Квантизация весов обученной нейронной сети методом максимизации корреляции

М.М. Пушкарева[1], Э.М. Хайров[1], Я.М. Карандашев[1]

[1]Центр оптико-нейронных технологий Институт системных исследований РАН, Нахимовский просп. 36, к.1, Москва, Россия, 117218

Abstract. Для уменьшения требуемой памяти для хранения весов нейронной сети были применены различные методы квантизации, основанные на максимизации корреляции между исходными и дискретизованными значениями весов. Квантизация проводилась на предварительно обученных сетях (post learning). Первый метод предполагает разбиение распределений весов на линейные и экспоненциальные (экспоненциально- возрастающие) отрезки, второй — определение отрезков разбиений с помощью аппроксимации нормальным и лапласовым распределением и градиентного спуска. Было проведено сравнение двух предлагаемых методов на предварительно обученных сетях, таких как VGG-16, MobileNet-v2, ResNet50 и Inception-v3.

1. Введение

Существующие модели нейронных сетей имеют миллионы параметров, что затрудняет их применение на мобильных устройствах. Одним из способов уменьшения необходимой памяти для хранения весов является квантизация. Данный метод предусматривает разбиение области значений весов на отрезки и присвоение весам, принадлежащим определенному отрезку, некоторого дискретизованного значения. Для сокращения вычислений будем рассматривать квантизацию без последующего дообучения. Данный подход применялся в работе [1] для линейной и экспоненциальной дискретизаций и был доработан в статье [2] с учетом максимизации корреляции между исходными значениями весов и дискретизованными. Так как веса в большинстве слоев различных нейронных сетей имеют распределение похожее на нормальное или лапласово распределение, рассмотрим случаи максимизации корреляции для таких распределений и проведем сравнение с результатами полученными в работе [2].

2. Описание процесса дискретизации

2.1. Максимизация корреляции

Максимизируя корреляцию между дискретизованными значениями и исходными, можно получить оптимальное дискретизованное значение на фиксированном отрезке, используя плотность распределения весов в слое.

$$\rho = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \to \max$$

где x - исходные значения, y - дискретизованные. Максимизируя ρ при заданных концах отрезков $x_0, x_1, \ldots, x_{N-1}, x_N$, где N - количество градаций, получим оптимальные значения

для y. Для заданных концов отрезков оптимальное дискретизованное значение y:

$$y_i = \frac{\int_{x_i}^{x_{i+1}} xp(x)dx}{\int_{x_i}^{x_{i+1}} p(x)dx}$$

где $p(x)$ - плотность распределения. Учитывая оптимальные y, получим градиент для процедуры градиентного спуска для поиска оптимального распределения на отрезки.

$$\frac{\partial \rho}{\partial x_i} = \frac{p(x)(y_{i+1} - y_i)(y_{i+1} + y_i - 2x_i)}{2\sigma_x}$$

Так как распределние весов в сети имеют распределение близкое к нормальному или лапласову 'Рисунок 1', использовались плотности этих распределений для получения оптимального разбиения и дискретизованных значений.

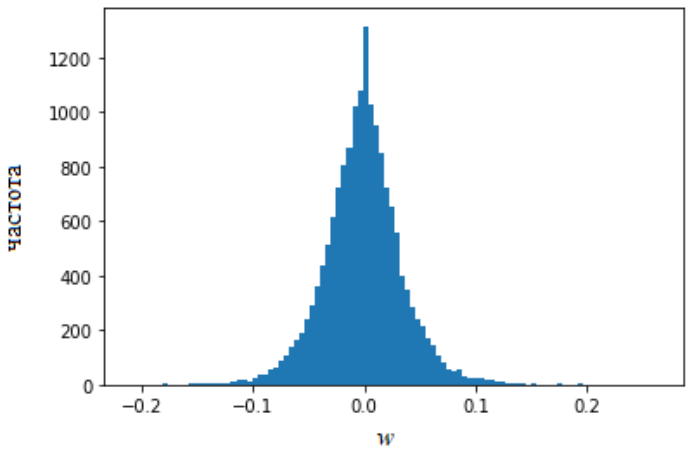

Рисунок 1. Распределение весов в слое ResNet50.

2.2. Результаты

Алгоритмы квантизации были протестированы на нейронных сетях, предобученных на задаче распознавания фотографий из датасета ImageNet [3]. Для выявления обобщающей способности алгоритма были использованы 1000 случайных изображений из полного датасета, включающего в себя 50000 изображений. Для исследования были выбраны готовые обученные модели Keras: VGG16, MobileNet-v2, Inception-v3, и ResNet-50. Проведено сравнение Top-1 и Top-5 accuracy для данных сетей с линейной и экспоненциальной квантизацией 'Таблица 1-4'.

Таблица 1. Accuracy Top-1 и Top-5 для различного числа градаций для сети ResNet50.

	Top-1			Top-5		
	Linear	Exponential	Gauss	Linear	Exponential	Gauss
2bit	0.00074	0.00074	0.00104	0.0052	0.0052	0.00516
3bit	0.0011	0.01536	0.001	0.00644	0.04918	0.00514
4bit	0.0063	0.5718	0.00846	0.02574	0.80408	0.02426
5bit	0.49212	0.7075	0.33928	0.73672	0.89876	0.57624
6bit	0.70674	0.72922	0.53656	0.89644	0.90994	0.78336
7bit	0.74006	0.74578	0.55042	0.91578	0.91904	0.79208
8bit	0.74622	0.7472	0.55382	0.91932	0.91964	0.79528

Таблица 2. Accuracy Top-1 и Top-5 для различного числа градаций для сети MobileNet-v2.

	Top-1			Top-5		
	Linear	Exponential	Gauss	Linear	Exponential	Gauss
2bit	0.00112	0.00112	0.00136	0.00492	0.00492	0.00546
3bit	0.00116	0.00244	0.00134	0.00528	0.00818	0.00504
4bit	0.0021	0.06408	0.04086	0.00782	0.15074	0.12578
5bit	0.16464	0.53818	0.42684	0.34422	0.77892	0.6728
6bit	0.57504	0.667	0.63238	0.80468	0.87298	0.8495
7bit	0.68512	0.69648	0.65246	0.88288	0.89142	0.86272
8bit	0.70144	0.70508	0.65036	0.89424	0.89622	0.8618

Таблица 3. Accuracy Top-1 и Top-5 для различного числа градаций для сети VGG16.

	Top-1			Top-5		
	Linear	Exponential	Gauss	Linear	Exponential	Gauss
2bit	0.00796	0.00796	0.0014	0.0257	0.0257	0.0063
3bit	0.00132	0.36036	0.01786	0.0068	0.61138	0.07888
4bit	0.00136	0.67372	0.397	0.00954	0.87872	0.65268
5bit	0.16492	0.70442	0.54954	0.35506	0.89632	0.7882
6bit	0.68202	0.71024	0.61656	0.88164	0.89872	0.83758
7bit	0.70822	0.71156	0.6241	0.89858	0.90044	0.84318
8bit	0.712	0.7118	0.62342	0.89982	0.89978	0.84382

3. Заключение

Результаты работы алгоритма лучше линейной дискретизации при небольшом количестве бит, но показал худшие результаты по сравнению с экспоненциальной. Предполагается, что это связано с тем, что реальное распределение весов в слое не является нормальным и оптимальное дискретизованное значение для нормального распределения аппроксимирует веса на отрезке хуже чем среднее значение весов.

4. Благодарности

Исследование выполнено при финансовой поддержке РФФИ грант № 18-07-00750.

Таблица 4. Accuracy Top-1 и Top-5 для различного числа градаций для сети Inception-v3.

	Top-1			Top-5		
	Linear	Exponential	Gauss	Linear	Exponential	Gauss
2bit	0.00112	0.00112	0.00096	0.00622	0.00622	0.0049
3bit	0.00084	0.00216	0.001	0.00492	0.008	0.005
4bit	0.00106	0.1971	0.00116	0.00634	0.35788	0.00492
5bit	0.06078	0.68508	0.02352	0.1372	0.88238	0.07592
6bit	0.68978	0.74028	0.03998	0.88692	0.91558	0.15442
7bit	0.75478	0.76524	0.4897	0.92502	0.9297	0.7233
8bit	0.76468	0.7658	0.48436	0.92986	0.93072	0.71828

5. Литература

[1] Cai, J. A Deep Look into Logarithmic Quantization of Model Parameters in Neural Networks / J. Cai, M. Takemoto, H. Nakajo // The 10th International Conference on Advances in Information Technology (IAIT) - ACM, New York, NY, USA, 2018. – P. 8. DOI: 10.1145/3291280.3291800.

[2] Malsagov, M.Yu. Exponential Discretization of Weights of Neural Network Connections in Pre-Trained Neural Networks / M.Yu. Malsagov, E.M. Khayrov, M.M. Pushkareva, I.M. Karandashev // Optical Memory and Neural Networks. – 2019. – Vol. 28(4). – P. 262-270.

[3] ImageNet – gataset [Electronic resource]. – Access mode: http://www.image-net.org.

Post-training quantization of neural network through correlation maximization

M.M. Pushkareva[1], E.M. Khayrov[1], I.M. Karandashev[1]

[1]SRISA RAS The Center of Optical Neural Technologies, Nakhimovskiy prospect 36/1, Moscow, Russia, 117218

Abstract. To reduce random access memory (RAM) requirements we apply different quantization methods for post-learning networks based on correlation maximization between initial and discrete weight values. First method assumes division of the weight distribution interval into linear and exponential segments. Second technique determines this segments using approximations by the normal and Laplace distributions and gradient descent. We provide the comparison of our implementations on different post-training networks: VGG-16, ResNet50, MobileNet-v2, Inception-v3, etc.