Understanding Data Augmentation in Neural Machine Translation: Two Perspectives towards Generalization

Guanlin Li*, Lemao Liu†, Guoping Huang†, Conghui Zhu*, Tiejun Zhao*, Shuming Shi†

*Harbin Institute of Technology, †Tencent AI Lab
{epsilonlee.green}@gmail.com, {chzhu, tzhao}@hit.edu.cn,
{redmondliu, donkeyhuang, shumingshi}@tencent.com

Abstract

Many Data Augmentation (DA) methods have been proposed for neural machine translation. Existing works measure the superiority of DA methods in terms of their performance on a specific test set, but we find that some DA methods do not exhibit consistent improvements across translation tasks. Based on the observation, this paper makes an initial attempt to answer a fundamental question: what benefits, which are consistent across different methods and tasks, does DA in general obtain? Inspired by recent theoretic advances in deep learning, the paper understands DA from two perspectives towards the generalization ability of a model: input sensitivity and prediction margin, which are defined independent of specific test set thereby may lead to findings with relatively low variance. Extensive experiments show that relatively consistent benefits across five DA methods and four translation tasks are achieved regarding both perspectives.

1 Introduction

Data Augmentation (DA) is a training paradigm that has been proved to be very effective in many modalities (Park et al., 2019; Perez and Wang, 2017; Sennrich et al., 2016a), especially for classification (Perez and Wang, 2017). In structured domain, Neural Machine Translation (NMT) is the frontier of DA research (Sennrich et al., 2016a; Norouzi et al., 2016; Zhang and Zong, 2016; Fadaee et al., 2017; Wang et al., 2018; Zhang et al., 2019; Edunov et al., 2018; Fadaee and Monz, 2018). However, by investigating a variety of DA methods, we find that their test performance across different translation tasks does not exhibit consistent improvement, and this phenomenon can be initially observed in (Wang et al., 2018) as well. The reason might be the evaluation metric on a specific test set when compared to the whole data population, which generates all possible data, has large variance so that leads to the inconsistency. This evaluation dilemma is also recognized and explored by Recht et al. (2018, 2019); Werpachowski et al. (2019), and is especially notorious for language generation tasks (Chaganty et al., 2018; Hashimoto et al., 2019) where the evaluation metrics, e.g. BLEU (Papineni et al., 2001), are extrinsic and heavily relies on the reference provided. Therefore, we ask a fundamental question: what benefits, which are more consistent across different DA methods and translation tasks, can DA in general obtain?

A direct answer to the above question is to use generalization gap (Kawaguchi et al., 2018) defined by the difference between population risk and empirical risk. This measure does not rely on any specific test set, accurately depicts generalization but is intractable to compute. So recently, many theorists have proposed either non-vacuous generalization bound (Dziugaite and Roy, 2017; Zhou et al., 2019) or novel generalization measures (Novak et al., 2018; Bartlett et al., 2017; Neyshabur et al., 2017; Jiang et al., 2019) to roughly reflect the gap. Inspired by them, we propose to understand the benefits of DA from two perspectives: input sensitivity and prediction margin. The proposed underlying two measures are well adapted from Novak et al. (2018) and Bartlett et al. (2017) and can be computed only on the train samples to unveil the consistent benefits of DA. Under a carefully designed fair setting over four different translation tasks, we examine five methods from two main categories of DA and compare them with a model trained without DA. The empirical experiments demonstrate the following findings: a). DA methods exhibit more consistent effects across different translation tasks in terms of both measures. b). DA methods can either alleviate input sensitivity or promote prediction margin.
By and large, our main contributions are:

- We make an initial attempt to understand the essence of DA in NMT by investigating its benefits which are relatively consistent across five DA methods and four translation tasks.
- We highlight two perspectives towards generalization to measure the benefits of DA in NMT and study them with carefully designed fair experiments.

## 2 DA Methods in NMT

### 2.1 Training Objective Decomposition

Given the train set \( T \) the baseline NMT model \( p_b(y|x) \) without using DA is trained under the empirical data distribution \( \hat{p}(X, Y|T) \) through maximum likelihood estimation:

\[
J_{\text{MLE}} = \mathbb{E}_{x,y \sim \hat{p}(X, Y|T)}[\log p_b(y|x)], \tag{1}
\]

where \( \hat{p} \) is a mixture of Dirac distribution concentrated around each training instance with uniform mixture coefficients \( (1/|T|) \). Then we define the augmentation (AUG) model as a conditional distribution over the train set, \( q(X, Y|T) \). \(^1\) Under the AUG model, the training objective becomes:

\[
J_{\text{AUG}} = \mathbb{E}_{x,y \sim q(X, Y|T)}[\log p_b(y|x)]. \tag{2}
\]

More realistically, for any DA method in any training run, we can collect the augmented instances to form a set \( A \) distinguishing \( T \), when considering the curriculum of mixing \( A \) with the original train \( T \). Since we would like to derive a conceptual framework that reflects this form of importance weighting, we further decompose AUG model into a linear interpolation (\( \alpha \)) of \( \hat{p}(X, Y|T) \) and an augmentation distribution \( q_{\text{AUG}}(X, Y|T) \):

\[
\alpha \cdot \hat{p}(X, Y|T) + (1 - \alpha) \cdot q_{\text{AUG}}(X, Y|T), \tag{3}
\]

where \( \alpha \) controls the mixture ratio within a batch during SGD training. The ratio has been founded as an important factor influencing final performance (Sennrich et al., 2016a; Fadaee et al., 2017; Edunov et al., 2018; Fadaee and Monz, 2018).

**Key factors** Through Eq. 3, we can identify two key factors for conducting fair experiments: a) the number of SGD updates on every original training instance means how much the model learns from \( T \); b) the mixture ratio means how much the model learns from \( A \) online, with which together balance the learning of the translation knowledge.

### 2.2 Settings and Main Performance

**Settings** By carefully controlling the above two factors, we conduct fair and extensive experiments with Transformer (Vaswani et al., 2017) on four translation tasks for five DA methods. Fairseq (Ott et al., 2019) is used as our codebase. We use standard benchmarks IWSLT17 En-Fr, WMT19 Zh-En, WMT19 En-De, where we train both translation directions on the IWSLT corpus. The five DA methods are briefly summarized as follows:

- **RAML:** reward-augmented maximum likelihood training, which augment the target-side with a sampling distribution \( P(Y|Y^*) \) concentrated around \( Y^* \) (Norouzi et al., 2016).
- **Switchout (SO):** similar to RAML, but also adds the some kind of augmentation to the source-side (Wang et al., 2018).
- **Self-training (ST):** fix the source-side, uses an forward NMT model to generate the target-side (Zhang and Zong, 2016).
- **Target-agree (TA):** similar to ST, but uses a forward NMT model with right-to-left decoder (Zhang et al., 2019).
- **Back-translation (BT):** fix the target-side, uses an backward NMT model to generate the source-side (Sennrich et al., 2016a).

The implementation of RAML and SO are borrowed from the Appx. of Wang et al. (2018). \(^2\)

<table>
<thead>
<tr>
<th>Method</th>
<th>Fr⇒En</th>
<th>En⇒Fr</th>
<th>Zh⇒En</th>
<th>En⇒De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>38.38 (5)</td>
<td>38.88 (6)</td>
<td>17.25 (6)</td>
<td>26.19 (4)</td>
</tr>
<tr>
<td>RAML</td>
<td>+0.22 (3)</td>
<td>+0.67 (3)</td>
<td>+0.23 (4)</td>
<td>-0.16 (6)</td>
</tr>
<tr>
<td>SO</td>
<td>+0.01 (4)</td>
<td>+0.62 (4)</td>
<td>+0.02 (5)</td>
<td>-0.15 (5)</td>
</tr>
<tr>
<td>ST</td>
<td>-0.13 (6)</td>
<td>+0.46 (5)</td>
<td>+1.51 (2)</td>
<td>+0.83 (2)</td>
</tr>
<tr>
<td>TA</td>
<td>+0.62 (2)</td>
<td>+1.13 (1)</td>
<td>+2.41 (1)</td>
<td>+1.01 (1)</td>
</tr>
<tr>
<td>BT</td>
<td>+0.82 (1)</td>
<td>+0.99 (2)</td>
<td>+1.06 (3)</td>
<td>+0.39 (3)</td>
</tr>
</tbody>
</table>

| Table 1: Main BLEU results (CTC=0.62). |

\(^1\) In the paper, we do not consider using monolingual data for DA thus conditioning only on bilingual data since this will further bring monolingual data selection discussed in Fadaee and Monz (2018) as a factor to influence the performance of different DA methods; we leave this factor for future study.

\(^2\) We categorize and analyze how we choose or train DA methods with a generative formulation in Appx. A and B.
Identify the best DA method because their rankings across the four translation tasks vary a bit. To measure the degree of consistency, we use a correlation measure called Kendall’s coefficient of concordance (Kendall and Smith, 1939; Mazurek, 2011) to evaluate the correlation of the rankings produced on the four translation tasks (apx. C). The value shows strong consistency (correlation) of different rankings when it is close to 1. We call the correlation value Cross-Task Consistency measure or CTC. The CTC for the BLEU measure is 0.62, which is of weak consistency. This phenomenon might be a result of the intrinsic nature of using a single specific test as a substitute of the whole data population for evaluation. In the next section, we introduce two measures that are more consistent (with close-to-1 CTC value). They in some extent reflect the model generalization and are easy-to-compute as well.

3 Two Measures Towards Generalization

We attempt to understand the benefits that DA can obtain through the quantification of input sensitivity and prediction margin. The two measures are adapted from Novak et al. (2018) and Bartlett et al. (2017); Neyshabur et al. (2017); Jiang et al. (2019). They have been proved through massive experiments to be correlated with model generalization. Our main purpose here is to utilize them to unveil the consistency property (measured by CTC) of DA across different methods and translation tasks. The next two subsections define the two measures and report their statistics on subsamples of the train set respectively.

3.1 Input Sensitivity

Input sensitivity is the sensitivity of the loss computed from the model towards a minor change of input representation. Given a point of interest \( x \), the original form in Novak et al. (2018) is defined as the expected Jacobian norm of the loss vector evaluated at \( x \) and only compute the \( L^2 \) norm of a slice of the Jacobian matrix indexed by the label. We adopt the later measure which is the gradient norm of the loss scalar indexed by \( y \) to \( x \):

\[
\mathbb{E}_x \| J(x_y) \|_F.
\]

If \( x \in \mathbb{R}^d \) lies in a space with differential structure, we can apply Eq. 5 directly. But in NMT the naive representation of an instance \((x, y)\) is the token index given by the vocabulary, so we cannot compute the gradient of the loss with respect to \( x \). We follow Sundararajan et al. (2017) and use the result of \( x \) after embedding lookup as its learned representation, denoted as \( \text{Emb}(x) \in \mathbb{R}^{L_x \times d} \) where \( L_x \) is the length of the input and \( d \) the size of the embedding. By regarding the translation model \( p_y(y|x) \) as a function that decomposes at each step of \( y \) given \( \text{Emb}(x) \) as input to get a scalar average log likelihood, denoted as \( \mathcal{L}_{x,y} = \frac{1}{T_y} \sum_t \log p_y(y_t|y_{<t}, x) \). Moreover, initial experiments on just using the single original \( x_i \) to evaluate the gradient will still result in inconsistency, due to the non-equivalence of the localness concept compared with the continuous setting, i.e. for language input, the localness is between discrete inputs in the neighbor of \( x_i \). So we evaluate the sensitivity of \( x_i \) by averaging gradient norms over its \( k \) nearest neighbor \( s_{k(i)} \in \text{KNN}[x_i] \) through cosine similarity between word embed-

<table>
<thead>
<tr>
<th>Method</th>
<th>Fr→En</th>
<th>En→Fr</th>
<th>Zh→En</th>
<th>En→De</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.565 (6)</td>
<td>0.623 (4)</td>
<td>0.422 (6)</td>
<td>0.682 (5)</td>
</tr>
<tr>
<td>RAML</td>
<td>-0.056 (2)</td>
<td>+0.003 (5)</td>
<td>-0.008 (4)</td>
<td>-0.003 (4)</td>
</tr>
<tr>
<td>SO</td>
<td>-0.082 (1)</td>
<td>-0.099 (1)</td>
<td>-0.053 (2)</td>
<td>-0.143 (1)</td>
</tr>
<tr>
<td>ST</td>
<td>-0.034 (3)</td>
<td>-0.099 (3)</td>
<td>-0.054 (1)</td>
<td>-0.116 (2)</td>
</tr>
<tr>
<td>TA</td>
<td>-0.023 (5)</td>
<td>-0.010 (2)</td>
<td>-0.043 (3)</td>
<td>-0.013 (3)</td>
</tr>
<tr>
<td>BT</td>
<td>-0.026 (4)</td>
<td>+0.035 (6)</td>
<td>-0.007 (5)</td>
<td>+0.232 (6)</td>
</tr>
</tbody>
</table>

Table 2: Sensitivity measure (CTC=0.72).
ings. We set $k$ to 5 in our experiment to guarantee words in the $k$ nearest neighbor has similar semantic meaning. Formally, we define the input sensitivity of an NMT model as:

$$
\mathbb{E}_{(x,y)} \frac{1}{L_x} \sum_i \frac{1}{k} \sum_{x_{(j)} \in \text{ENN}[x_i]} \left\| \frac{\partial L_{x,y}}{\partial \text{Emb}(x)_i(j)} \right\|_F,
$$

(6)

where the $\text{Emb}(x)_i$ is the embedding lookup of the $i^{th}$ token index, so we compute the average token-wise gradient of each instance.

We use subsamples of the train set to approximately compute the expectation in Eq. 6 and the overall statistics are shown in Table 2. Similar to Table 1, the DA methods are shown in their Δ value respect to the baseline. A first thing to notice is that the ranking is more steady across tasks (CTC=0.72). It also shows that for input x, DA in general can reduce the gradient norm of the prediction loss on $\text{Emb}(x)_i$, which shows that DA can obtain more stable model towards data corruption.

To further understand what effect DA in general has on each input token type, we compute the Δ sensitivity between the baseline and one DA method on the same token type and sort them according to the Δ with positive value (which means DA reduces the sensitivity of that token type). We then divide the sorted types into ten bins and compute the average token type frequency of that bin. As shown in Figure 1, DA in general, improves the sensitivity of token types with relatively low frequency more than those with high frequency, thus may somehow improve the translation quality of low frequency token types.

3.2 Prediction Margin

Margin is a classic concept in support vector machine (Vapnik, 2013), which is defined as the geometric distance between the support vectors and the decision boundary. Larger margin implies better generalization. In nonlinear case, it reflects the distance of a correctly classified input representation with class $i$ to move towards the decision boundary between $i$ and any other class $j$ (Jiang et al., 2019). However, since the decision boundary does not have analytical form due to nonlinearity, computing the geometric distance is intractable. In our setting we regard NMT model as doing step-wise classification with $z = (x, y_{<t})$ as input feature and $y_t$ as the label. In Bartlett et al. (2017), the original definition of the margin of correctly predicted input is:

$$
p_{\theta}(y_{t}|z) - \max_{v' \neq y_t} p_{\theta}(v'|z)
$$

(7)

where $y_t$ is the ground-truth label, $v'$ another class type, $R$ the spectral complexity of the model and $N$ the number of training instance in the train subsamples for computing their margins. We simplify Eq. 7 to only consider the numerator. The reasons are: a) under the same model architecture, $R$s are very close across different DA methods; b) we can omit $||x||_2/N$ since it remains unchanged as well. In this way, we can map every $z, y_t$ to a margin with label type $y_t = v$, where $v \in V_{tgt}$:

$$
m_{z,y_t}^v := p_{\theta}(y_t|z) - \max_{v' \neq y_t} p_{\theta}(v'|z).
$$

(8)

So for every target token type $v$ we can collect a set of margins $\{m_{z,y_t}^v\}$, and the margin sets of all token types are combined as the total margin set $\cup_v \{m_{z,y_t}^v\}$. Following Neyshabur et al. (2017), we do not compute the minimum margin of the total set which can be highly sensitive to outliers. Instead, if the total set has cardinality $N'$, we obtain the $\epsilon N'$-th smallest margin from the set as the overall prediction margin, with a tolerant coefficient $\epsilon \in [0, 0.1]$ ($\epsilon$ is set to 0.001). 3 We can also obtain token-wise prediction margin from $\{m_{z,y_t}^v\}$, which is the prediction margin of a specific token type $v$.

The overall prediction margins are listed in Table 3. The relative rank is highly consistent across the four translation tasks (CTC=0.98). Although RAML and SO seem to be inferior to the baseline, other DA methods improve the margin in general. We give a possible explanation for this in the next subsection. Similar to the previous subsection, we also report the average token type frequency of each margin binned token groups in Figure 2 and find that DA, in general, brings larger margin improvement over low frequency tokens.

Table 3: Margin measure (CTC=0.98).
3.3 Discussion

**Why use these two measures?** We have conducted a relatively complete survey of the recent measures towards measuring generalization ability proposed by the deep learning community, such as model complexity (Zhang et al., 2016; Neyshabur et al., 2017), flatness (Dinh et al., 2017), stiffness (Fort et al., 2019) and second order Hessian of the input or the number of linear regions in hiddens (Novak et al., 2018; Montufar et al., 2014). Some of those measures have complex definition such as linear regions, others are very expensive to compute for models as large as Transformer such as Hessian and stiffness. However, we compute weight norm with different forms proposed in Neyshabur et al. (2017) and find no regularity which suggests that the complexity measure through norms for networks architectures like Transformer or convolutional/recurrent neural networks might be very different from simple feed-forward ones which might be still an open problem in theoretic deep learning community. As a matter of fact, due to computational easiness and large-scale empirical evidence, we choose sensitivity and margin the measures.

**Why no absolute consistency between the two measures?** In Section 3.1 and 3.2, the two measures do not show well consistency between them: under the margin based measure, RAML and SO do not exhibit superiority over the baseline like they do under the sensitivity measure. One reason might be: despite the measures are empirically proved to reflect generalization, they are only one view towards generalization respectively. Specifically, in recent generalization theory (Novak et al., 2018; Bartlett et al., 2017), the measures are evaluated between models with extremely evident difference in generalization ability (measured by test performance difference), for example, between models trained with random labels and true labels. Instead, our comparison is among models with similar capacity and are well-trained, which rises challenge for us to get very consistent statistics through a single view. This may inspire us to combine multiple views of model training to design better measures with stronger correlation.

4 Conclusion and Future Work

In this paper, we aim at delivering relatively consistent benefit measures of DA based on the phenomenon of inconsistent BLEU improvement across translation tasks. To our expect, the proposed two measures exhibit relative consistency on five DA methods across four translation tasks, which demonstrate that DA can benefit the model with improved sensitivity or prediction margin especially for low frequency words.

However, the problem of intrinsic evaluation or better understanding of the unreasonable effective of DA should just be at the beginning. One natural extension of the work is to consider using monolingual data within the controlled experimental setting. Basically, DA is a trade-off between noise-injection and knowledge augmentation, so it could be a nice theoretic direction to think about DA under statistical query model (Kearns, 1998) with translation between formal or constrained natural languages (ws-, 2019). This may also shed light on another essential question: what is the intrinsic properties of the augmented data (Branchaud-Charron et al., 2019) that make a difference especially for discrete data with no geometry interpretation. Applications like active data selection (Coleman et al., 2019) guided with margin or sensitivity can be derived. In general, understanding NMT model’s behavior (not only with DAs) beyond BLEU (Neubig et al., 2019) should be taken seriously, e.g. to design a beavior suite like Osband et al. (2019) is most valuable.

Acknowledgements

The authors are very grateful to all the anonymous reviewers for pointing out the topic of intrinsic evaluation of DA methods to summarize our work as well as proposing several critical issues regarding limitations of the proposed measures. We thank Xintong Li for detailed discussion and paper formatting; Xintong, Chaoqun Duan, Yong Jiang for proof-reading. Tiejun Zhao is supported by National Key R&D Program of China (Project No. 2017YFB1002102).
References


Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In EMNLP.


