The Evolved Transformer

David R. So Chen Liang Quoc V. Le

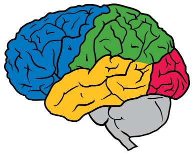
Motivation

Can we apply **Neural Architecture Search** to **feedforward sequence models**?

Methods

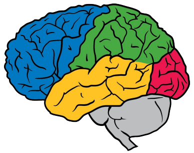
- Evolution
    - simple
    - works well in vision domain

Methods



- Evolution
  - simple
  - works well in vision domain
- **Obstacles**
  - large search space
  - high compute task

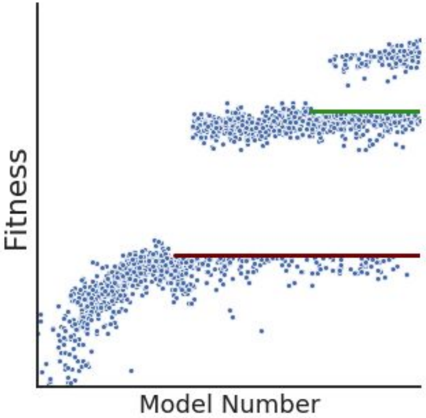
Methods



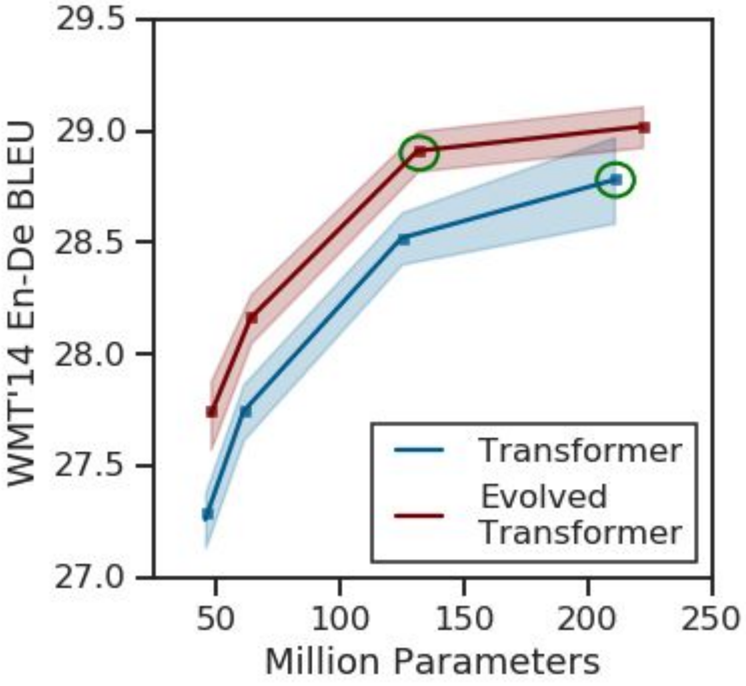
- Evolution
  - simple
  - works well in vision domain
- **Obstacles**
  - large search space
  - high compute task
- **Solutions:**
  - First Warm Start NAS

Methods

- Evolution
  - simple
  - works well in vision domain
- **Obstacles**
  - large search space
  - high compute task
- **Solutions:**
  - First Warm Start NAS
  - Progressive Dynamic Hurdles (PDH): discard bad models for cheap



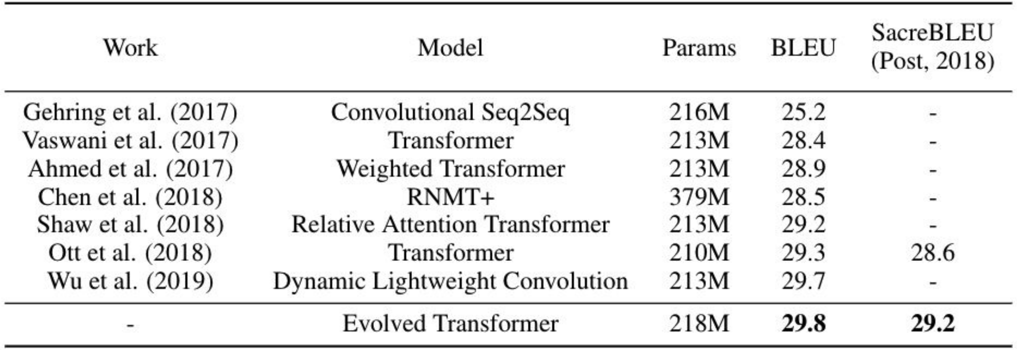
Evolved Transformer Performance



Model	Embedding Size	BLEU	Δ BLEU
Transformer	128	21.3 ± 0.1	-
ET	128	**22.0** ± 0.1	**+ 0.7**
Transformer	432	27.3 ± 0.1	-
ET	432	**27.7** ± 0.1	+ 0.4
Transformer	512	27.7 ± 0.1	-
ET	512	**28.2** ± 0.1	+ 0.5
Transformer	768	28.5 ± 0.1	-
ET	768	**28.9** ± 0.1	+ 0.4
Transformer	1024	28.8 ± 0.2	-
ET	1024	**29.0** ± 0.1	+ 0.2

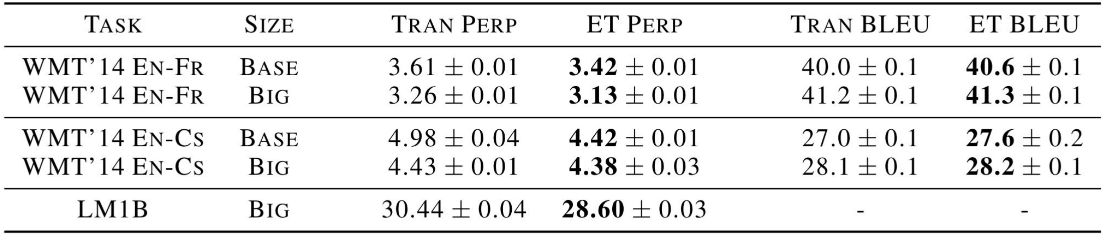
Evolved Transformer Performance

- **State of the Art on WMT En-De**



Work	Model	Params	BLEU	SacreBLEU (Post, 2018)
Gehring et al. (2017)	Convolutional Seq2Seq	216M	25.2	-
Vaswani et al. (2017)	Transformer	213M	28.4	-
Ahmed et al. (2017)	Weighted Transformer	213M	28.9	-
Chen et al. (2018)	RNMT+	379M	28.5	-
Shaw et al. (2018)	Relative Attention Transformer	213M	29.2	-
Ott et al. (2018)	Transformer	210M	29.3	28.6
Wu et al. (2019)	Dynamic Lightweight Convolution	213M	29.7	-
-	Evolved Transformer	218M	**29.8**	**29.2**

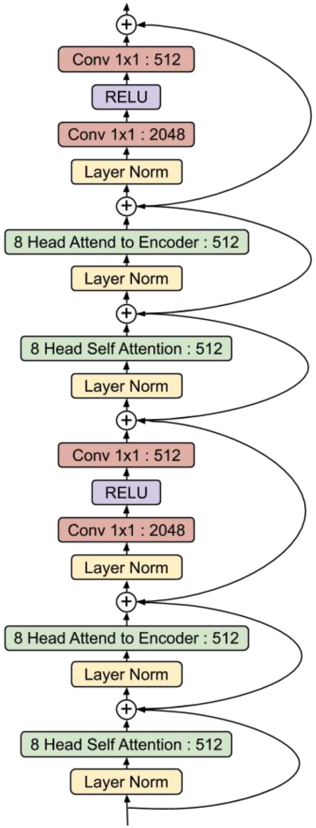
- **Generalizes to Other Tasks**



Task	Size	Tran Perp	ET Perp	Tran BLEU	ET BLEU
WMT’14 En-Fr	Base	3.61 ± 0.01	**3.42** ± 0.01	40.0 ± 0.1	**40.6** ± 0.1
WMT’14 En-Fr	Big	3.26 ± 0.01	**3.13** ± 0.01	41.2 ± 0.1	**41.3** ± 0.1
WMT’14 En-Cs	Base	4.98 ± 0.04	**4.42** ± 0.01	27.0 ± 0.1	**27.6** ± 0.2
WMT’14 En-Cs	Big	4.43 ± 0.01	**4.38** ± 0.03	28.1 ± 0.1	**28.2** ± 0.1
LM1B	Big	30.44 ± 0.04	**28.60** ± 0.03	-	-

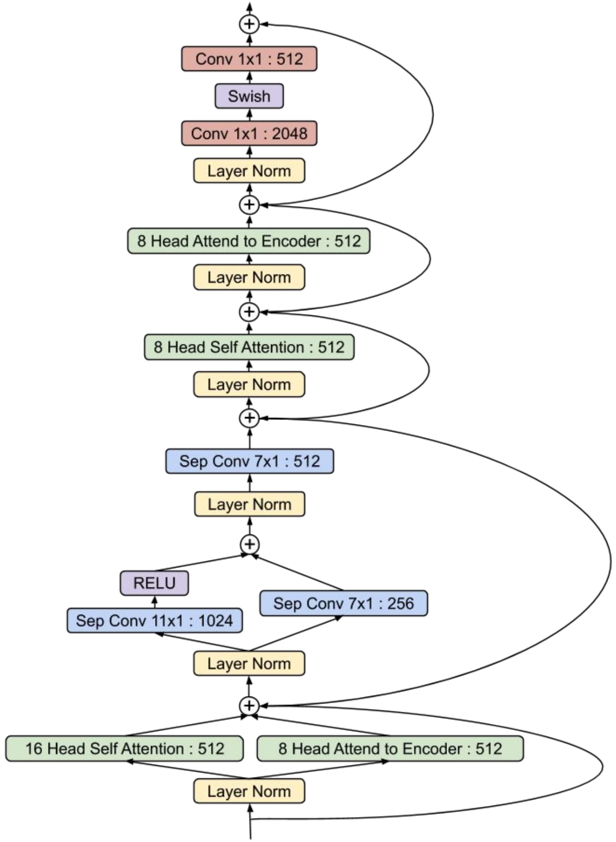
Architecture Comparison

Transformer



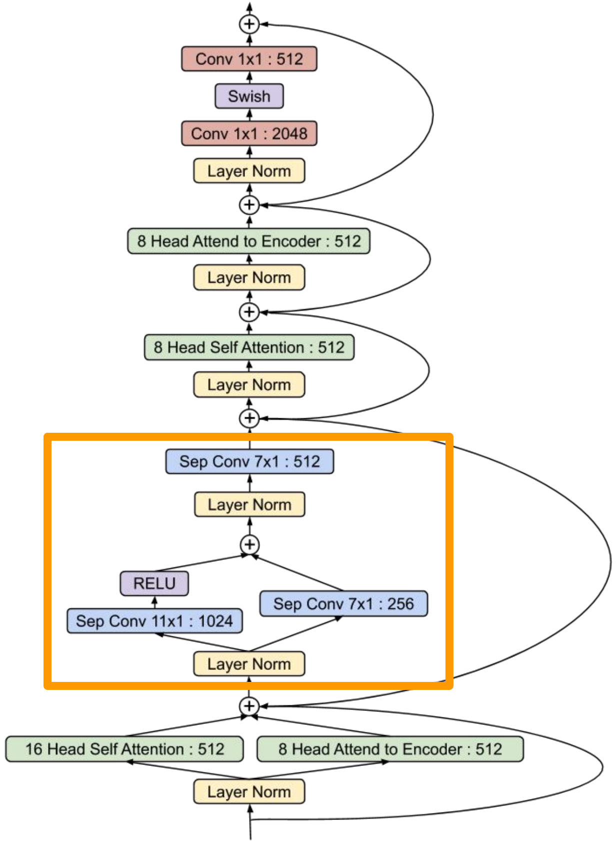
Architecture Comparison

Evolved Transformer



Architecture Comparison

Evolved Transformer



Summary

- First work applying NAS on feedforward sequence model.
- Discovered the Evolved Transformer, which shows better efficiency.
- Open sourced in Tensor2Tensor.

Scan to see the paper and code.

Poster: Pacific Ballroom 6:30 pm to 9:00 pm