Efficient Content-Based Sparse Attention with Routing Transformers

Aurko Roy and **Mohammad Saffar** and **Ashish Vaswani** and **David Grangier**
Google Research
`{aurkor, msaffar, avaswani, grangier}@google.com`

Abstract

Self-attention has recently been adopted for a wide range of sequence modeling problems. Despite its effectiveness, self-attention suffers from quadratic compute and memory requirements with respect to sequence length. Successful approaches to reduce this complexity focused on attending to local sliding windows or a small set of locations *independent* of content. Our work proposes to learn dynamic sparse attention patterns that avoid allocating computation and memory to attend to content unrelated to the query of interest. This work builds upon two lines of research: it combines the modeling flexibility of prior work on *content-based* sparse attention with the efficiency gains from approaches based on *local, temporal* sparse attention. Our model, the Routing Transformer, endows self-attention with a sparse routing module based on online k-means while reducing the overall complexity of attention to $O(n^{1.5}d)$ from $O(n^2 d)$ for sequence length n and hidden dimension d. We show that our model outperforms comparable sparse attention models on language modeling on `Wikitext-103` (15.8 vs 18.3 perplexity), as well as on image generation on `ImageNet-64` (3.43 vs 3.44 bits/dim) while using fewer self-attention layers. Additionally, we set a new state-of-the-art on the newly released `PG-19` data-set, obtaining a test perplexity of 33.2 with a 22 layer Routing Transformer model trained on sequences of length 8192.

1 Introduction

Generative models of sequences have witnessed rapid progress driven by the application of attention to neural networks. In particular, (Bahdanau et al., 2014; Cho et al., 2014; Vaswani et al., 2017) relied on attention to drastically improve the state-of-the art in machine translation. Subsequent research (Radford et al., 2018; Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019) demonstrated the power of self-attention in learning powerful representations of language to address several natural language processing tasks. Self-attention also brought impressive progress for generative modeling outside of language, e.g. image (Parmar et al., 2018; Menick and Kalchbrenner, 2018; Child et al., 2019) and music generation (Huang et al., 2018; Child et al., 2019).

Self-attention operates over sequences in a step-wise manner: at every time-step, attention assigns an *attention weight* to each previous input element (representation of past time-steps) and uses these weights to compute the representation of the current time-step as a weighted sum of the past input elements (Vaswani et al., 2017). Self-attention (Shaw et al., 2018) is a particular case of attention (Bahdanau et al., 2014; Chorowski et al., 2015; Luong et al., 2015).

Self-attention is commonly used in auto-regressive generative models. These models generate observations step-by-step, modeling the probability of the next symbol given the previously generated ones. At every time step, self-attentive generative models can directly focus on any part of the previous context. In contrast, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have direct interactions with only a local neighborhood of context around the current time step.

This advantage however comes at a price: unlike recurrent networks or convolution networks, the time and space complexity of self-attention is quadratic in n, the length of the sequence. Specifically, for every position $i \leq n$, self-attention computes weights for its whole context of length i, which induces a complexity of $\sum_{i \leq n} i = n(n-1)/2$. This makes it difficult to scale attention based models to modeling long sequences. However, long sequences are the norm in many domains, including music, image, speech, video generation and document level machine translation.

Therefore, an important research direction is to investigate sparse and memory efficient forms of attention in order to scale to tasks with long sequence lengths. Previous work has proposed *data independent* or fixed sparsity patterns bounding temporal dependencies, such as local or strided attention. At each time step, the model attends only to a fix number of time steps in the past (Child et al., 2019). Extensions to local attention have suggested learning the length of the temporal sparsity for each attention module in the network (Sukhbaatar et al., 2019). These strategies draw their inspiration from RNNs and CNNs and bound their complexity by attending only to representations summarizing a *local* neighborhood of the current time step. Their attention matrices (matrices containing the attention weights for every pair of previous, current timestep) are natively sparse and requires instantiating only non-zero entries. While these approaches have achieved good results, fixing the sparsity pattern of a content based mechanism such as self-attention can limit its ability to pool in information from large contexts.

As an alternative to local attention, (Correia et al., 2019) considers content-based sparsity, an approach allowing for arbitrary sparsity patterns. This formulation however does require instantiating a full dense attention matrix prior to sparsification through variants of L_0-sparsity or sparsemax approximations (Blondel et al., 2019).

The present work builds upon these two lines of research and proposes to retain the modeling flexibility of content-based sparse attention while leveraging the efficiency of natively sparse attention matrices. Our formulation avoids sparsemax variants and relies on clustering of attention instead. Each attention module considers a clustering of the space: the current time-step only attends to context belonging to the same cluster. In other words, the current time-step query is *routed* to a limited number of context through its cluster assignment. This strategy draws inspiration from the application of k-means clustering to Non-negative Matrix Factorization (NMF) (Lee and Seung, 2001; Ding et al., 2005; Kim and Park, 2008), which is relevant to the sparsification of non-negative matrices like attention matrices.

Our proposed model, Routing Transformer, combines our efficient clustered-based sparse attention with classical local attention to reach excellent performance both for language and image generation. These results are obtained without the need to maintain attention matrices larger than batch length which is the case with the segment level recurrence mechanism used in (Dai et al., 2019; Sukhbaatar et al., 2019). We present experimental results on language modeling (`Wikitext-103` and `enwik-8`) and unconditional image generation (`ImageNet-64`). Routing Transformer sets new state-of-the-art while having comparable or fewer number of self-attention layers and heads, both on `Wikitext-103` (15.8 vs 18.3 perplexity) and on `ImageNet-64` (3.43 vs 3.44 bits/dim). We also report competitive results on `enwik-8` (0.99 vs 0.98 perplexity).

2 Related Work

Attention with Temporal Sparsity: Research on efficient attention neural models parallels the advent of attention-based architectures. In the context of speech recognition, (Jaitly et al., 2015) proposed the Neural Transducer which segments sequences in non-overlapping chunks and attention is performed in each chunk independently. Limiting attention to a fixed temporal context around the current prediction has also been explored in (Chorowski et al., 2015), while (Chiu* and Raffel*, 2018) dynamically segment the sequence into variable sized-chunks.

Hierarchical attention strategies have also been explored: the model first considers which part of the inputs should be attended to before computing full attention in a contiguous neighborhood of the selected area (Gregor et al., 2015; Xu et al., 2015; Luong et al., 2015). Later, hierarchical attention has been simplified by (Liu et al., 2018) that alternates coarse layers (attending to the whole sequence at a lower temporal resolution) with local layers (attending to a neighborhood of the current prediction).

This alternating strategy is also employed by (Child et al., 2019), which introduces bounded and strided attention, i.e. attending to a fixed context in the past at a sub-sampled temporal resolution. This work formalizes such a strategy using a sparse attention formalism, showing how it relates to full attention with a specific sparsity pattern in the attention matrix. It shows that sparse attention is sufficient to get state-of-the-art results in modeling long sequences over language modeling, image generation and music generation. (Sukhbaatar et al., 2019) builds upon this work and shows that is it is possible to obtain further sparsity by letting the model learn the length of the temporal context for each attention module. This work also makes use of the attention cache introduced in (Dai et al., 2019), a memory mechanism to train models over temporal contexts which extend beyond the length of the training batches.

Attention with Content-Based Sparsity: The above work mainly relies on two efficient ideas: attending to less elements by only considering a fixed bounded local context in the past, and attending to less elements by decreasing the temporal resolution of context. These ideas do not allow arbitrary sparsity patterns in attention matrices. Content-based sparse attention has been introduced to allow for richer patterns and more expressive models. (Martins and Kreutzer, 2017; Malaviya et al., 2018) propose to compute attention weights with variants of sparsemax. (Correia et al., 2019) generalizes this approach to every layer in a Transformer using entmax which allows for more efficient inference. This line of work allows for learning arbitrary sparsity attention patterns from data, based on the content of the current query and past context. However, sparsity here cannot be leveraged to improve space and time complexity since sparsemax/entmax formulations require instantiating the full attention matrix prior to sparsification. This is a drawback compared to temporal sparsity approaches. Our work is motivated by bridging this gap and allows for arbitrary sparsity patterns while avoiding to instantiate non-zero entries of attention matrices. Contemporaneous to our work, (Kitaev et al., 2020) proposed to use Locality Sensitive Hashing (LSH) using random hyper-planes to infer content based sparsity patterns for attention: tokens that fall into the same hash bucket, get to attend to each other. While similar in spirit to our approach, the approach of (Kitaev et al., 2020) keeps the randomly initialized hyper-planes fixed throughout, while we use mini-batch spherical k-means to learn the space-partitioning centroids. The motivation in both approaches is to approximate Maximum Inner Product Search (MIPS) in the context of dot product attention, for which both LSH and spherical k-means have been used in literature. However, typically spherical k-means is known to out-performs LSH for MIPS (see e.g. (Auvolat et al., 2015)). This is borne out in the common task of `Imagenet-64` generation, where Reformer gets around 3.65 bits/dim (Figure 3), while the Routing Transformer gets 3.43 bits/dim.

Sparse Computation beyond Attention: Learning models with sparse representations/activations for saving time and computation has addressed in the past in various context. Previous work often refers to this goal as *gating* for conditional computation. Gating techniques relying on sampling and straight-through gradient estimators are common (Bengio et al., 2013; Eigen et al., 2013; Cho and Bengio, 2014). Conditional computation can also be addressed with reinforcement learning (Denoyer and Gallinari, 2014; Indurthi et al., 2019). Memory augmented neural networks with sparse reads and writes have also been proposed in (Rae et al., 2016) as a way to scale Neural Turing Machines (Graves et al., 2014). In the domain of language modeling, a related work is the sparsely gated Mixture-of-experts (MOE) (Shazeer et al., 2017) where sparsity is induced by *experts* and a trainable gating network controls the routing strategy to each sub-network. Another related work is (Lample et al., 2019) who use product quantization based key-value lookups to replace the feed forward network in the Transformer. Our work differs from theirs in that we make use of dynamic key-value pairs to infer sparsity patterns, while their key-value pairs are the same across examples.

3 Self-Attentive Auto-regressive Sequence Modeling

Auto-regressive sequence models decompose the probability of a sequence $\mathbf{x} = (x_1, \ldots, x_n)$ as

$$p(\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_{i+1}|x_{\leq i}). \tag{1}$$

In neural models, the conditional distribution $p_\theta(x_{i+1}|x_{\leq i})$ is modeled by a neural network with learned parameters θ and these parameters are typically learned to maximize the likelihood of the training data. In particular, Transformer architectures have shown to reach state-of-the-art accuracy in several domains, including language modeling (Vaswani et al., 2017; Radford et al., 2018), image generation (Parmar et al., 2018) and music generation (Huang et al., 2018). Transformer models compose a series of attention modules. Each module refines the input representation by taking a weighted average of the representations from the previous modules.

For every module, the input representation is a sequence of n vectors $\mathbf{x} = (x_1, \ldots, x_n)$ from a continuous space of dimension d. Thus one may actually treat the input sequence as a $n \times d$ matrix X. A self-attention layer operates on this representation. It first applies three linear projections,

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \tag{2}$$

where Q, K and V are referred to as *keys*, *queries* and *values*, while W_Q, W_K, W_V are learned projection matrices.

The key and the query matrices determine the $n \times n$ attention matrix $A = \mathrm{softmax}\left(QK^\top\right)$, where the softmax operator over matrices denotes that the softmax function has been applied to each row. A may be interpreted as a matrix of weights in $[0, 1]$ where A_{ij} denotes how much query position i at the next layer must pay attention to key position j at the previous layer. In the case of self-attention for auto-regressive models, queries attend only over keys from previous time-steps, i.e.

$$A = \mathrm{softmax}\left(\mathrm{ltr}(QK^\top)\right) \tag{3}$$

where ltr denotes the lower triangular operator. Given the attention matrix A, the next layer representation X' is computed simply as AV. In summary,

$$X'_i = \sum_{j \leq i}^{n} A_{ij} V_j, \tag{4}$$

In practice, Transformer (Vaswani et al., 2017) adds several extensions to this basic self-attention mechanism. In particular, the result X' of performing self-attention is scaled by $1/\sqrt{d}$. Moreover, each layer relies on multiple attention *heads*, i.e. each layer performs multiple projections onto triplet (queries, keys, values) and attention is performed for each head. The attention results from all heads are then concatenated. This strategy allows each head to specialize on different aspects of the input sequence. In addition, Transformer further processes the result of attention through a learnable non-linear transformation (multi-layer perceptron, mlp) followed by a residual connection and a normalization step, i.e.

$$X' = \text{layernorm}(X' + X) \tag{5}$$

$$X'' = \text{layernorm}(\text{mlp}(X') + X), \tag{6}$$

where layernorm denotes the parameterized normalization step from (Ba et al., 2016a). A full Transformer model is therefore a chain of attention modules (Eq. 6) preceded by an embedding module (learnable representation for symbols and their positions) and followed by a logistic classification module (learnable linear classifier to predict the next symbol).

Our work is interested in the application of the Transformer to long sequences, a challenging problem since space and time complexity of attention is quadratic in sequence length n. We describe various approaches to sparse attention including ours in the next section.

4 Efficient Content-Dependent Sparse Attention

Attention-based models can be problematic for long sequences. For a sequence of length n, the full attention matrix A, as introduced in Section 3, is $n \times n$-dimensional and can be prohibitive to instantiate. This motivates sparse attention models, i.e. models relying on attention matrices which have a majority of zero entries.

For each query, a sparse attention model defines a set of keys which can be attended to. In the following, we introduce the set S_i as the set of key positions that the query at position i can attend to, i.e.

$$X'_i = \sum_{j \in S_i} A_{ij} V_j. \tag{7}$$

For example, classical causal self attention can attend to every key prior to the current query, which translates to $S_i = \{j \mid j \leq i\}$. Most previous work on attention sparsity defined such sets purely based on positions, independently of actual query and key vectors. For example, local attention (Luong et al., 2015) considers attending only to a k-long time window prior to the current query, $S_i = \{j \mid i - k \leq j \leq i\}$. (Child et al., 2019) propose block sparse attention where half the heads perform local attention, and half the heads perform *strided attention* given by $S_i = \{j \mid i - j \pmod k = 0, j \leq i\}$. (Sukhbaatar et al., 2019) is also a variant of local attention where the cardinality of $|S_i|$ is learned from data with an L_1 penalty to trade-off sparsity with modeling accuracy.

These *local* attention sparsity variants are effective in practice since correlation between observations naturally decrease with time for many problems. In our experiments, we actually find that local attention is a surprisingly strong baseline in both image generation and language modeling: for e.g., a scaled up ImageTransformer (Parmar et al., 2018) gets 3.48 bits/dim compared to the 3.44 bits/dim reported in (Child et al., 2019). Similarly, scaled up versions of Transformer with local attention and the relative positional encoding scheme of (Shaw et al., 2018) are able to get 19.8 perplexity on `Wikitext-103`, 1.10 bits per byte on `enwik-8` and 39.3 on `PG-19`, while Transformer-XL (Dai et al., 2019) gets 18.3, 0.99 and 36.3 respectively. From an efficiency perspective, local attention is also interesting since sparsity patterns are regular, contiguous in memory and known in advance.

In this work, however, we are interested in a more generic formulation of attention sparsity and would like the sparsity pattern to be informed by the data, i.e., $\mathcal{S} = f(\mathbf{x})$. This approach has several modeling advantages: it can accommodate data without a clear ordering over observations. For temporal data, it can also discover patterns with greater sparsity if some types of queries have a longer lasting effect on future observations than others. Content-based sparse attention should however be carefully implemented if we need to avoid instantiating full attention matrices at any point in time. For instance, (Correia et al., 2019) infer sparsity from data but their formulation instantiates a full attention matrix before finding its sparse counterpart. Next section explains how a natively sparse approach can actually be devised inspired by non-negative matrix factorization (NMF).

4.1 Routing Attention with Clustering

Our strategy follows the motivation we delineated in the previous section: we model sparse attention matrices with a low rank sparsity patterns relying on k-means clustering. Our strategy first assigns queries and keys to clusters. Then only queries and keys from the same cluster are considered for attention.

Precisely, our model clusters both keys K and queries Q undergo mini-batch k-means clustering on the same set of centroid vectors $(\mu_1, \cdots, \mu_k) \in \mathbb{R}^{k \times d}$. These centroid parameters are model parameters and are shared across sequences. They are learned online along with the rest of the parameters, as delineated in (Bottou and Bengio, 1995). Once cluster membership for each position i in the sequence is determined, we denote with μ the cluster corresponding to the query vector Q_i. This allows us to define our sparse attention strategy as

$$X'_i = \sum_{K_j \in \mu, j \leq i} A_{ij} V_j \tag{8}$$

In summary, queries are routed to keys belonging to the same cluster. Therefore, our attention sparsity pattern is of rank k, i.e. it can be represented as $FG^\top$ where F and G are binary matrices denoting cluster memberships of queries and keys respectively. It is important to note that this low rank property only concerns the sparsity pattern, while the resulting attention matrix $\mathrm{ltr}(FG^\top * A)$ can however be of higher rank ($*$ denotes element-wise product).

We work with queries and keys which are unit vectors, projecting them onto the unit ball, immediately before computing them. In practice, instead of normalizing by the ℓ_2 norm, we use Layer Normalization (Ba et al., 2016b) with the scale and bias terms disabled. This has the benefit of projecting vectors in $\mathbb{R}^d$ to the d-ball and prevents its entries from becoming too small. These layer normalized keys and queries are also used subsequently for computing the dot product attention. Note that performing k-means algorithm on unit vectors is equivalent to the *spherical* k-means algorithm. Projecting queries and keys to the unit ball implies that:

$$\begin{aligned} &\|Q_i - K_j\|^2 && (9)\\ &= \|Q_i\|^2 + \|K_j\|^2 - 2Q_i^\top K_j && (10)\\ &= 2 - 2\left(Q_i^\top K_j\right). && (11)\end{aligned}$$

Thus if Q_i and K_j belong to the same cluster center μ, then it follows that there is some $\varepsilon > 0$, such that $\|Q_i - \mu\|, \|K_j - \mu\| < \varepsilon$. This implies via triangle inequality that:

$$\|Q_i - K_j\| \leq \|Q_i - \mu\| + \|K_j - \mu\| < 2\varepsilon. \tag{12}$$

Thus from Equation 11 it follows that, $Q_i^\top K_j > 1 - 2\varepsilon^2$. Therefore, when two time steps $i \geq j$ are assigned the same cluster due to a small $\|Q_i - \mu\|, \|K_j - \mu\|$ distance, it also means that their attention weight $Q_i^\top K_j$ is high. This analysis shows that our clustering routing strategy preserves large attention weights as non-zero entries.

Since, we route attention via spherical k-means clustering, we dub our model *Routing Transformer*. A visualization of the attention scheme and its comparison to local and strided attention is given in Figure 1. The computational complexity of this variant of sparse attention is $O(nkd + n^2d/k)$. Cluster assignments correspond to the first term, i.e. it compares n routing vectors to all k centroids in a space of size d. Query/key dot products corresponds to the second term, i.e. assuming balanced clusters, each of the n queries is compared to n/k in its cluster through a dot product of dimension d. Therefore the optimal choice of k is $\sqrt{n}$ as in (Child et al., 2019), thereby reducing overall memory and computational cost to $O\left(n^{1.5}d\right)$ instead of $O(n^2d)$ (Vaswani et al., 2017).

In practice, we apply mini-batch k-means to train the cluster centroids. However, in order to infer balanced routing patterns, we define the sets C_i to be of equal size roughly $n/k \sim \sqrt{n}$, i.e. for every centroid μ_i we sort tokens by distance to μ_i and cluster membership is determined by this threshold (top-k). This adds an additional $O(n \log n)$ term to the cost, however note that this is eclipsed by the dominating term of $O(n^{1.5}d)$. This strategy is simple and efficient. In particular, it guarantees that all clusters have the same size, which is extremely important in terms of computational efficiency on parallel hardware like graphic cards. As a downside, this assignment does not guarantee that each point belongs to a single cluster. In the future, we want to investigate using balanced variants of k-means (Banerjee and Ghosh, 2004; Malinen and Fränti, 2014) which is not common in an online setting.

During training, we update each cluster centroid μ by an exponentially moving average of all the keys and queries assigned to it:

$$\mu \leftarrow \lambda\mu + \frac{(1-\lambda)}{2} \sum_{Q_i \in \mu} Q_i + \frac{(1-\lambda)}{2} \sum_{K_j \in \mu} K_j,$$

where λ is a decay parameter which we usually set to 0.999. Additionally, we also exclude padding tokens from affecting the centroids.

There is an additional nuance regarding clustering queries and keys that comes into play when using causal attention (i.e. left to right masking), as is usually the case in language models. When grouping queries and keys belonging to a certain cluster centroid μ, we may get as members queries Q_i for keys K_j where time-step $i < j$. This therefore requires an additional masking strategy in addition to the lower triangular mask used for causal attention. One solution that avoids having to use an additional mask, is to simply share keys and queries. Empirically, we have found that this works at par or better than separate keys and queries together with an additional masking strategy in the causal attention setting. For encoder self attention and encoder-decoder cross-attention, additional masking or sharing queries and keys is not necessary.

5 Experiments

We evaluate our sparse attention model on various generative modeling tasks including text and image generation. The following sections report our results on Wikitext-103 (Merity et al., 2016), enwik-8 (Mahoney, 2011), PG-19 (Rae et al., 2020), as well as ImageNet-64. We find that local attention is a surprisingly strong baseline and that our Routing Transformer outperforms Transformer-XL (Dai et al., 2019) and the Sparse Transformer model of (Child et al., 2019) on all tasks. On the recently released PG-19 data-set, the Routing Transformer model out-performs both Transformer-XL and Compressive Transformer (Rae et al., 2020), setting a new state-of-the-art result.

In all our models except the one used for PG-19, we allocate half the heads to do local attention and the other half to route attention as in Equation 8. For all our experiments except for PG-19, we use the Adam optimizer (Kingma and Ba, 2014) with learning rate 2×10^{-4} with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ following the learning rate schedule described in (Vaswani et al., 2017). We train all models on 256 TPUv3 cores. The setup used for PG-19 is described in Section 5.4.

5.1 Wikitext-103

Wikitext-103 (Merity et al., 2016) is a large public benchmark data-set for testing long term dependencies in word-level language models. It contains over 100 million tokens from 28K articles extracted from Wikipedia with an average of 3.6K tokens per article, which makes it a reference data-set to model long-term textual dependencies. We train a 10 layer Routing Transformer with 16 heads using the relative position encoding of (Shaw et al., 2018) and with attention and ReLU dropout rate of 0.3 each. For routing attention as in Section 4.1 we choose $k = 16$ and attention window to be 256 during both training and evaluation. We describe our results in Table 2 and compare it to other recent work on sparse or recurrent attention such as Adaptive Inputs (Baevski and Auli, 2019) and TransformerXL (Dai et al., 2019) as well as a local attention with relative position encoding baseline (Huang et al., 2018). We find that local attention is a great inductive bias for sparse attention and is better than the adaptive methods proposed in (Baevski and Auli, 2019; Sukhbaatar et al., 2019). Moreover, our Routing Transformer model is able to get a test perplexity of 15.8 improving on the 18.3 obtained by TransformerXL (Dai et al., 2019) while having fewer self-attention layers, and without the need for segment level recurrence.

5.2 enwik-8

The enwik-8 (Mahoney, 2011) is a data-set to benchmark text compression algorithms in the context of the Hutter prize. This data-set consists of the first 100M bytes of unprocessed Wikipedia. It is typically used to evaluate character-level language models. Similar to the prior work of (Dai et al., 2019; Child et al., 2019) we use a sequence length $n = 8192$ and benchmark our results against various baselines including local attention. We train a 24 layer model with 8 attention heads with an attention and ReLU dropout rate of 0.4 each and using the relative position encoding of (Shaw et al., 2018). For routing attention as in Section 4.1 we set $k = 32$ and attention window 256. We report perplexity of 0.99 like TransformerXL and Sparse Transformer, slightly under 0.98 from Adaptive Transformer.

5.3 ImageNet 64×64

In order to evaluate the ability of our model to capture long term dependencies on a modality other than text, we report results on the ImageNet 64×64 data-set as used in (Child et al., 2019). For autoregressive image generation, this data-set consists of images of $64 \times 64 \times 3$ bytes represented as long sequences of length $12,288$ presented in raster scan, red-green-blue order. We train a 24 layer model with 16 attention heads, with half the heads performing local attention, and the other half routing attention as in Section 3. For routing attention we set $k = 8$, attention window 2048, batch size 1 and train our model for roughly 70 epochs as in (Child et al., 2019). We compare our model to a scaled-up ImageTransformer model with local attention (Parmar et al., 2018) and the SparseTransformer model of (Child et al., 2019).

We find that local attention (Parmar et al., 2018) is a strong baseline for image generation, obtaining 3.48 bits/dim when scaled up to 24 layers and 16 heads, compared to later work like Sub-scale Pixel Networks (SPN) (Menick and Kalchbrenner, 2018). Our Routing Transformer model achieves a performance of 3.425 bits/dim (see Table 1) compared to the previous state-of-the-art of 3.437 bits/dim (Child et al., 2019), thereby showing the advantage of the content based sparsity formulation of Section 4.1.

5.4 PG-19

PG-19 is a new data-set released by (Rae et al., 2020) which is larger and longer than previous language modeling data-sets. The data-set is created from approximately $28,000$ Project Gutenberg books published before 1919, consisting of 1.9 billion tokens and comprises an average context size of roughly $69,000$ words. This is text that is 10× longer in context than all prior data-sets such as Wikitext-103, with minimal pre-processing and an open vocabulary that makes it extremely challenging for long text modeling tasks. We train a 22 layer Routing Transformer model with 8 heads with a sequence length of 8192 and set a new state-of-the-art result on this data-set, improving on both Compressive Transformers (Rae et al., 2020), as well as Transformer-XL (Dai et al., 2019). For this data-set we change our training setup in three ways. Firstly, we use only 2 routing heads instead

Model	Layers	Heads	Bits/dim
Glow (Kingma and Dhariwal, 2018)	-	-	3.81
PixelCNN (Van den Oord et al., 2016)	-	-	3.57
PixelSNAIL (Chen et al., 2018)	-	-	3.52
SPN (Menick and Kalchbrenner, 2018)	-	-	3.52
ImageTransformer (Parmar et al., 2018)	24	16	3.48
Sparse Transformer (Child et al., 2019)	48	16	**3.44**
Routing Transformer	24	16	**3.43**

Table 1: Results on image generation on ImageNet 64 × 64 in bits/dim.

Model	Layers	Heads	Perplexity
LSTMs (Grave et al., 2016)	-	-	40.8
QRNNs (Merity et al., 2018)	-	-	33.0
Adaptive Transformer (Sukhbaatar et al., 2019)	36	8	20.6
Local Transformer	16	16	19.8
Adaptive Input (Baevski and Auli, 2019)	16	16	18.7
TransformerXL (Dai et al., 2019)	18	16	18.3
Routing Transformer	10	16	**15.8**

Table 2: Results on language modeling on `Wikitext-103` data-set. Local Transformer refers to Transformer (Vaswani et al., 2017) with relative position encoding (Shaw et al., 2018) together with local attention. Perplexity is reported on the test set.

Model	Layers	Heads	Bits per byte
T64 (Al-Rfou et al., 2019)	64	2	1.13
Local Transformer	24	8	1.10
TransformerXL (Dai et al., 2019)	24	8	0.99
Sparse Transformer (Child et al., 2019)	30	8	0.99
Adaptive Transformer (Sukhbaatar et al., 2019)	24	8	**0.98**
Routing Transformer	12	8	0.99

Table 3: Results on language modeling on `enwik-8` data-set. Local Transformer refers to Transformer (Vaswani et al., 2017) with relative position encoding (Shaw et al., 2018) together with local attention. Bits per byte (bpc) is reported on the test set.

of sharing it equally with local heads. Secondly, we use routing heads only in the last two layers of the model instead of having them present in every layer. This is motivated by our empirical finding that long range attention is only needed in the last few layers - see also (Rae and Razavi, 2020). Finally, we use the Adafactor optimizer (Shazeer and Stern, 2018) which is more memory efficient than Adam in training larger models. We use a learning rate constant of 0.01 with a linear warmup over $10,000$ steps followed by a *rsqrt_normalized_decay*. We do not make use of any dropout, or weight decay. The hidden dimension of our model is 1032 and the batch size is 8192 tokens.

From Table 4, we see that Local Transformer again sets a very strong baseline, with a 24-layer local attention model obtaining a test set perplexity of 39.3, while a 36-layer Transformer-XL gets 36.3. Moreover, a 22-layer Routing Transformer model improves on the 36-layer Compressive Transformer, obtaining a test set perplexity of 33.2 compared to 33.6, while being able to generate sequences of length 8192.

6 Analysis

6.1 Local vs Global

We evaluate the difference in attention patterns between local and routed attention and compute the Jensen-Shannon divergence between local attention and routed attention for a random subset of heads in our network on the `Wikitext-103` data-set. The divergence is computed over the entire sequence length of 4096. We average over 10 runs and report means and standard deviations of the JSD in Table 5. Note that the JSD is always non-negative and is upper-bounded by 0.6931 when computed using the natural logarithm. We observe that the divergence between the different local heads is always very low compared to the divergence between local and routing attention heads, which is almost always very close to the upper-bound of 0.6931. Divergence between different routing attention heads falls somewhere in between, being closer to the upper-bound. This shows that the attention distribution inferred by the routing attention of Section 4.1 is highly non-local in nature and different heads specialize in attending to very different parts of the input.

Qualitatively, the reason for the strong performance of the Routing Transformer is due to the fact that it approximates full dot-product attention by performing an approximate Maximum Inner Product Search (MIPS) over the entire set of tokens, and selecting pairs that have a high dot product for attention. This allows various entities such as gender, nouns and names of places to be consistent throughout the entire sequence, since on expectation the dot product similarity between similar entities are high, while for differing entities they are expected to be low. Essentially, we hypothesize that for every time step, the prediction depends on a small support of *high value* tokens: local attention facilitates local consistency and fluency, while a full dot product attention would facilitate global consistency. However, for long sequences since full attention is intractable, we believe that using spherical k-means to perform a MIPS search over the global set of tokens and performing attention between these high dot product items is a good proxy for *full attention*.

6.2 Memory vs Sparse Attention

We also note that sparse attention is an orthogonal approach to that of Transformer-XL and Compressive Transformer, which train on small sequences and by performing careful cross attention across different chunks of these small sequences hope to generalize to longer sequences. By contrast, we directly train on long sequences from the beginning - e.g., the Compressive Transformer trains on chunks of size 512 for `PG-19`, while we train on sequences of length 8192. The benefit of the Transformer-XL like approach is that it is less memory consuming and thus is able to scale to 36 layers. Sparse attention (including local attention) on the other hand is more memory expensive and therefore can scale to fewer layers for the same problem. However, as we demonstrate, it is competitive with the Transformer-XL like approaches even when using fewer layers and is guaranteed to generalize to the long sequence length that it was trained on.

6.3 Wall-clock time

We also compare the step times of the Local Transformer and the Routing Transformer on a TPUv3 for the `PG-19` data-set in Table 6. Since sequence lengths are 8192, a wall-clock time comparison with full attention is infeasible in this setting. From Table 6, we see that the Routing Transformer is roughly 0.54× slower to train compared to the Local Transformer in wall-clock time on a TPUv3. This is due to the lack of support for sparse operations on the TPU; on the GPU various sparse kernels have been proposed which promise to significantly speed up training of these models (Gale et al., 2020). Note that our goal in this work is a memory efficient version of sparse attention that can well approximate full attention for long sequences - wall-clock time efficiency is only a secondary goal. Indeed, one can make attention very efficient in the wall-clock time sense by simply projecting the $n \times d$ sequence embedding to a $k \times d$ vector for some constant k, but this trades off performance for wall-clock efficiency.

7 Conclusion

Transformer models constitutes the state-of-the-art in auto-regressive generative models for sequen-

Model	Layers	Heads	Perplexity
Local Transformer	24	8	39.3
TransformerXL (Dai et al., 2019)	36	-	36.3
Compressive Transformer (Rae et al., 2020)	36	-	33.6
Routing Transformer	22	8	**33.2**

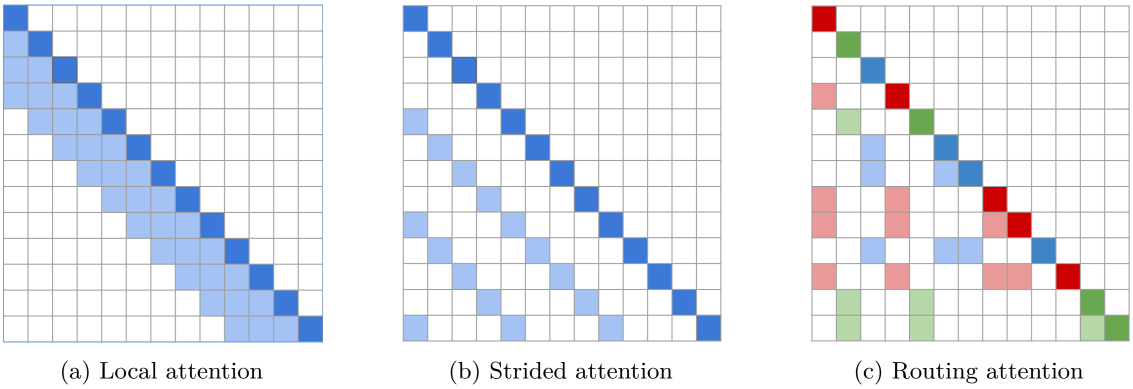
Table 4: Results on language modeling on `PG-19` data-set. Local Transformer refers to Transformer (Vaswani et al., 2017) with relative position encoding (Shaw et al., 2018) together with local attention. Perplexity is reported on the test set.

JSD($local \| local$)	JSD($local \| routing$)	JSD($routing \| routing$)
0.1776 ± 0.0649	0.6044 ± 0.0181	0.4181 ± 0.0415

Table 5: Jensen-Shannon divergence between the attention distributions of a random local attention head and a random head that routes attention as in Section 4.1 per layer on the `Wikitext-103` data-set. We report means and standard deviations computed over 10 runs and use the natural logarithm so that divergences are upper-bounded by 0.6931.

Model	Layers	Heads	Steps/sec
Local Transformer	24	8	1.231
Routing Transformer	22	8	0.669

Table 6: Step time comparsion between Local Transformer and Routing Transformer on a TPUv3 for the `PG-19` data-set with sequence length 8192.


(a) Local attention

(b) Strided attention

(c) Routing attention

Figure 1: Figures showing 2-D attention schemes for the Routing Transformer compared to local attention and strided attention of (Child et al., 2019). The rows represent the outputs while the columns represent the inputs. For local and strided attention, the colored squares represent the elements every output row attends to. For attention routed as in Section 4.1, the different colors represent cluster memberships for the output token.

tial data. Their space-time complexity is however quadratic in sequence length, due to their attention modules. Our work proposes a sparse attention model, the Routing Transformer. It relies on content-based sparse attention motivated by non-negative matrix factorization. Compared with local attention models, it does not require fixed attention patterns but enjoys similar space-time complexity. In contrast with prior work on content-based sparse attention, it does not require computing a full attention matrix but still selects sparsity patterns based on content similarity.

Our experiments over text and image generation draw two main conclusions. First, we show that a carefully tuned local attention model establishes a strong baseline on modern benchmark, even compared to recent state-of-the-art models. Second, we show that the Routing Transformer redefines the state-of-the-art in large long sequence benchmarks of `Wikitext-103`, `PG-19` and `ImageNet-64`, while being very close to do so on `enwik-8` as well. Our analysis also shows that routed attention modules offer complementary attention patterns when compared to local attention.

Overall, our work contributes an efficient attention mechanism that applies to the modeling of long sequences and redefines the state of the art for auto-regressive generative modeling. Our approach could prove useful in domains where the inputs are naturally sparse, such as 3D point clouds, social networks, or protein interactions.

8 Acknowledgments

The authors would like to thank Phillip Wang and Aran Komatsuzaki for a Pytorch implementation of Routing Transformer. The authors would also like to thank Yonghui Wu, Weikang Zhou and Dehao Chen for helpful feedback in improving our work.

References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.

Alex Auvolat, Sarath Chandar, Pascal Vincent, Hugo Larochelle, and Yoshua Bengio. 2015. Clustering is efficient for approximate maximum inner product search. *arXiv preprint arXiv:1507.05910*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016a. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016b. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexei Baevski and Michael Auli. 2019. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Arindam Banerjee and Joydeep Ghosh. 2004. Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres. *IEEE Transactions on Neural Networks*, 15(3):702–719.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. 2019. Learning classifiers with fenchel-young losses: Generalized entropies, margins, and algorithms. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 606–615.

Leon Bottou and Yoshua Bengio. 1995. Convergence properties of the k-means algorithms. In *Advances in neural information processing systems*, pages 585–592.

Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. 2018. Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, pages 864–872.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Chung-Cheng Chiu* and Colin Raffel*. 2018. Monotonic chunkwise attention. In *International Conference on Learning Representations*.

Kyunghyun Cho and Yoshua Bengio. 2014. Exponentially increasing the capacity-to-computation ratio for conditional computation in deep learning. *arXiv preprint arXiv:1406.7362*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Ludovic Denoyer and Patrick Gallinari. 2014. Deep sequential neural network. *arXiv preprint arXiv:1410.0510*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Ding, Xiaofeng He, and Horst D Simon. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen. 2020. Sparse gpu kernels for deep learning. *arXiv preprint arXiv:2006.10901*.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2016. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*.

Sathish Reddy Indurthi, Insoo Chung, and Sangha Kim. 2019. Look harder: A neural machine translation model with hard attention. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3037–3043.

Navdeep Jaitly, David Sussillo, Quoc V Le, Oriol Vinyals, Ilya Sutskever, and Samy Bengio. 2015. A neural transducer. *arXiv preprint arXiv:1511.04868*.

Jingu Kim and Haesun Park. 2008. Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2019. Large memory layers with product keys. In *Advances in Neural Information Processing Systems*, pages 8548–8559.

Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, pages 4487–4496.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Matt Mahoney. 2011. Large text compression benchmark. *URL: http://www. mattmahoney. net/text/text. html*.

Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. 2018. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–376, Melbourne, Australia. Association for Computational Linguistics.

Mikko I Malinen and Pasi Fränti. 2014. Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 32–41. Springer.

André F. T. Martins and Julia Kreutzer. 2017. Learning what's easy: Fully differentiable neural easy-first taggers. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 349–362, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Menick and Nal Kalchbrenner. 2018. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *International Conference on Learning Representations*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. An analysis of neural language modeling at multiple scales. *arXiv preprint arXiv:1803.08240*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Jack Rae, Jonathan J Hunt, Ivo Danihelka, Timothy Harley, Andrew W Senior, Gregory Wayne, Alex Graves, and Timothy Lillicrap. 2016. Scaling memory-augmented neural networks with sparse reads and writes. In *Advances in Neural Information Processing Systems*, pages 3621–3629.

Jack Rae and Ali Razavi. 2020. Do transformers need deep long-range memory? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7524–7529, Online. Association for Computational Linguistics.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604.

Sainbayar Sukhbaatar, Édouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015.

Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

A Samples from Routing Transformer

In the following sections we present a few samples generated from the Routing Transformer trained on the `PG-19` data-set with sequence length 8192. We use nucleus sampling (Holtzman et al., 2019) with $p = 0.8$ and temperature of 1.0 to generate these samples.

A.1 Sample - I

During the early days of the Council of Nice and the subsequent existence of the Council of Basle, one section of the Council of Nice made a marked opposition to the pretensions of the Council of Basle. Some of them were men of high rank, others members of the lower classes. They had been formed into a union which was called the Papal Council, and which for the time being proved to be of the utmost importance to that Church in which it met. The first session of the Council of Nice took place on September 2, 1487. The two bodies met in solemn assembly and made arrangements with one another. It was decided that a considerable portion of the Council should proceed to Zurich and lay before the Council of Constance the proposals of the Church of Basle for settling their disputes. It was also resolved that a meeting of the representatives of the Christian Emperors of Germany, France, Portugal, Hungary, England and France should be held. Three bishops were commissioned to despatch ambassadors to each of the two Councils to urge their respective envoys to combine and come to some agreement regarding matters ecclesiastical. It was agreed that the Council of Basle should at once take steps for the reformation of the Church and the peace of Christendom; while the two meetings were to be united in one. Various questions of dispute were settled in a friendly way; but the whole subject of the relations of the Church to the Papacy was laid before the Council of Basel, and an agreement arrived at regarding the ecclesiastical and civil relations of the Church with the head of the Papacy.One important result of this Council was that it thus obtained two important concessions from the Popes: the first in making a papal establishment the natural basis of ecclesiastical authority on a great scale and yielding to the papal pretensions; and the second in providing for a Papal Council of Basle in which there should be ecclesiastical authority, and a bishop of the Roman Church, to meet the needs of the Churches of Europe. The Council of Basle likewise obtained the provision that the election of the Pope should be conducted by the same general council and by the head of the Church at Rome, and that no other form of appointment than that of a personal election to the Papacy should be in force. It was in effect a completion of the Council of Basle. It left without a head, indeed, but with an indication of its existence, the crowning work of the nineteenth century. The Council of Basle had not succeeded in bringing about the acceptance of the Papal headship; but there can be no question that the defeat of the Papal claim, at the Council of Lyons in the year following (December 17,1530), determined the attitude of the Papacy towards the Church, and prepared the way for the action of the Council of Trent. For at that time it seemed as though, after the Council of Lyons, the Council of Trent could no longer prevent the intrusion of the Papacy into the Church, and it was recognised that there was to be no more preaching in the Churches of Europe, for this once. Yet the fact remains that there was no Papal interference with Church government. From that time forward,however, the rule of the Church became more rigorous, and towards the end of the sixteenth century began the crisis in the Church which lasted until the general council of the Council of Trent.The organisation of the Swiss Church had been brought down to the time of Zwingli (1516-1531). It was based upon an organisation strictly clerical in character, as the Canons of the Roman Church insisted upon the clergy being for the most part clerics of the clerical order. In this respect this system was a reminiscence of that of the Roman Church, except that the mass of the people were clerics of the clerical order, who were liable to be deposed at any moment by the spiritual authorities. In the present instance we must recognise that Zwingli introduced a new conception of Church government; for although a great deal of the work of the Reformation was done under the direction of Zwingli, yet the organisation of the Swiss Church to some extent, and the connection of the civil with the ecclesiastical system, served as models for the organisation of the Church in all the Protestant lands.No doubt there was a great amount of copying of Rome, and some irregularities of arrangement were to be found. It is to be noted, however,that most of the reformation principles and practices of the Reformation were embodied in the Church organisation of the Swiss Protestants; the chief result being that, whereas the earlier system was still simple, the Church reformed more strongly and specifically, and was thereby destined to get more help in the direction of the Protestant reformation. So that even in the confusion arising from the change of Church organisation in the sixteenth century the Swiss Church was drawn much more closely to Rome than it would otherwise have been.The first work of the Reformation, however, to which the introduction of the Bible is to be attributed, was done in the early years of the sixteenth century. The era of the Reformation had begun; and this event was by no means likely to pass over without some indication of its influence in the world, for the Reformation had assumed the character of a great political event. The work of the reformation was in a large degree concerned with the national character of Protestantism. The reformation had been the work of religious philosophers, and it was a momentous and noteworthy step towards the winning of the political independence of the nations. But Luther had accomplished no permanent political revolution.Instead of that he had worked to establish that political absolutism of the kings which is the most distinctive characteristic of the Protestant polity. It was not in the modern European sense that he destroyed feudalism and other institutions based on tradition; for the victory was of the Gospel, and he hoped by its means to add another to the ten thousand proofs of the Divine origin of the kingdom of God. The power which he had created was in a large sense political power, and it was part of his function to secure such political power for the Church. He also worked, at an early stage, to further the establishment of the independence of the Church of the Brethren, but it was not until the Reformation became an aggressive factor in the life of the nation that the need for further political recognition of the Church was felt.The reformation movement was to have a most important effect on other aspects of the life of the people, and also upon the growth and extension of Protestantism. The great change which was thus produced, and which has been described as the direct and immediate outcome of the Reformation, was in effect essentially religious in its nature. The re-establishment of the Church of the Brethren has never been one of the least noteworthy phases in the history of the nation. During the next two centuries the popular Church of the Brethren increased in number, importance, and popularity. The king, the nobles, and the more educated portion of the people came more and more to regard it as the natural bulwark of Protestantism; and in a comparatively short time, and within comparatively short space of time, that work which Luther did for the establishment of the national life has been carried to a high degree of accomplishment by the English and other Protestant communities. The beginning of the Reformation, as already indicated, was a direct consequence of the effects which were brought about by the Reformation.It was not simply in the Church that the recognition of the Church of the Brethren made itself felt. The religious feelings which were aroused,and which were finally developed into a religious habit, have already been sufficiently dealt with in connection with the general history of the German nation; and the re-establishment of a purely spiritual faith and of a dominant religious life in the land, one which could not possibly have been attained save by the outpouring of the Holy Spirit and by the renewing and transforming influences of the Divine Spirit, was among the first results of the Reformation. To that work belongs the development of the German Reformation in its broadest and widest form; and the causes which determined the course of that development may be shortly stated as follows:In the first place, we have seen how the study of the Bible and of the Apocrypha, and of the Jewish conception of God and of the obligation to fast, created a desire for the study of the Bible in a larger and deeper manner than any before known; and, secondly, how the passion for writing profane history and for the writing of sacred history was fostered by the increase of the Roman Church; and,thirdly, how the study of the Scripture in a more liberal spirit–a great impulse to the study of the Old Testament in an earlier period in all its forms, and towards a development of the conception of God and of a more secular spirit in the life of the nation, helped to accelerate the spread of a new and healthier conception of the Christian life. This latter result, and this alone, tended to produce a new and productive condition of the nation in the matter of religion; but it also reacted on the missionary endeavours of the members of the Church of the Brethren to attain a deeper religious development. The desire to read the Bible, to adopt the principles of the Reformers, and to raise the standard of life and manners, not only stimulated the energy and assisted the zeal of the societies of the Brethren, but also stimulated their wider application to particular branches of the work which they had to do. In other words, the deeper study of the Bible as the study of the Old Testament became the religion of the people, and by the sheer force of the influence of these early studies the religious work of the German Reformation took shape, and became one of the most important political movements in Germany. The movement,thus inaugurated, was still later in reaching results in other countries.Before reaching Germany, however, the religious work of the Reformation had made a great impression upon one of the rulers of that country. Philip of Hesse, in 1495, was a child in years; but he was a man of religious instincts and aspirations, and his first utterances were destined to be the embodiment of that new religious idea which for so many years had been deeply implanted in the national mind. The importance of this movement will not be denied. It was an expression of the revival of the primitive and devout tendencies in the Lutheran Church; and in the Lutheran Reformation itself there was far less of scientific study than of poetic expression. It is plain, then, that the Reformation movement in Germany was in some respects influenced, as it was also in some respects modified, by the study of the Scriptures in their original languages, and with a more modern translation of the Bible into modern German.But the condition in which the Reformation found Germany had in a large measure changed. The enthusiasm which formerly animated men for the study of the Bible in all its original tongues was broken down. They did not recognize that the Bible for the understanding of God's Word, and for its guidance through life, was not only the best language in which it was written, but, as already noticed, it was the chief interpreter of all other languages. When we remember that Luther was a professor of Divinity at Wittenberg; that Luther had expounded, in the German tongue, his new faith and new life; and that this same translation had found its way into the minds of thousands and tens of thousands of people in other countries; and that the old German Bibles did not by any means constitute the translation generally used, and that, except for the selection of modern translations, the standard text of the German Bibles for our service was of course by no means the best, we can hardly fail to see that it became clear that the Scriptures as the Bible for the understanding of God's Word were inadequate for the elucidation of religious problems; and, further, that there was no substitute, no adequate translation of the Bible that was available.In order that the question of its complete translation might be understood, it was necessary to seek to adapt it to the spirit and needs of Germany, and this was the task which the Government of the German Empire set itself, and upon the result of which depended the situation under which the Reformation came about.The aim of the Reformation, in the words of Luther himself, was,primarily, the study of the Bible as a living interpreter of God's words and revealing God's will in them; and, secondarily, the acquisition of a living, active, self-interpreting, and God-glorifying Christian spirit. In order to study the Old Testament as a living revelation of God's character and as an example of what God's Spirit, as that Spirit of truth, is capable of doing, it was essential that they should have some historical contact with the Old Testament; and this contact was brought about by the introduction of commentaries on its text. It was in this way that the institution of the /Kleinpostille/ and the growth of a literature for it were due to the zealous and devoted efforts of German Christians at this period. It was because the /Kleinpostille/ and the/Kleinpostille-Lexicon/ were due to the vigorous, self-denying, energetic,and helpful German literature which sprang up in Germany during this period, that the celebrated /Lutherana/ was put forth in the sixteenth century.Nor did it remain for the Reformation to avail itself of the facilities which this literary form gave it in Germany. It had not been intended to continue its work without the aid of a translation, and before it was generally accepted as such an adequate one, a work of translation had to be done, and this was accomplished in a most able and painstaking fashion by /The Commentary on the Galatians/ in 1531. In that work, also, the advantages of translation, as well as the emphasis which the services which it rendered were warranted to lay upon, were well recognised, and it has always been thought that Luther's translation was the best rendering that was available for his readers.There is no need to dwell upon the fact that a work such as this,which for twenty years was in the hands of all the students of German theology, could not have found its way to a Christian home in a Protestant country like Germany without being a source of new and most valuable information. We find, indeed, in it the most valuable reflection on the extent of the religious life and the condition of culture in the countries which represented the belief and received the teachings of the Reformation, as well as the most remarkable revelation of the kind which the Lutheran Reformation contain

A.2 Sample - II

White-deer a pair of grey, northern Algonquin, also white-deer of a paler colour than common. Great babbler, the commonest of summer warblers,all these are found in a great number of localities in southern Ontario;but at Lake Erie and Lake Ontario, where they are few, they are quite common.Then, again, during the migration season they will often be seen consorting with their relatives the Canada Jay. On this account, a very large number of hawks that, though they are not regular songsters, are generally taken on the wing. But they are especially abundant in Newfoundland in the neighbourhood of the Little Fête and other great feasts, and are likewise met with in Newfoundland in winter, where they may be seen all the time, though they do not come in great numbers into the towns.Audubon tells us that although nearly all these birds spend the summer in Canada, yet they frequently winter in South America. Such have been frequently seen, but never described, by other observers. In studying any of these little northern warblers, we must go back to the winter quarters of these little birds, or at least see where they pass the summer.[Illustration: AMERICAN GOLDEN PLOVER, MALE AND FEMALE]How beautifully speckled are the breasts of these Golden Plovers! how beautifully spotted the upper parts of the head and breast, especially the under wing coverts. But on this account, their bright colours are particularly attractive, because the group is very abundant, and their close relative, the Golden Plover, is also frequently seen in the far north.This bird breeds sparingly in various parts of North America, but almost exclusively in Labrador. There it nests in small colonies of a dozen or more, making choice, I have no doubt, of some open, dry piece of ground,building their nests of grass and scraps of grass, placing them in the midst of grass on which, in company with their kindred, they pass the winter. The nest is built, in all probability, on the ground, or on the top of a tussock of grass or a tuft of oats, which has been dried, or rolled into a conical shape by birds, but which they have neglected to do for themselves; and after laying their eggs, they scrape down the soil upon which the nest is built, and together, with a few young, feed them all the summer. They pair about the end of April, and begin to breed so soon as the breeding season has passed, at the same time that the male bird may be seen sitting upon the outside of the nest.The nest of this species is not built as closely as that of the English species, and not being peculiar to America, a large number of its eggs has been obtained in Great Britain, and it is highly probable that it exists abundantly in the United States also.In breeding time, the Gulls and Terns, as well as the other birds, do not congregate in large flocks, but generally avoid flocks that are daily passing, and thereby contribute very much towards diminishing the number of their feathered associates, which, being fewer, would be more easily preserved. The same thing may be said of the very numerous young which come with the large migration northward, and, in a measure,counteract the tendency to overcrowding.But although the Gulls and Terns are thus apt to resort to the north in winter,how many of the same species are known to breed in the other parts of the world? The British Islands, indeed, are but thinly populated, and the season for breeding does not arrive so early as that for breeding in Europe. We find, therefore, in the British Islands only a few pairs or very few individuals. The Skuas and Petrels are probably more numerous,but such is the local distribution of this species, that it is difficult to find more than three or four of its breeding haunts. Our only figure of this species is in the "Manual" for the year 1858, in which it is figured under the name of _Crex pusilla_.[Illustration: BLACK GUILLEMOT]BLACK GUILLEMOT.* * * * *SPECIFIC CHARACTER.BLACK GUILLEMOT.–Bill, the base of the upper mandible and the tip of the ear black; legs, legs, toes, and feet, black; wings, blackish, the feathers margined with dull ash-grey; upper parts ash-grey; quills blackish, margined with greyish; tail blackish, the inner three feathers of the outer web tinged with brown, and the next tipped with white, except on the inner web; the two outer feathers of the outer web tipped with white.* * * * *The present species was discovered by Captain King at Sitka, in Russian America, and may be distinguished from the preceding by its black rump,beneath which are eight blackish-brown lines, beginning at the base of the feathers. In its haunts, it is rather tame, but in autumn it seldom perches on trees. On the coast the breed begins to breed in December, and by the end of April it will have laid about six eggs. It is somewhat gregarious, sometimes in large flocks. A female caught in Baffin's Bay in 1825 was of a sooty black colour above and light ash-grey below, with three of the tail-feathers of a blackish tinge.* * * * *TEMMINCK'S GUILLEMOT.TEMMINCK'S GUILLEMOT (_Haematopus bairdii_) is said to have been taken near the mouth of the Columbia, and by Captain Cook has been called the Common Guillemot.TEMMINCK'S HELMET.TEMMINCK'S HELMET. Plate XXI. fig. 3.* * * * *Adult Male. Plate XXII. fig. 1, 2.Bill, the base of the upper mandible and the tip of the ear black; legs,feet, toes, and feet black; upper part of the head and neck dark ash-grey;back, scapulars, wing-coverts, and quills black, the latter margined with pale greyish-white; tail of the same colour, the middle feathers of the outer web at the end tipped with white; three outer feathers of the same, and the next two very slightly tipped with the same; lower parts white.Total length 5 inches, extent of wings 5, depth of body 2 1/2 inches.This species is only two feet ten inches in length, and during the summer time, during which it can be seen floating on the ocean in autumn,resembles the preceding, but it is so extremely scarce, that it is rather a difficult matter to ascertain its haunts. I have no doubt that it migrates from Europe, across the Atlantic, to the north, even where it is now known to be extinct.* * * * *AMERICAN SEA-EAGLE.EIDER-BILLED BOOBY.* * * * *_HaliaA|etus leucogaster_, Wils.* * * * *AMERICAN WHITE-FRONTED BOOBY (_HaliaA|etus leucogaster_, TEMM.) is one of the smallest of the American species, measuring only five inches and three quarters in length. The bill is black, and the feet deep brown. It is a bird in the collection of the late Mr. John Cassin of New York, and was shot in the neighborhood of Lake

Erie. Length 5 inches and 3/4, extent of wings 3 inches and 1/4, depth of body 1 1/2.* * * * *I have been indebted for the above description of the Blue-headed Buzzards to my friend, Mr. Wm. L. Beal.* * * * *PALL MALL BLUE-HEADED BOOBY (_HaliA|etus pallens_, TEMM.) may be distinguished by the reddish band over the eye, and the brown patch on the primaries,which are longer and more attenuated, than the black ones of the last species, the bill being a little broader and red, and the legs lighter than those of the last species. It has been called the Alpine Blue-headed Booby, by the late Dr. Edward Smith, in his description of this bird. I believe that there is but little difference in its appearance, except the colour of the bill, which in the male is of a dark brown, in the female yellow.* * * * *HORNED OWL._Strix flammea_, LINN.* * * * *_Strix argemone_, LINN.* * * * *The habits of the Horned Owl are, like those of the Snow Owl and the Long-eared Owl, imperfectly known. They have long been familiar objects to the inhabitants of the northern parts of our country, who are accustomed to their appearance and mode of travelling in companies. They are most frequently seen in the night. It is often heard to hoot, or squeal,and at times is very noisy.It is found during the whole of the northern summer, on the pine plains and barrens, on the <DW72>s of the higher elevations of our country, and in the northern parts of Maine, Nova Scotia, Newfoundland, and in several parts of New England. It is one of the most common inhabitants of our villages, and is so extremely restless and active, that it is almost impossible to catch it. They are very bold and noisy, rising from the tops of the low bushes and branches, and making a terrible hissing, as they do when alarmed, which will draw on them the attention of the person who perceives them. They are generally seen in flocks, and at all times wary, giving notice of the approach of danger, by their peculiar crowing, and various notes, which are peculiar to themselves, and often mistaken for a call. Their note resembles that of the Owl, and is much louder, resembling the cry of the Great Horned Owl.* * * * *I have been thus particular in giving you the above description, as I believe this species to be the one I have already figured. You will readily believe that it would be impossible for me to decide in which of the two localities which I have described the bird is to be looked for. I only mention the latter, as the description agrees better with that of the present bird than with that of any other in which I have seen it.* * * * *CHIMÆOLURUS VIRGINIANUS, _Lath._ Ind. Ornith. vol. ii. p. 301.—_Ch.Bonaparte_, Synops. of Birds of the United States, p. 54.CHIMÆOLURUS VIRGINIANUS, _Nuttall_, Manual, part i. p. 215.AMERICAN CHIMÆOLURUS, CHIMÆOLURUS AMERICANUS, _Ch. Bonaparte_, Amer.Ornith. vol. ii. p. 39. pl. ii. fig. 2.—_Nuttall_, Manual, p. 209.Adult Male. Plate XXIII. Fig. 1.Bill rather long, slender, strong, compressed toward the end; upper mandible with the dorsal outline a little convex, the ridge rather wide and flat,the sides convex from the base, the edges overlapping, the tip declinate;lower mandible with the angle narrow and very long, the dorsal line rather convex, the sides rounded, the tip acute. Nostrils basal, lateral,round, covered by the reversed filaments of the frontal sinuses. Head rather large. Body moderate. Legs of ordinary length; tarsus very strong,scutellate anteriorly, acute behind; toes free, scutellate above, the lateral ones nearly equal, the hind toe larger; claws of ordinary length,compressed.Plumage soft, blended, somewhat blended, not glossy. Wings rather long,third quill longest, second and fourth equal. Tail of ordinary length,slightly emarginate, the two lateral feathers longest, the two lateral inferior with some small tips.Bill deep brown, black at the end, paler at the sides. Iris brown. Feet flesh-colour. Head and neck pale ash-grey. Back, scapulars, and rump dark umber-brown, reflecting into deep brown, the tail, secondary quills,and coverts, as well as the ends of the secondary quills, and tips of the larger ones, white. Wings dusky, their coverts margined externally with reddish-brown. Fore part of the back, breast, and abdomen deep brown, tinged with orange; the breast tinged with yellow, the abdomen with a tinge of dull red. On the breast a broad band of dusky red on each side.Length 7 inches, extent of wings 10; bill along the ridge 1-3/12, along the gap 1-1/12; tarsus 2-1/12.Adult Female. Plate XXIII. Fig. 2.The Female resembles the male, but somewhat resembles the white-headed Woodpecker, the head, neck, breast, and abdomen being pale ash-grey.The young resemble the female, and differ from the male, in having the chin and fore part of the breast light ash-grey, and the rest of the under parts ash-grey.THE COTTON PLANT.GOSSIUM GLYCYLLARUM, _Willd._ Sp. Pl. vol. ii. p. 779. _Pursh_,Flor. Amer. vol. ii. p. 422.—DECANDRIA MONOGYNIA, _Linn._DECANDRIA RHAMNACEAE, _Juss._This plant, from which the generic name of this genus is derived,is distinguished by its pendulous cymes of large, silky, terminal panicles, and by the sinuosities of the branches, which are mostly smooth. The leaves are cordate, downy, and attenuated at the base. The flowers are pale orange-, and exhale a strong and very pleasant odour.THE HIGH BERRIES OF THE NORTH.(_MAGNOLIA CANADENSIS_, DESK.) NORTH OF KINGSBRIDGE.[Illustration: THE HIGH BERRIES OF THE NORTH.]The highest trees in the county of Brunswick are found near the town of Kingston; but the low and more sheltered parts of the country have abundance of the lowgrowing aromatic, which grows there from seed, and is,consequently, of a superior quality. Not more than fifty or sixty miles below the town of St. John’s, this shrub attains a height of upwards of fifty feet, with spreading branches of beautiful spreading foliage.THE CRANE CRANE._CATHARTE CANADENSIS_, TEMM.PLATE XXIII. MALE AND FEMALE.This species has never, or very rarely, been observed on our seaboard during the spring and summer, unless I mistake not, as is said by the natives, in many parts of Newfoundland. It frequently comes within a few miles of the sea-shore, and after passing over the downs or beach, settles upon the marshes or small islands, erecting its nest on the summit of a large tree, and generally resting on the trunk. There is, at all times, a sufficient number of young ones to fill its nest, and, consequently,it seldom requires to be robbed. It generally dwells upon high and exposed situations, yet never in an open forest. As many as four or five nests of this species may often be observed on a single tree, situated on a level with the ground, or where the lower branches have been broken off by storms. It sits upright, with its neck or tail drawn in, and so rarely, on opening its mouth, that you may often look down into it, and take your bird out by the neck or tail. It is only during the autumn, and towards the close of that season, that it deserts the salt marshes, retires to its aerial breeding-places, and generally makes its nest on a swamp or river island. The habits of this bird are so like those of the common stone crane, that it would have escaped notice were it not for the variation in the colour of its bill. This is of a white colour, shading off towards the tips of the upper mandible, which are pale brown.So common is this species on the Atlantic seaboard, that few persons can fail to have seen it. While on board our ship at St. John’s, on the 30th of October 1828, I noticed many of these birds on a small pond that runs near our town. They were wading about and darting from one point of the shore to another, as if searching for a distant fish. They were rather shyer than the common white crane, but had the same abrupt note,so different from that of the red-necked species. They continued to hop about the pond, looking out for food, the whole time that the vessel remained there. The