Machine Translation Using Deep Learning

521H0034 Lương Chí Dũng

Abstract

Machine Translation is a crucial area in Natural Language Processing that enables computers to automatically translate text or speech from one language to another. Over the years, various techniques have been developed, from rule-based systems to statistical models, and most recently, deep learning methods. This paper focuses on translating English to Vietnamese and highlights recent advancements using state-of-the-art deep learning architectures.

CONTENTS

I. Introduction

Machine Translation (MT) is the process of automatically converting text from one human language to another using computer algorithms. Historically, MT systems have been applied in numerous domains including international communication, localization of content, and information retrieval. Early methods, such as rule-based and statistical techniques, laid the groundwork for what would eventually become modern deep learning approaches.

In recent years, the need to improve translation quality and contextual understanding has led to the application of deep learning techniques in MT. This research focuses on the English to Vietnamese translation task, demonstrating how these deep learning methods overcome the limitations of traditional approaches and deliver more fluent and context-aware translations. This work is intended to contribute to both academic research and practical applications in machine translation.

II. Motivation For Deep Learning Approach

Earlier approaches to MT, including rule-based systems and statistical machine translation (SMT), suffered from several limitations:

- **Limited Contextual Understanding:** Rule-based systems required extensive linguistic expertise and could not effectively handle the ambiguity inherent in human language.
- **Data Sparsity:** SMT approaches relied on co-occurrence statistics that often failed when encountering rare or unseen expressions.
- **Inadequate Handling of Long Sequences:** Both methods struggled with capturing long-term dependencies in sentences.

The advent of deep learning provided robust frameworks such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks that improved the sequence modeling capabilities. However, even these approaches had limitations in parallel processing and efficiency.

A significant breakthrough came with the introduction of the Transformer model in the paper *"Attention is All You Need"* [1]. This model replaced recurrent layers with a self-attention mechanism that allows the model to weigh the importance of different parts of the input sequence simultaneously. The self-attention mechanism is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q (queries), K (keys), and V (values) represent linear transformations of the input data and d_k is the dimension of the keys.

Building on these advancements, pretrained models such as T5 and BART have further pushed the boundaries by leveraging large-scale pretraining and fine-tuning techniques. The evolution from classical models to pretrained deep architectures demonstrates the progress from simpler neural networks to highly sophisticated, end-to-end learning systems.

III. Training Method

This study employs two primary training methodologies:

A. *Transformer Seq-Seq Model Trained from Scratch*

For the Transformer sequence-to-sequence model, the following methodology is applied:

1) **Dataset:** The model is trained on the public dataset "mt-en-vi" from Hugging Face, which contains **2,884,451** English-Vietnamese sentence pairs.
2) **Building the Vocabulary:** Use Byte Pair Encoding (BPE) to build vocabularies for both English and Vietnamese. This allows for a subword-level tokenization that efficiently handles rare and out-of-vocabulary words.
3) **Preprocessing:** Utilize dedicated tokenizers (e.g., `tokenizer_en` for English and `tokenizer_vi` for Vietnamese) to convert raw text into sequences of `input_ids`.
4) **Model Architecture:** The model is built with the following hyperparameters:
   - vocab_size_src $= 100\,000$
   - vocab_size_tgt $= 100\,000$
   - d_model $= 512$
   - num_heads $= 8$
   - num_layers $= 12$
5) **Training Process:** The training is performed using the **teacher forcing** strategy in an auto-regressive manner. This implies that during training, the target sequence is provided at each step as input to help guide the model in predicting the next word.
6) **Sequence Length:** Both source and target sequences are truncated or padded to a maximum length (`MAX_LEN = 50`).

B. *Fine-Tuning the Pretrained mBART-50-large Model*

The second approach involves fine-tuning a pretrained multilingual BART [2] (mBART-50-large) model:

1) **Dataset:** The model is fine-tuned on the same dataset as the previous approach.
2) **Preprocessing and Tokenization:** Preprocessing is similar to the Transformer model, relying on the existing tokenizers provided with mBART, which have been optimized for the languages involved.
3) **Model Initialization:** Since the mBART-50-large model comes with pretrained weights, it only requires reloading

these weights along with the associated tokenizers. Fine-tuning then adjusts the model parameters on the specific translation dataset.

4) **Training Parameter:** Consistent with the training process, `MAX_LEN = 50` is used for both source and target sequences.

The use of mBART leverages state-of-the-art pretrained representations, making fine-tuning a more straightforward process that often results in faster convergence and improved performance relative to training from scratch.

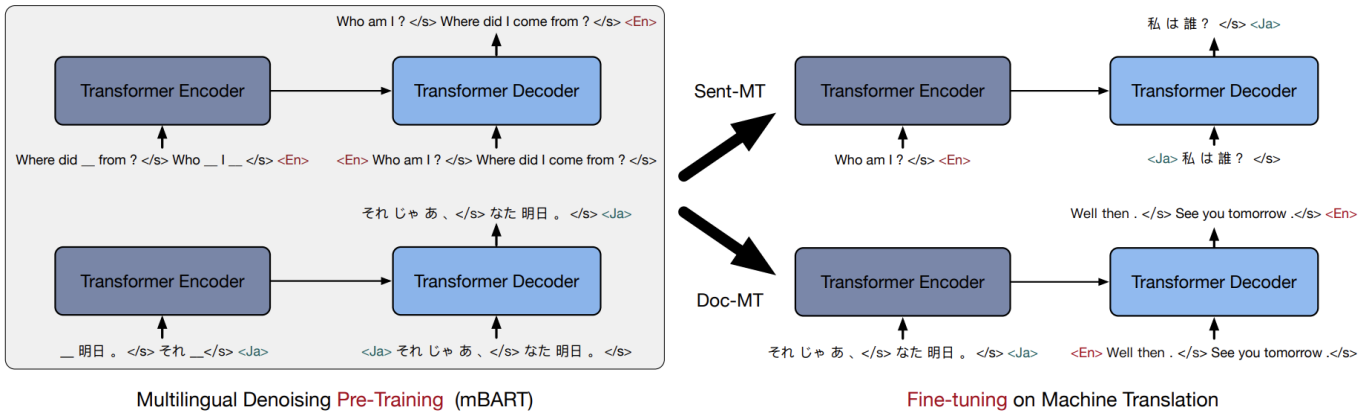

Figure 1: Framework for Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right). source

IV. Evaluation Metric

Evaluating machine translation performance requires careful consideration of both model inference techniques and metric selection.

A. *Inference Methods for Auto-Regressive Models*

During inference, the approach to generating translations differs from the training stage. While training utilizes teacher forcing (where the ground truth is fed as the next input), inference must rely on model predictions:

- **Greedy Search:** At each time step, the token with the highest probability is chosen. While simple, this method may lead to suboptimal sequences.
- **Beam Search:** This method maintains multiple candidate sequences (beams) throughout decoding, striking a balance between exploration and exploitation of the candidate space. Beam Search is the primary method employed in this work. It can better capture long-range dependencies and produce higher-quality translations compared to greedy search.

B. *BLEU and SacreBLEU Metrics*

The BLEU (Bilingual Evaluation Understudy) score is widely used for assessing machine translation outputs due to:

- **N-gram Overlap:** BLEU measures the n-gram overlap between the candidate translation and a set of reference translations, providing an indication of fluency and adequacy.
- **Simplicity:** Despite its limitations, BLEU offers a straightforward and automated way to quantify translation quality.

Additionally, SacreBLEU is used in this research for evaluation. SacreBLEU standardizes the BLEU calculation by ensuring consistency across datasets and experiments. It simplifies the process of comparing results with a common implementation, making it a robust metric for real-world machine translation evaluation.

C. *Evaluation (SacreBLEU)*

The evaluation of the models is conducted using the SacreBLEU metric, and the results are summarized in Table I.

Table I: Evaluation (SacreBLEU) Scores

Model	SacreBLEU Score
Transformer	**28.71**
mBART-50-large	**42.38**

Analysis: The mBART-50-large model significantly outperforms the baseline Transformer, achieving a **+13.67 BLEU** improvement with only 1 epoch of fine-tuning. This substantial gap highlights the benefits of multilingual pretraining and transfer learning, especially in scenarios where the target language (Vietnamese) has relatively fewer resources. While the Transformer model provides a solid foundational approach, pretrained models such as mBART leverage broader language knowledge and exhibit superior generalization capabilities.

References

[1] Vaswani, Ashish, et al. “Attention is All You Need.” Advances in Neural Information Processing Systems, 2017.

[2] Liu, Yinhan, et al. “Multilingual Denoising Pre-training for Neural Machine Translation.” arXiv preprint arXiv:2001.08210, 2020.