		IWSLT EN→DE	WMT'17 EN→DE		WMT'17 LV→EN	
Encoder	Decoder	BLEU	BLEU	METEOR	BLEU	METEOR
self-att	self-att	25.4 ± 0.2	27.6 ± 0.0	47.2 ± 0.1	18.3 ± 0.0	51.1 ± 0.1
self-att	RNN	25.1 ± 0.1	27.4 ± 0.1	47.0 ± 0.1	18.4 ± 0.2	51.1 ± 0.1
self-att	CNN	25.4 ± 0.4	27.6 ± 0.2	46.7 ± 0.1	18.0 ± 0.3	50.3 ± 0.3
RNN	self-att	25.8 ± 0.1	27.2 ± 0.1	46.7 ± 0.1	17.8 ± 0.1	50.6 ± 0.1
CNN	self-att	25.7 ± 0.1	26.6 ± 0.3	46.3 ± 0.1	16.8 ± 0.4	49.4 ± 0.4
RNN	RNN	25.1 ± 0.1	26.7 ± 0.1	46.4 ± 0.2	17.8 ± 0.1	50.5 ± 0.1
CNN	CNN	25.3 ± 0.3	26.9 ± 0.1	46.1 ± 0.0	16.4 ± 0.2	47.9 ± 0.2
self-att	*combined*	25.1 ± 0.2	27.6 ± 0.2	47.2 ± 0.2	18.3 ± 0.1	51.1 ± 0.1
self-att	*none*	23.7 ± 0.2	25.3 ± 0.2	43.1 ± 0.1	15.9 ± 0.1	45.1 ± 0.2

Table 5: Different variations of the encoder and decoder self-attention layer.

In addition to that, we try a combination where the first and fourth block use self-attention, the second and fifth an RNN, the third and sixth a CNN (*combined*).

Replacing the self-attention on both the encoder and the decoder side with an RNN or CNN results in a degradation of performance. In most settings, such as WMT'17 EN→DE for both variations and WMT'17 LV→EN for the RNN, the performance is comparable when replacing the decoder side self-attention. For the encoder however, except for IWSLT, we see a drop in performance of up to 1.5 BLEU points when not using self-attention. Therefore, self-attention seems to be more important on the encoder side than on the decoder side. Despite the disadvantage of having a limited context window, the CNN performs as well as self-attention on the decoder side on IWLT and WMT'17 EN→DE in terms of BLEU and only slightly worse in terms of METEOR. The combination of the three mechanisms (*combined*) on the decoder side performs almost identical to the full Transformer model, except for IWSLT where it is slightly worse.

It is surprising how well the model works without any self-attention as the decoder essentially looses any information about the history of generated words. Translations are entirely based on the previous word, provided through the target side word embedding, and the current position, provided through the positional embedding.

5 Conclusion

We described an ADL for specifying NMT architectures based on composable building blocks. Instead of committing to a single architecture, the language allows for combining architectures on a granular level. Using this language we explored how specific aspects of the Transformer architecture can successfully be applied to RNNs and CNNs. We performed an extensive evaluation on IWSLT EN→DE, WMT'17 EN→DE and LV→EN, reporting both BLEU and METEOR over multiple runs in each setting.

We found that RNN based models benefit from multiple source attention mechanisms and residual feed-forward blocks. CNN based models on the other hand can be improved through layer normalization and also feed-forward blocks. These variations bring the RNN and CNN based models close to the Transformer. Furthermore, we showed that one can successfully combine architectures. We found that self-attention is much more important on the encoder side than it is on the decoder side, where even a model without self-attention performed surprisingly well. For the data sets we evaluated on, models with self-attention on the encoder side and either an RNN or CNN on the decoder side performed competitively to the Transformer model in most cases.

We make our implementation available so that it can be used for exploring novel architecture variations.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-