Neural Machine Translation

Aniket Ashok, Suraj Raghavendra Vadvadgi, Vinit Kumar

Abstract

The proposed model in this paper, uses the encoder-decoder machine translation mechanism to translate long sentences with higher probability than the previous works. Also, in previous works, the encoder used to encode a sentence in a fixed length vector, but this paper describes a new task which tends to reduce the load on the encoder by encoding the input sequence generated from the input sentence into a sequence of vectors. Finally, for the result, a subset of these vectors is chosen adaptively as we will be describing it later in the paper. This task happens with the translation process. Since we tend to apply encoder-decoder machine translation using recurrent neural network technique, it builds on the same technique used in other works and then advances to correct the issues in encoding and decoding the longer sentences efficiently.

1 Introduction

The fundamental part of being human is the ability to communicate with one another. There are roughly 6,600 different languages around the world today. Machine translation gives people from different countries and ethnic groups to convey the original tone or intent of message, considering cultural and regional differences between source and target languages. The major categories for machine translation applications are for industries in business use (like Government, Software and technology, Military and defense, Healthcare, Finance, E-commerce, Education, Media), online/ app for consumers use (Text-to-text, Text-to-speech, Speech-to-text, Speech-to-speech, Image-to-text). To meet these demands, many technological companies are investing on this.

The recent advances in deep learning and neural network technologies like Deep Neural Networks, Convolution Neural Networks, and Recurrent Neural Networks, Feedforward Neural Networks has led to increased quality of translations. These technologies find applications in language modeling, speech and object recognition, paraphrase detection etc. The shift from phrase-based approach (2006) used in google translator to the deep learning technologies (2016) resulted in 60% increased accuracy. Google now supports over 100 languages for translation.

Language translation has become very common with Google translation pioneering in this field. This project aims to use machine translation model to translate English language sentences to French for the native people to understand. The idea is to use Recurrent Neural Network over the encoder-decoder NLP model to attain the result of translating a sentence from English to French. The application will read sentences that will have <EOS> tagging and will be encoded to an internal representation of lengths that are bounded. The decoder model will then be used to collect the words from the encoded input. The decoder model stops whenever the <EOS> tagging is reached. The encoder-decoder model uses the Long short-term memory (LSTM) for first encoding and then decoding.

2 Related Work

2.1 Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Task: Google's Neural Machine Translation system is an end-to-end learning approach for automated translation, which attempts to address issues like robustness (input sentences with rare words), computational expenses in training and

translation inference on large data sets and the weakness of traditional methods. This is done by a model consisting of deep LSTM network with 8 encoder and 8 decoder layers using residual and attention connections. And employs low-precision arithmetic during inference computations and beam search techniques for normalization procedures to increase the translation quality.

Data: The datasets used where publicly available corpora WMT'14 En->Fr(training set includes 36M sentence pairs) and En->De (training set includes M sentence pairs) for GNMT models with work, character, word-piece based vocabularies. In addition to this, the GNMT was tested on Google's translation production corpora (magnitude which is 2-3 decimal order of magnitude bigger than WMT corpora).

Development set = newstest2012+newstest2013
Test set = newtest2014

Approach: In the implementation, the model consists of encode (bidirectional LSTM at bottom layer), decode (RNN+softmax layer) and attention networks. the recurrent networks are deep stack (better then shallow) Long Short-Term Memory (LSTM) RNNs. Our LSTM RNNs have 8 layers, with residual connections between layers to encourage gradient flow. For parallelism (model and data), they connect the attention from the bottom layer of the decoder network to the top layer of the encoder network and train n model replicas concurrently using a Downpour SGD algorithm. To improve inference time, they employ low-precision arithmetic for inference, which is further accelerated by special hardware (Google's Tensor Processing Unit, or TPU). To effectively deal with rare words, they use sub-word units (also known as "wordpieces") for inputs and outputs in our system. Using wordpieces gives a good balance between the flexibility of single characters and the efficiency of full words for decoding and sidesteps the need for special treatment of unknown words. Their beam search technique includes a length normalization procedure to deal efficiently with the problem of comparing hypotheses of different lengths during decoding, and a coverage penalty to encourage the model to translate all the provided input.

Evaluation: The model GNMT was evaluated using BLEU score matrix. As it doesn't capture the quality of translation fully, side by side (S*S) evaluation was carried on where the human raters evaluate and compare the quality of two translations presented side by side for a given source sentence. And it ranges from 0(completely non sense translation) to 6(perfect translation).

The models in experiments are word-based, character-based, mixed word-based or several wordpiece models with varying vocabulary sizes (BLEU increases with increase in vocabulary size and decreasing speed for training). The above step resulted in 38.95 BLEU score with WMP-32K(En->Fr). The second step, RL training which was used to fine-tune sentence BLEU after normal MLE training increased the score by 1. Outcome of ensembling 8 RL-refined models in the next step was 41.16 BLEU.

Finally, to better understand the quality of model and the effect of RL refinement, 4-way side-by-side human evaluation and BLEU scores were compared on phrase based statistical translation and NMT with or without RL refined models. All the GNMT models are wordpiece models, without model ensembling, and use a shared source and target vocabulary with 32K wordpieces. The results show that our model reduces translation errors by more than 60% compared to the PBMT model on these major pairs of languages. And the results between GNMT and human cases are indistinguishable.

2.2 Neural Machine Translation in Linear Time

Task: Neural networks for machine translation either have running time that is super linear in the length of the source and target sequences, or they process the source sequence into a constant size representation, burdening the model with a memorization step. Both drawbacks grow more severe as the length of the sequences increases. To overcome this, they presented a neural translation model, the ByteNet, and a neural language model, the ByteNet Decoder, that aimed at addressing these drawbacks. The ByteNet used convolutional neural networks with dilation for both the source network and the target network. The ByteNet connected the source and target networks via stacking and unfolded the target network dynamically to generate variable length output sequences. They viewed the ByteNet as an

instance of a wider family of sequence-mapping architectures that stack the sub-networks and used dynamic unfolding. The sub-networks themselves may be convolutional or recurrent.

Data: They did raw character-level machine translation on the NewsTest English-German WMT data. The number of sentences were 3003, number of german-words were 63078, number of English-words were 67624, number of distinct german-words were 13930 and number of distinct English words were 10458.

Approach: ByteNet architecture is composed of a target network that is stacked on a source network and generates variable-length outputs via dynamic unfolding. The target network, referred to as the ByteNet Decoder, is a language model that is formed of one-dimensional convolutional layers that use dilation and are masked. The source network processes the source string into a representation and is formed of one-dimensional convolutional layers that use dilation but are not masked.

Each sentence is padded with special characters to the nearest greater multiple of 25. Each pair of sentences is mapped to a bucket based on the pair of padded lengths for efficient batching during training. Sub-BN learns bucket-specific statistics that cannot easily be merged across buckets, this was tackled by circumventing this issue by simply searching over possible target intervals as a first step during decoding with a beam search; each hypothesis uses Sub-BN statistics that are specific to a target length interval.

Evaluation: They took BLEU points to compare the results that they achieved.
They evaluated the ByteNet on raw character-level machine translation on the English-German WMT benchmark. The ByteNet achieved a score of 18.9 and 21.7 BLEU points on, respectively, the 2014 and the 2015 test sets; these results approach the best results obtained with other neural translation models that have quadratic running time.

The ByteNet used in the experiments had 15 residual blocks in the source network and 15 residual blocks in the target network. As in the ByteNet Decoder, the residual blocks were arranged in sets of five with corresponding dilation rates of 1,2,4,8 and 16. They used residual blocks with ReLUs and Sub-BN. The number of hidden units d was 892. The size of the kernel in the source network was 1×5, whereas the size of the masked kernel in the target network was 1×3. They used bags of character n-grams as additional embeddings at the source and target inputs: for n>2 they pruned all n-grams that occur less than 500 times. For the optimization they used Adam with a learning rate of 0.003.

3 Data

The dataset used to work on this paper was from WMT'14. Specifically, we used the English French parallel corpora: Europarl containing 61 Million words, news commentary (5.5 Million) words and two crawled corpora of 90 Million and 272.5 Million words. But we resized the data to 348 Million words using a data selection method by Axelrod et al. (2011).

Europarl corpus is a set of documents that consists of proceedings of the European parliament. The data are multiple sentences written in one of the European Languages and English in several files. The input is going to be sentences that will be further tokenized. The gold standard output will be translated language of the input file sentences. The data was collected by a group of researchers led by Philip Koehn at University of Edinburgh for some research purposes in statistical /machine Translation [www.statmt.org/europarl]. It has the European parliament proceedings data from 1996 to present. The data came from the website of the European Parliament Proceedings. There is a gold standard to assess the effectiveness of the data. They have been provided by Philip Koehn and are part of the standard dataset available at europarl page. We have 2007723 number of sentences which have their French version and English translated version. These are the true labels that are available and will be helping us out with our model.

Sentences	2,007,723
French Words	51, 388, 643
English Words	50, 196, 035
Input/output Pairs	(I/P: 2,007,723; O/P: 2,007,723)

If all the 21 languages are considered:

Train	1.93GB
Test	15.6MB

Training Dataset 2, 007, 723 Input/Output Pairs for all words. Test Dataset 10896 pairs for test dataset for the French to English words. The model is not using validation dataset. No, this dataset doesn't have a standard split.

4 Methodology

We have used bi-directional recurrent neural networks for both the encoder and decoder with 1000 hidden units. Also, a single maxout hidden layer is used to compute the conditional probability of each target word. Besides in order to train the model we have used stochastic gradient descent (SGD) algorithm with Adadelta to train the model. Finally, after the model is trained, we have used a beam search to translate source sentences with maximum probability. The forward RNN receives the input sequence and generates the forward hidden states of the RNN. Similarly, the backward RNN generates backward hidden states. These hidden states are used to obtain an annotation for every word. Hence, every target word which is to be translated uses these annotations which contains summaries of both the preceding and following words. As RNN focuses on recent inputs largely, these annotations obtained have more information of the words close to the target word which is to be translated. So, this approach tries to align the target word to the word to be translated with maximum probability using these annotations.

Our baseline model is RNN which uses Encoder and Decoder (2 models) for machine translation. We have used Supervised learning where 2 sets of training data (English sentences and their French translated sentences) are used to train the model. The model uses Encoder and decoder based on RNN to train the model. Our baseline model calls these two models.

The hyper-parameters used are:
•max_length=20 #max number of words per sentence
•num_epochs=10 #number of epochs for the model to run
•vocab_size=15000 #total vocab size
•hidden-size=100 #number of neurons used in encoder and decoder
•embedding_size=2000 #size to embed the input data

Various combinations of epochs, hidden size and embedding used have been used to keep the training time and efficiency of the model optimised. We ran the model on CPU and it took more than 10 hours to train the data.

When we used the entire training set (more 2 million sentences) it took forever to train the model and thus we have reduced the size of the training data to 4000 sentences. Training the model on CPU (as we didn't have any GPUs) took a lot of time and hence the reduction in the training size.

5 Experiments

Model	Trained On	BLEU
Baseline	Europarl (4000 sentences)	0
Baseline+2Layered RNN	Europarl (4000 sentences)	0
Baseline+3Layered RNN	Europarl (4000 sentences)	0
Baseline+4Layered RNN	Europarl (4000 sentences)	0

The BLEU score of 0 for Baseline is due to the fact that the baseline was implemented with the use of 12 GPU and 24 computers and trained for over 3 days. We didn't have such computational power and hence reduced the training dataset to 4000 sentences from over 2million sentences that it was originally trained on.

Baseline model+RNN with 5 layers performed better than all the other improvements. Although the BLEU score is 0 for it as well, but the output sentences (French) from input sentences (English) showed better correlation to its English counterparts. The improved model with multi-layered RNN attached better weights to the vectorized sentences and improved the conditional probability for individual words and phrases as well.

With models where we used lesser number of layers, the correlation between the predicted French words and their English counterparts aren't as much or doesn't appear to be as related as with 5 layers.

5 layers of neural network has outperformed the other models, due to better conditional probabilities for each vectorized word as it added better weights to them.

6 Conclusion

The machine translation using encoder and decoder methodology and training the model using recurrent neural network is a recent approach which gives better results as we increase the number of layers in the neural network. We received 0.0 BLEU score primarily because of the availability lower computation power since we trained our model on 4000 and 10000 sentences with embedding size as 2000 as opposed to 2,500,000 sentences and embedding size of 25000 which the paper tried to achieve. Yet, as we moved from 4000 to 10000 sentences number of word predictions increased still it was not enough to predict correct sentences. Yet, this technique offered a better approach to phrase based translations which have been the prime focus up to now.