Knowledge Distillation for Sustainable Neural Machine Translation

KantanAI
A Keywords Studio

ML Labs

DCU

DUBLIN

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SFI

ADAPT

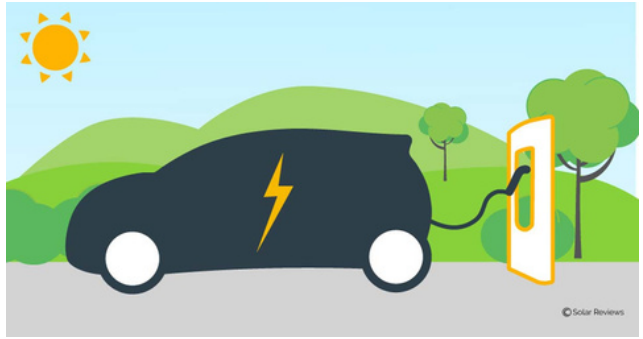
Engaging Content
Engaging People

Overview

- Introduction
  - Motivation
  - Sequence-level Knowledge Distillation
- Background
  - Related Work
  - Extension of Related Work
- Results
- Conclusion
- Future Work

Introduction

Motivation

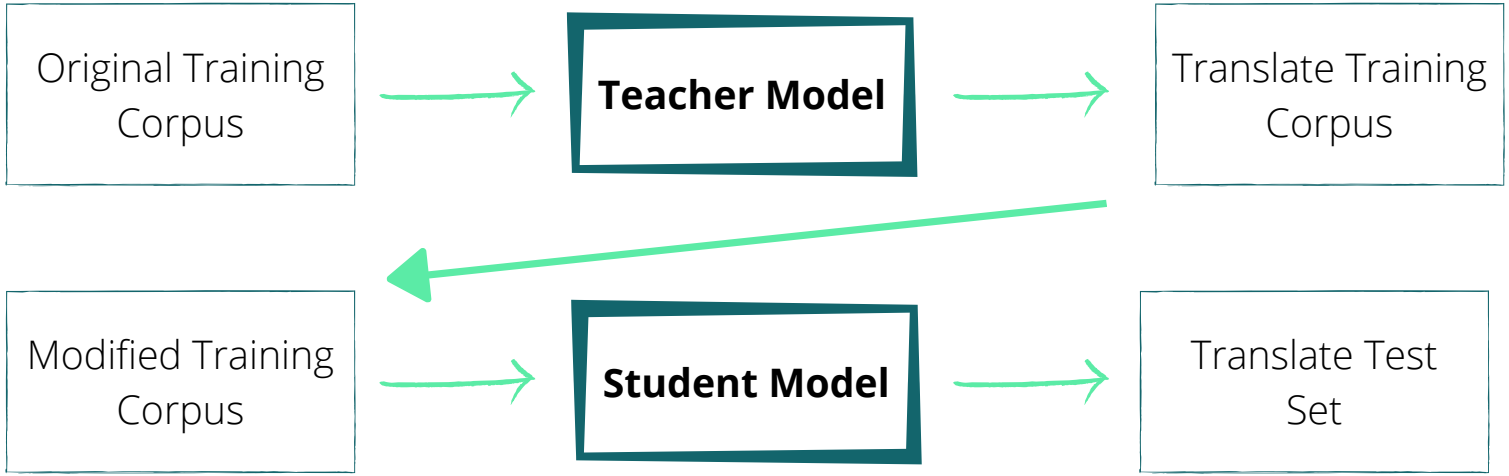


https://unctad.org/news/developing-countries-pay-environmental-cost-electric-car-batteries

Sequence-Level Knowledge Distillation

- The usual training criteria for multi-class classifiers (Bucila et al., 2006; Hinton et al., 2015) can be used to develop a function for knowledge distillation and expanded to use for sequence-level knowledge distillation (Kim and Rush (2016)).
- The aim is to minimise the cross-entropy between the data distribution and model distribution.
- The loss function for sequence-level knowledge distillation uses sequence distributions instead of word distributions.
- Knowledge is distilled by generating a new training set by translating a data set with the teacher model using beam search.

Sequence-Level Knowledge Distillation



Background

Related Work

- We use the Europarl (Tiedemann, 2012) corpus with the parallel sentences in German and English for our experiments.

- The corpus is randomly divided into three subsets. The training set consists of roughly 2 million sentences and the validation and test sets of 3000 sentences, respectively.

- We make use of the MarianNMT toolkit (Junczys-Dowmunt et al., 2018) and Transformer (Vaswani et al., 2017) architecture to train the models for 20 epochs.

Model Type	Corpus	# of enc/dec layers
Baseline	EuroParl	3
Teacher	Europarl	6
Student I	KD	3
Student II	KD + EuroParl	3

Related Work

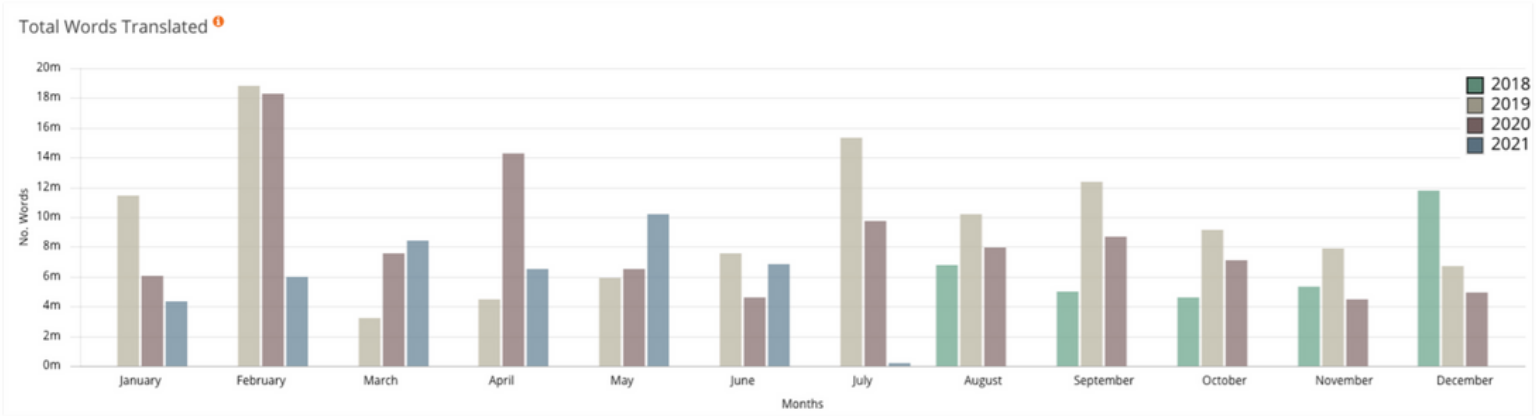
Baseline	# of GPUs	Training time	BLEU ↑
1	1	07:54:15	26.47
2	2	04:51:38	**26.62**
3	4	04:10:44	26.39

Teacher	# of GPUs	Training time	BLEU ↑
1	1	13:34:29	26.71
2	2	08:01:14	26.71
3	4	06:28:52	**26.76**

Related Work

Student	Training set	# of GPUs	Training time	BLEU ↑
1	KD	1	07:17:39	26.68
2	KD	2	04:22:14	26.49
3	KD	4	04:05:28	26.22
4	EuroParl+KD	1	16:10:45	26.82
5	EuroParl+KD	2	09:23:42	26.92
6	EuroParl+KD	4	08:33:50	**27.01**

Related Work



Model	M_{avg} (Translation Time)	Cost (USD)	CO_2 Emissions ($kgCO_2$-eq)
Teacher-1GPU	85.84	1,140.44	13.31
Student-KD+EP-1GPU	46.14	624.86	7.15
Teacher-2GPU	55.21	743.84	8.56
Student-KD+EP-2GPU	30.10	413.34	4.67
Teacher-4GPU	71.24	955.36	11.04
Student-KD+EP-4GPU	37.82	505.88	5.87

Extension of Related Work

- We tested a number of variants of a teacher model for translating the source sentences of our training data.
- The quality of translations by a quantised teacher model would naturally be worse than that of the translations by non-quantized teacher model.
- In other words, you are likely to obtain a worse student model when you use a quantised teacher model to distil knowledge.
- Our investigation focused on examining the magnitude of quality drop of the student models when using the different variants of the teacher models for KD, and in return how faster, cheaper and environmental friendly the KD training process would be.

Extension of Related Work

Setup	Beam Size	Mini/Maxi Batch	Quantisation
Original	12	10/100	fp32
Beam	1	10/100	fp32
Quantisation	12	10/100	fp16
Combined	1	128/256	fp16

Results

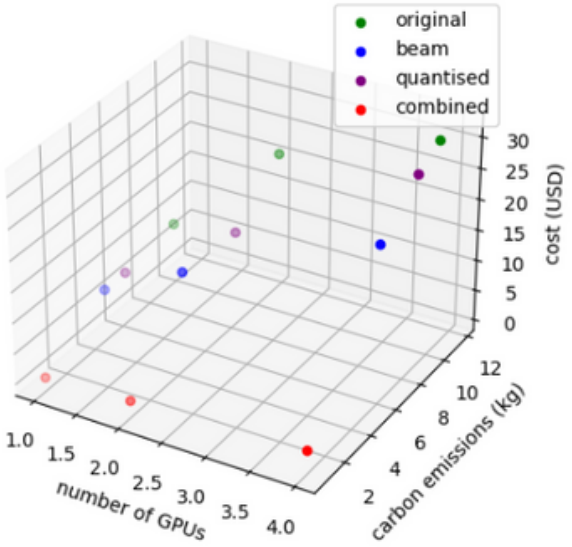
Results

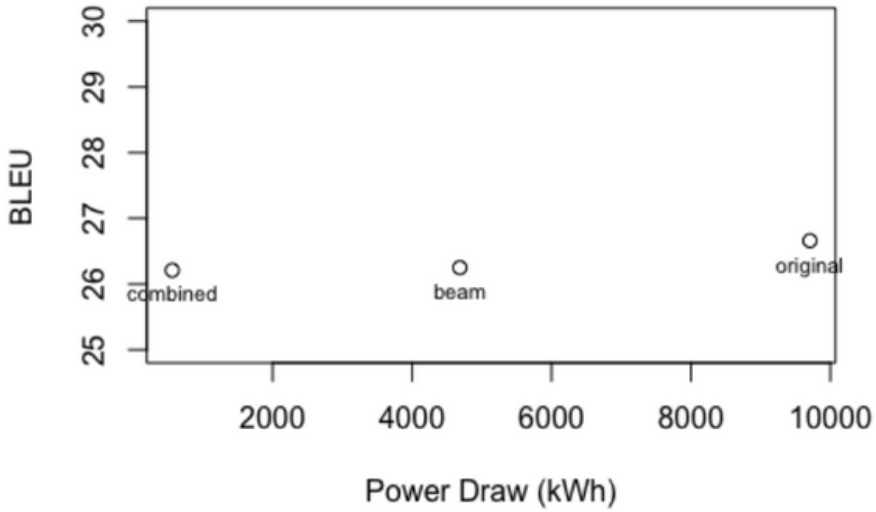
Setup	# of GPUs	Training time	BLEU	TER	chrF
Original	1	07:17:39	26.66	49.5	59.37
	2	04:22:14	26.49	49.3	59.51
	4	04:05:28	26.22	50.1	59.11
Beam	1	06:42:33	26.25	48.6	60.51
	2	04:23:33	26.38	48.5	60.52
	4	04:38:00	26.49	50.3	59.44
Combined	1	06:18:08	26.21	48.7	60.32
	2	04:25:53	26.53	48.7	60.49
	4	04:38:00	26.21	49.5	60.02

Results

Setup	# of GPUs	Time	Power (kW)	CO_2 (kg)	Cost in (USD)
Original	1	13:35:28	9,705.34	9.85 ± 3.32	8.21
	2	10:34:02	11,531.92	11.71 ± 3.94	20.46
	4	11:21:34	10,467.93	10.63 ± 3.58	31.90
Beam	1	11:07:05	4,685.21	4.76 ± 1.60	6.72
	2	07:26:50	4,337.43	4.40 ± 1.48	14.42
	4	08:31:21	5,893.63	5.98 ± 2.01	23.93
Quantisation	1	11:10:30	6,146.91	6.24 ± 2.10	6.75
	2	06:59:42	8,207.46	8.33 ± 2.80	13.54
	4	10:31:47	8,779.62	8.91 ± 3.00	29.57
Combined	1	00:47:36	560.59	0.57 ± 0.19	0.48
	2	00:30:26	618.17	0.63 ± 0.21	0.98
	4	00:42:58	646.55	0.66 ± 0.22	2.01

Results





Conclusion

- We described various methods in which the distillation of knowledge can be made more efficient and in turn more sustainable.
- The impact of batch sizes when using quantisation and a smaller beam size result in a less than 1 BLEU point drop in accuracy.
- At the same time decoding time is reduced by at least 10 hours.
- In terms of efficiency, the combined setup is found to be the best method for distilling knowledge from the teacher to student models.
- The CO2 emissions of our combined setup is on average 10kg less than the original setup while accuracy decreases only slightly.

Conclusion

- The environmental impact of distilling knowledge from a teacher model to a student model is encouraging.
- Taking only the end result (student model) into account is not sustainable and more consideration needs to be put into the whole process.
- During the process of distilling knowledge from a teacher model to a student model, using just 2 GPUs can result in the fastest translation time
- Using 1GPU is the most cost-effective and in most cases the most environmentally friendly as well.
- When taking CO2 emissions and cost into account, using 4 GPUs is much less efficient compared to using only 1 GPU.

Future Work

- In future we will investigate the efficiency of using CPUs during the distillation process as well as during inference when the student models are deployed.
- We also aim to develop a composite metric that takes carbon emissions, accuracy and access to resources into account in order to rate the performance of MT Systems.
- Furthermore, we aim to investigate to what extent these decoding methods work on different language pairs, especially their effect on low-resource languages where access to data is considerably more problematic.