UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Streaming neural machine translation systems from English into European languages

Guillem Calabuig Domenech

Jorge Civera Saiz

Javier Iranzo Sánchez

MLLP | Machine Learning and Language Processing

VRAIN
Valencian Research Institute for Artificial Intelligence

21 September, 2022

Contents

1 Introduction

Goals

- To understand the theoretical developments and technological advances that have led neural machine translation (NMT) to be the current state of the art.
- To learn and showcase the different components and processes involved in the development of NMT systems, as well as the importance of data and how it is compiled for these systems.
- To improve the offline NMT models for the English to French translation task that already exist in the MLLP research group.
- To explore, compare and apply different methods to adapt NMT systems to a specific domain in order to significantly improve the translation quality in in-domain evaluation tasks.
- To understand the challenges that exist in building streaming and real time NMT models, how they are evaluated, and construct a system for such task based on offline system results.
- To develop MT systems ready to be deployed and used in real scenarios for the English to French translation task.

Introduction

Machine Translation

- In MT, we search for the best translation $\hat{\mathbf{y}}$ of $\mathbf{x}$ given by

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}^*} p(\mathbf{y}|\mathbf{x})$$

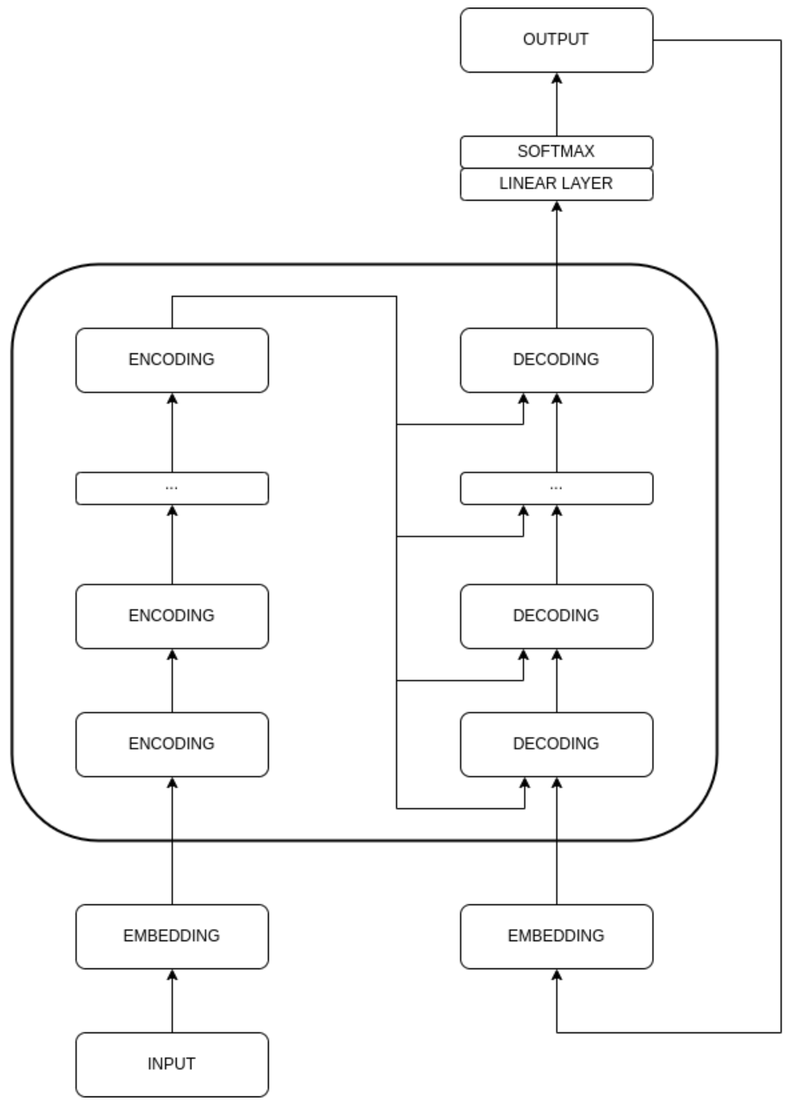
- $\mathbf{x}$ is the *source* sentence and $\mathbf{y}$ is the *target* sentence

- The neural models we will use approximate $p(\mathbf{y}|\mathbf{x})$ directly

Introduction

Transformer

- Deep learning architecture based on the *attention* mechanism
- State-of-the-art in many NLP tasks
- Softmax giving output probabilities from which output phrase is computed



Introduction

Evaluation

- Manual evaluation is very costly
- Automatic evaluation
  - ***BLEU: Bilingual Evaluation Understudy*** $\rightarrow$ Higher is better

Framework

- Research internship at the VRAIN MLLP research group
- Technology-transfer contract between MLLP and CERN
  - En$\rightarrow$Fr MT systems for offline and real-time scenarios

2 Data

Training Dataset

Source	Corpus	Bilingual pairs	Words	
			English	French
Internet	WikiMatrix	2.7 M	57.8 M	63.1 M
	WikiMedia	1.0 M	24.1 M	25.8 M
	Giga Fr-En	22.5 M	575.8 M	672.2 M
	ParaCrawl	216.6 M	3.7 G	4.1 G
	CCAligned	15.6 M	156.7 M	171.1 M
	CommonCrawl	0.1 M	4.1 M	4.7 M
	EUBookshop	10.8 M	224.6 M	244.5 M
	UNPC	30.3 M	658.4 M	816.4 M
	News Commentary	3.2 M	70.7 M	76.6 M
Parliamentary Meetings	DGT-TM	4.9 M	86.3 M	95.4 M
	Europarl	1.2 M	28.6 M	29.9 M
	Europarl-ST	96.5 K	2.3 M	2.6 M
	Total	**309.0 M**	5.6 G	6.3 G

Data

Data Processing

- Filtering → Remove low quality sentences / Reduce noise in data
  - Language identification
  - Source-to-Target length ratio
- Tokenization → Divide text into *tokens*
- Truecasing → Maintain the most frequent version of each token
- Subword Segmentation → Mimic an open vocabulary using token segments
  - Byte-Pair Encoding
  - SentencePiece

Original sentence	Mrs Plooij-van Gorsel, I can tell you that this matter is on is on the agenda for the Quaestors' meeting on Wednesday.
Truecased and tokenized	Mrs Plooij @-@ van Gorsel , I can tell you that this matter is on the agenda for the Quaestors ' meeting on Wednesday .
Truecased, tokenized and BPE encoded sentence	Mrs P@@ loo@@ i@@ j @-@ van Gor@@ sel , I can tell you that this matter is on the agenda for the Qu@@ a@@ est@@ ors ' meeting on Wednesday .
Truecased and SPM encoded sentence	Mrs_ P loo ij - van_ Gor sel ,_ I_ can_ tell_ you_ that_ this_ matter_ is_ on_ the_ agenda_ for_ the_ Qu a est ors '_ meeting_ on_ Wednesday ._

3 CERN News Corpus

Corpus Compilation

- Crawling → Source contents from CERN website
- Alignment → Transform raw text to parallel documents
  - Split lines (MOSES)
  - Hunalign
- Manual Revision → Assure semantic meaning matches in both languages

Bilingual CERN News Corpus

	Sentence pairs	English words	French words
CN21	2200	53.4K	60.8K
CN22	1799	44K	49.9K
CNTraining	55943	1230K	1395K
CNTraining90	50340	1090K	1228K
CNTraining70	39150	891K	1000K
CNTraining50	27971	615K	688K

4 Offline systems

Fairseq

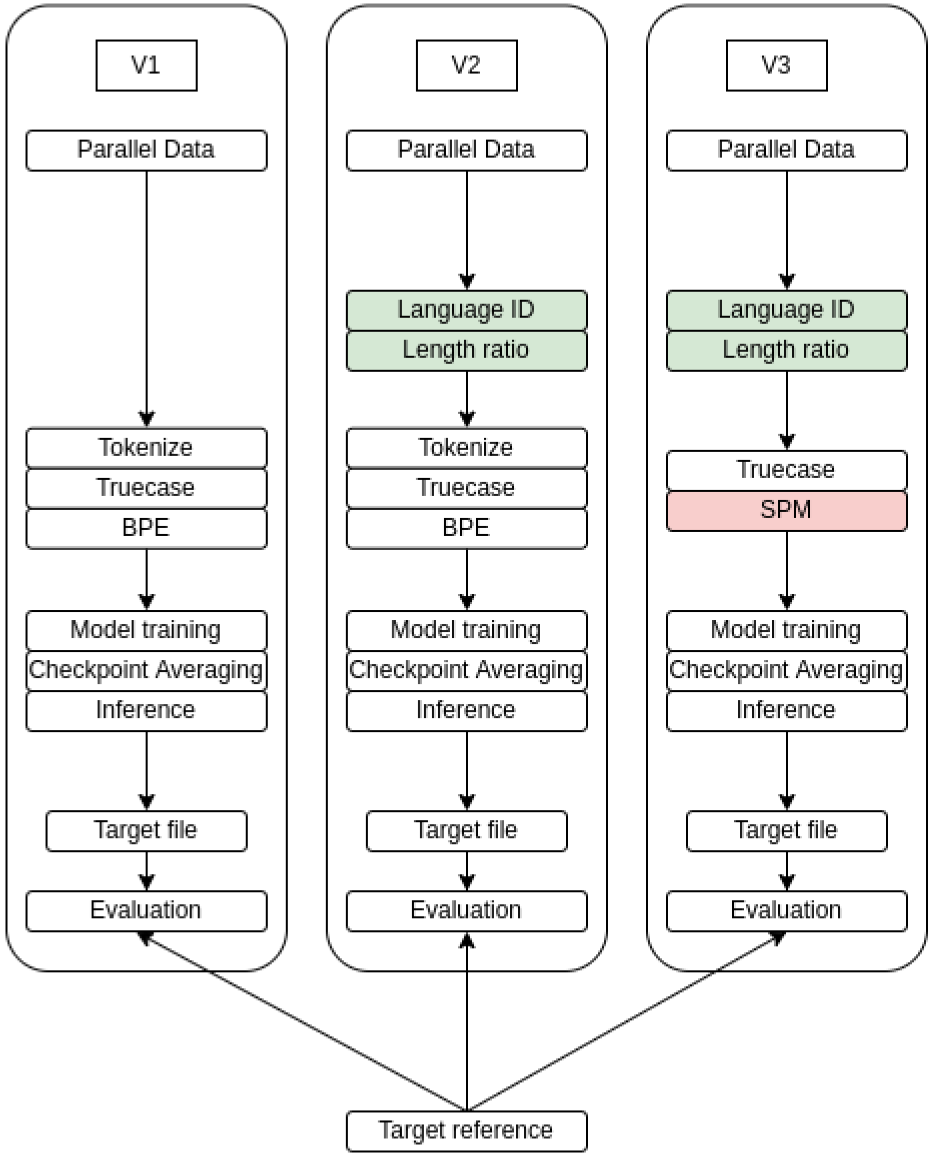
- Facebook AI research implementation of Transformer model (binarize, training, averaging, inference)

Offline scenario

- Consider the full input sentence when computing target sentence
- No time cost is considered

Offline systems

Versions



Offline systems

General Domain Results

System	BLEU	
	WMT13	WMT14
V0	**34.8**	40.8
V1	32.1	39.1
V2	32.6	39.4
V3	34.0	**41.0**

In-Domain Results

System	BLEU	
	CN21	CN22
V0	37.2	37.7
V3	**38.3**	**38.7**

5 Domain Adaptation

Backtranslations

- Translate monolingual text in the target language to the source language
  - Construct a synthetic parallel corpus
- Leverage monolingual data in the target language and domain of interest
- CERN Document Server monolingual resource → 1.4M French sentences
- Add syntethic parallel data to training corpus

General and In-Domain Results

System	WMT13	WMT14	CN21	CN22
V0	**34.8**	40.8	37.2	37.7
V3	34.0	**41.0**	38.3	38.7
V4	34.0	40.9	**38.6**	**38.8**

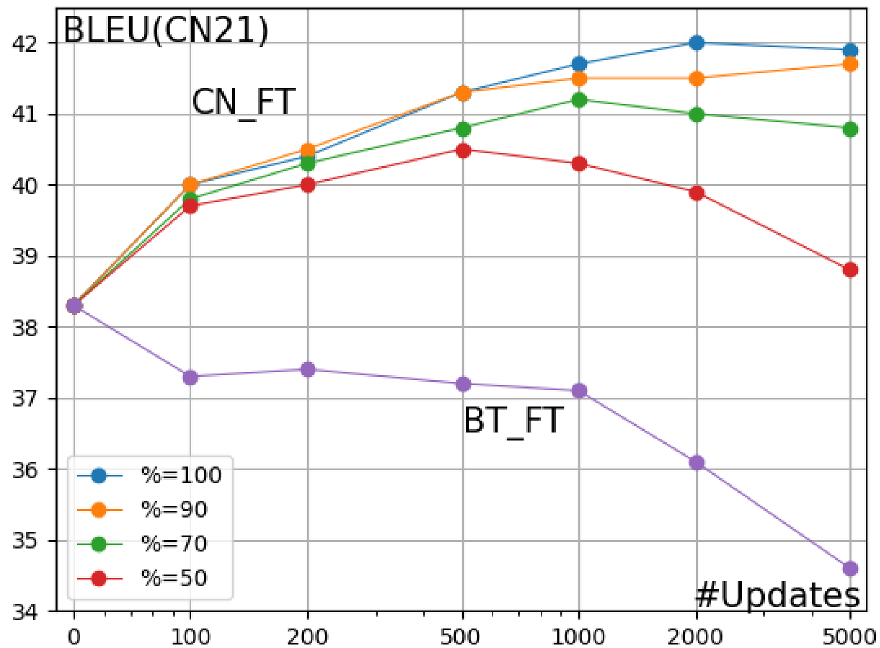
Domain Adaptation

Fine-Tuning

- A model trained for a general task is used in a specific task or domain
- We modify the model parameters to *adapt* it to the domain using in-domain data
- Two different fine-tunings of our models
  - CERN News training data
  - CDS Backtranslations

Domain Adaptation

Finetuning Results



- Best in-domain results at 2000 finetuning updates achieved with CERN News training set

System	CN21	CN22
V0	37.2	37.7
V3FTk100	42.0	**43.1**
V4FTk100	**42.3**	42.9

6 Streaming MT

- To simultaneously translate $\mathbf{x}$ into the target $\hat{\mathbf{y}}$, we find the best translation by

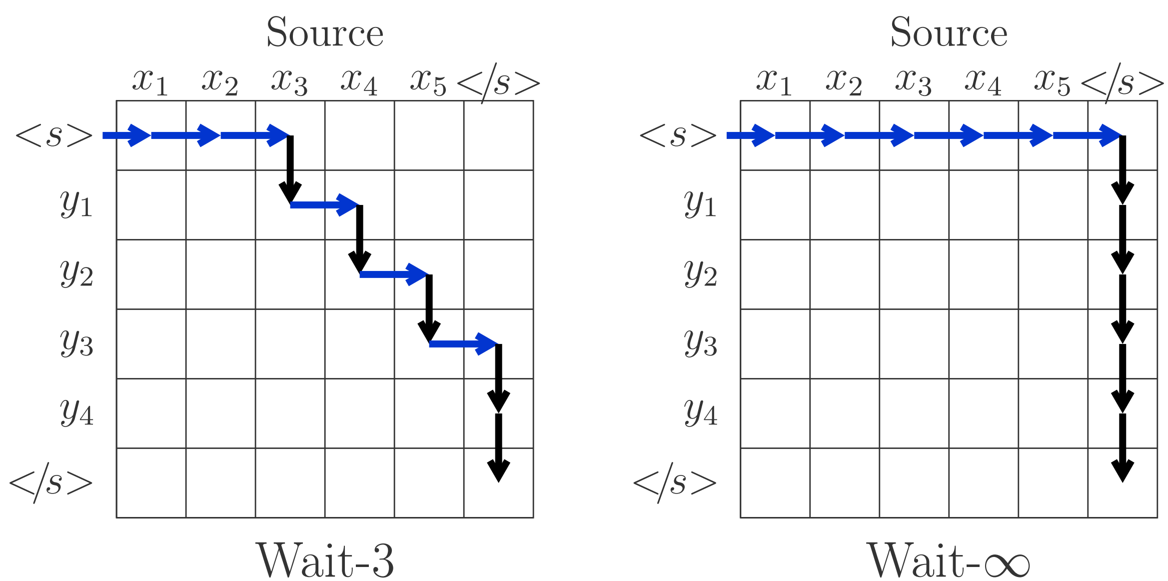
$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}^*} p_g(\mathbf{y} \mid \mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}^*} \prod_t p(y_t \mid \mathbf{x}_{\leq g(t)}, \mathbf{y}_{<t}).$$

- The model only has access to a prefix of the full source to translate
- It needs a *policy* g to decide when to perform a reading or writing action

Offline	Hay libros que valen la pena volver a leer.											
	wait whole sentence	There are books that are worth reading again.										
Simultaneous	Hay libros		que		valen		la	pena	volver		a	leer.
	wait 2 words	There	are		books	that		are		worth	reading	again.

Wait-k Policy

- The model reads k tokens before emitting translations
  - Afterwards, it alternates between writing and reading operations



- ***Multi-Path Wait-k***: simultaneous MT training scheme that considers different values for k

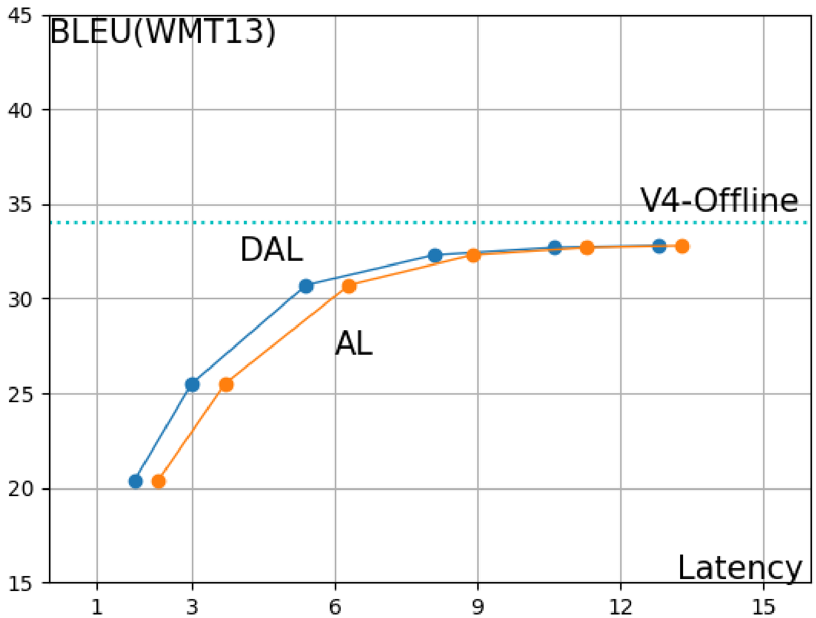
Streaming MT

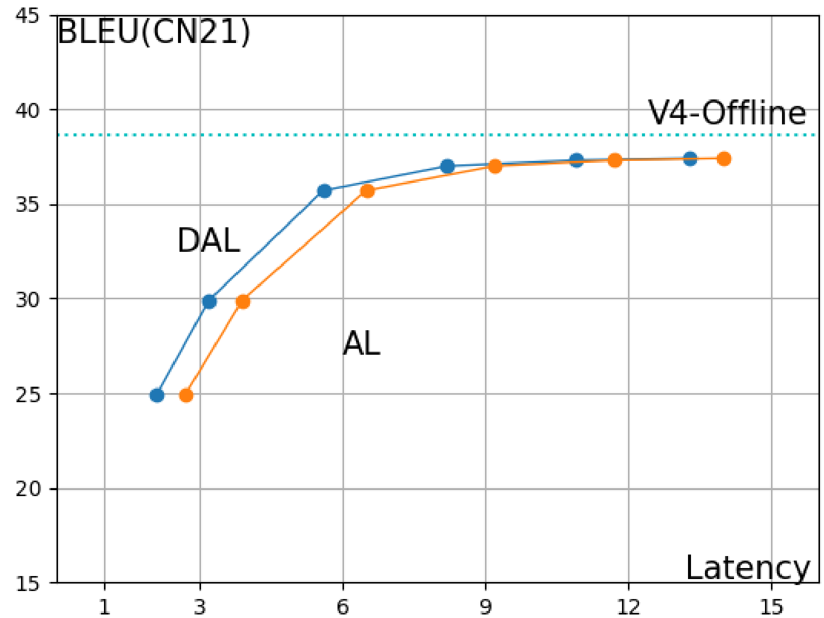
Latency Evaluation

- Automatic evaluation of latency, independent from hardware and environment
- Based on the number of source words available when producing the t'th target word
  - ***AP: Average Proportion*** $\rightarrow$ Average policy value among all writing times
  - ***AL: Average Lagging*** $\rightarrow$ Does not account for the cost of writing operations
  - ***DAL: Differentiable Average Lagging*** $\rightarrow$ Accounts for the cost of writing operations
- ***AL*** and ***DAL*** are grounded on the idea of counting how many words the system falls behind a speaker being live translated

Streaming MT

Latency Results





- V4 Multi-Path Wait-k with $k = 6$ at inference time

Multi-k	WMT13	WMT14	CN21	CN22
BLEU	30.7	36.2	35.7	35.9
AP	0.8	0.7	0.7	0.7
AL	5.4	5.5	5.6	5.5
DAL	6.3	6.4	6.5	6.5

7 Conclusions

Achieved Goals

- State of the art offline and simultaneous NMT were studied
- Different data processing techniques were assessed
- A parallel dataset to train and evaluate NMT systems was compiled from CERN News
- In-domain MT systems improved general-domain systems by a relative 12.9%
- Streaming MT systems were developed and evaluated in terms of the trade-off between accuracy and latency
- Offline and online MT were built systems to be integrated on CERN premises

Future work

- Deployment of MT systems on CERN premises
- Domain adaptation for simultaneous MT systems
- Study of alternative fine-tuning techniques to perfor domain-adaptaion
- Compilation of new in-domain datasets

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Streaming neural machine translation systems from English into European languages

Guillem Calabuig Domenech

Jorge Civera Saiz

Javier Iranzo Sánchez

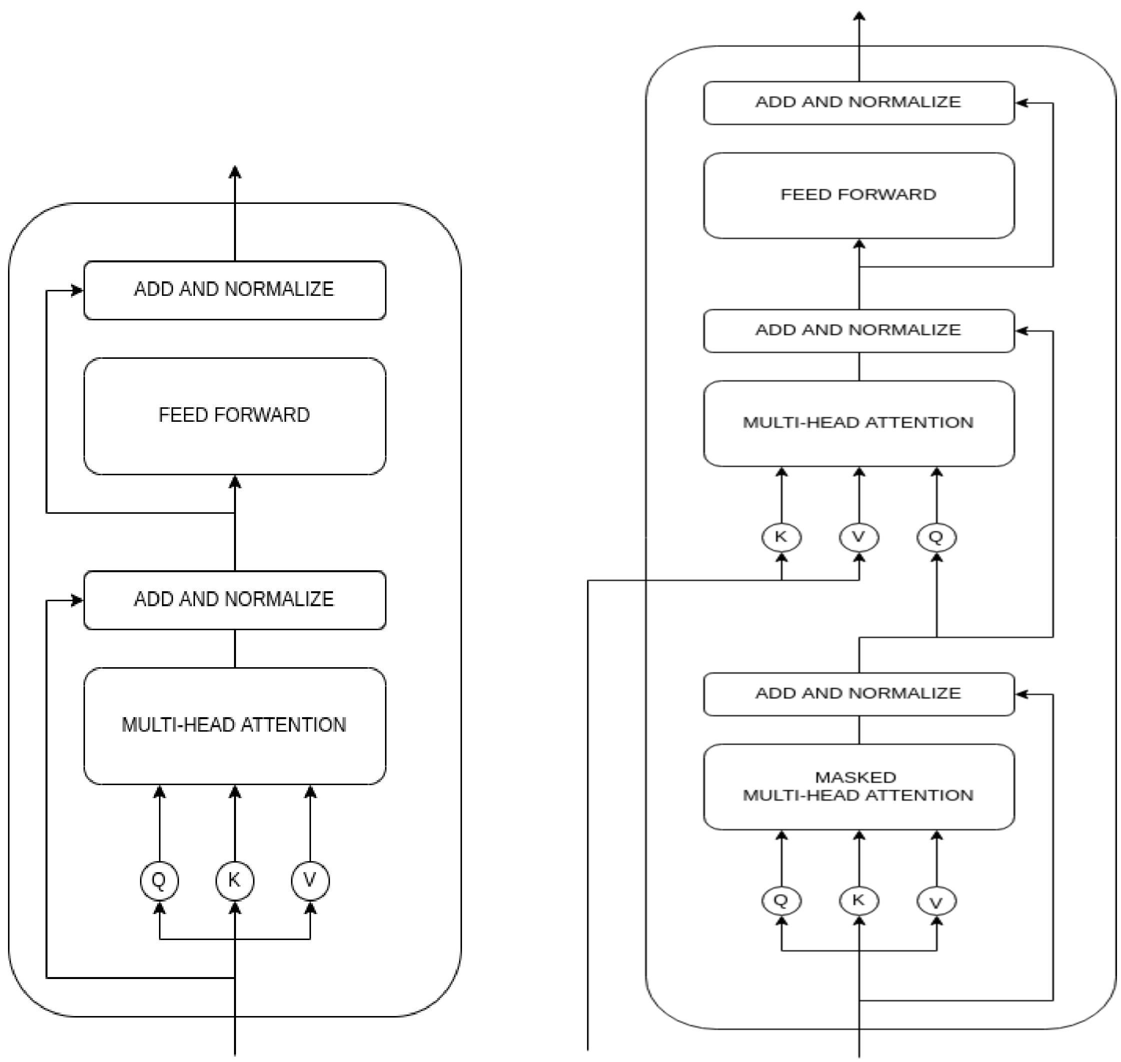
MLLP Machine Learning and Language Processing

VRAIN Valencian Research Institute for Artificial Intelligence

21 September, 2022

Appendix

Transformer Encoder-Decoder Units



(a) (b)

Appendix

Quality Evaluation

BLEU

$$BLEU(4) = BrevityPenalty \times AveragePrecision(4)$$

$$BrevityPenalty = \begin{cases} 1 & |output| > |reference| \\ \exp\left(1 - \frac{|output|}{|reference|}\right) & |output| \leq |reference| \end{cases}$$

$$AveragePrecision(N) = \frac{1}{N}\sum_{n=1}^{N} log p_n, \qquad p_n = \frac{\text{matching n-grams}}{\text{total n-grams in output}}$$

Appendix

Simultaneous Translation

- $g(i)$ is the number of source tokens read when writing a translation at position i.

Wait-k

$$g_{\text{wait}-k}(i) = \left\lfloor k + \frac{i-1}{\gamma} \right\rfloor, \qquad \gamma = \mathbb{E}\big[\gamma_n\big], \qquad \gamma_n = \frac{|\mathbf{y}_n|}{|\mathbf{x}_n|}.$$

Multi-Path Wait-k

For one wait-k path $\mathbf{z}^k_{<i}$:

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k) = \prod_i p(y_i \mid \mathbf{x}_{\leq \mathbf{z}^k_i}, \mathbf{y}_{<i}, \mathbf{z}^k_{<i}).$$

We optimize over multiple wait-k paths:

$$\mathbb{E}_K[p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k)] \approx \prod_{k \sim \mathcal{U}(K)} p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^k).$$

Appendix

Average Proportion

$$AP = \frac{1}{|\mathbf{x}||\mathbf{y}|} \sum_{i=1}^{\mathbf{y}} g(i)$$

Average Lagging

$$AL_g = \frac{1}{\tau} \sum_{i=1}^{\tau} \left(g(i) - \frac{i-1}{\gamma} \right), \qquad \tau = \tau_g(|\mathbf{x}|) = \min_{i:g(i)=|\mathbf{x}|} i$$

Differentiable Average Lagging

$$DAL_d = \frac{1}{|\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} \left(g_{d'}(i) - (i-1)d \right), \qquad d = \frac{1}{\gamma} = \frac{|\mathbf{x}|}{|\mathbf{y}|}$$

$$g_{d'}(i) = \begin{cases} g(i) & i = 1 \\ \max\left(g(i), g_{d'}(i-1) + d \right) & i > 1 \end{cases}$$