. Оценка влияния битности эмбеддинга нейронной сети на изменение ошибок первого и второго рода, происходит на англоязычном наборе данных VoxCeleb - 1 [12] – содержит более ста тысяч аудиофайлов для семи тысяч дикторов, длина записи варьируется от 3 до 10 секунд, частота дискретизации записей 16 кГц с разрядностью 16 бит. Для проведения процесса верификации, разработчиками датасета составлены пары

аудиозаписей, в котором сочетаются аудиофайлы диктора самим с собой или с другими. Этот набор данных состоит из трёх наборов VoxCeleb-E – содержит 37611 пару записей без фонового шума, VoxCeleb-O – содержит 550894 пары записей без фонового шума, VoxCeleb-H – содержит 579818 пар записей с фоновым шумом, который аугментирован шумами разного рода. В таблице 1 представлены данные о количестве памяти, требуемой для хранения внутренних представлений сети для всех пар дикторов каждого набора данных при использовании различных типов чисел. В скобках указано число, показывающее на сколько, уменьшилось количество требуемых МБ на хранение квантованной базы данных относительно полной точности.

Таблица 1. Объём наборов данных VoxCeleb-1 при разном типе чисел с плавающей запятой

	Наборы данных		
	VoxCeleb-E, (МБ)	**VoxCeleb-O, (МБ)**	**VoxCeleb-H, (МБ)**
Float32	73,46	1075,96	1132,46
Float16	36,73 (↓36,73)	537,98 (↓537,98)	566,23 (↓566,23)
Float8	18,36 (↓55,09)	268,99 (↓806,97)	283,11 (↓849,34)
Float6	13,77 (↓59,69)	201,74 (↓874,22)	212,34 (↓920,12)
Float4	9,18 (↓64,28)	134,50 (↓941,47)	141,56 (↓990,90)

4. Методика оценки. Для оценки схожести между выходными тензорами нейронной сети до и после уменьшения количества бит используется метрика косинусного сходства:

$$\cos(\alpha) = \frac{(\vec{a}, \vec{b})}{|\vec{a}| \times |\vec{b}|}, \tag{4}$$

где $\cos(\alpha)$ – коэффициент косинусного сходства; $\vec{a}$, $\vec{b}$ – векторное представление двух аудиозаписей; $|\vec{a}|$, $|\vec{b}|$ – евклидова норма векторов; $(\vec{a}, \vec{b})$ – скалярное произведение.

Оценка изменений показателей коэффициентов косинусного сходства осуществляется путём вычисления ошибок первого и второго рода и равновероятной ошибки. Математическое определение представлено:

$$EER|_{FAR=FRR} = FAR = FRR;$$
$$FAR = \frac{FP}{FP + TN}; \quad (5)$$
$$FRR = \frac{FN}{FN + TP},$$

где EER – равновероятная ошибка, пересечение ошибок первого и второго рода; FAR – ошибка первого рода; FRR – ошибка второго рода; FP – ложно принятые записи; TP – верно принятые записи; TN – верно отвергнутые записи; FN – ложно отвергнутые записи.

При использовании различных типов чисел в выходном тензоре модели, значения ошибок первого и второго рода позволяют оценить тенденцию изменений значений метрики схожести относительно базовой версии сети. На рисунке 1 представлена схема, описывающая процесс оценки распознавания дикторов, которая используется в рамках исследования.

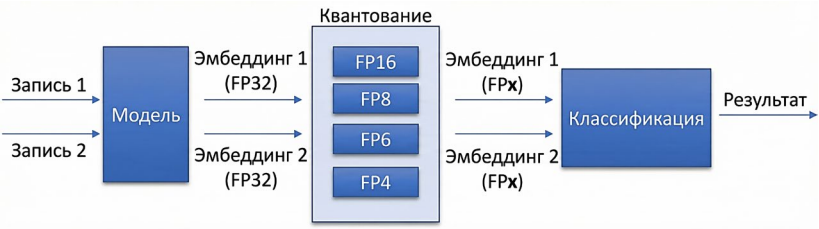


Рис. 1. Распознавание дикторов при разном типе чисел с плавающей запятой в эмбеддинге

Результатом классификации является коэффициент косинусного сходства, при сравнении со значением порога принимается решение о соответствии двух записей к одному и тому же диктору.

5. Context Aware Masking. CAM++ [13] – свёрточная нейронная сеть, которая содержит 7,2 миллиона параметров. Относится к моделям, которые работают с представлением диктора на базе x-вектора [14]. Для извлечения признаков из внутреннего пространства, десятисекундной записи, модели требуется 264,43 МБ видеопамяти. Опубликованная версия сети обучена на англоязычном наборе данных VoxCeleb-2 [15]. Разработчиками выложен код и необходимые инструкции для проведения дополнительного

обучения, однако доступное кодовое сопровождение предназначено для работы только с одним набором данных VoxCeleb-2.

Перед извлечением признаков, входной сигнал преобразовывают в блок фильтров с размером окна 25 мс и шагом 10 мс. Сформированный блок фильтров подаётся на модуль, состоящий из четырёх пропускающих блоков, которые используются для извлечения признаков по времени и частоте. Полученные признаки поступают на три последовательных модифицированных D-TDNN [16] блока. Каждый блок состоит из двух внутренних модулей CAM [17] и TDNN.

Для проведения оценки изменения ошибок первого и второго рода, произведён анализ распределений косинусных коэффициентов схожести между выходными тензорами нейронной сети до и после выделения различного количества бит на число. Первичная оценка влияния битности внутренних представлений осуществляется путём визуального анализа изменения графиков для разных типов чисел относительно полученных результатов для полной точности. Более широкие области графика соответствуют большему количеству коэффициентов косинусного сходства. Полученные результаты представлены на рисунке 2.

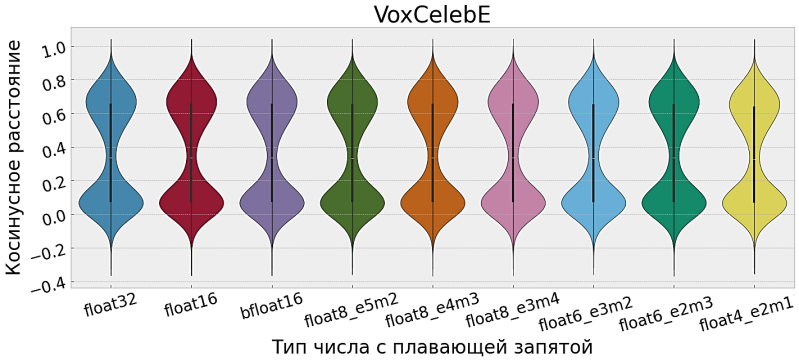


Рис. 2. Изменение разброса косинусного сходства при квантовании эмбеддинга CAM++

Из графика, представленного на рисунке 2, следует, что выделение меньшего количества бит на число выходного пространства CAM++ незначительно изменило общее распределение косинусных коэффициентов сходства. Полученные результаты позволяют сделать заключение об устойчивости эмбеддинга сети при проведении процедуры квантования. Подтверждением полученных выводов

являются данные, показатели ошибок первого и второго рода при значении решающего порога, представленного разработчиками метода, равного 0,381 [13]. Полученные результаты представлены в таблице 2. Возле каждого значения ошибки в таблице будет предоставлено условное обозначение: ↓ – если получившийся показатель улучшился, по сравнению полученными в полной точности, что привело к уменьшению ошибки; ↑ – если получившийся показатель ухудшился, по сравнению полученным в полной точности, что привело к увеличению метрики; или = – значит, что показатели метрик совпадают для обоих типов чисел и значения ошибки не изменился.

Таблица 2. Изменение ошибок первого и второго рода при квантовании внутреннего пространства CAM++

	Наборы данных					
	VoxCeleb-E		**VoxCeleb-O**		**VoxCeleb-H**	
	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)
Float32	0,93	3,61	1,39	2,94	3,86	3,47
Float16	0,93 (=)	3,61 (=)	1,39 (=)	2,94 (=)	3,86 (=)	3,47 (=)
BFloat16	0,93 (=)	3,61 (=)	1,39 (=)	2,94 (=)	3,86 (=)	3,47 (=)
Float8 e5m2	0,91 (↓0,02)	3,66 (↑0,05)	1,36 (↓0,03)	2,97 (↑0,03)	3,79 (↓0,02)	3,53 (↑0,08)
Float8 e4m3	0,93 (=)	3,62 (↑0,01)	1,37 (↓0,02)	2,95 (↑0,01)	3,84 (↓0,02)	3,49 (↑0,02)
Float8 e3m4	0,94 (↑0,01)	3,61 (=)	1,39 (=)	2,93 (↓0,01)	3,86 (=)	3,48 (↑0,01)
Float6 e3m2	0,91 (↓0,02)	3,67 (↑0,06)	1,36 (↓0,03)	2,97 (↑0,03)	3,79 (↓0,07)	3,53 (↑0,08)
Float6 e2m3	0,92 (↓0,01)	3,63 (↑0,02)	1,37 (↓0,02)	2,96 (↑0,02)	3,83 (↓0,03)	3,51 (↑0,04)
Float4 e2m1	0,80 (↓0,13)	4,10 (↑0,49)	1,19 (↓0,20)	3,29 (↑0,35)	3,37 (↓0,02)	3,95 (↑0,52)

Из данных представленных в таблице 2 следует, что значения ошибок первого и второго рода незначительно меняются, приблизительно, на 0,03% и 0,05% при использовании типов данных с восьмью или шестью битами; на 0,2% и 0,52% при выделении четырёх бит на число. Для получения детальной информации о влиянии квантования внутренних представлений нейронной сети на результаты распознавания дикторов, необходимо определить значения решающих порогов через нахождение равновероятной ошибки. Было замечено, что при оценке EER, обычно, графики ошибок первого и второго рода квантованных эмбеддингов практически совпадают с результатами,

полученными при использовании полной точности. Поэтому, в статье представлены только те результаты, в которых отклонение графиков ошибок визуально заметны. На рисунке 3 представлены графики пересечений ошибок первого и второго рода при использовании типа чисел данных float4-e2m1 и float32.

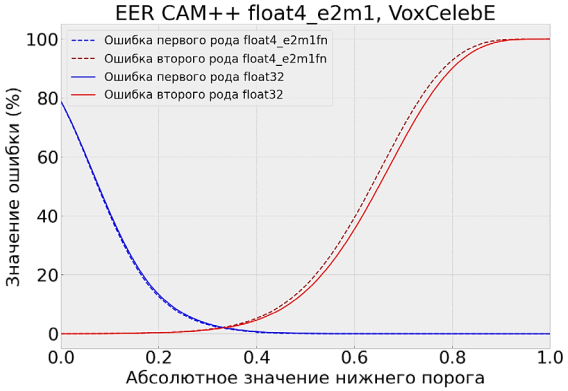


Рис. 3. Пересечение ошибок первого и второго рода при выделении четырёх бит на число для эмбеддинга CAM++

Из результатов, представленных на графике рисунка 3 следует, что при использовании float4-e2m1 точка пересечения ошибок первого и второго рода совпадает с точкой, полученной при полной точности.

6. Reshape Dimensions Network. ReDimNet [18] – свёрточная нейронная сеть, отличительной особенностью, которой является извлечение уникальных речевых признаков диктора при комбинировании двух типов свёрточных слоёв: одномерного и двумерного. Модель содержит 5 миллионов параметров. Для извлечения признаков из внутреннего пространства, десятисекундной записи, требуется 4615,97 МБ видеопамяти. Потребление такого большого количества памяти по соотношению к малому количеству параметров, происходит из-за использования шести блоков многоголового внимания на каждый из них приходится по 663 МБ (используются тензоры с высокой размерностью в качестве основных компонентов внимания: ключ, значение, запрос). Опубликованная версия модели обучена на нескольких наборах данных VoxBlink-2 [19] и VoxCeleb-2. Разработчиками метода не выложена необходимая инструкция для обучения модели на новом языковом домене или наборе данных. На рисунке 4 представлен график для набора данных VoxCeleb-E в зависимости от количества выделенных бит на число в эмбеддинге.

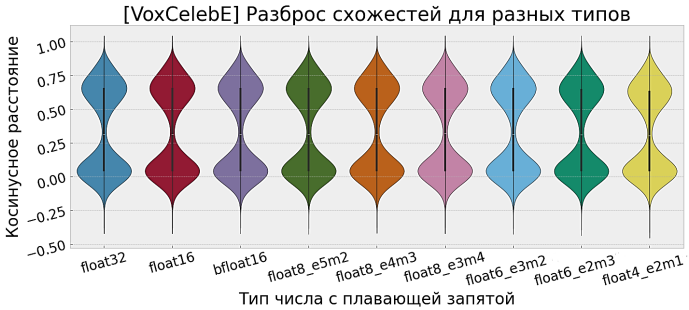


Рис. 4. Изменение разброса коэффициентов косинусного сходства при квантовании эмбеддинга ReDimNet

Из графика, представленного на рисунке 4, следует, что распределение коэффициентов косинусного сходства при выделении меньшего количества бит схоже с распределением при выделении тридцати двух бит. В таблице 3 представлены значения ошибок первого и второго рода при значении решающего порога равного 0,351, получен в ходе исследования.

Таблица 3. Изменение ошибок первого и второго рода при переквантовании внутреннего пространства ReDimNet

	Наборы данных					
	VoxCeleb-E		**VoxCeleb-O**		**VoxCeleb-H**	
	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)
Float32	1,57	0,41	2,59	0,19	4,97	0,42
Float16	1,57 (=)	0,41 (=)	2,52 (↓0,07)	0,19 (=)	4,97 (=)	0,42 (=)
BFloat16	1,57 (=)	0,41 (=)	2,61 (↑0,02)	0,19 (=)	4,97 (=)	0,42 (=)
Float8 e5m2	1,56 (↓0,01)	0,42 (↑0,01)	2,55 (↓0,04)	0,19 (=)	4,91 (↓0,06)	0,43 (↑0,01)
Float8 e4m3	1,57 (=)	0,41 (=)	2,58 (↓0,01)	0,18 (↓0,01)	4,96 (↓0,01)	0,43 (↑0,01)
Float8 e3m4	1,44 (↓0,13)	9,47 (↑9,06)	2,47 (↓0,12)	7,32 (↑7,13)	4,53 (↓0,44)	0,48 (↑0,06)
Float6 e3m2	1,56 (↓0,01)	0,42 (↑0,01)	2,55 (↓0,04)	0,19 (=)	4,91 (↓0,06)	0,43 (↑0,01)
Float6 e2m3	1,47 (↓0,10)	0,46 (↑0,05)	2,40 (↓0,19)	0,23 (↑0,04)	4,63 (↓0,34)	0,46 (↑0,04)
Float4 e2m1	1,32 (↓0,25)	0,56 (↑0,15)	2,17 (↓0,01)	0,34 (↑0,15)	4,17 (↓0,80)	0,55 (↑0,13)

По данным в таблице 3 видно, что изменение при выделении четырёх бит значение ошибки первого рода улучшилась на 0,32 %, значение ошибки второго рода ухудшилась на 0,15 %. В других случаях значения ошибок, практически, идентичны показателям ошибок, полученной в полной точности. На рисунке 5 представлен график пересечения ошибок первого и второго рода при выделении четырёх бит на число.

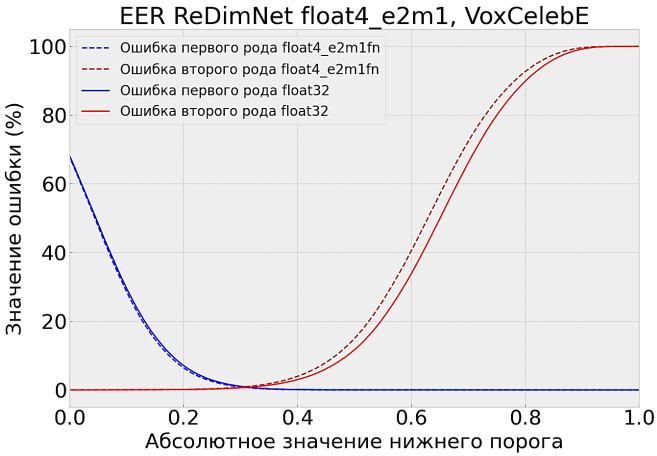


Рис. 5. Пересечение ошибок первого и второго рода при выделении четырёх бит на число для эмбеддинга ReDimNet

Из графика, представленного на рисунке 5, следует, что при выделении четырёх бит на число не приводит к сдвигу значения решающего порога, относительно базовой версии сети. Что соответствует использованию необходимого диапазона чисел каждого типа данных для сохранения исходных показателей качества распознавания.

7. WavLM. WavLM [20] – нейронная сеть, базирующая на архитектуре трансформер [21], содержит 100 миллионов параметров. Модель разработана для решения ряда задач: верификация дикторов, автоматическое распознавание речи, диаризация дикторов. Вторичной параметризацией речевого сигнала является блок фильтров с размером окна 25 мс и шагом 10 мс. Архитектурно модель делится на две сети: проецирующий вторичные признаки речевого сигнала во внутреннее пространство сети, состоит из семи свёрточных блоков (одномерная свёртка с функцией активации GELU [22]); аппроксиматор, который состоит из двенадцати блоков механизма внимания, архитектурно

является кодировщиком языковой модели. Для извлечения признаков из внутреннего пространства, десятисекундной записи, модели требуется 487,73 МБ видеопамяти. Рассматриваемая версия модели обучена на нескольких англоязычных наборах данных Libri-Light [23], GigaSpeech [24], VoxPopuli [25]. Разработчиками метода не выложена необходимая инструкция для обучения модели на новый языковой домен или набор данных.

Процедура оценки распределения коэффициентов косинусного сходства при разных квантованиях выходного представлений модели аналогична предыдущему разделу. На рисунке 6 представлен соответствующий график.

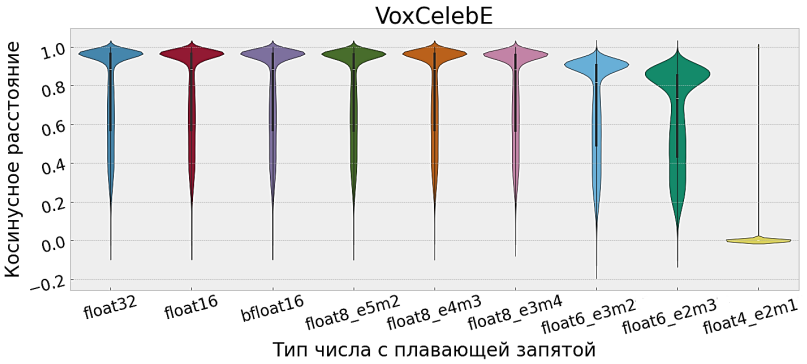


Рис. 6. Изменение разброса коэффициентов косинусного сходства при квантовании эмбеддинга WavLM

Из графика, представленного на рисунке 6, следует, что при выделении шести бит на число выходного тензора, наблюдается более выраженное распределение косинусных коэффициентов на отрезке от одного до нуля. При выделении четырёх бит на число большая часть коэффициентов косинусного сходства распределяется возле нуля, что свидетельствует об изменении положения точки пересечения ошибок первого и второго рода от 0,86 к 0,1. Подобное изменение распределения, позволяет сделать вывод об изменении решающего порога относительно базовой версии нейронной сети.

Информация об изменении ошибок первого и второго рода, при значении решающего порога, представленного разработчиками метода, 0,86 [20], продемонстрирована в таблице 4.

Таблица 4. Изменение ошибок первого и второго рода при квантовании внутреннего пространства WavLM

	Наборы данных					
	VoxCeleb-E		**VoxCeleb-O**		**VoxCeleb-H**	
	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)	FAR, (%)	FRR, (%)
Float32	4,57	1,10	6,23	2,96	2,01	1,04
Float16	4,57 (=)	1,10 (=)	6,23 (=)	2,96 (=)	2,01 (=)	1,04 (=)
BFloat16	4,57 (=)	1,10 (=)	6,23 (=)	2,95 (=)	2,01 (=)	1,04 (=)
Float8 e5m2	4,34 (↓0,17)	1,17 (↑0,07)	5,99 (↓0,24)	3,22 (↑0,26)	19,26 (↑17,25)	1,10 (↑0,06)
Float8 e4m3	4,10 (↓0,47)	1,12 (↑0,02)	6,14 (↓0,09)	3,04 (↑0,08)	19,90 (↑17,89)	1,05 (↑0,01)
Float8 e3m4	4,10 (↓0,47)	1,27 (↑0,17)	5,68 (↓0,55)	3,58 (↑0,62)	18,31 (↑15,35)	1,18 (↑0,14)
Float6 e3m2	0,05 (↓4,52)	9,60 (↑8,50)	1,10 (↓5,13)	22,83 (↑19,87)	3,20 (↑1,19)	9,20 (↑8,16)
Float6 e2m3	0,06 (↓4,51)	55,03 (↑54,07)	0,20 (↓6,03)	65,57 (↑62,61)	0,04 (↓1,97)	54,68 (↑53,64)
Float4 e2m1	0,04 (↓4,53)	97,55 (↑96,45)	0,0 (↓6,23)	99,63 (↑96,67)	0,01 (↓2,00)	97,54 (↑96,50)

По результатам, представленным в таблице 4, построены графики пересечения ошибок первого и второго рода при выделении четырёх бит на число эмбеддинга WavLM, рисунок 7.

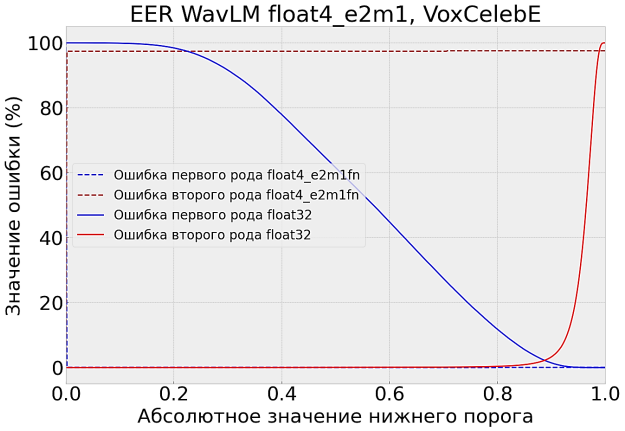


Рис. 7. Пересечение ошибок первого и второго рода при выделении четырёх бит на число для эмбеддинга WavLM

Из данных, представленных в таблице 4 и на рисунке 7, следует, что квантование выходного тензора WavLM во float4 привело к смещению пересечения ошибок первого и второго рода от 0,9 к 0. Подобный результат получен из-за использования недостаточного количества допустимых значений диапазона во float4, по сравнению с другими типами чисел. Однако, для CAM++ и ReDimNet не произошло сильных изменений относительно полной точности, поскольку используется необходимый диапазон значений в скрытом пространстве сети, для получения точности распознавания близкой к исходной версии сети.

Рассмотрим нормированный периодический сигнал с частотой дискретизации 16 кГц и длительностью две секунды, который содержит только два предельных значения -1 и 1. Синтезированный сигнал позволяет получить максимально допустимые значения выходного внутреннего пространства нейронной сети, соответствующие реальному речевому сигналу. При проведении анализа значений активаций нейросетевых моделей используются следующие показатели: максимального, минимального, среднего и стандартного отклонения. Структуры рассматриваемых сетей, на этапе обучения, можно обобщить тремя блоками: третичная параметризация, аппроксимация уникальных речевых признаков нейронной сетью, классификация (верификации дикторов из обучающей выборки). На этапе применения не используется блок классификации. Уникальные признаки речи, содержащиеся в скрытом пространстве нейронной сети, получают с одного из последних слоёв нейросетевого аппроксиматора. Несмотря на архитектурное различие сетей, у CAM++ и ReDimNet есть одна общая деталь, использование в составе последнего блока, слой нормализации мини-пакетов (batch normalization, BN) [26]. Поскольку, информация извлекается с последних слоёв аппроксиматора, целесообразно рассмотреть значения активаций до и после прохождения через слой BN. Дополнительно изучить показатели входного тензора.

Для входного тензора: 11,445; -15,942; 5,615; -12,437. Для CAM++, показатели в эмбеддинге равны: 0,179; -0,245; -0,002; 0,072 и 5,010; -6,509; -0,067; 1,979; до и после прохождение через слой BN, соответственно.

Для входного тензора: 1; -1; 0,333; 0,943. Для ReDimNet, показания в эмбеддинге равны: -3,639; 4,017; -0,229; 0,891 и -14,850; 14,575; 0,287; 4,875; до и после прохождение через слой BN, соответственно.

У WavLM в последних слоях не используется слой нормализации. Поэтому, рассмотрены значения активаций до и после прохождения через последний слой, значения которого используются для дальнейшей классификации. Для входного тензора: 1; -1; 0,943; 0,333. Для эмбеддинга получены следующие значения: 0; 0,274; 0,002; 0,014 и -0,316; 0,1790; -0,0289 до и после прохождения через последней слой сети.

По полученным результатам можно выдвинуть гипотезу об устойчивости квантованного эмбеддинга к квантованию в зависимости от составляющих блоков архитектуры нейронной сети. В состав архитектур CAM++ и ReDimNet входит ResNet блок, 4 и 17 соответственно. На рисунке 8 представлена структура блока.

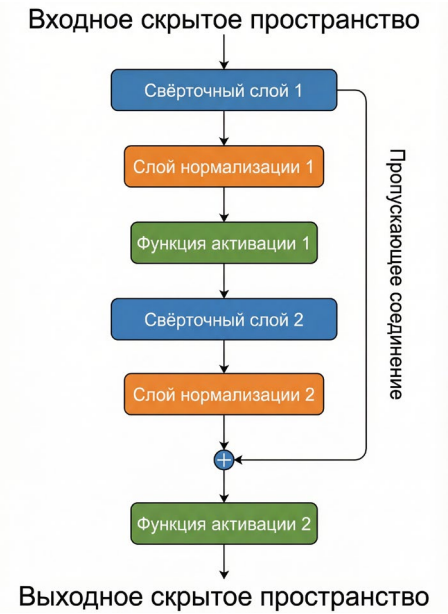
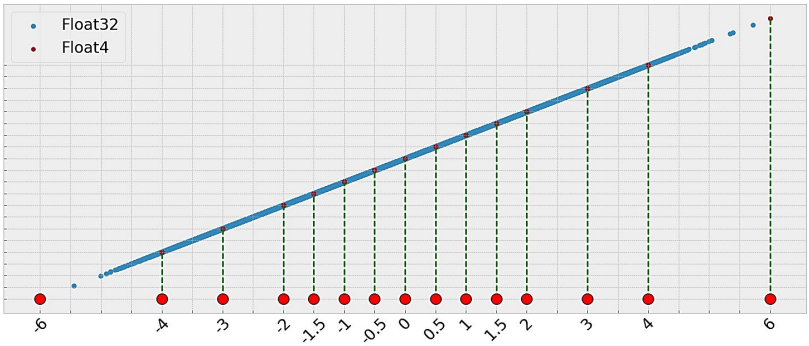


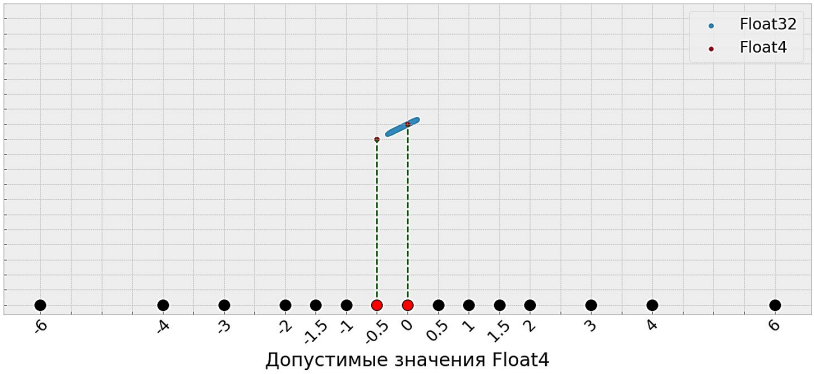
Рис. 8. Структура ResNet блока

Если исследуемая архитектура содержит ResNet блоки, то диапазон значений выходного тензора после прохождения через последний BN будет достаточен для применения квантования, например, float4 без потери качества относительно значений эмбеддинга в полной точности. В других случаях необходимо провести дополнительную процедуру оценки выходного тензора из основного аппроксимирующего блока сети.

Поскольку, полученные результаты характерны только синтезированному сигналу. Проведено исследование для десяти тысяч выходных скрытых пространств, соответствующие реальным речевым сигналам, моделей CAM++ и WavLM при выделении тридцати двух и четырёх бит на число эмбеддинга. Для исследования полученных результатов квантованных чисел построены графики со значениями чисел внутреннего пространства нейронных сетей для float32 и float4. На рисунке 9 в пункте а) – представлены результаты квантования для CAM++, в пункте б) – представлены результаты квантования для WavLM.



а) квантование эмбеддинга CAM++ во Float4



б) квантование эмбеддинга WavLM во Float4

Рис. 9. Распределение значений эмбеддингов а) для CAM++ б) для WavLM при выделении 4 бит на число

Из результатов графиков, представленных на рисунке 9, следует, что в квантованном эмбеддинге CAM++ может содержаться весь доступный диапазон чисел во float4. Из-за того, что значения эмбеддингов для WavLM в полной точности лежат на отрезке от -0,4 до 0,25. Поэтому, в квантованном во float4 эмбеддинге содержатся только два числа -0,5 и 0, при применении метрики косинусного расстояния в большинстве случаев будет получен нулевой коэффициент схожести, что не соответствует исходным показателям, полученным в базовой версии сети. Для расширения используемого диапазона допустимых значений float4 необходимо воспользоваться процедурой масштабирования каждого числа в выходном тензоре (E_s):

$$E_s = \frac{F_{\max}}{E_{\max}} \times E, \tag{6}$$

где $F_{\max}$ – максимальное допустимое значение во float4; $E_{\max}$ – максимальное значение в эмбеддинге; E – эмбеддинг WavLM (одномерный тензор с размерностью 512).

На рисунке 10 представлены графики пересечений ошибок первого и второго рода до и после проведения операции масштабирования.

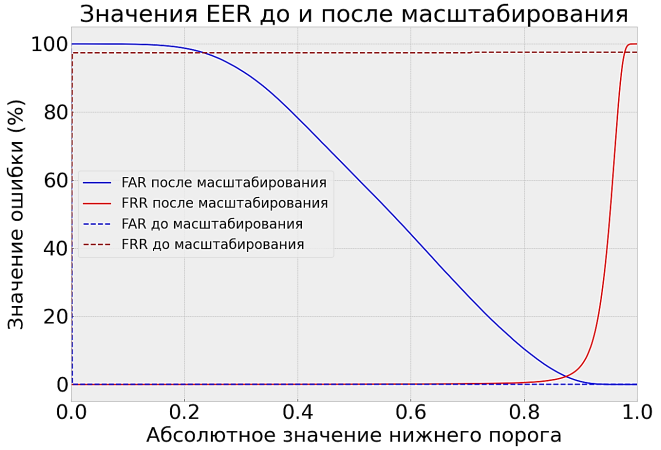


Рис. 10. Пересечение ошибок первого и второго рода до и после масштабирования для эмбеддинга WavLM

Из графика, представленного на рисунке 10, следует, что после применения операции масштабирования изменились значения ошибок первого и второго рода с 0,035% и 97,55 % до 3,5 % и 1,6 %, при исходном значении решающего порога 0,86. Что свидетельствует об использовании необходимого диапазона допустимых значений во float4, для достижения показателей ошибок близких к полученным в полной точности.

8. Заключение. В рамках статьи проведён анализ изменения ошибок первого и второго рода в задаче распознавания дикторов, при выделении разного количества бит числа с плавающей запятой выходного тензора нейронной сети. Исследованы три SOTA нейросетевых метода CAM++, WavLM и ReDimNet, которые обладают различными архитектурными особенностями. Значения ошибок первого и второго рода базируются на показателях косинусного сходства между эмбеддингами нейронных сетей. На основе полученных значений в таблицах 2-4 построен график, который представлен на рисунке 11.

Рис. 11. Значения True Positive, True negative, False Positive и False Negative в зависимости от количества выделенных бит на число

Из данных, представленных на рисунке 11, видно, что при использовании minifloat, площади покрытия метрик, влияющие на показатели ошибок первого и второго рода, практически, совпадают

с результатами, полученными при использовании полной точности. Стоит отметить, что для достижения подобных результатов у WavLM, при выделении четырёх бит на числа эмбеддинга, потребовалось произвести операцию масштабирования на максимальное допустимое значение float4. Результаты, полученные в ходе исследования, показывают, что значения решающего порога, полученные для эмбеддингов в полной точности инвариантны к их квантованию (разница между исходными значениями и квантованными, обычно, не превышает 0,05). Этот факт позволяет сэкономить время при исследовании новых нейросетевых решений. Дальнейшим направлением исследования является разработка нейросетевого метода, устойчивостью к использованию операции квантования к весам, активациям и выходному тензору сети, при этом сохраняя максимально близкие показатели ошибок первого и второго рода, получаемые при работе с полной точностью.

Литература

1. ГОСТ Р 54412-2019 (ISO/IEC TR 24741:2018). Информационные технологии. Биометрия. Общие положение и примеры применения // М.: Госстандарт России. 2019.
2. Morin F., Bengio Y. Hierarchical Probabilistic Neural Network Language Model // Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. 2005. pp. 246–252.
3. IEEE Standard for Floating-Point Arithmetic. IEEE Std 754TM-2019 (Revision of IEEE Std 754-2008). 2019. DOI: 10.1109/IEEESTD.2019.8766229.
4. Google Cloud Blog. BFloat16: The secret to high performance on Cloud TPUs // URL: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus (дата обращения: 28.06.2025).
5. NVIDIA TF32 – DeepRec latest documentation. URL: https://deeprec.readthedocs.io/en/latest/NVIDIA-TF32.html (дата обращения: 01.06.2025).
6. Rouhani B.D., et al. OCP Microscaling Formats (MX) Specification. Open Compute Project. 2023.
7. Liu B., Wang H., Qian Y. Towards Lightweight Speaker Verification via Adaptive Neural Network Quantization // IEEE/ACM Trans. Audio Speech Lang. Process. 2024. vol. 32. pp. 3771–3784.
8. Hong Y., Chung W.-J., Kang H.-G. Optimization of DNN-based speaker verification model through efficient quantization technique. arXiv preprint arXiv:2407.08991. 2024.
9. Wang H., Liu B., Wu Y., Chen Z., Qian Y. Lowbit Neural Network Quantization for Speaker Verification // IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Rhodes Island, Greece: IEEE, 2023. pp. 1–5.
10. Jouppi N.P., et al. In-Datacenter Performance Analysis of a Tensor Processing Unit // Proceedings of the 44th Annual International Symposium on Computer Architecture. New York, NY, USA: Association for Computing Machinery, 2017. pp. 1–12. DOI: 10.1145/3079856.3080246.

11. Micikevicius P., et al. FP8 Formats for Deep Learning. arXiv preprint arXiv:2209.05433. 2022.
12. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
13. Wang H., Zheng S., Chen Y., Cheng L., Chen Q. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2023. pp. 5301–5305. DOI: 10.21437/Interspeech.2023-1513.
14. Колмаков Н.П., Толстых А.А. Устойчивые признаки для распознавания дикторов на дискретных аудиосигналах. Сборник трудов XXIX Международной научно-технической конференции «Радиолокация, Навигация, Связь». 2023. С. 387–394.
15. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
16. Huang G., Liu Z., van der Maaten L., Weinberger K.Q. Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. С. 2261–2269. DOI: 10.1109/CVPR.2017.243.
17. Yu Y.-Q., Zheng S., Suo H., Lei Y., Li W.-J. Cam: Context-Aware Masking for Robust Speaker Verification // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021. pp. 6703–6707. DOI: 10.1109/ICASSP39728.2021.9414704.
18. Yakovlev I., et al. Reshape Dimensions Network for Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2024. pp. 3235–3239.
19. Lin Y., Cheng M, Zhang F, Gao Y, Zhang S, Li M. VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2024. pp. 4263–4267.
20. Chen S., et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing // IEEE J. Sel. Top. Signal Process. 2022. vol. 16. № 6. pp. 1505–1518.
21. Vaswani A., et al. Attention Is All You Need // Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017. pp. 6000–6010.
22. Hendrycks D., Gimpel K. Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415. 2023.
23. GitHub. facebookresearch/libri-light: dataset for lightly supervised training using the librivox audio book recordings. https://librivox.org/ // URL: https://github.com/facebookresearch/libri-light (дата обращения: 02.02.2025).
24. Chen G., et al. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio // Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2021. pp. 3670–3674. DOI: 10.21437/Interspeech.2021-1965.
25. Wang C., et al. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80.

26. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // Proceedings of the 32nd International Conference on International Conference on Machine Learning (PMLR). 2015. vol. 37. pp. 448–456.

Колмаков Никита Павлович — младший научный сотрудник, Института проблем передачи информации им. А.А. Харкевича Российской академии наук. Область научных интересов: цифровая обработка сигналов, машинное обучение. Число научных публикаций — 6. kolmakov.np@iitp.ru; Большой Каретный переулок, 19/1, 127051, Москва, Россия; р.т.: +7(495)650-4225.

Голубинский Андрей Николаевич — д-р техн. наук, доцент, Российский научный фонд. Область научных интересов: машинное обучение, нейросетевое моделирование, автоматизированные системы управления с элементами искусственного интеллекта, обработка речевых сигналов. Число научных публикаций — 250. annikgol@mail.ru; улица Солянка, 19-1, 109028, Москва, Россия; р.т.: +7(495)650-4225.

DOI 10.15622/ia.25.1.6

N. KOLMAKOV, A. GOLUBINSKIY

ASSESSING THE INFLUENCE OF FLOATING-POINTS BIT DEPTH ON SPEAKER RECOGNITION ACCURACY

Kolmakov N., Golubinskiy A. **Assessing the Influence of Floating-Points Bit Depth on Speaker Recognition Accuracy.**

Abstract. The article analyzes the impact of varying the bit depth (quantization) of a neural network's output tensor on speaker recognition accuracy. This tensor represents the neural network's latent space, containing the latent features utilized for speaker recognition tasks. Typically, thirty-two bits are allocated per value in the output space (the output tensors of the methods under study contain 512 values), resulting in significant memory requirements for maintaining a continuously updated database. Consequently, the "minifloat" floating-point format is of particular interest, as it enables numerical representations using only eight, six, or four bits. To ensure comprehensive results, three neural network models demonstrating superior recognition performance on the test set were selected: CAM++, WavLM, and ReDimNet. These models possess unique architectural characteristics, facilitating the assessment of how bit depth reduction affects recognition accuracy across different neural network architectures. Recognition accuracy is evaluated using the Equal Error Rate (EER). The evaluation employs the English-language VoxCeleb-1 dataset, the audio characteristics of which correspond to those of a small-scale biometric system database. The relevance of this study is underscored by the increasing volume of research proposing the use of voice as a verification key. Therefore, managing large biometric datasets requires substantial storage capacity and RAM. Modern databases are continuously updated and expanded, leading to increased resource demands for their maintenance. Applying quantization to the neural network's output tensor offers a potential solution. However, excessive reduction of the bit depth in the output tensor can lead to a significant degradation in recognition quality compared to the baseline network. The primary focus of this research is to minimize the resources required to support a biometric system without the need for additional neural network training.

Keywords: neural networks, speaker recognition, floating point, embedding quantization.

References

1. GOST R 54412-2019 (ISO/IEC TR 24741:2018). [Information Technology. Biometrics. General Position and Examples of Application] M.: Gosstandart Rossii. 2019. (In Russ.).
2. Morin F., Bengio Y. Hierarchical Probabilistic Neural Network Language Model. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. 2005. pp. 246–252.
3. IEEE Standard for Floating-Point Arithmetic. IEEE Std 754TM-2019 (Revision of IEEE Std 754-2008). 2019. DOI: 10.1109/IEEESTD.2019.8766229.
4. Google Cloud Blog. BFloat16: The secret to high performance on Cloud TPUs. Available at: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus (accessed 28.06.2025).
5. NVIDIA TF32 – DeepRec latest documentation. Available at: https://deeprec.readthedocs.io/en/latest/NVIDIA-TF32.html (accessed 01.06.2025).
6. Rouhani B.D., et al. OCP Microscaling Formats (MX) Specification. Open Compute Project. 2023.

7. Liu B., Wang H., Qian Y. Towards Lightweight Speaker Verification via Adaptive Neural Network Quantization. IEEE/ACM Trans. Audio Speech Lang. Process. 2024. vol. 32. pp. 3771–3784.
8. Hong Y., Chung W.-J., Kang H.-G. Optimization of DNN-based speaker verification model through efficient quantization technique. arXiv preprint arXiv:2407.08991. 2024.
9. Wang H., Liu B., Wu Y., Chen Z., Qian Y. Lowbit Neural Network Quantization for Speaker Verification. IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Rhodes Island, Greece: IEEE, 2023. pp. 1–5.
10. Jouppi N.P., et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture. New York, NY, USA: Association for Computing Machinery, 2017. pp. 1–12. DOI: 10.1145/3079856.3080246.
11. Micikevicius P., et al. FP8 Formats for Deep Learning. arXiv preprint arXiv:2209.05433. 2022.
12. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
13. Wang H., Zheng S., Chen Y., Cheng L., Chen Q. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2023. pp. 5301–5305. DOI: 10.21437/Interspeech.2023-1513.
14. Колмаков Н.П., Толстых А.А. [Robust Speaker Features Recognition on Discrete Audio Signals]. Sbornik trudov XXIX Mezhdunarodnoj nauchno-tehnicheskoj konferencii «Radiolokacija, Navigacija, Svjaz'» [Proceedings of the XXIX International Scientific and Technical Conference «Radiolocation, Navigation, Communication»]. 2023. pp. 387–394. (In Russ.).
15. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
16. Huang G., Liu Z., van der Maaten L., Weinberger K.Q. Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. C. 2261–2269. DOI: 10.1109/CVPR.2017.243.
17. Yu Y.-Q., Zheng S., Suo H., Lei Y., Li W.-J. Cam: Context-Aware Masking for Robust Speaker Verification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021. pp. 6703–6707. DOI: 10.1109/ICASSP39728.2021.9414704.
18. Yakovlev I., et al. Reshape Dimensions Network for Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2024. pp. 3235–3239.
19. Lin Y., Cheng M, Zhang F, Gao Y, Zhang S, Li M. VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark. Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2024. pp. 4263–4267.
20. Chen S., et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE J. Sel. Top. Signal Process. 2022. vol. 16. № 6. pp. 1505–1518.

21. Vaswani A., et al. Attention Is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017. pp. 6000–6010.
22. Hendrycks D., Gimpel K. Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415. 2023.
23. GitHub. facebookresearch/libri-light: dataset for lightly supervised training using the librivox audio book recordings. https://librivox.org/. Available at: https://github.com/facebookresearch/libri-light (accessed 02.02.2025).
24. Chen G., et al. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio.Proceedings of the Annual Conference of the International Speech Communication Association Interspeech. 2021. pp. 3670–3674. DOI: 10.21437/Interspeech.2021-1965.
25. Wang C., et al. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021. pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80.
26. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning (PMLR). 2015. vol. 37. pp. 448–456.

Kolmakov Nikita — Junior researcher, Institute for Information Transmission Problems (Kharkevich Institute) Russian Academy of Sciences. Research interests: digital signal processing, machine learning. The number of publications — 6. kolmakov.np@iitp.ru; 19/1, Bolshoy Karetny Lane, 127051, Moscow, Russia; office phone: +7(495)650-4225.

Golubinskiy Andrey — Ph.D., Dr.Sci., Associate professor, Russian Science Foundation. Research interests: machine learning, neural network modeling, automated control systems with artificial intelligence elements, speech signal processing. The number of publications — 250. annikgol@mail.ru; 19-1, Solyanka St., 109028, Moscow, Russia; office phone: +7(495)650-4225.