# Recurrent Stacking of Layers for Compact Neural Machine Translation Models

Raj Dabre Atsushi Fujita
{raj.dabre,atsushi.fujita}@nict.go.jp
National Institute of Information and Communications Technology (NICT), Kyoto, Japan

## 1. Summary

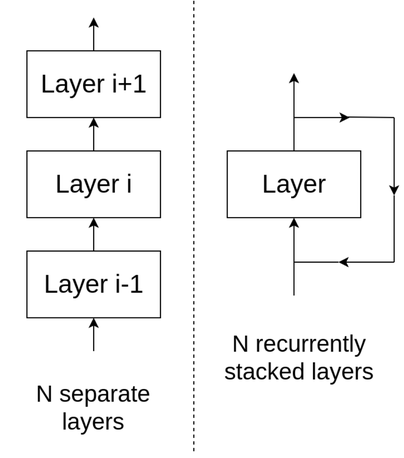
### Multi-layer encoders and decoders in NMT

- **Vanilla NMT:** Each layer has independent parameters
- **RS-NMT:** Use the same parameters for all layers
- **Shared encoder-decoder NMT:** Minimal NMT model

### Key points

- Extensive experiments for several datasets
  - WMT: Turkish↔English
  - ALT, GCP, KFTT, ASPEC: Japanese↔English
- Model size reduction: 50–70%
- Acceptable loss of translation quality: 1–2 BLEU points
- Increase of attention entropy visualized with heat-maps

## 2. Recurrently Stacked NMT




- Enable reuse of layer parameters for depth $> 1$
  - With modification of only 1 line of code
- What happens? Gradual calibration of attention (Panel 5)
- Extension: Encoder-decoder parameter sharing (Panel 4)

## 3. Experimental Settings

### Datasets

| Langs | Dataset | Train | Dev | Test | Vocab | Steps |
|---|---|---|---|---|---|---|
| Ja-En | ALT | 18K | 1000 | 1018 | 8k | 40k |
| | GCP | 400K | 2000 | 2000 | 16k | 60k/120k |
| | KFTT | 440K | 1166 | 1160 | 8k | 160k |
| | ASPEC | 1.50M | 1790 | 1812 | 32k | 400k |
| Tr-En | WMT | 208K | 3000 | 3007<br>3010 | 16k | 50k<br>4 GPUs |

### Implementation and training details

- Implementation in tensor2tensor
  - Modification of Transformer (Vaswani+, 17)
  - Internal sub-word segmentation
- Model training and decoding settings
  - Vanilla NMT, RS-NMT, Shared encoder-decoder NMT
  - Training iterations chosen based on time for convergence
  - Default model and hyperparamter settings
  - Last 10 checkpoints averaged for decoding
  - Beam size of 4 and length penalty of 1.0 (En↔Ja), 0.6 (En↔Tr)

## 4. Translation Performance in BLEU

### RS-NMT vs. vanilla NMT

| #recurrently stacked layers | ALT | | GCP | | KFTT | | ASPEC | | WMT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ja-En | En-Ja | Ja-En | En-Ja | Ja-En | En-Ja | Ja-En | En-Ja | Tr-En test17 | Tr-En test18 | En-Tr test17 | En-Tr test18 |
| **1** | 7.59 | 10.59 | 21.95 | 23.89 | 21.64 | 25.00 | 23.28 | 32.19 | 13.08 | 13.75 | 12.45 | 11.94 |
| **2** | 7.60 | 10.92 | 23.24 | 24.47 | 24.50 | 28.53 | 27.84 | 38.54 | 15.19 | 15.95 | 15.07 | 14.62 |
| **3** | 7.99 | 11.14 | 23.42 | 25.02 | 25.84 | 29.90 | 28.05 | 39.26 | 15.80 | 16.39 | 15.98 | 14.68 |
| **4** | 7.91 | 11.30 | 24.33 | 25.28 | 26.23 | 30.36 | **28.08** | 39.31 | 16.38 | 17.05 | 16.52 | 14.93 |
| **5** | **8.28** | 11.30 | 23.95 | 25.38 | 26.42 | 30.78 | 28.02 | 38.86 | 16.63 | 17.12 | **16.60** | 15.51 |
| **6** | 8.26 | **11.37** | **24.36** | **25.84** | **26.51** | **30.83** | 27.20 | **40.04** | **16.68** | **17.31** | 16.59 | **15.77** |
| **2-layer model** | 8.35 | 11.90 | 24.23 | 25.62 | 24.14 | 30.05 | 28.06 | 38.91 | 16.61 | 17.17 | 16.55 | 15.37 |
| **6-layer model** | 8.47 | 13.21 | 24.67 | 26.22 | 27.19 | 32.72 | 28.77 | 41.32 | 17.80 | 18.36 | 17.99 | 16.29 |

- RS-NMT is only up to 2 BLEU behind vanilla NMT
  - Compensated by back-translation (Sennrich+, 16)

### Shared encoder-decoder NMT

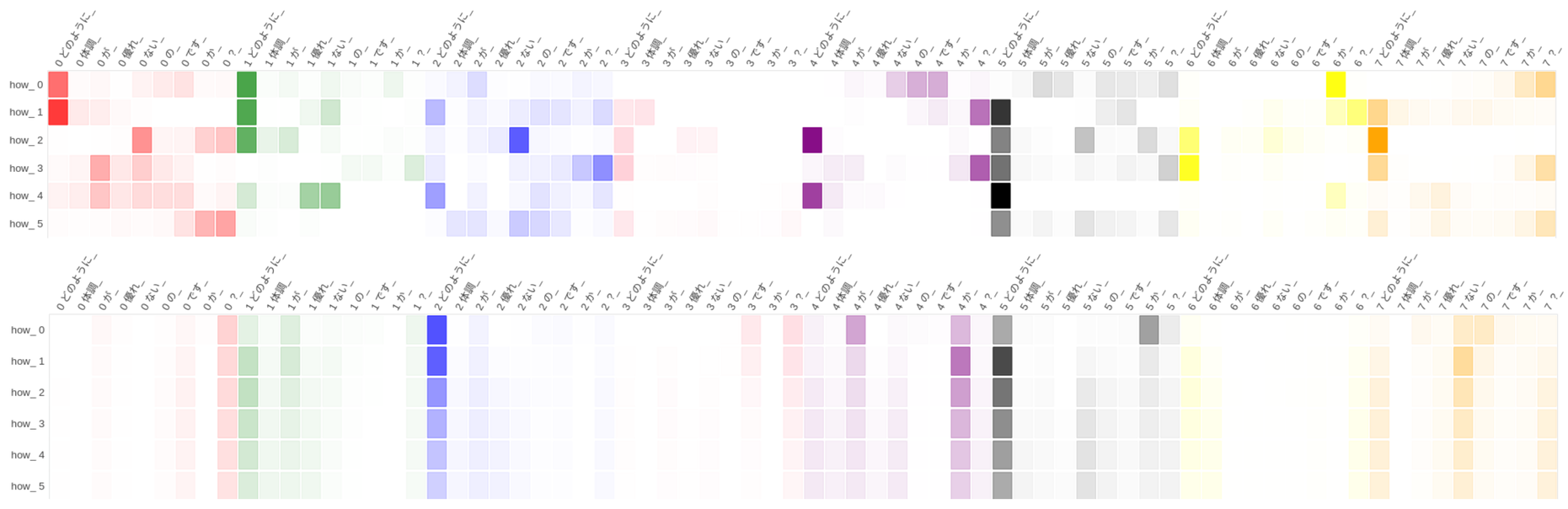
| Recurrent Stacking | Shared EncDec | GCP, En-Ja | | | WMT, Tr-En | | | |
|---|---|---|---|---|---|---|---|---|
| | | #Params | %Params Reduced | BLEU | #Params | %Params Reduced | BLEU test17 | BLEU test18 |
| | | 207.9M | 0 | 26.22 | 158.8M | 0 | 17.80 | 18.36 |
| | ✓ | 151.2M | 37.50 | 26.32 | 102.1M | 35.68 | 17.51 | 18.48 |
| ✓ | | 97.6M | 53.02 | 25.84 | 48.6M | 69.38 | 16.68 | 17.31 |
| ✓ | ✓ | 88.2M | 57.57 | 24.98 | 39.1M | 75.33 | 16.14 | 16.67 |

- 1 BLEU loss for RS-NMT / No BLEU loss for vanilla NMT
- Reduction of 70% parameters for acceptable BLEU reduction

## 5. Visual Analysis of RS-NMT Attention

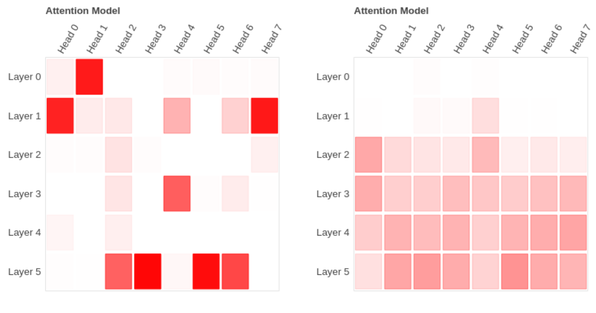
### Transformer's Cross-attention

- 8 heads with different colors
- 6 layers × input tokens
- Darker colors
  → Stronger attention
- Calibration of attention by RS layers




### Attention entropy

- 6 layers × 8 heads
- Darker colors
  → Higher Entropy
  → More uncertainty
  → Attend more words
- Towards average attention?




## 6. Future Work

- Maximum compression with RS-NMT & knowledge distillation
- Enabling use of fewer levels of RS decoding
- Exploration of limits of RS layers
- Further analysis of the nature of RS-NMT