Abstract

Transformers have outperformed recurrent neural networks (RNNs) in natural language generation. But this comes with a significant computational cost, as the attention mechanism’s complexity scales quadratically with sequence length. Efficient transformer variants have received increasing interest in recent works. Among them, a linear-complexity recurrent variant has proven well suited for autoregressive generation. It approximates the softmax attention with randomized or heuristic feature maps, but can be difficult to train and may yield suboptimal accuracy. This work aims to convert a pretrained transformer into its efficient recurrent counterpart, improving efficiency while maintaining accuracy. Specifically, we propose a swap-then-finetune procedure: in an off-the-shelf pretrained transformer, we replace the softmax attention with its linear-complexity recurrent alternative and then finetune. With a learned feature map, our approach provides an improved tradeoff between efficiency and accuracy over the standard transformer and other recurrent variants. We also show that the finetuning process has lower training cost relative to training these recurrent variants from scratch. As many models for natural language tasks are increasingly dependent on large-scale pretrained transformers, this work presents a viable approach to improving inference efficiency without repeating the expensive pretraining process.\footnote{Work was done during an internship at Microsoft.}

1 Introduction

Transformer models (Vaswani et al., 2017) have advanced the state of the art beyond recurrent neural network models (e.g., LSTMs, Hochreiter and Schmidhuber, 1997; GRUs, Cho et al., 2014) across a wide range of natural language processing tasks. In particular, the transformer architecture has been widely used in autoregressive modeling such as language modeling (Baevski and Auli, 2019) and machine translation (Vaswani et al., 2017). The transformer makes crucial use of interactions between feature vectors over the input sequence through the attention mechanism (Bahdanau et al., 2015). However, this comes with significant computation and memory footprint during generation. Since the output is incrementally predicted conditioned on the prefix, generation steps cannot be parallelized over time steps and require quadratic time complexity in sequence length. The memory consumption in every generation step also grows linearly as the sequence becomes longer. This bottleneck for long sequence generation limits the use of large-scale pretrained transformers, such as GPT-3 (Brown et al., 2020), Image Transformer (Parmar et al., 2018), and DALL-E (Ramesh et al., 2021).

Recent work aims at reducing the overhead of autoregressive transformers (Child et al., 2019; Kitaev et al., 2020; Beltagy et al., 2020, \textit{inter alia}). Among them are recurrent alternatives that approximate the standard softmax attention (Katharopoulos et al., 2020; Peng et al., 2021; Choromanski et al., 2021; Schlag et al., 2021). Similar to recurrent neural networks (RNNs), those models represent the context by a recurrent state with a fixed size, thereby achieving linear time and constant memory complexity in generation sequence length. When the recurrent state size is smaller than the sequence length, these variants provide substantial speed and memory advantages over the transformer. A small state size, however, tends to deteriorate the generation quality (Peng et al., 2021), leading to a tradeoff between efficiency and accuracy.

This work improves the balance between efficiency and accuracy by a conversion approach: instead of training a recurrent alternative from scratch, we develop a method to convert a pretrained transformer into an efficient RNN that speeds up generation and reduces memory foot-
prints. Our conversion proceeds with a swap-then-finetune process. Specifically, we change the exponential similarity function in the attention mechanism to the dot product after a single-layer MLP feature mapping. We then finetune the MLP parameters and the other network parameters. Our experiments in language modeling and machine translation show that the conversion can compress the context into a much smaller recurrent state than the sequence length (e.g., 1/16 of the sequence length in WikiText-103 language modeling) while retaining high accuracy. In addition, this conversion requires much less GPU time than training randomly initialized models from scratch.

State-of-the-art models in many natural language tasks are increasingly dependent on large-scale pretrained transformer models (e.g., GPT-2, Radford et al., 2019; BERT, Devlin et al., 2019; RoBERTa, Liu et al., 2019; T5, Raffel et al., 2020; BART, Lewis et al., 2020; DeBERTa, He et al., 2021). Converting a large off-the-shelf transformer to a lightweight inference model without repeating the whole training procedure is particularly useful in many downstream applications. Our work focuses on text generation and presents a viable approach towards efficient inference with high accuracy.

2 Convert a Transformer into an RNN

The transformer architecture consists of multihead attention, feedforward, and layer normalization modules (Vaswani et al., 2017). When a transformer is trained for a sequence generation task with teacher forcing (Williams and Zipser, 1989), the attention can be parallelized over positions because the target sequence is fully available. During generation, on the other hand, the output is incrementally constructed. As a result, the attention becomes an inference bottleneck for long sequences. We present a method to eliminate this bottleneck by converting a pretrained transformer into an efficient RNN of linear time and constant space complexity. We provide a detailed complexity analysis in terms of the sequence length and model dimensions.

2.1 Multihead Attention

The attention module takes as input sequences of source and target vectors. The source vectors are used to produce key and value features, while the target vectors are mapped to query vectors. More formally, denote by \( \{x^{tgt}_i\}_{i=1}^N \) and \( \{x^{src}_j\}_{j=1}^M \) the target and source vectors, where \( x^{tgt}_i, x^{src}_j \in \mathbb{R}^h \) and \( h \) is the model dimensionality. We assume \( r \) attention heads of \( d \) dimensions (\( h = dr \)). For each head, the input vectors are first mapped to \( d \) dimensional query, key, and value features by learned affine transformations with \( W_q \in \mathbb{R}^{d \times h} \) and \( b_q \in \mathbb{R}^d \).

\[
q_i = W_q x^{tgt}_i + b_q, \quad (1a)
\]
\[
k_j = W_k x^{src}_j + b_k, \quad v_j = W_v x^{src}_j + b_v. \quad (1b)
\]

The similarities of each query vector \( q_i \) with all \( M \) key vectors are computed and normalized to produce attention coefficients, which are then used to output a weighted average of the value vectors (Vaswani et al., 2017):

\[
x_{out}^i = \frac{1}{M} \sum_{j=1}^M \frac{\exp\left(x_q \cdot x_k/\sqrt{d}\right)}{\sum_{k=1}^M \exp\left(x_q \cdot x_k/\sqrt{d}\right)} v_j, \quad (2a)
\]
\[
\text{sim}(x, y) = \exp\left(x \cdot y/\sqrt{d}\right). \quad (2b)
\]

Multihead attention runs this procedure for each of the \( r \) heads in parallel and concatenates \( r \) output vectors to get the final \( h \) dimensional vector.\(^2\)

Generation Speed Overhead Fig. 1 depicts the transformer computation steps from input vectors and their time complexity. We assume that the time complexity of multiplying an \( n \times m \) matrix by an \( m \times k \) is \( O(nmk) \) as implemented in cuBLAS (NVIDIA, 2014).\(^3\) It consists of the following two stages.

- **Feature Mapping**: computation of \( \{q_j\}_{j=1}^N \), \( \{k_j\}_{j=1}^M \), and \( \{v_j\}_{j=1}^M \) for all \( r \) heads from input vectors (Eqs. 1a-1b). Time complexity of \( O(Nh^2) \), \( O(Mh^2) \), and \( O(Mh^2) \).
- **Attention**: weighted average over the value vectors (Eq. 2a). \( O(MNh) \), quadratic in sequence length \( (M, N) \).

Generation Memory Overhead In autoregressive generation, query, key, and value vectors consume space complexity of \( O(h) \), \( O(Mh) \), and \( O(Mh) \) in every generation step. Every step’s attention weight (Eq. 2a) spans over \( M \) source positions, taking \( O(Mr) \) space, linear in sequence length \( M \).

2.2 Converting Transformers to RNNs

To address this generation bottleneck of quadratic time and linear space, we propose Transformer-to-RNN (T2R), a method to convert a pretrained

\(^2\)Layer normalization (Ba et al., 2016), residual connection (He et al., 2016), and projection are suppressed for brevity.

\(^3\)If the batch size is small enough, parallelization can speed up matrix multiplication.
We first replace the dot-then-exponential similarity function in a pretrained transformer (Eq. 2b) by
\[
\tilde{\text{sim}}(x, y) = \phi(x) \cdot \phi(y), \quad (3a)
\]
where \( \phi(x) = \text{relu}(W_\phi x + b_\phi). \) (3b)

Here \( W_\phi \in \mathbb{R}^{k \times d} \) and \( b_\phi \in \mathbb{R}^k \) are learned parameters of a single-layer MLP. They map a \( d \) dimensional vector to a \( k \) dimensional kernel feature space. The relu activation (Fukushima, 1980) ensures that the features are non-negative. Different MLP parameters are used for different attention heads, and thus we add a total of \( r k (d + 1) \) learnable parameters per layer (less than 0.2% parameter increase in our language model, §3). We then finetune all parameters in this modified network, including the MLP parameters, with the original task objective.\(^4\)

During inference generation, we reformulate the attention computation (Eq. 2a) as
\[
\tilde{x}_i^{\text{out}} = \sum_{j=1}^{M} \text{sim}(q_i, k_j) v_j = \left( \phi(q_i) \cdot \sum_{j=1}^{M} \phi(k_j) \cdot v_j \right)^T, \quad (4)
\]

by the associativity of matrix multiplication. This formulation lends itself to recurrent computation. In causal attention where each query only attends to its prefix to predict the next word (\( M = i \)), define states:
\[
S_i = \sum_{j=1}^{i} \phi(k_j) \otimes v_j, \quad z_i = \sum_{j=1}^{i} \phi(k_j) \quad (5)
\]

where \( S_i, z_i \in \mathbb{R}^{k \times d}, \mathbb{R}^k \). These states can be computed recurrently (Katharopoulos et al., 2020):
\[
S_i = S_{i-1} + \phi(k_i) v_i^T, \quad z_i = z_{i-1} + \phi(k_i) \quad (6)
\]

In the self-attention or encoder-to-decoder (cross) attention of a sequence-to-sequence model, \( S_i \) and \( z_i \) are constant with respect to \( i \) and only need to be computed once. Given the two states at position \( i \), we can obtain the output vector:
\[
\tilde{x}_i^{\text{out}} = \left( \frac{\phi(q_i)^T S_i}{\phi(q_i)^T z_i} \right)^T \quad (7)
\]

This avoids quadratic computation with respect to the input sequence length. We also speed up inference by merging the MLP feature map with the affine feature maps that produce queries and keys.
\[
\phi(q_i) = \text{relu}(\tilde{W}_q x_i^{\text{tgt}} + \tilde{b}_q), \quad (8a)
\]
\[
\phi(k_j) = \text{relu}(\tilde{W}_k x_j^{\text{src}} + \tilde{b}_k), \quad (8b)
\]

where \( \tilde{W}_q = W_\phi W_q, \tilde{W}_k = W_\phi W_k \) (8c)
\[
\tilde{b}_q = b_\phi + W_\phi b_q, \quad \tilde{b}_k = b_\phi + W_\phi b_k. \quad (8d)
\]

After the model is trained, Eqs. 8c–8d are computed once before generation; the intermediate features of \( q_i \) and \( k_j \) are never computed during inference.

**Generation Speed Overhead** The time complexity of each step in a T2R model is shown in Fig. 1. Similar to the transformer, it proceeds over

---

\(^4\)We found that relu stabilized training by prohibiting negative similarities \( \phi(q) \cdot \phi(k) \). Other activation functions, such as cos, tanh, and elu, did not improve performance.

\(^5\)We tried training the MLP parameters only, but this setting resulted in degraded development performance.
We present extensive experiments on standard benchmarks for language modeling and machine translation. Our results show that T2R achieves efficient autoregressive generation while retaining high accuracy.

3.1 Baselines and Comparison

We compare performance with previous transformer models for autoregressive generation with linear time and constant space complexity in input sequence length. As discussed in §2.3, those prior methods correspond to two different untrainable feature maps $\phi$. We experiment with two types of feature maps for comparisons: ELU ($\phi(x) = \text{elu}(x) + 1$, Katharopoulos et al., 2020); RFA (random feature approximation with softmax temperature reparameterization, Peng et al., 2021). Each feature map is evaluated in two settings: random initialization and pretrain. Random initialization is our reimplementation of the experiments in Katharopoulos et al. (2020) and Peng et al. (2021). The pretrain setting follows the same protocol as T2R except that we use different feature maps $\phi$ than our proposed one-layer MLP with relu activation. Positive orthogonal random features (Performer, Choromanski et al., 2021) provide similar random approximation to RFA and were evaluated in the biology domain, but we found that this method caused training divergence in the language modeling task.\(^7\)

3.2 Setup and Implementations

We apply our method to causal attention in language models and both cross and causal attention in machine translation. For language modeling, we use a 32-dimensional feature map function. We do not modify the encoder in machine translation as its generation speed overhead is much less significant than the decoder (Kasai et al., 2021). Our exploration showed that reducing the feature size of causal attention tends to have less impact on the final translation accuracy as opposed to cross attention; we use feature sizes of 32 and 4 for cross and causal attention, respectively. This observation \(^6\)

6. See §5 for our discussion on more transformer variants with linear time complexity, but most of those variants need modifications for autoregressive modeling and have yet to be empirically evaluated in autoregressive generation tasks.

7. Our implementation closely follows the code released by the authors (https://github.com/lucidrains/performer-pytorch/blob/main/performer_pytorch/performer_pytorch.py#L75-L81), but does not subtract the maximum logit; otherwise it would disallow the linear complexity in causal attention. We conjecture that this is the reason why Performer becomes less stable in our experiments. We suspect that some techniques are necessary to improve numerical stability in language modeling and machine translation.

### 2.3 Autoregressive Linear Transformers

In principle, any kernel function can be used as the similarity function in Eq. 2a (Tsai et al., 2019). Previous work proposed several untrainable feature map functions $\phi$ and developed autoregressive transformer variants with linear time and constant space complexity in sequence length (Katharopoulos et al., 2020; Peng et al., 2021; Choromanski et al., 2021). While those models follow similar computation steps to T2R, there are several differences in generation efficiency. Since the feature map in Katharopoulos et al. (2020) preserves input dimensions, the feature size is always the same as the head dimensions ($k = d$). This means that the speedup and memory savings from using a small feature size are restricted by design. In our experiments (§3.3), our T2R models gain further efficiency by using a feature size that is even smaller than the head dimensions ($k = 32$ and $d = 128$ for language modeling). Peng et al. (2021) and Choromanski et al. (2021) scale query and key vectors by their norms before the random approximation to bound the error. Consequently, the feature mapping stage needs additional steps of producing intermediate $q$ and $k$ and scaling them. T2R suppresses these steps and speeds up generation further (§3.3).

### 3 Experiments

We present extensive experiments on standard benchmarks for language modeling and machine translation. Our results show that T2R achieves two stages.

- **Feature Mapping**: computation of $\{\phi(q_i)\}_{i=1}^{N}$, $\{\phi(k_j)\}_{j=1}^{M}$, and $\{v_j\}_{j=1}^{M}$ for all $r$ heads (Eqs. 8a–8b). Time complexity of $O(Nhk)$, $O(Mhk)$, and $O(Mh^2)$.

- **Attention**: the RNN states and the outputs for $r$ heads (Eqs. 5–7) are computed with $O(Mhk)$ and $O(Nhk)$.

Comparing this with the pretrained transformer, we see that if the feature size is much smaller than input sequence lengths ($k \ll M, N$), the change in the attention stage from $O(MNh)$ to $O(hk(M + N))$ in T2R brings a substantial speedup.

### Generation Memory Overhead

T2R only needs to store the RNN state, and thus its space complexity is $O(hk)$, constant in sequence length. This implies reduction in memory footprint when $k \ll M$, compared to the transformer’s $O(Mh)$.
is consistent with previous work that showed that causal attention can be more drastically simplified than cross attention in transformer machine translation models (You et al., 2020; Tay et al., 2021).

3.2.1 Language Modeling

We use the WikiText-103 benchmark, which consists of 103M tokens sampled from English Wikipedia (Merity et al., 2017). We choose similar hyperparameters to prior work (Baevski and Auli, 2019; Fan et al., 2020): 32 layers, 8 heads, 128 head dimensions, 1024 model dimensions, 4096 fully connected dimensions and dropout (Srivastava et al., 2014) and layer dropout rates of 0.2. We partition the training data into non-overlapping blocks of 512 contiguous tokens ignoring document boundaries and train the model to predict each token from left to right (Baevski and Auli, 2019). Validation and test perplexity are measured by predicting the last 256 words out of the input of 512 consecutive words to avoid evaluating tokens in the beginning with limited context (early token curse, Press et al., 2021). We generally follow the optimization method from Baevski and Auli (2019), but some hyperparameters, such as the learning rate for the T2R finetuning, are adjusted for better convergence than randomly initialized training. See Appendix A.1 for more details.

3.2.2 Machine Translation

We experiment with 3 translation benchmarks: WMT14 EN-DE (4.5M train pairs, Bojar et al., 2016), WMT14 EN-FR (36M, Bojar et al., 2014), and WMT17 ZH-EN (20M, Bojar et al., 2017). We follow the preprocessing and data splits by previous work (EN-DE: Vaswani et al., 2017; EN-FR: Gehring et al., 2017; EN-ZH: Hassan et al., 2018). We use the hyperparameters of the large sized transformer (Vaswani et al., 2017): 6 layers, 16 attention heads, 1024 model dimensions, and 4096 hidden dimensions for both the encoder and decoder. We apply dropout with 0.3 and label smoothing with $\varepsilon = 0.1$. Following Ott et al. (2018), we use an increased batch size of approximately 460K tokens. Each randomly initialized model is trained for 30K (60K for the large EN-FR dataset) steps using Adam with a learning rate of $5 \cdot 10^{-4}$ and $\beta = (0.9, 0.98)$ (Kingma and Ba, 2015). We observed that convergence of the T2R conversion can be achieved with 20K (40K for EN-FR) steps and a reduced learning rate of $2 \cdot 10^{-4}$. We average the checkpoints from the last five epochs to obtain the final model (Vaswani et al., 2017). In inference, we apply beam search with size 5 and length penalty 0.6. Consistent with previous practice, we evaluate with tokenized BLEU (Papineni et al., 2002). Further details are described in Appendix A.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>k</th>
<th>ppl. train</th>
<th>dev. test</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELU + Random Init.</td>
<td>128</td>
<td>22.0</td>
<td>22.8</td>
<td>470h</td>
</tr>
<tr>
<td>RFA + Random Init.</td>
<td>32</td>
<td>20.4</td>
<td>21.3</td>
<td>512h</td>
</tr>
<tr>
<td>T2R + Random Init.</td>
<td>32</td>
<td>20.1</td>
<td>20.8</td>
<td>474h</td>
</tr>
<tr>
<td>ELU + Pretrain</td>
<td>128</td>
<td>21.5</td>
<td>22.2</td>
<td>97h</td>
</tr>
<tr>
<td>RFA + Pretrain</td>
<td>32</td>
<td>20.8</td>
<td>21.6</td>
<td>104h</td>
</tr>
<tr>
<td>T2R + Pretrain</td>
<td>32</td>
<td>19.0</td>
<td>19.6</td>
<td>98h</td>
</tr>
<tr>
<td>T2R 75% + Pretrain</td>
<td>32</td>
<td>17.9</td>
<td>18.5</td>
<td>95h</td>
</tr>
<tr>
<td>Pretrained Transformer</td>
<td>–</td>
<td>17.9</td>
<td>18.5</td>
<td>–</td>
</tr>
<tr>
<td>Baevski and Auli (2019)</td>
<td>–</td>
<td>18.7</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: WikiText-103 language modeling results (perplexity). Train time is measured in GPU hours. The top two rows are our reimplementations of Katharopoulos et al. (2020) and Peng et al. (2021). Pretrain indicates initialization with a pretrained transformer for language modeling. T2R 75% indicates a model where every fourth layer from the top is kept as the original transformer layer. Perplexity (ppl.) is measured by predicting the last 256 words out of the input of 512 consecutive words. All models use 128 head dimensions. We assume access to a pretrained transformer model and measure the finetuning time in GPU hours.

3.3 Results

Language Modeling  Seen in Table 1 are language modeling results in perplexity. We observe that T2R with the learnable MLP feature map outperforms the other two linear transformer models by more than 2.0 perplexity points in the pretrain setting. Unlike the other linear transformer models, T2R greatly benefits from pretraining (T2R + Pretrain: 19.6 vs. T2R + Random Init.: 20.8 test perplexity points). We attribute this advantage of T2R to the fact that the MLP feature map is able to learn attention patterns that are similar to those of the pretrained transformer, as evidenced in §4. Notice also that the T2R conversion is ~5x faster (measured in GPU hours) than training a model from scratch. These results illustrate that a lightweight model can be obtained without repeating the expensive training of large-scale pretrained language models such as GPT-2 and GPT-3 (Radford et al., 2019; Brown et al., 2020). T2R’s generation speedup (~4x when producing 512 consecutive words) and
<table>
<thead>
<tr>
<th>Model</th>
<th>Feature Size $k$</th>
<th>WMT14</th>
<th>WMT17</th>
<th>Train Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cross</td>
<td>causal</td>
<td>EN-DE</td>
<td>EN-FR</td>
</tr>
<tr>
<td>ELU + Random Init.</td>
<td>64</td>
<td>64</td>
<td>28.4</td>
<td>*</td>
</tr>
<tr>
<td>RFA + Random Init.</td>
<td>32</td>
<td>4</td>
<td>28.1</td>
<td>41.7</td>
</tr>
<tr>
<td>T2R + Random Init.</td>
<td>32</td>
<td>4</td>
<td>27.5</td>
<td>39.8</td>
</tr>
<tr>
<td>ELU + Pretrain</td>
<td>64</td>
<td>64</td>
<td>28.4</td>
<td>41.8</td>
</tr>
<tr>
<td>RFA + Pretrain</td>
<td>32</td>
<td>4</td>
<td>27.6</td>
<td>41.8</td>
</tr>
<tr>
<td>T2R + Pretrain</td>
<td>32</td>
<td>4</td>
<td>28.7</td>
<td>42.1</td>
</tr>
<tr>
<td>Pretrained Transformer Large</td>
<td>–</td>
<td>–</td>
<td>28.9</td>
<td>42.2</td>
</tr>
<tr>
<td>Vaswani et al. (2017)</td>
<td>–</td>
<td>–</td>
<td>28.4</td>
<td>41.8</td>
</tr>
</tbody>
</table>

Table 2: Machine translation test results in BLEU scores. The top two rows are our reimplementations of Katharopoulos et al. (2020) and Peng et al. (2021). Pretrain indicates initialization with a trained transformer-large model. *: diverged even when running with multiple random seeds and smaller learning rates. We assume access to a pretrained transformer model and measure the finetuning time in GPU hours.

memory savings are later benchmarked with varying sequence lengths. There remains a gap of 1.1 perplexity points between the T2R and pretrained transformer models (19.6 vs. 18.5). However, the gap can be closed when every fourth layer from the top is kept as the original transformer layer and the model is finetuned in the same way (T2R 75%). This suggests that keeping a small fraction of the quadratic attention layers can provide an effective middle ground between efficiency and accuracy.\(^8\)

**Machine Translation** Seen in Table 2 are machine translation results in BLEU from various configurations. Departing from the language modeling experiments, the T2R model underperforms the other two linear transformer models when initialized randomly. However, consistent with language modeling, the T2R model substantially benefits from pretraining (e.g., 28.7 vs. 27.5 BLEU points in EN-DE). As a result, the T2R model achieves similar BLEU scores to the original transformer across all language pairs. ELU trained from the pretrained transformer yields comparable performance to T2R, but the feature size is much larger (64 vs. 32 and 64 vs. 4 in cross and causal attention), thus leading to increased overhead, as shown later. Note that the T2R finetuning time is only moderately smaller than that of randomly initialized training here, but further speedup in conversion can be potentially achieved with more extensive hyperparameter tuning.\(^9\)

**Speedup and Memory Savings in Generation**

We run a conditional generation experiment to compare the decoding speed of the models in Table 2 (Fig. 2). Here we assume the input and output sequences are of the same length. All models are tested using greedy decoding with the same batch size of 16 on a TPU v2 accelerator.\(^10\) We see that indeed the linear transformer models can generate an almost constant number of tokens per second regardless of the sequence length and outpace the transformer model dramatically as the sequence becomes longer. The T2R model achieves a 15%+ speedup that the computational cost for conversion can be much lighter than training from scratch, and T2R is advantageous when only a limited number of GPUs are available.

\(^8\)Concurrent work (Lei, 2021) also explores reducing the number of attention layers for efficiency.

\(^9\)We found that the batch size could be reduced for T2R conversion without hurting accuracy, while randomly initialized models deteriorate with small batch sizes. This suggests

\(^10\)https://opensource.google/projects/jax.
4 Analysis and Ablations

We presented T2R, a method to convert a pretrained transformer into an efficient RNN. In this section, we analyze our conversion approach by examining the impact of the feature size and induced attention weight distributions. Our analysis shows that T2R implicitly learns attention distributions similar to the original transformer.

Feature Size and Pretraining We saw that T2R benefits substantially from transformer pretraining. Fig. 4 compares T2R with pretraining and random initialization in terms of the relation between the validation perplexity from WikiText-103 and the feature sizes. We see that as the feature size (RNN state size) becomes smaller, pretraining becomes particularly important to achieve low perplexity. Transformer pretraining achieves a Pareto improvement over random initialization in the tradeoff between efficiency (small feature size) and accuracy (low perplexity).

Attention Distribution T2R is not explicitly trained to mimic the original attention distributions, and there is no guarantee that the MLP feature map approximates the exponential similarity function, unlike previous approximation approaches (Peng et al., 2021; Choromanski et al., 2021). Here, we analyze the properties of the attention weight dis-
tributions that are induced by finetuning. We use the validation data from WikiText-103 and run language models to predict the next word given the input of 512 contiguous words. We compute the attention weight distribution over the 512 words for each attention head in the model layers.

Fig. 5 compares the attention distributions from T2R in various configurations. T2R MLP frozen indicates a model that is finetuned with the MLP parameters frozen. Euclidean distances in attention distributions between the original transformer and each model are averaged across validation samples, model layers, and attention heads. Comparing T2R before finetuning and the full T2R model, we see that the finetuning process induces much more similar attention distributions, and the distance diminishes as the feature size increases (and the perplexity approaches the original transformer, Fig. 4). We also observed that when the MLP parameters are not trained (T2R MLP frozen), the distance from the original attention distributions increases. These results suggest that finetuning of the whole network in T2R implicitly develops similar attention distributions to the original transformer even though the training supervision comes solely from language modeling.

5 Further Related Work

In addition to the work we already discussed, we highlight related methods from prior work that make transformer models efficient.

5.1 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) is closely related to our T2R conversion and uses a similar pipeline: a teacher model with large capacity is first trained and is used to generate silver training data for a new lightweight inference model. It has been successfully applied to machine translation (e.g., Kim and Rush, 2016; Gu et al., 2018) to make generation efficient. In particular, several prior works distill a transformer translation model to an RNN (Senellart et al., 2018; Kim et al., 2019). We share the same motivation toward fast generation with light memory, but our approach differs in two ways: the original training data are used for finetuning an RNN model, and its model parameters are initialized with the “teacher” transformer.

Our method does not use the computationally expensive teacher model to generate new training data. While data generation is a one-time computational cost, it becomes expensive as the teacher model size and training data increase. Moreover, since the pretrained parameters can be directly used, conversion requires fewer GPU hours than training a brand new lightweight model from scratch (§3.3).

5.2 Efficient Transformers

Prior work suggested many other strategies to improve efficiency in transformers, such as weight sharing and factorization (Dehghani et al., 2019; Lan et al., 2020), weight and layer pruning (Michel et al., 2019; Fan et al., 2020), quantization (Zafir, et al., 2019; Shen et al., 2020), and modifying the combination of sublayers (Press et al., 2020; Mandava et al., 2020). Some of these methods present orthogonal design choices and can be integrated into our T2R model to gain further efficiency. For a more comprehensive survey, see Tay et al. (2020b). Below we describe several prior works along two major strategies: compressing the attention context and sparsifying the attention patterns.

Attention Context Compression This strand of methods compresses the context that is attended to, thereby reducing the time and memory overhead in the attention. RNN models that we converted pretrained transformers into compress the context into a recurrent state. Other approaches include low rank approximation of the attention computation (Wang et al., 2020; Tay et al., 2021) and adding a memory module that can access multiple tokens at once (Liu et al., 2018; Dai et al., 2019; Lee et al., 2019; Ainslie et al., 2020; Rae et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020).

Sparse Attention Patterns Another approach to reducing the time and memory overhead from the attention computation is to limit the tokens that are attended to by sparsifying the attention patterns. These patterns can be set in advance or learned during training (Tay et al., 2020b). For example, prior works introduced fixed patterns of blockwise attention (Qiu et al., 2020) and strided attention (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020). Other previous works presented methods to learn attention patterns from data (Sukhbaatar et al., 2019; Roy et al., 2020; Tay et al., 2020a).

It should be noted that significant modifications are necessary to apply many of these methods to autoregressive generation tasks such as language...
modeling and machine translation, and their empirical evaluation in these generation settings has yet to be conducted (Peng et al., 2021). This work presents extensive empirical evaluation in autoregressive generation settings.

6 Conclusion and Future Work

We present T2R, a method that converts a pre-trained transformer to a recurrent neural network that reduces the time and memory cost of autoregressive generation. Our experiments in language modeling and machine translation demonstrated that our model produces an improved trade-off between efficiency and accuracy over randomly initialized training and previous models with lightweight attention. Our work provides further support for the claim that large-scale pretrained models can be compressed into efficient inference models that facilitate downstream applications.

Acknowledgments

We thank Ofir Press, Bill Dolan, Lei Li, and the anonymous reviewers for their valuable feedback and discussion on this work. Nikolaos Pappas was supported by the Swiss National Science Foundation grant P400P2_183911.

References


Tao Lei. 2021. When attention meets fast recurrence: Training language models with reduced compute.


Swetha Mandava, Szymon Migacz, and Alex Fit Florea. 2020. Pay attention when required.


NVIDIA. 2014. The NVIDIA CUDA basic linear algebra subroutines (CUBLAS).


Ofir Press, Noah A. Smith, and Omer Levy. 2020. Im-


Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2020. Efficient content-based sparse attention with routing transformers. TACL.


A Appendix

A.1 Hyperparameters and Setting

All training is implemented in fairseq (Ott et al., 2019) and run with PyTorch 1.7.1 (Paszke et al., 2019), 8 Tesla V100 GPUs, and CUDA 11.0. We used mixed precision and distributed training over 8 GPUs (Micikevicius et al., 2018; Ott et al., 2018). Apart from EN→ZH where we used separate BPE operations and only tied the decoder input and output embeddings, we tie all embeddings (Press and Wolf, 2017; Inan et al., 2017). We experimented with feature sizes of [16, 32, 64] and [4, 8, 16, 32] for language modeling and machine translation respectively, and chose the smallest feature sizes that retained the development performance compared to the standard transformer.

A.1.1 Language Modeling

We generally follow the optimization method from Baevski and Auli (2019). For optimizing a model from random initialization, the learning rate is linearly warmed up from 10^{-7} to 1 for the initial 16K steps and then annealed using a cosine learning rate schedule with cycles (Loshchilov and Hutter, 2017). Each period lasts for twice the number of updates than the previous cycle, and we lower the maximum and minimum learning rates by 25% compared to the previous cycle. The initial minimum and maximum learning rates are 10^{-5} and 1 respectively (Baevski and Auli, 2019). We train the model with a batch size of about 74K tokens with a total of 286K steps (Baevski and Auli, 2019). When we convert a pretrained transformer to an RNN model by finetuning, we found that we could speed up training by reducing the warm-up steps, total update steps, maximum and minimum rates, and batch size to 8K steps, 142K steps, 5 \times 10^{-6}, 0.5, and 25K tokens without loss in validation perplexity.

Randomly Initialized Training We generally follow the hyperparameters chosen in Baevski and Auli (2019); Fan et al. (2020). Specifically, we list the hyperparameters in Table 3 for easy replication. All other hyperparameter options are left as default values in fairseq.

Finetuning Pretrained Transformer Seen in Table 4 are the hyperparameters for finetuning a pretrained transformer to RNN models. The learning rates, the max number of updates, and the learning period length are all reduced.
Architecture transformer_lm_wiki103
Criterion adaptive_loss
tokens-per-sample 512
sample-break-mode none
# max tokens 3072
dropout rate 0.2
layer dropout rate 0.2
decoder embed dim 1024
decoder fnn dim 4096
# decoder attn heads 8
optimizer nag
lr-scheduler cosine
lr-period-updates 270K
lr-shrink 0.75
t-mult 2
max-lr 1
min-lr 1e-9
lr 1e-4
clip-norm 0.1
warm-up lr 1e-7
# warmup updates 16K
# max updates 286K
# GPUs 8
update-freq 3

<table>
<thead>
<tr>
<th>Architecture</th>
<th>transformer_lm_wiki103</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>adaptive_loss</td>
</tr>
<tr>
<td>Tokens-per-sample</td>
<td>512</td>
</tr>
<tr>
<td>Sample-break-mode</td>
<td>none</td>
</tr>
<tr>
<td># Max Tokens</td>
<td>3072</td>
</tr>
<tr>
<td>Dropout Rate</td>
<td>0.2</td>
</tr>
<tr>
<td>Layer Dropout Rate</td>
<td>0.2</td>
</tr>
<tr>
<td>Decoder Embed Dim</td>
<td>1024</td>
</tr>
<tr>
<td>Decoder Fnn Dim</td>
<td>4096</td>
</tr>
<tr>
<td># Decoder Attn Heads</td>
<td>8</td>
</tr>
<tr>
<td>Optimizer</td>
<td>nag</td>
</tr>
<tr>
<td>Lr-Scheduler</td>
<td>cosine</td>
</tr>
<tr>
<td>Lr-Period-Updates</td>
<td>270K</td>
</tr>
<tr>
<td>Lr-Shrink</td>
<td>0.75</td>
</tr>
<tr>
<td>T-Mult</td>
<td>2</td>
</tr>
<tr>
<td>Max-Lr</td>
<td>1</td>
</tr>
<tr>
<td>Min-Lr</td>
<td>1e-9</td>
</tr>
<tr>
<td>Lr</td>
<td>1e-4</td>
</tr>
<tr>
<td>Clip-Norm</td>
<td>0.1</td>
</tr>
<tr>
<td>Warm-Up Lr</td>
<td>1e-7</td>
</tr>
<tr>
<td># Warmup Updates</td>
<td>16K</td>
</tr>
<tr>
<td># Max Updates</td>
<td>286K</td>
</tr>
<tr>
<td># GPUs</td>
<td>8</td>
</tr>
<tr>
<td>Update-Freq</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: Language modeling hyperparameters when randomly initialized in the fairseq library.

A.1.2 Machine Translation
We experiment with 3 translation benchmarks: WMT14 EN-DE (4.5M train pairs, Bojar et al., 2016), WMT14 EN-FR (36M, Bojar et al., 2014), and WMT17 ZH-EN (20M, Bojar et al., 2017). We follow the preprocessing and data splits by previous work (EN-DE: Vaswani et al., 2017; EN-FR: Gehring et al., 2017; EN-ZH: Hassan et al., 2018; Wu et al., 2019). These datasets are all encoded into subwords by BPE (Sennrich et al., 2016). We run joint BPE on all language pairs except EN-ZH. We use the hyperparameters of the large sized transformer (Vaswani et al., 2017): 6 layers, 16 attention heads, 1024 model dimensions, and 4096 hidden dimensions for both the encoder and decoder. We apply dropout with 0.3, weight decay with 0.01 and label smoothing with $\varepsilon = 0.1$. Following Ott et al. (2018), we use an increased batch size of approximately 460K tokens by accumulating gradients without updating parameters.

Randomly Initialized Training We generally follow the hyperparameters chosen in Vaswani et al. (2017); Ott et al. (2018). Specifically, we list the hyperparameters in Table 5 for easy replication. All other hyperparameter options are left as default values in fairseq. The parameters from the last five epochs were averaged to obtain the final model.

Finetuning Pretrained Transformer Seen in Table 6 are the hyperparameters for finetuning a pretrained transformer to RNN models. The learning rate and the max number of updates are reduced. The parameters from the last five epochs were again averaged to obtain the final model.

A.2 Attention Distribution
Peakiness of Attention Fig. 6 plots the average entropy of the T2R models with and without pretraining. Entropy is averaged across validation samples, layers, and attention heads. Comparing Figs. 4 and 6, we see that there is a strong correlation between validation perplexity and entropy. The entropy decreases (and thus the attention distribution gets peakier) when a large feature size is used or the transformer pretraining is applied. This observation hints at potential future improvement of linear transformer models by introducing an inductive bias towards peaky attention distributions.

Table 5: Machine translation hyperparameters when randomly initialized in the fairseq library. *: we reduced the learning rate for T2R to avoid training divergence.

Table 6: Finetuning machine translation hyperparameters. The learning rate is smaller than randomly initialized training.