Paper Reading: The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation (ACL 2018)

Mia Xu Chen, Orhan Firat, Ankur Bapna

Google AI

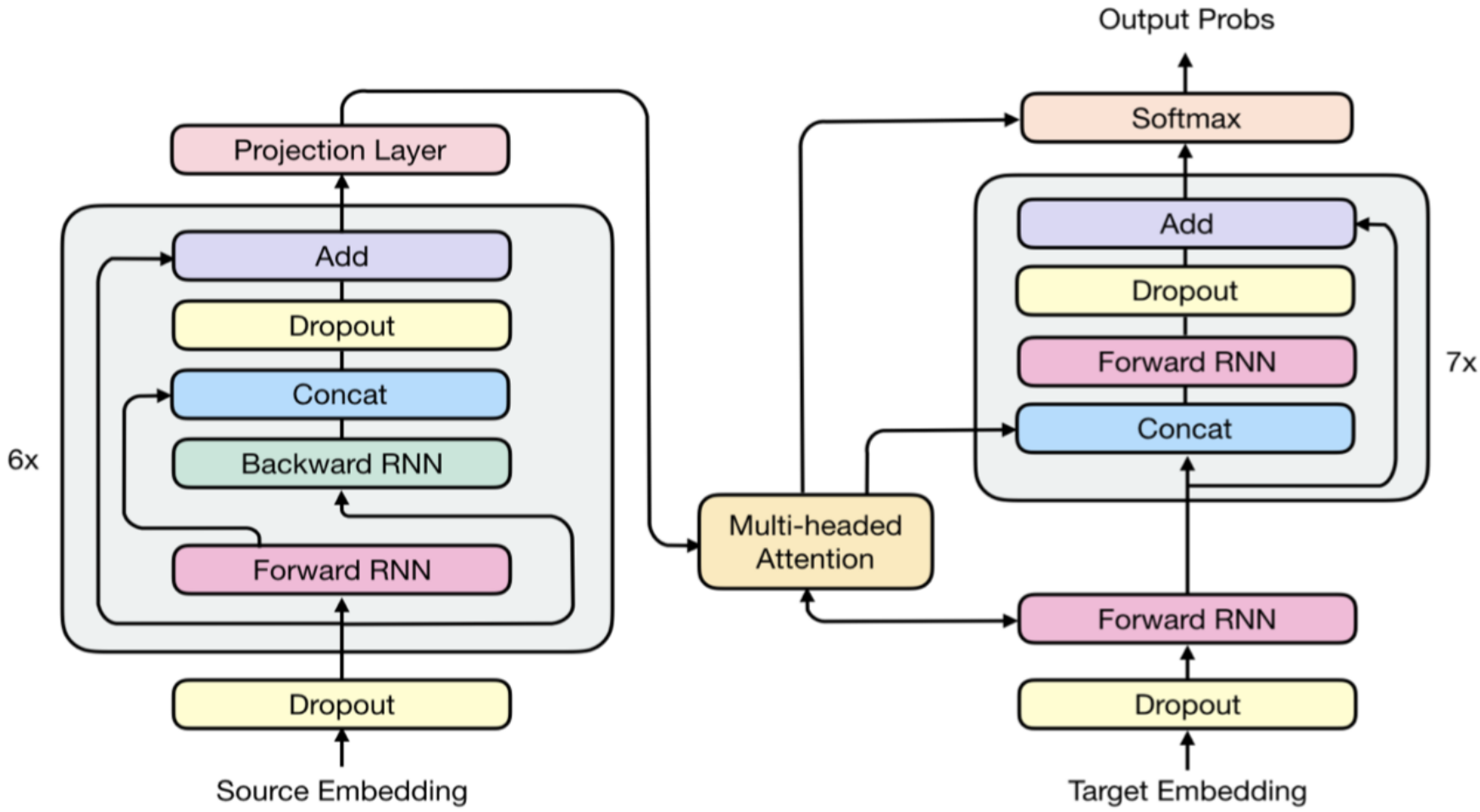
Motivation

- Each NMT models consists of a fundamental architecture accompanied by a set of modeling and training techniques .
- Enhance RNMT model by using many components from other models.

Background: previous NMT models

- Google-NMT: Encoder: one bi-directional LSTM layer followed by 7 uni-directional LSTM layer. Decoder: single attention network and 8 uni-directional LSTM layers.
- Convolutional NMT Models-ConvS2S: Both the encoder and decoder are constructed by stacking multiple convolutional layers.
- Conditional Transformation-based NMT Models-Transformer

Model Architecture of proposed model: RNMT+



Model Analysis and Comparison

Model	Test BLEU	Epochs	Training Time
GNMT	38.95	-	-
ConvS2S [7]	39.49 ± 0.11	62.2	438h
Trans. Base	39.43 ± 0.17	20.7	90h
Trans. Big [8]	40.73 ± 0.19	8.3	120h
RNMT+	41.00 ± 0.05	8.5	120h

Figure: WMT14 En-Fr

Model	Test BLEU	Epochs	Training Time
GNMT	24.67	-	-
ConvS2S	25.01 ± 0.17	38	20h
Trans. Base	27.26 ± 0.15	38	17h
Trans. Big	27.94 ± 0.18	26.9	48h
RNMT+	28.49 ± 0.05	24.6	40h

Figure: WMT14 En-De

Model	Examples/s	FLOPs	Params
ConvS2S	80	15.7B	263.4M
Trans. Base	160	6.2B	93.3M
Trans. Big	50	31.2B	375.4M
RNMT+	30	28.1B	378.9M

Ablation Experiments

Model	RNMT+	Trans. Big
Baseline	41.00	40.73
- Label Smoothing	40.33	40.49
- Multi-head Attention	40.44	39.83
- Layer Norm.	*	*
- Sync. Training	39.68	*

Figure: Ablation results of RNMT+ and the Transformer Big model on WMT'14 En-Fr

Assessing Individual Encoders and Decoders

Encoder	Decoder	En→Fr Test BLEU
Trans. Big	Trans. Big	40.73 ± 0.19
RNMT+	RNMT+	41.00 ± 0.05
Trans. Big	RNMT+	$\mathbf{41.12 \pm 0.16}$
RNMT+	Trans. Big	39.92 ± 0.21

Assessing Encoder Combinations

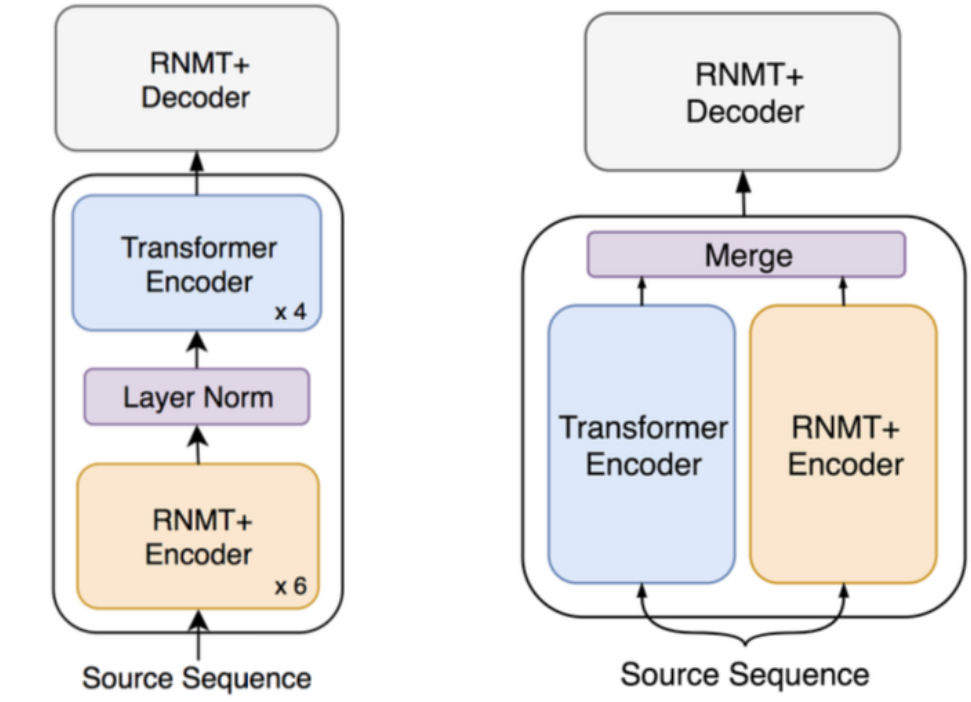


Figure: Vertical and Horizontal mixing of Transformer and RNMT+ components in an encoder

Assessing Encoder Combinations

Model	En→Fr BLEU	En→De BLEU
Trans. Big	40.73 ± 0.19	27.94 ± 0.18
RNMT+	41.00 ± 0.05	28.49 ± 0.05
Cascaded	$\mathbf{41.67 \pm 0.11}$	28.62 ± 0.06
MultiCol	41.66 ± 0.11	$\mathbf{28.84 \pm 0.06}$

Figure: Results for hybrids with cascaded encoder and multi-column encoder

Conclusion

- Enhancing RNMT $\longrightarrow$ RNMT+, outperforms the three fundamental architectures.
- Exploring the efficacy of several architectural in recent studies on Seq2Seq models for NMT, they are broadly applicable to multiple model architectures.