TRANSFER LEARNING FOR ADVERSARIAL MACHINE TRANSLATION

Sharada Murali
sharadam@usc.edu

ABSTRACT

Neural Machine Translation (NMT) is the process of mapping a segment of words from a source language to a target language using neural networks. However, NMT systems rely on large datasets for the source and target languages, and perform poorly on low-resource languages where there is insufficient parallel data. An effective method for improving NMT on low-resource languages is to employ transfer learning, where a model trained on a high-resource language pair is used to initialize training for the low-resource language pair. In this work, the effect of employing transfer learning methods on an adversarial machine translation model is studied. Apart from directly initializing the parent GAN model to train the low-resource language pair, the effect of freezing parameters from the parent model during transfer learning has also been tested. The results of these experiments show a consistent increase in the BLEU scores of the child model upon transfer from a parent model, and give rise to several avenues of future work.

1 Introduction

Machine translation is a field that has grown exponentially over the past decade, with certain translators now providing near-human level translation accuracies. This enhancement in performance has largely been due to improvements in neural network architectures, giving rise to models such as the Recurrent Neural Network (RNN), and LSTM-RNN[1]. These neural machine translation systems greatly outperform traditional MT methods in cases where large amounts of parallel data are available, such as from French to English. However, the performance of this method suffers greatly when the amount of data is insufficient, and in these cases traditional methods tend to work better than neural networks.

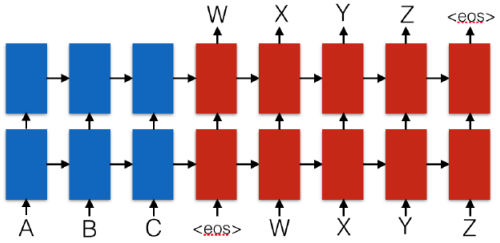

Figure 1: The Encoder-Decoder architecture used in NMT, from [2]

Transfer Learning is the process of using the knowledge gained from solving one task to solve other, related tasks. This approach has previously been successful in domains like speech recognition[3], and is now being used to improve NMT performance on low-resource languages. The process consists of a “pre-training" phase, where a network is trained on a high-resource parent dataset. This parent network is then used to initialize training on the low-resource language pair. In some cases[2], parameters of the parent model are selectively frozen during the training phase of the child model to improve performance even further.

Such works have been undertaken fairly recently, starting from [2], where the effects of parameter freezing and language similarity were studied. It also uses the transfer learning method to re-score standard statistical MT algorithms, showing increased BLEU[4] scores in all cases. The research done in [5] only focuses on performance improvements across similar parent-child languages, while [6] explores simpler transfer learning techniques and also attempts to address

effects of language similarity on the transfer learning results. In all of these works, transfer learning has been shown to improve the baseline BLEU scores for low-resource language pairs.

That being said, existing research on this topic is very sparse, and the results and conclusions presented in current works do not show much agreement and sometimes even contradict each other[2, 6]. For example, [2] claims that similar parent and child languages produce better improvements in the BLEU scores, while [6] states that the performance improvements from related and unrelated languages are not very different.

2 The GAN Architecture

A generative adversarial network[7] consists of two competing networks: a generator and a discriminator. The function of the discriminator is to predict whether the input given to it is real or fake via a scalar 0/1 output. The output of the discriminator can be expressed as $D(x;\theta_d)$, where x is the input of the discriminator and θ_d represents its parameters. Thus, the discriminator is trained to maximize the probability of assigning the correct label to the input data.

The input for the generator is typically sampled from white noise, and the objective of the generator is to learn a mapping $G(z;\theta_g)$, where θ_g represents the parameters of G, from the noise distribution $p_z(z)$. The generator simultaneously learns to improve its output so that it is labeled by the discriminator as a real output. One of the main challenges when designing and implementing GANs is stability: neither one of the generator or discriminator should overpower the other, and their learning rates should be chosen accordingly.

The GAN loss function can be characterized as the value function $V(G,D)$ of a two-player min-max game:

$$\min_G \max_D V(G,D) = \mathbb{E}_{y\sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

Thus, D is trained to minimize $\log D(y)$, while the parameters of G are simultaneously adjusted to minimize $\log(1 - D(G(z)))$.

Traditionally, both the generator and discriminators use Convolutional Neural Networks[8], which have been widely used in image processing applications. However, in applications involving time series data, or in this case, sentence translations, using a Recurrent Neural Network in the generator would yield better results. This variant of the GAN is called RNN-GAN, and has been used in applications such as music generation[9].

2.1 The Adversarial NMT

The aim of the Adversarial Neural Machine Translation model[10] is to learn a target language translation y' that is as close to the human translation y as possible. Thus, the RNN-generator attempts to learn a mapping $G(y|x)$ for a given source sentence x, and produces a target translation y' sampled from $G(.|x)$. The discriminator, which compares the generator output (x, y') to the ground-truth human translation (x, y), employs a CNN[11].

A significant challenge is in designing the training process for the generator, as it difficult to backpropagate errors from the discriminator. This is due to y' being discretely sampled, thus making $V(G,D)$ non-differentiable with respect to θ_g. To address this issue, the authors of [10] use the REINFORCE[12] algorithm, which employs a Monte-Carlo policy gradient method. Thus, the parameters of G are updated as:

$$\theta_g = \theta_g - \alpha\hat{\nabla}_{\theta_g} \tag{2}$$

where α is the learning rate of the generator, and $\hat{\nabla}_{\theta_g}$ is the approximate gradient, specified as:

$$\hat{\nabla}_{\theta_g} = \log(1 - D(x,y'))\hat{\nabla}_{\theta_g}\log G(y'|x) \tag{3}$$

$\hat{\nabla}_{\theta_g}\log G(y'|x)$ are the gradients specified in the standard sequence-to-sequence NMT. The Adversarial NMT architecture is shown in Fig. 2.

3 Experiments

The experiments carried out in this work are similar to those in [2] and [6], except they are undertaken with the Adversarial NMT architecture. In order to test the hypothesis that similar parent-child language pairs produce better transfer learning results, the experiments are carried out with similar and dissimilar language datasets.

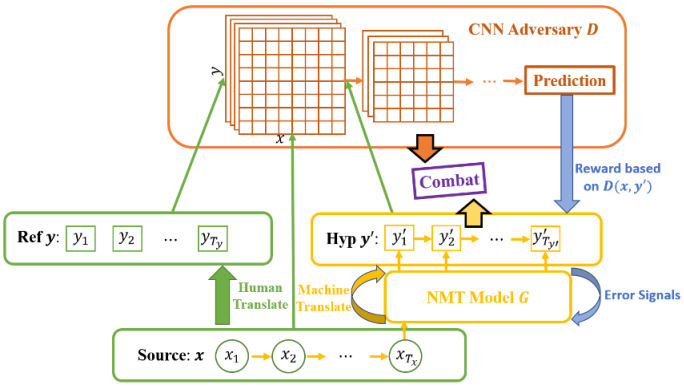

Figure 2: The RNN-GAN architecture, from [10]

Datasets: All the datasets used in this research work have been obtained from the Web Inventory of Transcribed and Translated Talks (WIT^3)[13]. These datasets contain multilingual transcriptions of TED talks, and a majority of them translate between a different language and English. The high-resource and low-resource languages chosen are listed below:

1. **High-resource languages**: Three high-resource language pairs have been chosen for the experiments presented in this work: German-English, Russian-English, and Czech-English. Each of these datasets contain over 100,000 sentences for training.
2. **Low-resource languages**: The Slovenian-English dataset with less than 15,000 sentence pairs has been chosen for transfer learning from the parent language pairs.

As Sloevnian is a Slavic language, it is related to both Czech and Russian. However, it is dissimilar to German, which is a Germanic language.

Training: Training was carried out on a virtual machine instance with a Tesla T4 GPU in Google Cloud. In all the experiments, the learning rate of both the generator and discriminator was set to 0.001. In order to compare the results of different experiments, each language pair was trained for 10 epochs. Training took approximately 4-6 hours on the high-resource language pairs, and 1 hour on the low-resource pair. The baseline results of all the language pairs are presented in Table 1.

Table 1: Baseline Results

Dataset	Training size	BLEU Score
German-English	160k	28.35
Russian-English	155k	19.81
Czech-English	94k	23.24
Slovenian-English	15k	7.76

BLEU Score: In order to evaluate the model's performance, the BLEU score is used, as it correlates closely with human judgment of translated sentences. A higher BLEU score corresponds to a greater degree of similarity with the "true" (human) translation. As seen in Table 1, the BLEU score of the low-resource language is significantly lower than the others.

3.1 Trivial transfer learning

For this experiment, the network parameters used to train the high-resource languages are used to initialize that of the low-resource model as-is. During the training of the high-resource languages, it was evident that the discriminator network achieved lower loss very early on in the process. Thus, only the generator model of the high-resource language was used to initialize the child model, and the discriminator was trained from scratch to prevent overfitting.

3.2 Transfer Learning with Parameter Freezing

In [2], parameters of the generator were selectively frozen during transfer learning; i.e. gradient updates to these parameters were not allowed while training the child model with the trained parent model. The authors reported that this method showed better improvements to the BLEU score than naively initializing the child model. In cases where the target languages were similar, or the same language, freezing the target output embeddings showed the best performance improvements, as these parameters are not expected to change much during re-training.

A similar approach has been taken in this experiment, as all the datasets have the same target language (English). The target embeddings are frozen during transfer learning, and similar to the previous experiment, the discriminator for the low-resource dataset is trained from scratch. Results of both the experiments are summarized in Table 2.

Table 2: Transfer Learning results on Slovenian-English. The numbers within brackets show the increase in the BLEU score from the baseline.

Parent Language Pair	Trivial transfer BLEU	Parameter Freezing BLEU
German-English	10.61 (+2.85)	10.80 (+3.04)
Russian-English	**10.80 (+3.04)**	9.87 (+2.11)
Czech-English	10.67 (+2.91)	**11.13 (+3.37)**

4 Results and Discussion

The results of Experiment 3.1 show improvements in BLEU scores in all cases. All of the parent language pairs, similar and dissimilar, showed increased BLEU scores when training the child network initialized with the parent model. Of the three parent language pairs, initializing with “similar" parent language pairs (Russian-English and Czech-English) showed greater improvement in performance compared to initializing with German-English, the dissimilar language. This seems to reflect the claim made in [2] that similarity in source parent language results in better performance in the child model.

From the two similar parent languages, the Russian-English model showed the highest improvement in the BLEU score (+3.04). One on hand, this seems unsurprising, as the Russian-English dataset consists of more training examples than the Czech-English one. However, the baseline BLEU of Czech-English is significantly higher than that of Russian-English and yet does not result in a proportionate increase in the BLEU of Slovenian-English on transfer learning. Additionally, there is no significant difference in the improved BLEU scores on transfer learning for both similar and dissimilar languages, as seen in Table 2.

The parameter freezing experiment also produced BLEU improvements with no apparent pattern. While Russian-English parent resulted in the best performance improvement in experiment 3.1, freezing the parameters of the parent model *decreased* BLEU scores in the child model relative to trivial transfer learning. (However, it still showed improvements compared to the Slovenian-English baseline.) On the other hand, parameter freezing in the dissimilar German-English parent model increased BLEU scores of the low-resource dataset even further, thus performing better than the case of the Russian-English parent model. Freezing the target embeddings in the Czech-English model for transfer learning produced the highest improvement in BLEU scores among all datasets and experiments.

The results of these experiments clearly show that transfer learning augments the performance of low-resource languages for similar and (relatively) dissimilar parent language-pairs. However, the claim made in [2] that similar parent languages produce better performance improvements does not seem apparent in all cases. Similarly, freezing the target embeddings of the parent generator while training the low-resource language does not always produce better results than naively re-training on the child dataset. Nonetheless, in order to establish a pattern in these results, or lack thereof, further trials need to be performed on more varied language pairs in both the parent and child models. For example, though we claim that German is a dissimilar language, it still belongs to the family of Eurasian languages and will share more commonalities with Slovenian than, say, Mandarin. In additional, it would be beneficial to perform these experiments with more low-resource languages while retaining a set of common parent language pairs to give more credence to previous claims.

5 Conclusion

While Neural Machine Translation has shown profound success in cases where large datasets are available, there is a need to develop methods to improve poor performance in low-resource cases. Using transfer learning methods has

shown a lot of promise, and there is enormous scope for additional research in this field. This is especially true for transfer learning in *adversarial* machine translation, for which there is currently no known literature.

The experiments performed in this work further confirm that transfer learning is a viable method to improve performance on low-resource languages, and freezing parameters can provide further performance enhancements for the same target language in some cases. However, there seems to be no consistent pattern or correlation between the "similarity" in parent and child languages, or the BLEU score improvements. Further work needs to be done in performing these experiments on more varied datasets, with different parent and child languages. It is also worthwhile to research the effect of changing the RNN-GAN architecture, and modifying parameters such as the learning rate. The outcomes of such experiments would provide a deeper insight, and potentially result in a more complete framework for transfer learning in adversarial machine translation.

References

[1] H. Sak, Andrew Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 338–342, 01 2014.

[2] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation, 2016.

[3] Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. Transfer learning for speech recognition on a budget, 2017.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[5] Toan Q. Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation, 2017.

[6] Tom Kocmi and Ondřej Bojar. Trivial transfer learning for low-resource neural machine translation, 2018.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[9] Olof Mogren. C-rnn-gan: A continuous recurrent neural network with adversarial training. In *Constructive Machine Learning Workshop (CML) at NIPS 2016*, page 1, 2016.

[10] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. Adversarial neural machine translation, 2017.

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[13] WIT3. `https://wit3.fbk.eu`.