Binarizing Translations

Akshat Shrivastava, Kevin Bi, Sarah Yu

Abstract

Here we explore ways to reduce computation and model size for neural machine translation. With the development of binary weight networks and xnor networks in vision, we wanted to extend that work to machine translation. In particular, we want to evaluate how binary convolutions can be used in machine translation and what the effects are.

Experiment & Structure

Data

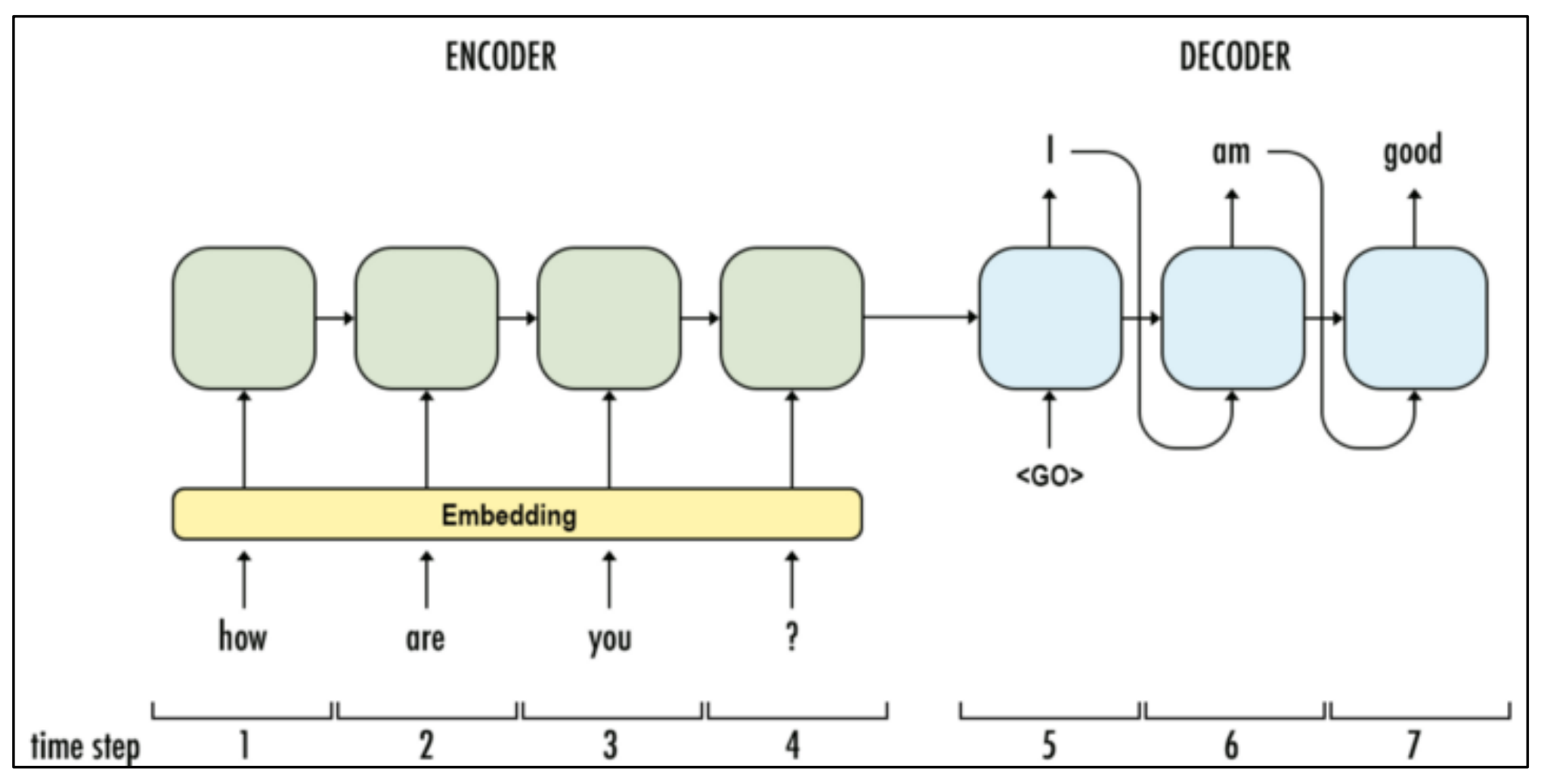
We evaluate our models on the Multi30k dataset, which contains:

- 30k training examples
- 1k valid examples
- 1k test examples

We use this dataset, due to the time constraints of our project, since it is smaller it allows us to tune our models rather quickly and compare against SOTA.

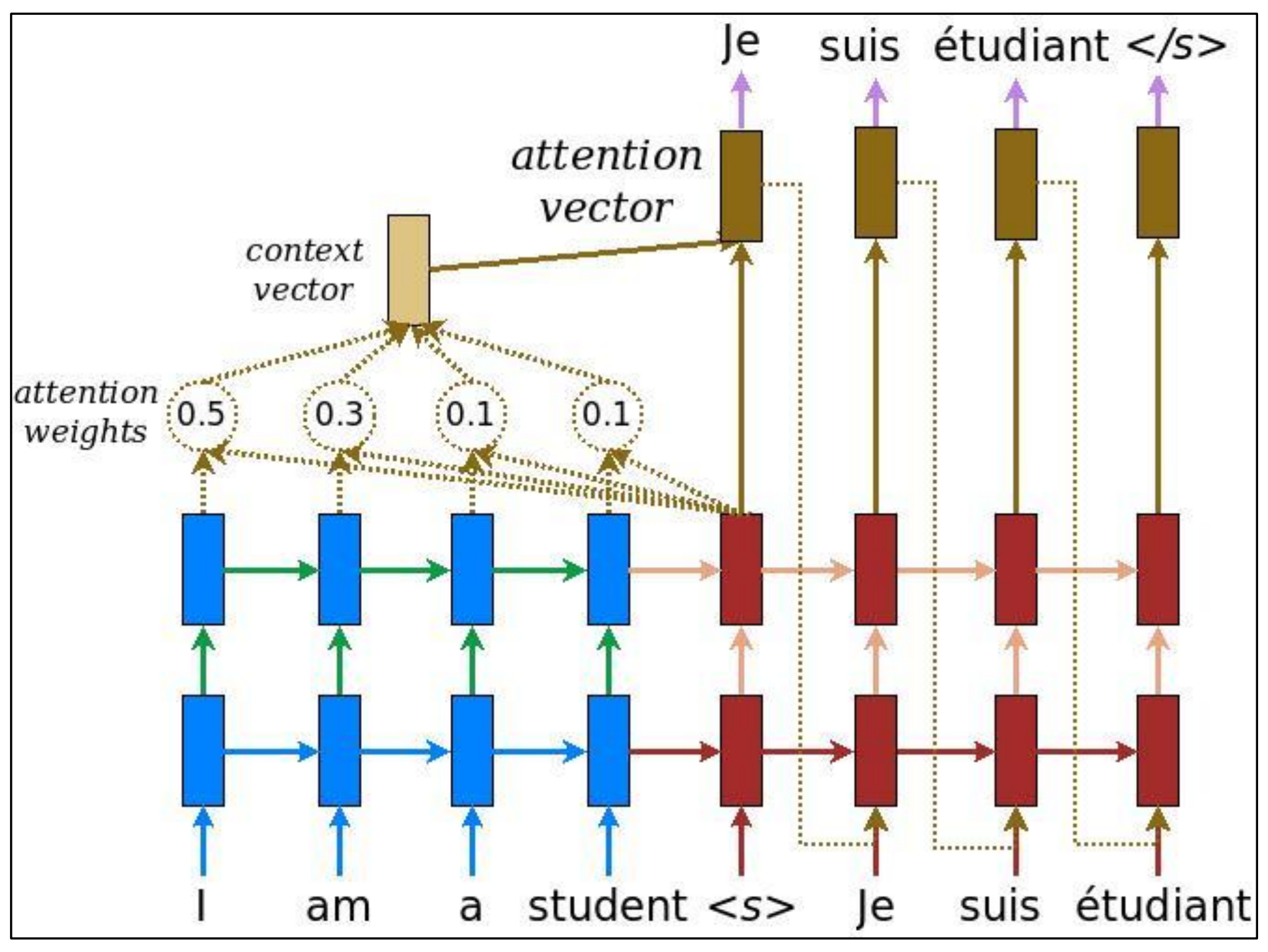
Models

We evaluate 6 different models and compare various metrics on translation performance, runtime, and model size:



- **Simple LSTM**

An encoder decoder model, where the encoder embeds the src tokens and runs them through a 4 layer LSTM encoder then use the final hidden state to run through the decoder

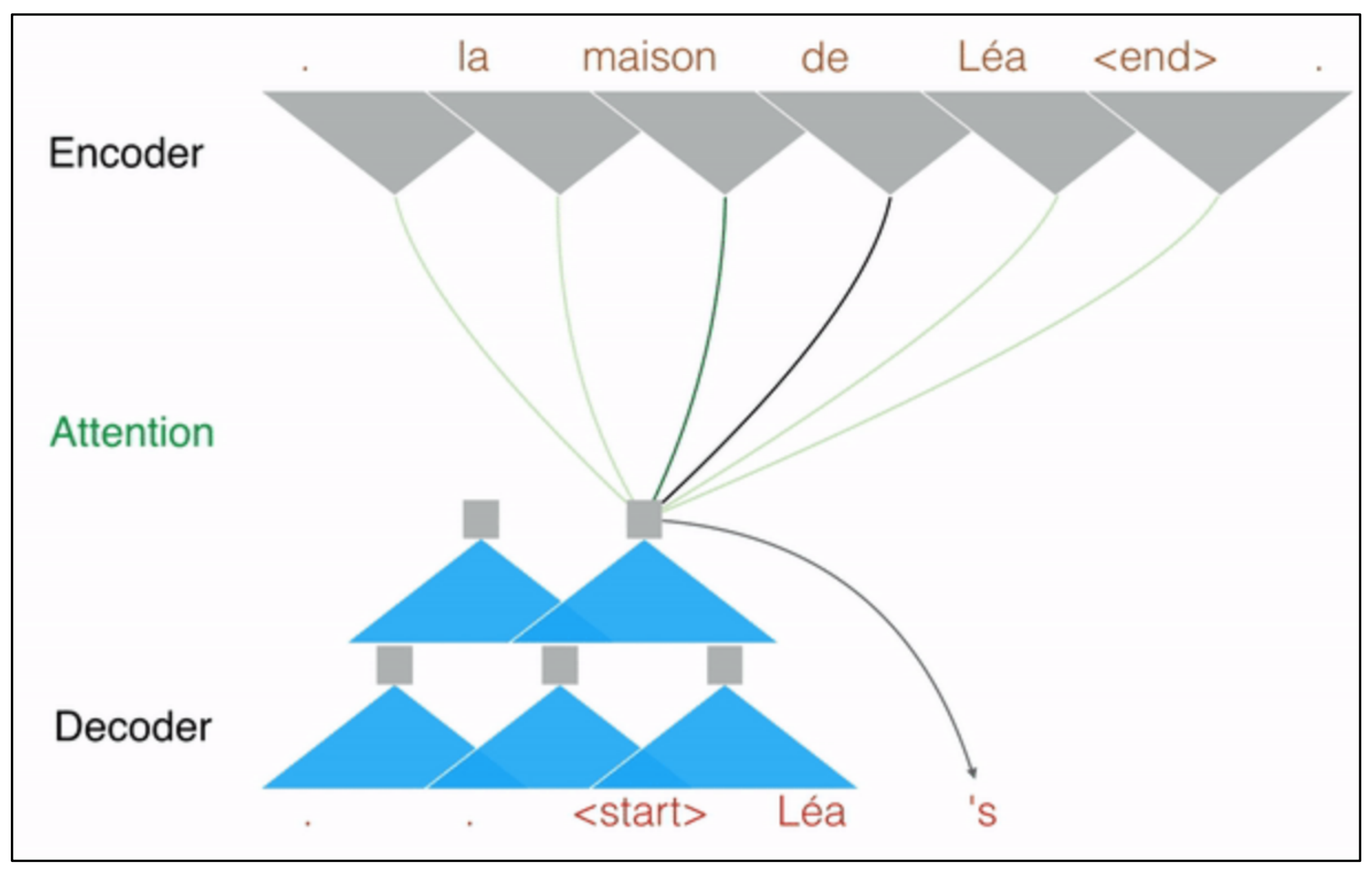


- **Attention RNN**

An encoder decoder model, similar to LSTM, but at every decoder step applies an attention mechanism over all the encoder outputs conditioned on the current hidden state.

- **Attention QRNN**

Attention RNN, but but using QRNN (Quasi Recurrent Neural Network) instead of LSTMs



- **ConvSeq2Seq**

The ConvSeq2Seq models creates a series of convolutional layers that are used for the encoder, and decoder along with attention.

- **ConvSeq2Seq Binarized Weight Network**

ConvSeq2Seq w/binarized weights

- **ConvSeq2Seq Xnor**

ConvSeq2Seq w/ binarized weights and inputs

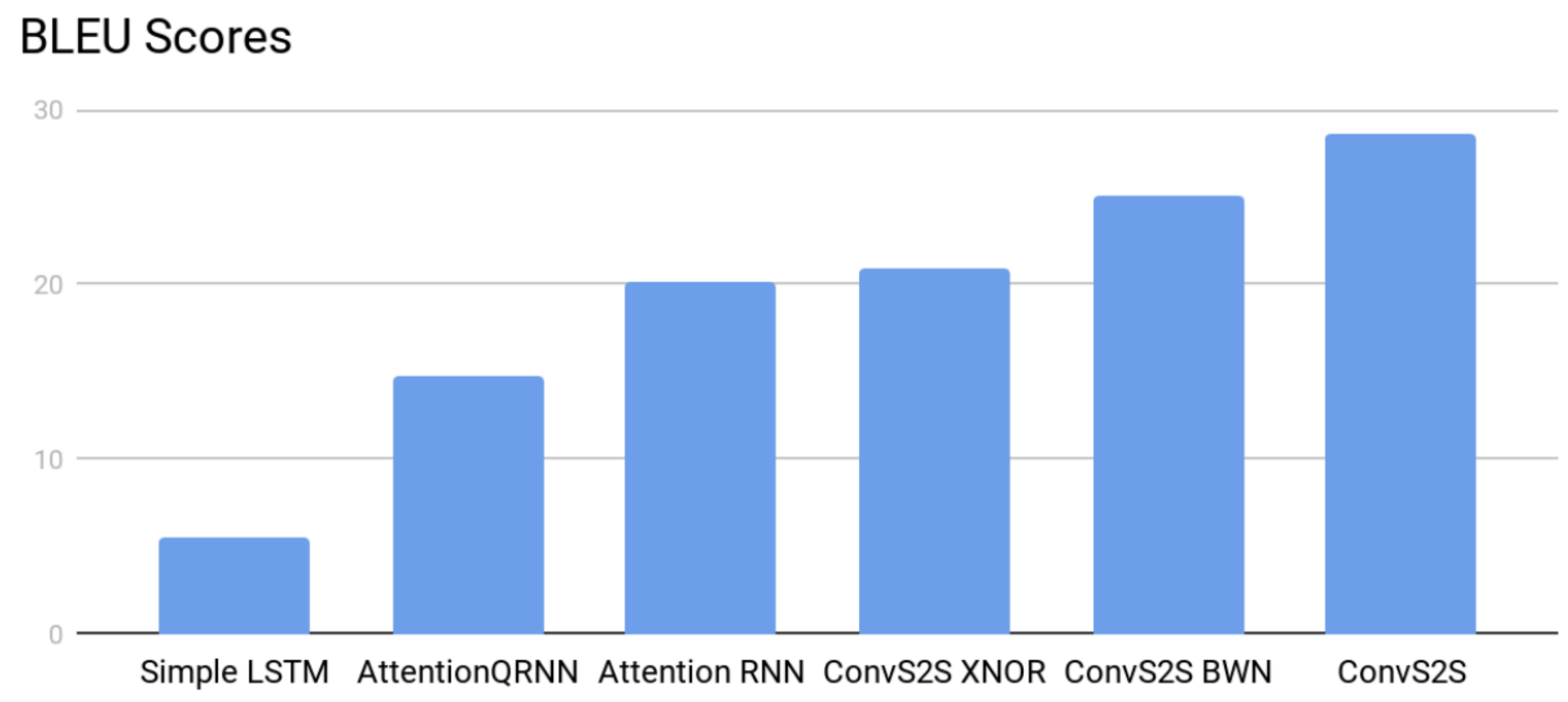
Results

Generated Translations

Source Sentence (English)	*A boy and an old man with a cane are talking.*											
Target Translation (German)	**Ein**	**Junge**	**und**	**ein**	**alter**	**Mann**	**mit**	**einem**	**Stock**		**unterhalten**	**sich**
Attention RNN (English)	*A boy and a bearded man are talking*											
Attention RNN Out	Ein	Junge	und	ein		Mann	mit	Bart			unterhalten	
ConvS2S (English)	*A boy and an old man with a cane are talking*											
ConvS2S Out	Ein	Junge	Und	ein	alter	Mann	mit	einem	Stock		unterhalten	sich
ConS2S BWN (English)	*A boy and an old man with a stick and talk*											
ConvS2S BWN Out	Ein	Junge	Und	ein	alter	Mann	mit	einem	Stock	und	unterhalt	sich

**We show the three most semantically similar examples from the 6 models*

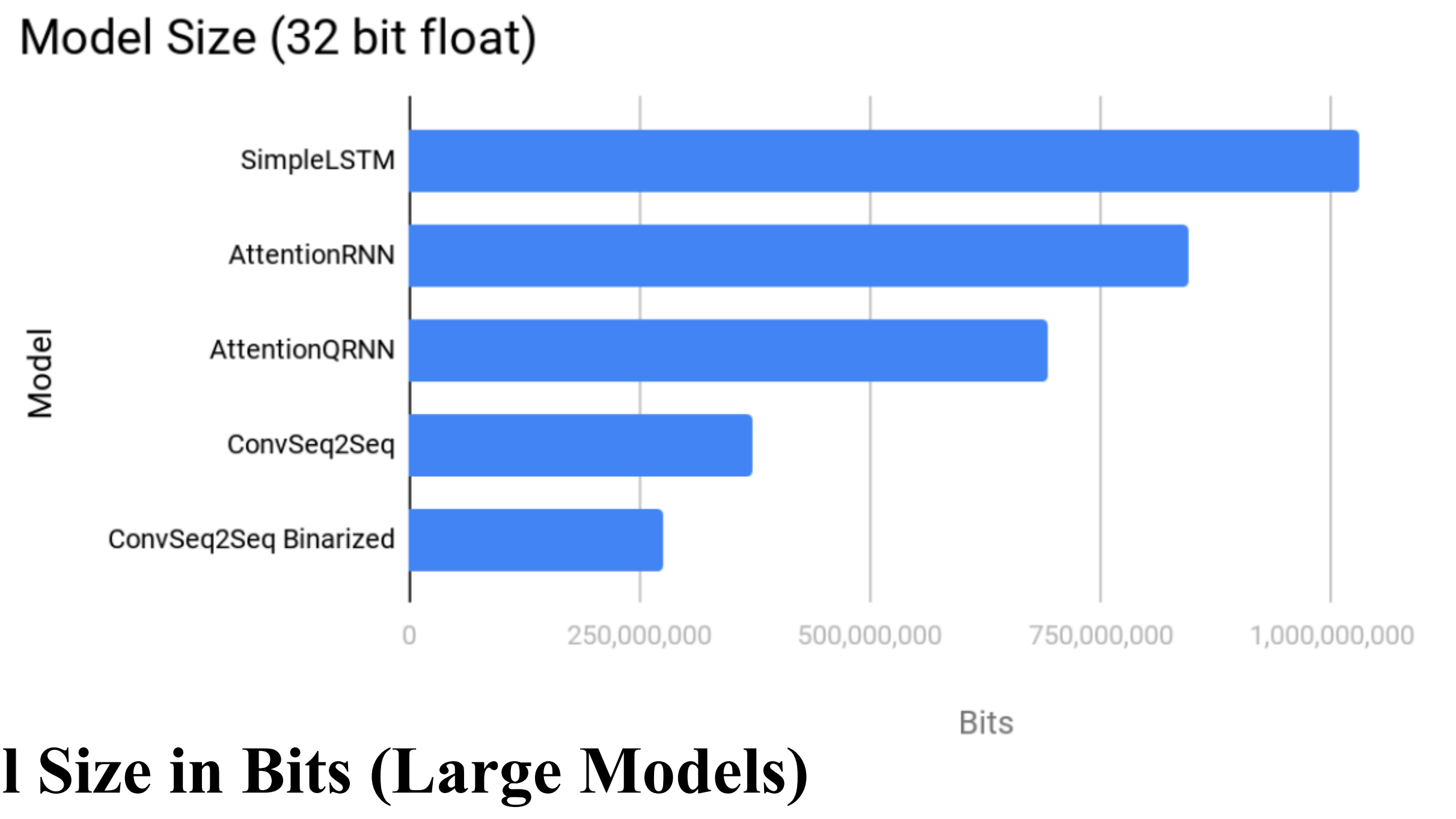
Performance



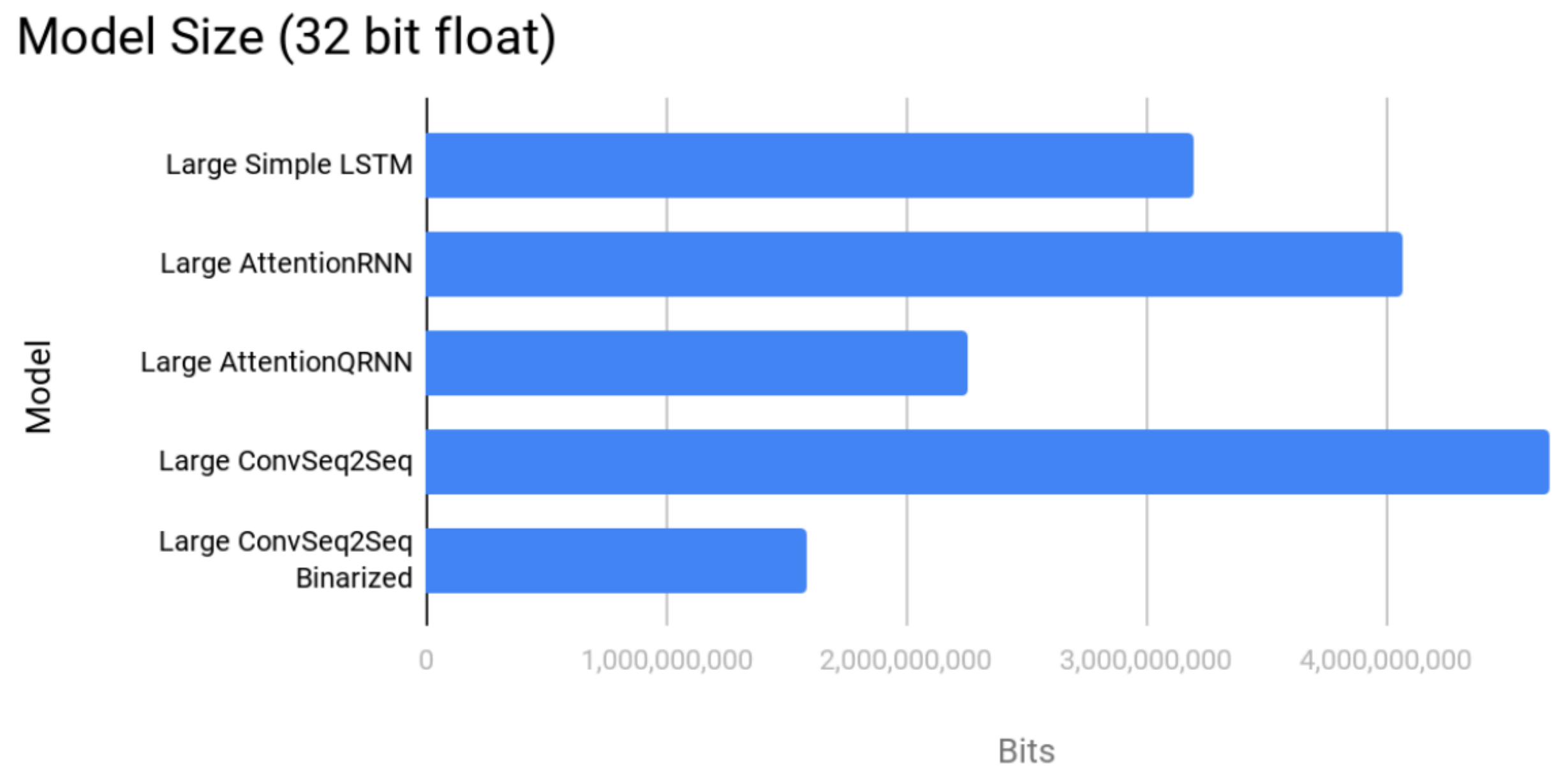
BLEU (Bilingual Evaluation Understudy) evaluates the quality of machine translations; a BLEU score gives the quality of the translation compared to human translation. We care more about this measure in relation to each other, rather than the absolute values.

System Usage

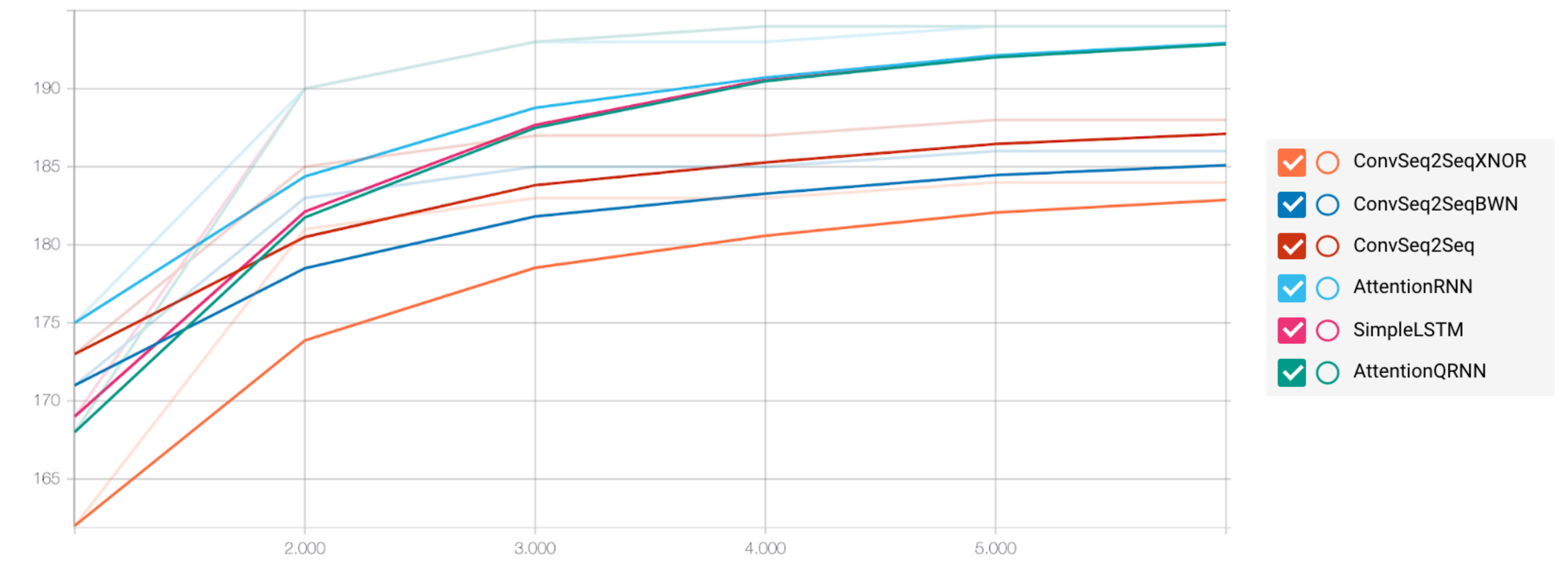
Metric 1: Model Size in Bits



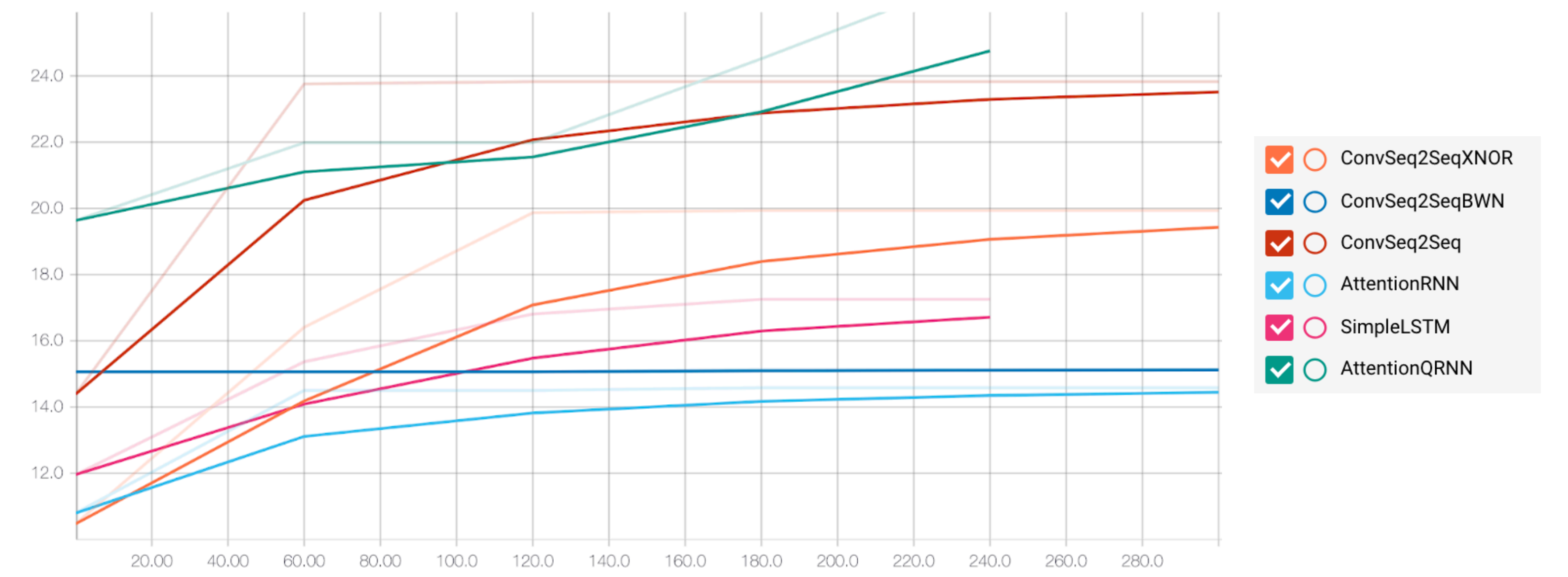
Metric 2: Model Size in Bits (Large Models)



Metric 3: CPU Percentage



Metric 4: GPU Percentage



Conclusion

- We analyze the performance of various different neural architectures for machine translation on both GPU and CPU.
- We show that binarizing a translation network can result in a much smaller model size, while taking a relatively small hit in translation performance.
- This work also shows that translation models can become much quicker with XNOR convolutions with a larger hit to accuracy.