Data Diversification Analysis on Data Preprocessing

Taeyoon Kim[*] *ChanHo Park*[*] *Heelim Hong*[*] *Minseok Kim*[*]
Ze Jin[†] *Changdae Kim*[‡] *Ji-Yong Shin*[§] *Myeongjae Jeon*[*]
UNIST[*], *Meta*[†], *ETRI*[‡], *Northeastern University*[§]

In this paper, we conduct a statistical analysis to examine the diversity distribution resulting from two different approaches: The first one, standard approach, is a baseline augmentation approach where a random augmentation is applied to each sample in each epoch independently; The second one, random batch approach, is another new augmentation approach designed where a random augmentation is applied to each tiny-batch in each epoch independently and which samples are in the same tiny-batch is random and independent across all epochs.

The diversity from augmentation is measured by the number of unique augmented samples from all samples in all epochs. Under the same assumptions from [1], the expectation of diversity is the same across standard approach and random batch approach. However, the variance of diversity from random batch approach is higher than that from standard approach, which is the cost of these mini-batch approaches to reduce compute time. The intuition behind this cost is, when the same augmentation is unfortunately applied to the same sample in different epochs, diversity loss is amplified due to the fact that the same augmentation is applied to all samples in the same mini-batch.

Following the same assumptions from [1], we assume K epochs, N samples, augmentation set A, batch size n.

Define X_{it} as the indicator of whether augmentation A_t is applied to the i-th sample.

$$X_{it} = \begin{cases} 1, & \text{if augmentation } A_t \text{ is applied to the } i\text{-th sample} \\ & \text{at least once in all epochs,} \\ 0, & \text{otherwise.} \end{cases}$$

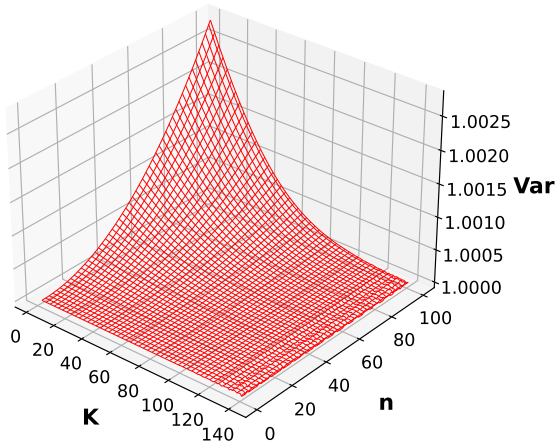
Define X_i as the number of unique augmented samples from the i-th sample in all epochs.

$$X_i = \sum_t X_{it}$$

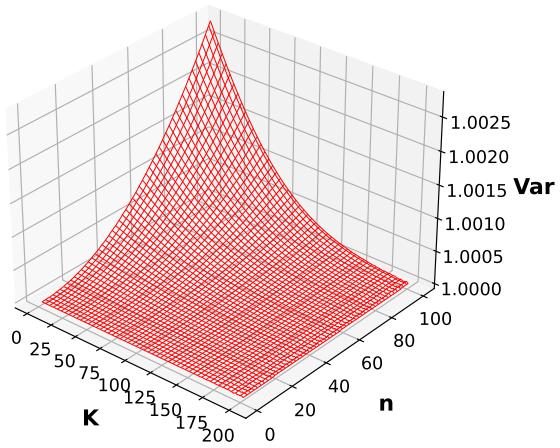
Define X as the number of unique augmented samples from all samples in all epochs.

$$X = \sum_i X_i$$

In order to visualize the variance of diversity in our approach vs per-sample augmentation, **??** presents the ratio of the standard deviation of diversity in our approach over the standard deviation of diversity in per-sample augmentation, given different parameters for RandAugment ($K = 2 \sim 140$,



(a) RandAugment



(b) AutoAugment

Figure 1: Ratio of the standard deviation.

$N = 1743042$, $|A| = 16$, $n = 1 \sim 100$) and AutoAugment ($K = 2 \sim 200$, $N = 1743042$, $|A| = 25$, $n = 1 \sim 100$), respectively, where standard deviation is the square root of variance.

Based on detailed proof of the below sections, the variance of diversity in our approach is almost the same as that of per-sample augmentation in typical DNN training tasks where mini-batch size n « sample size N. As a result, our approach maintains the same level of diversity as per-sample augmentation, thus guaranteeing the same level of convergence in model training. This analysis confirms that we do not inadvertently increase the number of epochs for model convergence and can benefit from faster training time compared to standard approaches.

1 Standard Approach

To start with, we get the expectation of X_i as follows.

$$E(X_i) = \sum_t E(X_{it}) = \sum_t P(X_{it} = 1) = \sum_t (1 - P(X_{it} = 0))$$
$$= \sum_t \left(1 - \left(\frac{|A|-1}{|A|}\right)^K\right) = |A| \left(1 - \left(\frac{|A|-1}{|A|}\right)^K\right)$$

Then we get the expectation of diversity or equivalently X as the sum of expectations of all X_i.

$$E(X) = \sum_i E(X_i) = N|A| \left(1 - \left(\frac{|A|-1}{|A|}\right)^K\right)$$

To get the variance of X, $Var(X) = E(X^2) - (E(X))^2$, we need $E(X^2)$ and $E(X)^2$.

$$E(X^2) = E(\sum_i X_i)^2 = E(\sum_i X_i^2 + \sum_{i \neq j} X_i X_j)$$
$$= \sum_i E(X_i^2) + \sum_{i \neq j} E(X_i X_j) = \sum_i E(X_i^2) + \sum_{i \neq j} E(X_i) E(X_j)$$

$$(E(X))^2 = (\sum_i E(X_i))^2 = \sum_i (E(X_i))^2 + \sum_{i \neq j} E(X_i) E(X_j)$$

Note that $E(X_i X_j) = E(X_i) E(X_j)$ in the standard approach that a random augmentation is applied to each sample independently. Since the last term of $E(X^2)$ and $E(X)^2$ are same, $Var(X)$ can be expanded as follows.

$$Var(X) = E(X^2) - (E(X))^2 = \sum_i E(X_i^2) - \sum_i (E(X_i))^2$$

We need $E(X_i^2)$ and it can be expanded as follows.

$$E(X_i^2) = E(\sum_t X_{it} \sum_s X_{is}) = \sum_t \sum_s E(X_{it} X_{is})$$
$$= \sum_t \sum_s P(X_{it} = 1, X_{is} = 1)$$

To expand the above formula further, we need other formulas. We started from the probability of $X_{it} = 0$ and $X_{is} = 0$ where t and s refers different augmentations.

$$P(X_{it} = 0, X_{is} = 0), t \neq s$$
$$= P(X_{it} = 0) P(X_{is} = 0 | X_{it} = 0)$$
$$= \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{|A|-2}{|A|-1}\right)^K = \left(\frac{|A|-2}{|A|}\right)^K$$

Then, we get the sum of the probability of $X_{it} = 0$ and $X_{is} = 0$ in all combinations of two augmentations.

$$\sum_t \sum_s P(X_{it} = 0, X_{is} = 0)$$
$$= \sum_{t=s} P(X_{it} = 0, X_{is} = 0) + \sum_{t \neq s} P(X_{it} = 0, X_{is} = 0)$$
$$= |A| \left(\frac{|A|-1}{|A|}\right)^K + |A| (|A|-1) \left(\frac{|A|-2}{|A|}\right)^K \tag{1}$$

Now, we get the probability of $X_{it} = 1$ and $X_{is} = 1$.

$$P(X_{it} = 1, X_{is} = 1)$$
$$= 1 + P(X_{it} = 0, X_{is} = 0) - P(X_{it} = 0) - P(X_{is} = 0)$$
$$= 1 + P(X_{it} = 0, X_{is} = 0) - \left(\frac{|A|-1}{|A|}\right)^K - \left(\frac{|A|-1}{|A|}\right)^K$$
$$= 1 + P(X_{it} = 0, X_{is} = 0) - 2\left(\frac{|A|-1}{|A|}\right)^K \tag{2}$$

By plugging the fomula (1) and (2), we can calculate the expectation of the square of X_i.

$$E(X_i^2) = E(\sum_t X_{it} \sum_s X_{is}) = \sum_t \sum_s E(X_{it} X_{is})$$
$$= \sum_t \sum_s P(X_{it} = 1, X_{is} = 1)$$
$$= \sum_t \sum_s \left(1 + P(X_{it} = 0, X_{is} = 0) - 2\left(\frac{|A|-1}{|A|}\right)^K\right)$$
$$= \sum_t \sum_s \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) + \sum_t \sum_s P(X_{it} = 0, X_{is} = 0)$$
$$= |A|^2 \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) + |A| \left(\frac{|A|-1}{|A|}\right)^K$$
$$+ |A| (|A|-1) \left(\frac{|A|-2}{|A|}\right)^K$$

Note that $E(X_i^2)$ and $E(X_i)$ is same for all samples. Finally, we get the variance of diversity or equivalently X.

$$
\begin{aligned}
&Var(X) = \sum_i E(X_i^2) - \sum_i (E(X_i))^2 \\
&= NE(X_1^2) + N(E(X_1))^2 \\
&= N|A|^2 \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) + N|A|\left(\frac{|A|-1}{|A|}\right)^K \\
&+ N|A|(|A|-1)\left(\frac{|A|-2}{|A|}\right)^K - N|A|^2\left(1 - \left(\frac{|A|-1}{|A|}\right)^K\right)^2
\end{aligned}
$$

2 Random Batch Approach

Note that the distribution of X_i and X is exactly the same under Standard Approach and Random Batch Approach. Thus, we use same formulas for $E(X_i)$, $E(X)$, and $E(X_i^2)$ as follows.

$$
\begin{aligned}
&E(X_i) = |A|(1 - (\frac{|A|-1}{|A|})^K) \\
&E(X) = N|A|(1 - (\frac{|A|-1}{|A|})^K) \\
&E(X_i^2) = |A|^2(1 - 2(\frac{|A|-1}{|A|})^K) + |A|(\frac{|A|-1}{|A|})^K \\
&\qquad + |A|(|A|-1)(\frac{|A|-2}{|A|})^K
\end{aligned}
$$

However, $E(X_iX_j) = E(X_i)E(X_j)$ does not hold in random batch approach as samples in a tiny-batch have dependency on their augmentation. Thus, $Var(X)$ can be expanded as follows.

$$
\begin{aligned}
&Var(X) = E(X^2) - (E(X))^2 = E(\sum_i X_i)^2 - (E(X))^2 \\
&= \sum_i E(X_i^2) + \sum_{i \neq j} E(X_iX_j) - (E(X))^2
\end{aligned}
$$

The only unknown term is $\sum_{i \neq j} E(X_iX_j)$. The term can be expanded as follows.

$$
\begin{aligned}
E(X_iX_j), i \neq j = E(\sum_t X_{it} \sum_s X_{js}) &= \sum_t \sum_s E(X_{it}X_{js}) \\
&= \sum_t \sum_s P(X_{it} = 1, X_{js} = 1)
\end{aligned}
$$

To expand the above formula further, we need other formulas. Let A_t be the t-th augmentation in set A and C_{ik} be the augmentation applied to the i-th sample in k-th epoch. Then, we start from the probability that A_t is not applied on j-th sample, given that A_t is not applied on i-th sample in some epoch.

$$
\begin{aligned}
&P(C_{j1} \neq A_t | C_{i1} \neq A_t), i \neq j \\
&= P(i \neq j \text{ in same batch in 1st epoch} | C_{i1} \neq A_t) \\
&\times P(C_{j1} \neq A_t | i \neq j \text{ in same batch in 1st epoch}, C_{i1} \neq A_t) \\
&+ P(i \neq j \text{ in different batch in 1st epoch} | C_{i1} \neq A_t) \\
&\times P(C_{j1} \neq A_t | i \neq j \text{ in different batch in 1st epoch}, C_{i1} \neq A_t) \\
&= P(i \neq j \text{ in same batch in 1st epoch}) \\
&\times P(C_{j1} \neq A_t | i \neq j \text{ in same batch in 1st epoch}, C_{i1} \neq A_t) \\
&+ P(i \neq j \text{ in different batch in 1st epoch}) P(C_{j1} \neq A_t) \\
&= \frac{n-1}{N-1} 1 + \frac{N-n}{N-1} \frac{|A|-1}{|A|} \\
&= \frac{n-1}{N-1} + \frac{N-n}{N-1} \frac{|A|-1}{|A|}
\end{aligned}
\tag{3}
$$

With the formula (3), we can get the probability of $X_{it} = 0$ and $X_{jt} = 0$ where $i \neq j$ as follows.

$$
\begin{aligned}
&P(X_{it} = 0, X_{jt} = 0), i \neq j \\
&= P(X_{it} = 0) \cdot P(X_{jt} = 0 | X_{it} = 0) \\
&= \left(\frac{|A|-1}{|A|}\right)^K \cdot P(C_{jk} \neq A_t, k = 1, \ldots, K | C_{ik} \neq A_t, k = 1, \ldots, K) \\
&= \left(\frac{|A|-1}{|A|}\right)^K \cdot (P(C_{j1} \neq A_t | C_{i1} \neq A_t))^K \\
&= \left(\frac{|A|-1}{|A|}\right)^K \cdot \left(\frac{n-1}{N-1} + \frac{N-n}{N-1} \frac{|A|-1}{|A|}\right)^K
\end{aligned}
\tag{4}
$$

Then, we also calculate the probability that A_s is not applied on j-th sample, given that A_t is not applied on i-th sample in some epoch. Note that A_s and A_t are not same.

$$
\begin{aligned}
&P(C_{j1} \neq A_s | C_{i1} \neq A_t), t \neq s, i \neq j \\
&= P(i \neq j \text{ in same batch in 1st epoch} | C_{i1} \neq A_t) \\
&\times P(C_{j1} \neq A_s | i \neq j \text{ in same batch in 1st epoch}, C_{i1} \neq A_t) \\
&+ P(i \neq j \text{ in different batch in 1st epoch} | C_{i1} \neq A_t) \\
&\times P(C_{j1} \neq A_s | i \neq j \text{ in different batch in 1st epoch}, C_{i1} \neq A_t) \\
&= P(i \neq j \text{ in same batch in 1st epoch}) \\
&\times P(C_{j1} \neq A_s | i \neq j \text{ in same batch in 1st epoch}, C_{i1} \neq A_t) \\
&+ P(i \neq j \text{ in different batch in 1st epoch}) P(C_{j1} \neq A_t) \\
&= \frac{n-1}{N-1} \frac{|A|-2}{|A|-1} + \frac{N-n}{N-1} \frac{|A|-1}{|A|}
\end{aligned}
\tag{5}
$$

Using the formula (5), we get the probability of $(X_{it} = 0$

and $X_{js} = 0$ where $t \neq s$ and $i \neq j$ as follows.

$$
\begin{aligned}
&P(X_{it} = 0, X_{js} = 0), t \neq s, i \neq j \\
&= P(X_{it} = 0) P(X_{js} = 0 | X_{it} = 0) \\
&= \left(\frac{|A|-1}{|A|}\right)^K P(C_{jk} \neq A_s, k = 1, \ldots, K | C_{ik} \neq A_t, k = 1, \ldots, K) \\
&= \left(\frac{|A|-1}{|A|}\right)^K (P(C_{j1} \neq A_s | C_{i1} \neq A_t))^K \\
&= \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1}\frac{|A|-2}{|A|-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K
\end{aligned}
\tag{6}
$$

By plugging the formula (4) and (6), we can get the sum of probabilities of $X_{it} = 0$ and $X_{js} = 0$ where $i \neq j$.

$$
\begin{aligned}
&\sum_t \sum_s P(X_{it} = 0, X_{js} = 0), i \neq j \\
&= \sum_{t=s} P(X_{it} = 0, X_{js} = 0) + \sum_{t \neq s} P(X_{it} = 0, X_{js} = 0) \\
&= |A| \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K \\
&+ |A|(|A|-1) \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1}\frac{|A|-2}{|A|-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K
\end{aligned}
$$

In addition, the probability $X_{it} = 1$ and X_{js} where $i \neq j$, which is required to expand the unknown term, can be represented as follows.

$$
\begin{aligned}
&P(X_{it} = 1, X_{js} = 1), i \neq j \\
&= 1 + P(X_{it} = 0, X_{js} = 0) - P(X_{it} = 0) - P(X_{js} = 0) \\
&= 1 + P(X_{it} = 0, X_{js} = 0) - \left(\frac{|A|-1}{|A|}\right)^K - \left(\frac{|A|-1}{|A|}\right)^K \\
&= 1 + P(X_{it} = 0, X_{js} = 0) - 2\left(\frac{|A|-1}{|A|}\right)^K
\end{aligned}
$$

Now, we are ready to expand the term $E(X_i X_j)$ where $i \neq j$.

$$
\begin{aligned}
&E(X_i X_j), i \neq j = \sum_t \sum_s P(X_{it} = 1, X_{js} = 1) \\
&= \sum_t \sum_s \left(1 + P(X_{it} = 0, X_{js} = 0) - 2\left(\frac{|A|-1}{|A|}\right)^K\right) \\
&= \sum_t \sum_s \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) + \sum_t \sum_s P(X_{it} = 0, X_{js} = 0) \\
&= |A|^2 \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) \\
&+ |A| \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K \\
&+ |A|(|A|-1) \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1}\frac{|A|-2}{|A|-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K
\end{aligned}
$$

Finally, the variance of diversity or equivalently X with the random batch approach is as follows.

$$
\begin{aligned}
&Var(X) = \sum_i E(X_i^2) + \sum_{i \neq j} E(X_i X_j) - (E(X))^2 \\
&= N|A|^2 \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) + N|A| \left(\frac{|A|-1}{|A|}\right)^K \\
&+ N|A|(|A|-1) \left(\frac{|A|-2}{|A|}\right)^K + N(N-1)|A|^2 \left(1 - 2\left(\frac{|A|-1}{|A|}\right)^K\right) \\
&+ N(N-1)|A| \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K \\
&+ N(N-1)|A|(|A|-1) \left(\frac{|A|-1}{|A|}\right)^K \left(\frac{n-1}{N-1}\frac{|A|-2}{|A|-1} + \frac{N-n}{N-1}\frac{|A|-1}{|A|}\right)^K \\
&- N^2|A|^2 \left(1 - \left(\frac{|A|-1}{|A|}\right)^K\right)^2
\end{aligned}
$$

References

[1] Gyewon Lee, Irene Lee, Hyeonmin Ha, Kyunggeun Lee, Hwarim Hyun, Ahnjae Shin, and Byung-Gon Chun. Refurbish your training data: Reusing partially augmented samples for faster deep neural network training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 537–550, 2021.