Towards Efficient Speech Translation: Degradation-less 8-bit Quantization

Benedikt Hilmes[(1)], Nick Rossenbach[(1, 2)], Parnia Bahar[(2)] and Ralf Schlüter[(1, 2)]

(1) Machine Learning and Human Language Technology, Computer Science Department, RWTH Aachen University
(2) AppTek GmbH

This work presents a summary of quantized self-attention-based models [1, 2] for automatic speech recognition (ASR), machine translation (MT), and speech-to-text translation (ST) towards improved decoding efficiency and utilization of neuromorphic hardware properties. Compared to the state-of-the-art baselines running in full precision 32-bit float (FP32), the quantized models converted into 8-bit integer (INT8) are more compressed in memory size with almost no loss in accuracy. However, when it comes to decoding time and efficiency, quantization methods are hardware and implementation dependent.

Our experiments on Transformer-based machine translation, Conformer-based speech recognition and the combination of both lie on uniform, per-tensor, asymmetric static quantization [3]. For static quantization all interval calculations are done in a preprocessing step before deploying the model. A calibration set is necessary to calculate data ranges for each tensor in the network, from which the mapping into the lower bit space is calculated. In our work, we explored several calibration methods, and found that for two common calibration methods, *Min Max* [4] and *Percentile* [5] calibration, we can generate quantized models with nearly no degradation. For *Min Max* the observed minimum and maximum values are used as boundaries for the mapping range, while for *Percentile* calibration the values are grouped in a histogram and filtered to cover a given percentile. Nevertheless, the percentile is chosen relatively large, and the amount of data is kept low, which will be further object of analysis in our work. Thus, for successful static quantization we found that it is both equally important to choose a representative calibration set and a calibration method which suits this data set well.

For ASR, we use a Hybrid model that shows good robustness to quantization with an insignificant increase in WER going from 5.96% to 6.09% absolute for simple *Min Max* quantization, using only 5 randomly chosen data samples for calibration. Interestingly, with increased sample size the performance degrades up to an absolute WER of 6.23%, so it is sensitive to the calibration data. We show that this can be mitigated by using the percentile method, but choosing the calibration properly remains a challenge. Different to ASR, both the Transformer-based MT baseline and quantized models show the similar performance of around 29.4 points in BLEU by using the *Min Max* calibration over 30 random samples that is in line with [6]. A smaller calibration set does not lead to significant loss. As ASR and MT were both trained on TED talks, we run the base and quantized cascaded translation models where the speech utterance is first transcribed by the ASR model and then its output is translated into German. On the same test set, the quantized system loses only 0.1 in BLEU over a baseline value of 21.0 BLEU. By this, we show that even if the individual systems degrade slightly, the cascaded ST system is robust enough to show almost no degradation.

In summary, we give a short overview over our ongoing work for post-training quantization of ASR, MT, and ST systems. Based on the experiments on these 3 tasks, for the first time we achieved that quantization to 8 bits does not lead to a degradation of cascaded speech translation. We show that slight differences between ASR and MT are still visible but can be kept to a minimum by setting the calibration parameters carefully. Further studies involve the analysis of calibration behaviour, as a crucial step towards lower bit quantization.

Acknowledgment: This work was partially supported by NeuroSys which, as part of the initiative “Clusters4Future”, is funded by the Federal Ministry of Education and Research BMBF (03ZU1106DA). The work reflects only the authors’ views and the funding party is not responsible for any use that may be made of the information it contains.

References:

[1] A. Vaswani, et. al., in *NIPS, 1-11,* 2017
[2] A. Gulati, et al., in INTERSPEECH, 5036-5040, 2020
[3] M. Nagel, et. al., “A White Paper on Neural Network Quantization”, in arXiv, 2021
[4] V. Vanhoucke, et. al., in *NIPS*, 1-8, 2011

[5] J. McKinstry, et. al., in *EMC2-NIPS,* 6-9, 2019
[6] G. Prato, et. al, *in* EMNLP, 1-14, 2020