Google AI

The Best of Both Worlds

Combining Recent Advances in Neural Machine Translation

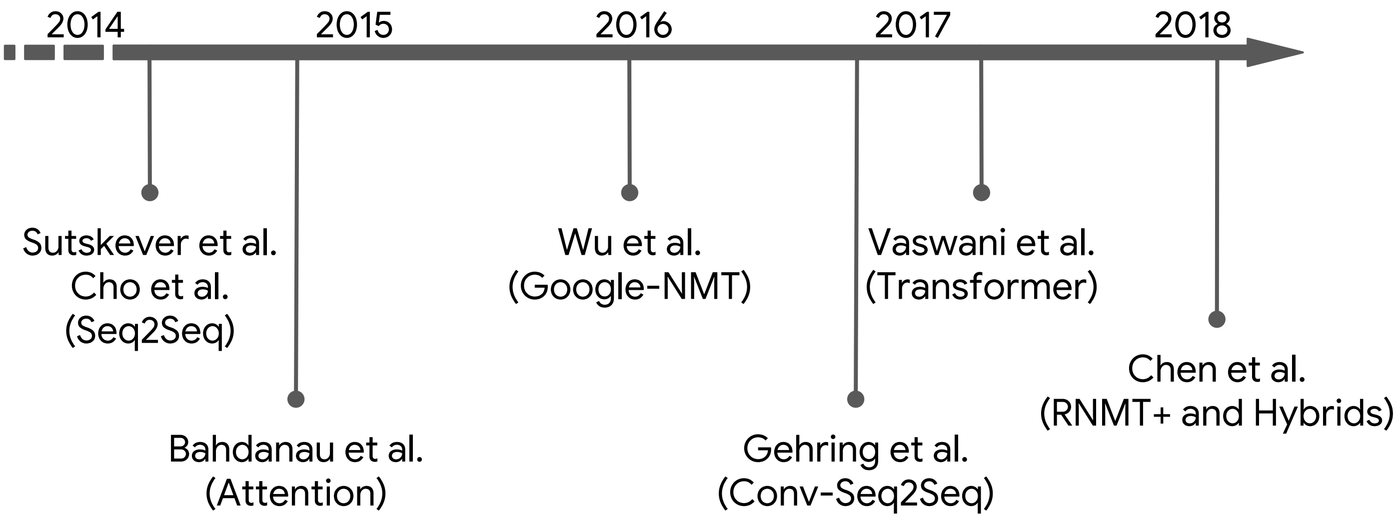
Mia Xu Chen* **Orhan Firat*** **Ankur Bapna***
Melvin Johnson **Wolfgang Macherey** George Foster Llion Jones Mike Schuster
Noam Shazeer Niki Parmar Ashish Vaswani Jakob Uszkoreit Lukasz Kaiser
Zhifeng Chen Yonghui Wu Macduff Hughes

July 16, 2018 ACL'18 Mebourne

*Equal Contribution

This is NOT an architecture search paper!

A Brief History of NMT Models



$$quality = f(X, \theta, \mu)$$

X : Data
θ : Model
μ : Hyperparameters

The Best of Both Worlds - I

Each new approach is:

- accompanied by a set of <u>modeling</u> and <u>training</u> techniques.

Goal:

1. Tease apart architectures and their accompanying techniques.
2. Identify key *modeling* and *training* techniques.
3. Apply them on RNN based Seq2Seq → **RNMT+**

Conclusion:

- **RNMT+** outperforms all previous three approaches.

The Best of Both Worlds - II

Also, each new approach has:

- a fundamental architecture (signature wiring of neural network).

Goal:

1. Analyse properties of each architecture.
2. Combine their strengths.
3. Devise new hybrid architectures ⟶ **Hybrids**

Conclusion:

- **Hybrids** obtain further improvements over all the others.

Building Blocks

- RNN Based NMT - **RNMT**
- Convolutional NMT - **ConvS2S**
- Conditional Transformation Based NMT - **Transformer**

GNMT - Wu et al.

- Core Components:
  - RNNs
  - Attention (Additive)
  - biLSTM + uniLSTM
  - Deep residuals
  - Async Training

- Pros:
  - De facto standard
  - Modelling state space

- Cons:
  - Temporal dependence
  - Not enough gradients

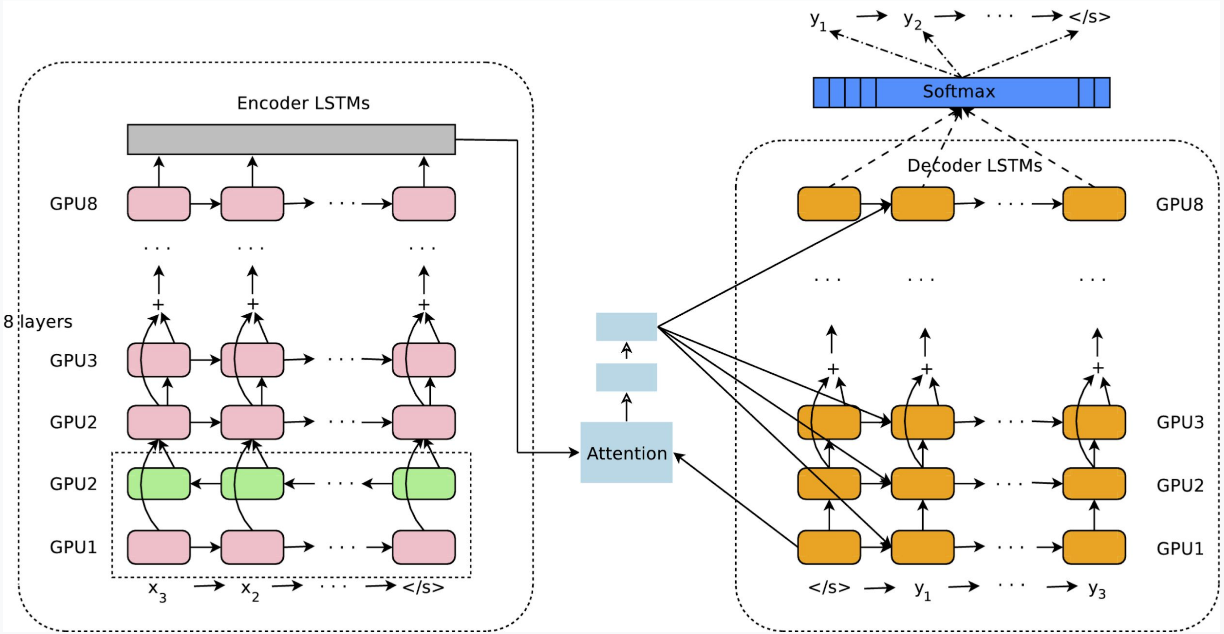


*Figure from "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" Wu et al. 2016

ConvS2S - Gehring et al.

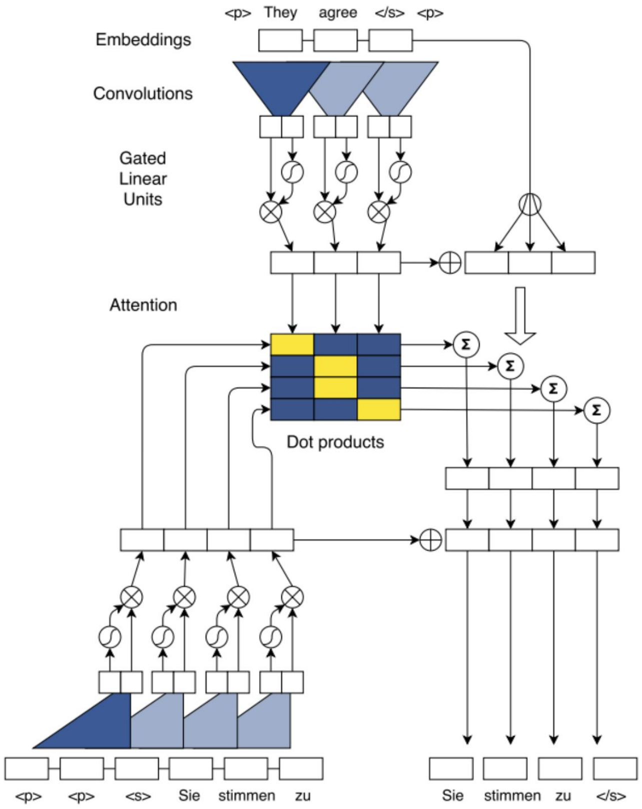


*Figure from “Convolutional Sequence to Sequence Learning” Gehring et al. 2017

- Core Components:
  - Convolution - GLUs
  - Multi-hop attention
  - Positional embeddings
  - Careful initialization
  - Careful normalization
  - Sync Training
- Pros:
  - No temporal dependence
  - More interpretable than RNN
  - Parallel decoder outputs during training
- Cons:
  - Need to stack more to increase the receptive field

Transformer - Vaswani et al.

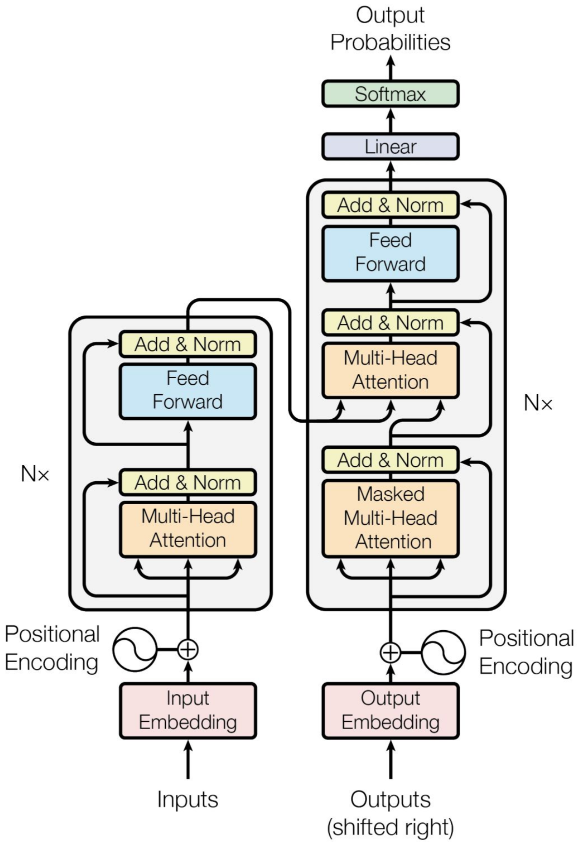


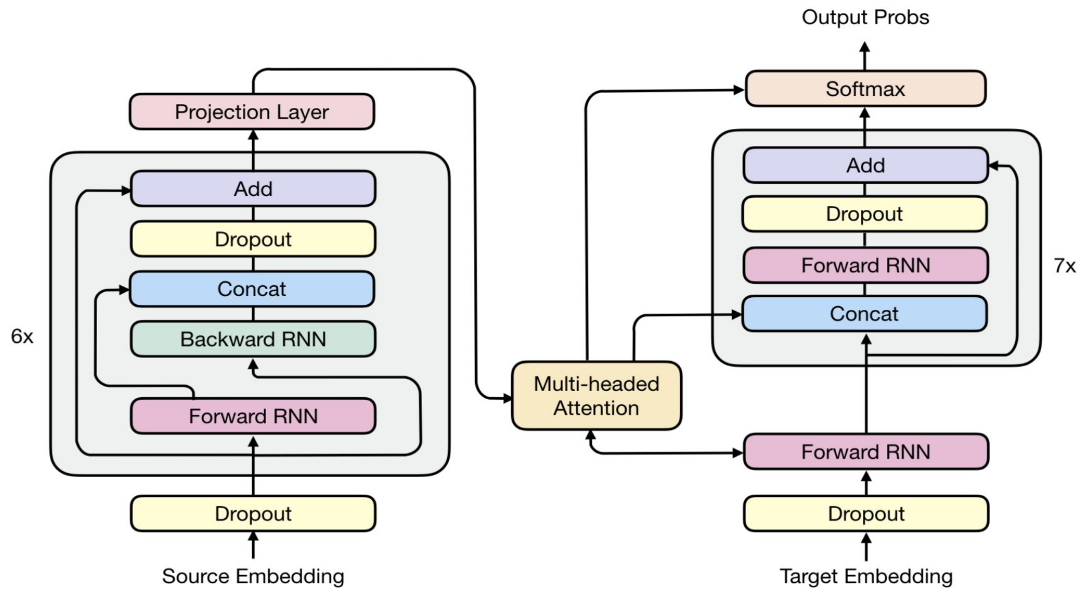
*Figure from "Attention is All You Need" Vaswani et al. 2017

- Core Components:
  - Self-Attention
  - Multi-headed attention
  - Layout: N->f()->D->R
  - Careful normalization
  - Careful batching
  - Sync training
  - Label Smoothing
  - Per-token loss
  - Learning rate schedule
  - Checkpoint Averaging

- Pros:
  - Gradients everywhere - faster optimization
  - Parallel encoding both training/inference

- Cons:
  - Combines many advances at once
  - Fragile

The Best of Both Worlds - I: RNMT+



- The Architecture:
  - Bi-directional encoder 6 x LSTM
  - Uni-directional decoder 8 x LSTM
  - Layer normalized LSTM cell
    - Per-gate normalization
  - Multi-head attention
    - 4 heads
    - Additive (Bahdanau) attention

Model Comparison - I : BLEU Scores

WMT'14 En-Fr
(35M sentence pairs)

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

WMT'14 En-De
(4.5M sentence pairs)

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

- RNMT+/ConvS2S: 32 GPUs, 4096 sentence pairs/batch.
- Transformer Base/Big: 16 GPUs, 65536 tokens/batch.

Model Comparison - II : Speed and Size

WMT'14 En-Fr
(35M sentence pairs)

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

WMT'14 En-De
(4.5M sentence pairs)

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

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

- RNMT+/ConvS2S: 32 GPUs, 4096 sentence pairs/batch.
- Transformer Base/Big: 16 GPUs, 65536 tokens/batch.

Stability: Ablations

WMT'14 En-Fr

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	*

* Indicates an unstable training run

Evaluate importance of four key techniques:

1. Label smoothing
   - Significant for both
2. Multi-head attention
   - Significant for both
3. Layer Normalization
   - Critical to stabilize training (especially with multi-head attention)
4. Synchronous training
   - Critical for Transformer
   - Significant quality drop for RNMT+
   - Successful only with a tailored learning-rate schedule

The Best of Both Worlds - II: Hybrids

Strengths of each architecture:

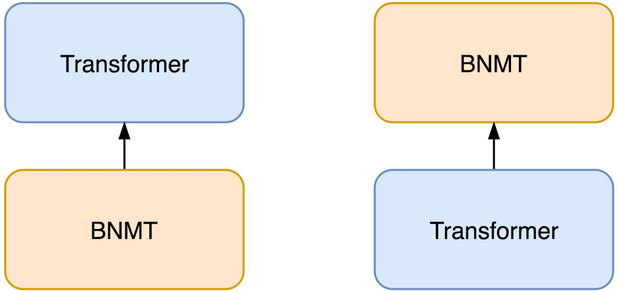
- **RNMT+**
    - Highly expressive - continuous state space representation.

- **Transformer**
    - Full receptive field - powerful feature extractor.

- Combining individual architecture strengths:
    - Capture complementary information - “Best of Both Worlds”.

- Trainability - important concern with hybrids
    - Connections between different types of layers need to be carefully designed.

Encoder - Decoder Hybrids



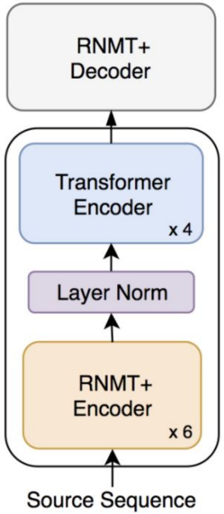
Encoder	Decoder	En→Fr Test BLEU
Trans. Big	Trans. Big	40.73 ± 0.19
RNMT+	RNMT+	41.00 ± 0.05
Trans. Big	RNMT+	$\mathbf{41.12} \pm \mathbf{0.16}$
RNMT+	Trans. Big	39.92 ± 0.21

Separation of roles:

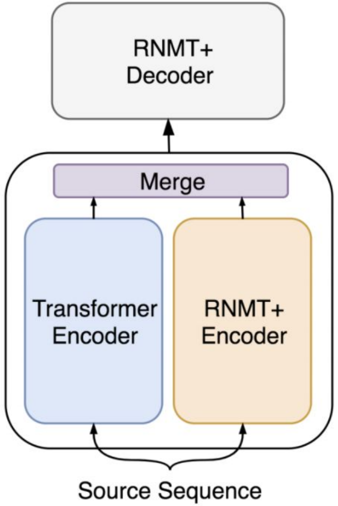
- Decoder - conditional LM
- Encoder - build feature representations

→ Designed to contrast the roles.
(last two rows)

Encoder Layer Hybrids



(a) Cascaded Encoder



(b) Multi-Column Encoder

Improved feature extraction:

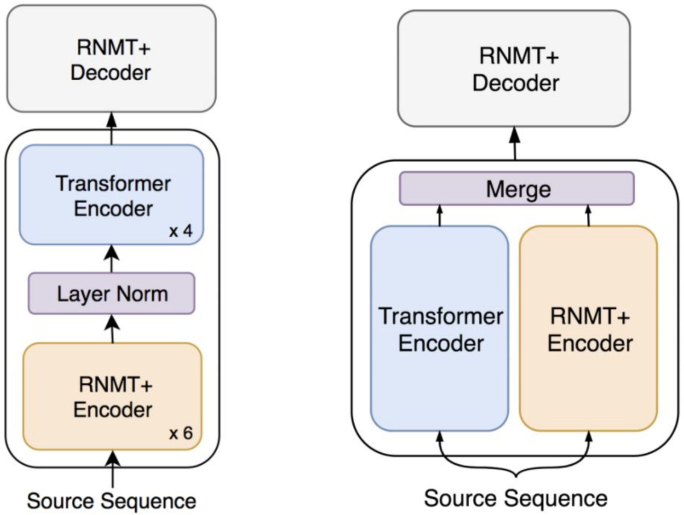
- Enrich stateful representations with global self-attention
- Increased capacity

Details:

- Pre-trained components to improve trainability
- Layer normalization at layer boundaries

Cascaded Hybrid - **vertical** combination
Multi-Column Hybrid - **horizontal** combination

Encoder Layer Hybrids



(a) Cascaded Encoder (b) Multi-Column Encoder

Model	En→Fr BLEU	En→De BLEU
Trans. Big	40.73 ± 0.19	27.94 ± 0.18
RNMT+	41.00 ± 0.05	28.59 ± 0.05
Cascaded	$\mathbf{41.67} \pm \mathbf{0.11}$	28.62 ± 0.06
MultiCol	41.66 ± 0.11	$\mathbf{28.84} \pm \mathbf{0.06}$

Lessons Learnt

Need to separate other improvements from the architecture itself:

- Your good ol' architecture may shine with new modelling and training techniques
- **Stronger baselines** (Denkowski and Neubig, 2017)

Dull Teachers - Smart Students

- "A model with a sufficiently advanced lr-schedule is indistinguishable from magic."

$$expressivity \not\propto trainability$$

Understanding and Criticism

- Hybrids have the potential, more than duct taping.
- Game is on for the next generation of NMT architectures

$$quality = f(X, \theta, \mu)$$

Google AI

Thank You

Open source implementation coming soon!

https://ai.google/research/join-us/

https://ai.google/research/join-us/ai-residency/