Supplementary materials of AlignQ: Alignment Quantization with ADMM-based Correlation Preservation

Ting-An Chen[1], De-Nian Yang[2,3], Ming-Syan Chen[1,3]
[1]Graduate Institute of Electrical Engineering, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
tachen@arbor.ee.ntu.edu.tw, dnyang@iis.sinica.edu.tw, mschen@ntu.edu.tw

A. Proof and theoretical analysis

A.1 CDF of an arbitrary continuous distribution is uniformly distributed

To address the issue that non-i.i.d. training and testing data induce a large quantization error (discussed in subsection 3.1.1 of the main paper), we propose a CDF alignment process to transform data in different distributions into the same space for quantization error minimization (introduced in subsection 3.1.2 of the main paper). In the following theorem, we demonstrate that the CDF of any continuous distributions follows uniform distribution.

Theorem 3.1 *Let X have the cumulative distribution function (cdf) of the continuous type that is strictly increasing on the support $a \leq x \leq b$. Then the function $Y = F(X)$ has a distribution $Uniform(0, 1)$.*

Proof. Since F(a) = 0 and F(b) = 1, the cdf of Y is

$$\begin{aligned} P(Y \leq y) &= P[F(X) \leq y] = P[X \leq F^{-1}(y)] \\ &= F[F^{-1}(y)] = y,\ 0 \leq y \leq 1, \end{aligned}$$

which is the cdf of a Uniform(0, 1) random variable. □

Theorem 3.1 proves that CDF of an arbitrary continuous distribution follows uniform distribution in the range (0, 1). Accordingly, we can ensure the training and testing data are in the same space (i.i.d.) by the alignment with their individual CDFs.

A.2 Theoretical analysis on the relation between data correlations and quantization error

In addition to the quantization error derived from non i.i.d. training and testing data, we further analyze the error induced by the changes of data after quantization. Existing research mainly focused on designing a quantization approach that reduced the discrepancy of individual data before and after quantization. However, they ignored the changes in data correlations after quantization. The following proposition and theorem demonstrate that not only the differences of individual values but also the significant changes in data correlations after quantization induce a large quantization error.

Proposition 1. The significant changes in the data correlations after quantization induces a larger quantization error (proved in Theorem 3.2).

In contrast to the quantization errors derived from the changes of the individual data $||\mathbf{X}_i - Q(\mathbf{X}_i)||_1, \forall i$, we consider the total quantization error $\sum_i ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1$, where $\mathbf{X}_i$ represents the i-th data, Q denotes the quantization function, and $||\cdot||_1$ is the $l1$-norm. To quantify the induced quantization error from our quantized model, we define a tolerated quantization error ϵ in the proof of Proposition 1. If the quantization error exceeds the tolerated error ϵ, then we regard the error is large, and vice and versa. Thereby, as the following theorem, we focus on analyzing the probability that the total quantization error is smaller than the tolerated error, denoted as $P(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1 < \epsilon)$.

Theorem 3.2 *Let $\boldsymbol{X}_i \in \mathbb{R}^d$ be the CNN representation of the i-th of n input image data. The function Q quantizes the values to the discete Uniform($-\alpha$, α), $\alpha \geq 0$. The quantized representation is denoted as $Q(\boldsymbol{X}_i)$, and the total quantization error of n data is $\sum_{i=1}^{n} ||\boldsymbol{X}_i - Q(\boldsymbol{X}_i)||_1$, where $||\cdot||_1$ represents the l1-norm. Now let the individual quantization error $\delta_i = \boldsymbol{X}_i - Q(\boldsymbol{X}_i)$, $\forall i = 1, 2, ..., n$, and the tolerated quantization error as ϵ. Then $P(\sum_{i=1}^{n} ||\boldsymbol{X}_i - Q(\boldsymbol{X}_i)||_1 < \epsilon) \geq 1 - \frac{n}{\epsilon^2}\mathbb{E}[||\delta_i||_1^2] - \frac{4\alpha}{\epsilon^2}\sum_{i,j=1;\ i<j}^{n}\mathbb{E}(||\delta_i||_1 + ||\delta_j||_1) - \frac{2}{\epsilon^2}\sum_{i,j=1;\ i<j}^{n}\mathbb{E}(|\boldsymbol{X}_i^T\boldsymbol{X}_j - Q(\boldsymbol{X}_i)^T Q(\boldsymbol{X}_j)|)$.*

Proof.

$$P(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1 \geq \epsilon)$$
$$\leq \frac{\mathbb{E}[(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1)^2]}{\epsilon^2}, \text{ by Markov's Inequality.}$$

Note that

Since we let $\delta_i = \mathbf{X}_i - Q(\mathbf{X}_i)$, i.e., $\mathbf{X}_i = Q(\mathbf{X}_i) + \delta_i$, the term $(\mathbf{X}_i)^T Q(\mathbf{X}_j)$ can be written as $(Q(\mathbf{X}_i) + \delta_i)^T Q(\mathbf{X}_j)$, and the term $Q(\mathbf{X}_i)^T \mathbf{X}_j$ can be reformulated as $Q(\mathbf{X}_i)^T (Q(\mathbf{X}_j) + \delta_j)$, which implies

$$\begin{aligned}
&\mathbb{E}[(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1)^2] \\
&= \sum_{i=1}^{n} \mathbb{E}(||\delta_i||_1)^2 \\
&+ 2 \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j + Q(\mathbf{X}_i)^T Q(\mathbf{X}_j) \\
&\qquad - (Q(\mathbf{X}_i) + \delta_i)^T Q(\mathbf{X}_j) \\
&\qquad - Q(\mathbf{X}_i)^T (Q(\mathbf{X}_j) + \delta_j)|) \\
&\leq n\mathbb{E}[||\delta||_1^2] \\
&+ 2 \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j) \\
&\qquad - \delta_i^T Q(\mathbf{X}_j) - Q(\mathbf{X}_i)^T \delta_j)|,
\end{aligned}$$

where $||\delta||_1 = \max_i ||\delta_i||_1$. Since

$$\begin{aligned}
&\sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j) \\
&\qquad - \delta_i^T Q(\mathbf{X}_j) - Q(\mathbf{X}_i)^T \delta_j)| \\
&\leq \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\delta_i^T Q(\mathbf{X}_j)| + |Q(\mathbf{X}_i)^T \delta_j)|) \\
&+ \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|),
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\delta_i^T Q(\mathbf{X}_j)| + |Q(\mathbf{X}_i)^T \delta_j)|) \\
&\leq \sum_{i,j=1;\ i<j}^{n} 2\alpha \cdot \mathbb{E}(||\delta_i||_1 + ||\delta_j||_1),
\end{aligned}$$

by $Q(\mathbf{X}_i)$ follows $Uniform(-\alpha, \alpha), \forall i$,

$$\begin{aligned}
&\sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j) \\
&\qquad - \delta_i^T Q(\mathbf{X}_j) - Q(\mathbf{X}_i)^T \delta_j)| \\
&\leq \sum_{i,j=1;\ i<j}^{n} 2\alpha \cdot \mathbb{E}(||\delta_i||_1 + ||\delta_j||_1) \\
&+ \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}[(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1)^2] \\
&\leq n\mathbb{E}[||\delta||_1^2] \\
&+2 \sum_{i,j=1;\ i<j}^{n} 2\alpha \cdot \mathbb{E}(||\delta_i||_1 + ||\delta_j||_1) \\
&+2 \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|),
\end{aligned}$$

Hence,

$$\begin{aligned}
&P(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1 \geq \epsilon) \\
&\leq \frac{n}{\epsilon^2} \mathbb{E}[||\delta||_1^2] + \frac{4\alpha}{\epsilon^2} \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(||\delta_i||_1 + ||\delta_j||_1) \\
&+ \frac{2}{\epsilon^2} \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|).
\end{aligned}$$

As a result, it is proved that

$$\begin{aligned}
&P(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1 < \epsilon) \\
&\geq 1 - \frac{n}{\epsilon^2} \mathbb{E}[||\delta||_1^2] - \frac{4\alpha}{\epsilon^2} \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(||\delta_i||_1 + ||\delta_j||_1) \\
&- \frac{2}{\epsilon^2} \sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T \mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|).
\end{aligned}$$

□

The last inequality of Theorem 3.2 derives a lower bound on the probability $P(\sum_{i=1}^{n} ||\mathbf{X}_i - Q(\mathbf{X}_i)||_1 < \epsilon)$. Accordingly, the probability increases, i.e., the total quantization error is relatively small in a high probability, when the lower bound becomes larger. Thereby, it indicates that a small quantization error can be acquired in a high probability when 1) the individual quantization errors $||\delta_i||_1, \forall i$ are

small, and 2) the total discrepancy of the data correlations before and after the quantization $\sum_{i,j=1;\ i<j}^{n} \mathbb{E}(|\mathbf{X}_i^T\mathbf{X}_j - Q(\mathbf{X}_i)^T Q(\mathbf{X}_j)|)$ is also small. To minimize the total quantization error, both the individual quantization errors and the changes in the data correlations need to be effectively reduced.

Existing research only focuses on the design of a quantization function correlated to the data distributions, for minimizing the individual quantization error. However, Theorem 3.2 proves that the significant changes in data correlations after quantization can induce a large quantization error. Therefore, in Sec. 3.2, we propose to minimize the discrepancy of the data correlations before and after the alignment-quantization process (detailed in Sec. 3.1).

B. Performance of quantized DSAN on Office-31 dataset

To validate AlignQ that can effectively minimize the quantization error derived from non-i.i.d. in training and testing data, we evaluate AlignQ and the state-of-the-art on the domain shift benchmark dataset, Office-31. In Sec. 4.3.2 in the main paper, we have employed a baseline transfer learning model DANN. Here we adopt a state-of-the-art model DSAN to verify the effectiveness of AlignQ. Table 1 presents the performances of the quantized DSAN on Office-31 dataset at different bitwidths. It manifests that AlignQ achieves outstanding performances compared with the rencent works, especially for the low bitwidths (4 bits). AlignQ applied on the 4-bit DSAN achieves more than 20% accuracy improvements compared with zero-shot quantization (ZSQ) [5, 9, 32], since ZSQ relies on the pretrained full-precision model on the training data, but the target testing data is in a different domain with a different distribution. In addition, compared with quantization-aware training (QAT) [29, 49], AlignQ obtains 1% to 6% accuracy increments. The result indicates that the learned quantization ranges and the gradient estimation according to the training data cannot be applied on the testing data with a minimal quantization error due to non-i.i.d. in training and testing data. In contrast, AlignQ addressing the issue of non i.i.d. data can further minimize the quantization error by aligning the distributions to the same domain (see Sec. 3.1 in the main paper) and preserving the data correlations during the alignment-quantization process (detailed in Sec. 3.2 in the main paper).

C. Effectiveness of CDF alignment and ADMM-based data correlation preservation in AlignQ

In Sec. 5 in the main paper, we have validated the individual effectiveness of the AlignQ components on CIFAR-10 and DANN on Office-31. Here we examine the quantization results of each component in AlignQ and the baseline uniform quantization in Table 2. Table 2 compares the quantization results of 1) "Uniform": the baseline uniform quantization (see Eq. (2) in the main paper), 2) **Ours (ADMM only)**: the ADMM-based data correlation preservation in AlignQ (illustrated in Sec. 3.1 in the main paper), 3) **Ours (CDF only)**: the CDF alignment quantization in AlignQ (detailed in Sec. 3.2 in the main paper), 4) **Ours (CDF+ADMM)**: considered with both CDF alignment quantization and ADMM-based data correlation preservation components in AlignQ, and 5) **Ours (Best of all)**: summarized with the best result from the above cases.

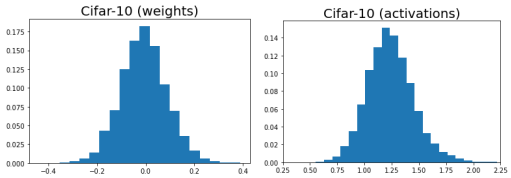
The results manifest that when we adopt the CDF alignment with quantization to transform the training and testing data in different domains to the same uniform space (case **CDF only**), we can achieve outstanding performances, especially for the low bitwidths. For the 4-bit quantization, case **CDF only** has an 85.1% accuracy over all domain shift tasks, significantly outperforms the uniform quanti-

Table 1. Accuracy (%) of quantized DSAN (ResNet-50) [50] on Office-31. Three data domains in Office-31 include Amazon (A), Webcam (W), and DSLR (D), thereby indicating six combinations of domain shift classification tasks. The average performance is denoted as "Avg.".

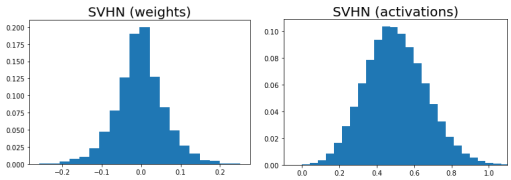
W/A bit	Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg.
32/32	Source only	81.3	98.0	100.0	86.6	62.9	62.2	81.8
	DSAN [49]	91.2	97.7	100.0	91.2	72.3	66.4	86.5
	AlignQ (Ours)	**93.0**	**98.2**	**100.0**	**92.9**	**73.5**	**67.6**	**87.5**
4/4	DoReFa [48]	90.1	97.7	100.0	89.3	65.8	62.0	84.2
	APoT [28]	86.0	97.1	97.1	76.8	61.7	57.0	79.3
	Choi *et al.* [9]	70.8	86.0	90.2	69.6	29.3	32.6	63.1
	ZeroQ [5]	70.8	86.0	89.3	67.9	25.6	34.8	62.4
	ZAQ [31]	71.2	87.1	90.3	68.2	28.5	35.9	63.5
	AlignQ (Ours)	**91.2**	**97.7**	**100.0**	**92.0**	**68.8**	**62.7**	**85.4**
5/5	DoReFa [48]	90.6	97.7	100.0	92.0	67.4	61.2	84.8
	APoT [28]	83.0	97.7	97.6	83.0	64.1	59.8	80.9
	Choi *et al.* [9]	90.6	98.2	100.0	89.3	66.6	63.2	84.7
	ZeroQ [5]	90.6	97.7	100.0	90.2	69.0	63.2	85.1
	ZAQ [31]	**91.2**	97.7	100.0	91.1	66.7	63.1	85.0
	AlignQ (Ours)	91.1	**98.2**	**100.0**	**92.0**	**69.2**	**63.2**	**85.6**
8/8	DoReFa [48]	90.6	97.7	100.0	91.0	67.9	62.6	85.0
	Choi *et al.* [9]	91.1	97.7	100.0	91.1	68.6	62.9	85.2
	ZeroQ [5]	91.0	97.7	100.0	91.1	68.4	62.9	85.2
	ZAQ [31]	91.2	97.7	100.0	91.1	68.2	62.8	85.2
	AlignQ (Ours)	**92.0**	**98.8**	**100.0**	**91.0**	**68.8**	**62.8**	**85.6**

Table 2. Effectiveness of AlignQ components. Accuracy (%) of quantized DSAN (ResNet-50) [50] on Office-31.

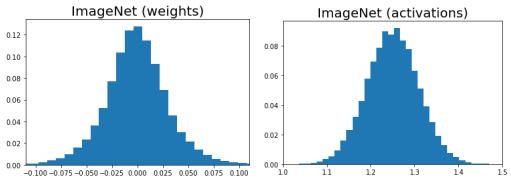
W/A bit	Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg.
4/4	Uniform	74.9	86.3	92.2	68.1	25.4	.43.0	63.3
	Ours (ADMM only)	76.0	87.1	92.9	68.8	26.3	.43.3	64.1
	Ours (CDF only)	91.1	97.7	100.0	90.1	67.2	62.4	85.1
	Ours (CDF+ADMM)	**91.2**	**97.7**	**100.0**	**92.0**	**68.8**	**62.4**	**85.4**
	Ours (Best of all)	**91.2**	**97.7**	**100.0**	**92.0**	**68.8**	**62.7**	**85.4**
5/5	Uniform	87.9	97.5	100.0	91.5	64.2	61.6	83.8
	Ours (ADMM only)	88.4	98.2	100.0	92.0	65.7	61.9	84.4
	Ours (CDF only)	**91.1**	97.7	100.0	92.0	67.6	61.9	85.2
	Ours (CDF+ADMM)	89.5	**98.2**	**100.0**	**92.0**	**69.2**	**63.2**	**85.4**
	Ours (Best of all)	**91.1**	**98.2**	**100.0**	**92.0**	**69.2**	**63.2**	**85.6**
8/8	Uniform	91.1	97.7	100.0	88.2	65.4	61.7	84.0
	Ours (ADMM only)	91.1	97.7	100.0	90.2	66.9	62.0	84.7
	Ours (CDF only)	91.1	97.7	100.0	91.0	68.8	62.0	85.1
	Ours (CDF+ADMM)	**92.0**	**98.8**	**100.0**	**91.0**	**68.8**	**62.7**	**85.6**
	Ours (Best of all)	**92.0**	**98.8**	**100.0**	**91.0**	**68.8**	**62.7**	**85.6**



(a) ResNet-20 on CIFAR-10 (b) ResNet-20 on CIFAR-10



(c) MobileNet-v2 on SVHN (d) MobileNet-v2 on SVHN



(e) ResNet-50 on ImageNet (f) ResNet-50 on ImageNet

Figure 1. Distributions of CNN weights and activations on the benchmark datasets, including ResNet-20 on CIFAR-10, MobileNet-v2 on SVHN, and ResNet-50 on ImageNet.

zation with only 63.3% (more than 20% accuracy increment). Furthermore, when we leverage ADMM optimization to minimize the changes in data correlations during the quantization, we can further obtain accuracy improvements in most of the tasks. For example, the 4-bit DSAN quantized by **ADMM only** compared with "Uniform" can obtain approximately 1% accuracy improvements in all tasks. In addition, The 5-bit DSAN quantized by **AlignQ** outperforms **ADMM only** in 1.5% to 2% accuracy improvements on the tasks such as $D \rightarrow A$ and $W \rightarrow A$. Accordingly, the results manifest that the proposed CDF alignment quantization can effectively address the issue of a large quantization error derived from non-i.i.d. in training and testing data. Furthermore, the ADMM optimization is also effective on minimizing the quantization error from the changes in data correlations during the quantization process.

D. Normality in CNN weights and activations

In Sec. 3.1.1 of the main paper, we consider using CDF of normal distributions as the data alignment function since the previous research has studied that the CNN weights and activations converge in normal by and large. Here, we conduct experiments on the benchmark image datasets and CNN architectures to examine the prior research studies. Fig. 1 visualizes the distributions of CNN weights and activations on the benchmark architectures and datasets. The results show the normality of the weights and activations

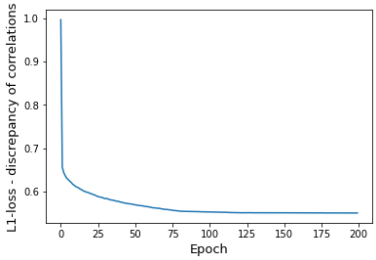


Figure 2. Training loss of quantized ResNet-20 on CIFAR-10 by AlignQ on minimizing the discrepancy of data correlations.

when CNN models converge, which is consistent with the previous studies [17, 29, 33, 49] and thereby appropriate for the CDF of normal distribution as a data alignment function.

E. Convergence analysis on ADMM optimization

Algorithm 1 in the main paper presents the training process of AlignQ. Sec. 3.2.3 illustrates that we update the ADMM parameters with the model weights in each training iteration for an efficient quantization process. To ensure the convergence, we examine the training loss on minimizing the discrepancy of data correlations during quantization as shown in Fig. 2. The decreasing loss with the training process manifests the convergence of the ADMM optimization process. Moreover, it also verifies that the discrepancy of data correlations is diminished.