Self-Paced Learning for Neural Machine Translation

Anonymous Authors[1]

Abstract

Recent studies have proven that the training of neural machine translation (NMT) can be fa- cilitated by mimicking the learning process of humans. Nevertheless, achievements of such kind of curriculum learning rely on the quality of artificial schedule drawn up with the hand- crafted features, e.g. sentence length or word rarity. We ameliorate this procedure with a more flexible manner by proposing self-paced learning, where NMT model is allowed to 1) automatically quantify the learning confidence over training examples; and 2) flexibly govern its learning via regulating the loss in each it- eration step. Experimental results over multi- ple translation tasks demonstrate that the pro- posed model yields better performance than strong baselines and those models trained with human-designed curricula on both translation quality and convergence speed.

1. Introduction

Neural machine translation (NMT) has achieved promising results with the use of various optimiza- tion tricks (Hassan et al., 2018; Chen et al., 2018; Xu et al., 2019; Li et al., 2020; Yang et al., 2020). In spite of that, these techniques lead to increased training time and massive hyper-parameters, making the development of a well-performed system expensive (Popel and Bojar, 2018; Ott et al., 2018). As an alternative mitigation, curriculum learn- ing (CL, Elman, 1993; Bengio et al., 2009) has shown its effectiveness on speeding up the con- vergence and stabilizing the NMT model training (Zhang et al., 2018; Platanios et al., 2019). CL teaches NMT model from easy examples to com- plex ones rather than equally considering all sam- ples, where the keys lie in the definition of "difficulty" and the strategy of curricula design (Krueger and Dayan, 2009; Kocmi and Bojar, 2017). Ex- isting studies artificially determine data difficulty according to prior linguistic knowledge such as sen- tence length (SL) and word rarity (WR) (Platanios et al., 2019; Zhang et al., 2019; Zhou et al., 2020), and manually tune the learning schedule (Liu et al., 2020; Fomicheva et al., 2020). However, neither there exists a clear distinction between easy and hard examples (Kumar et al., 2010), nor these human intuitions exactly conform to effective model training (Zhang et al., 2018). Instead, we resolve this problem by introducing self-paced learning (Kumar et al., 2010), where the emphasis of learning can be dynamically deter- mined by model itself rather than human intuitions. Specifically, our model measures the level of confi- dence on each training example (Gal and Ghahra- mani, 2016; Xiao and Wang, 2019), where an easy sample is actually the one of high confidence by the current trained model. Then, the confidence score is served as a factor to weight the loss of its corresponding example. In this way, the training process can be dynamically guided by model itself, refraining from human predefined patterns. We evaluate our proposed method on IWSLT15 En-Vi, WMT14 En-De, as well as WMT17 Zh-En translation tasks. Experimental results re- veal that our approach consistently yields better translation quality and faster convergence speed than TRANSFORMER (Vaswani et al., 2017) base- line and recent models that exploit CL (Platanios et al., 2019). Quantitative analyses further confirm that the intuitive curriculum schedule for a human does not fully cope with that for model learning.

2. Self-Paced Learning for NMT

As mentioned above, translation difficulty for hu- mans may not match that for neural networks. Even if these artificial supervisions are feasible, the long sequences or rare tokens are not always "difficult" as the model competence increases. From this view, we design a self-paced learning algorithm that offers NMT the abilities to 1) estimate the con- fidences over samples appropriated for the current training state; and 2) automatically control the fo- cus of learning through regulating the training loss, as illustrated in Fig. 1.

2.1. Confidence Estimation

We propose to determine the learning emphasis according to the model confidence (Ueffing and Ney, 2005; Soricut and Echihabi, 2010), which quantifies whether the current model

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

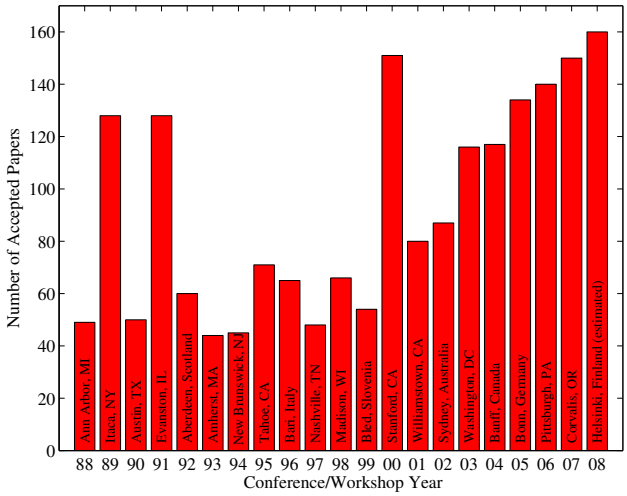

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

is confident or hesitant on translating the training samples. The model confidence can be quantified by Bayesian neural networks (Buntine and Weigend, 1991; Neal, 1996), which place distributions over the weights of network. For efficiency, we adopt widely used Monte Carlo dropout sampling (Gal and Ghahra- mani, 2016) to approximate Bayesian inference. Given current NMT model parameterized by and a mini-batch consisting of N sentence pairs (x1, y1), · · · , (xN, yN), we first perform M passes through the network, where the m-th pass ˆm randomly deactivates part of neurons. Thus, each example yields M sets of conditional probabilities. The lower variance of translation prob- abilities reflects higher confidence that the model has with respect to the instance (Dong et al., 2018; Wang et al., 2019). We propose multi-granularity strategies for confidence estimation:

2.2. Training Strategy

A larger confidence score indicates that the cur- rent model is confident on the corresponding exam- ple. Therefore, the model should learn more from the predicted loss. In order to govern the learning schedule automatically, we leverage the confidence scores as factors to weight the loss, thus control- ling the update at each time step. At the early stage of the study, the model learns more from confident samples, thus accelerating the training. The hesitant samples are not completely ignorant, but relatively few can be learned. As training proceeds, the loss of high-confidence sam- ples gradually reduce, and the model will pay more attention

Table 1. Overall experimental results of all approaches upon three translation tasks.

MODEL	IWSLT15	WMT14	WMT17
TRANSFORMER	95.9± 0.2	96.7± 0.2	24.11
+CL-SL	83.3± 0.6	80.0± 0.6	24.10
+CL-WR	61.9± 1.4	83.8± 0.7	25.42
SPL	74.8± 0.5	78.3± 0.6	25.86

on "complex" samples with low predic- tion accuracy, thus raising their confidence. In this way, the loss of different samples are dynamically revised, eventually balancing the learning. Contrast to related studies (Zhang et al., 2018, 2019; Kumar et al., 2019; Platanios et al., 2019) which adopt CL into NMT with predefined pat- terns, the superiority of our model lies in its flexi- bility on both learning emphasis and strategy. Sev- eral researchers may concern about the processing speed when integrating Monte Carlo Dropout sam- pling. Contrary to prior studies which estimate confidence during inference (Dong et al., 2018; Wang et al., 2019), we only perform forward propa- gation M = 5 times in training time, which avoids the auto-regressive decoding for efficiency.

3. Experiments

We evaluate our method upon TRANSFORMER- Base/Big model (Vaswani et al., 2017) and conduct experiments on IWSLT15 English-to-Vietnamese (EnVi), WMT14 English-to-German (EnDe) and WMT17 Chinese-to-English (ZhEn) tasks. For fair comparison, we use the same experimental setting as Platanios et al. (2019) for EnVi and follow the common configuration in Vaswani et al. (2017) for EnDe and ZhEn. During training, we apply 0.3 dropout ratio and batch size as 4,096 for EnVi task, and experi- ments are conducted upon one Nvidia GTX1080Ti GPU device. For EnDe and ZhEn task, we use 32,768 as batch size, and use four Nvidia V100 GPU devices for experiments. We use beam size as 4, 5, 10, and decoding alpha as 1.5, 0.6, 1.35 for each task, respectively (Vaswani et al., 2017). We compare our models with two baselines.

4. Conclusion

In this paper, we propose a novel self-paced learn- ing model for NMT in which the learning schedule is determined by model itself rather than being intu- itively predefined by humans. Experimental results on three translation tasks verify the universal ef- fectiveness of our approach. Quantitative analyses confirm that exploiting self-paced strategy presents a more flexible way to facilitate the model convergence than its CL counterparts. It is interesting to combine with other techniques (Li et al., 2018; Hao et al., 2019) to

further improve NMT. Besides, as this idea is not limited to machine translation, it is also interesting to validate our model in other NLP tasks, such as low-resource NMT model train- ing (Lample et al., 2018; Wan et al., 2020) and neural architecture search (Guo et al., 2020).