
MASKED MIXERS FOR LANGUAGE GENERATION AND RETRIEVAL

TECHNICAL REPORT

 **Benjamin L. Badger***

Guidehouse

1676 International Dr, McLean, VA 22102

bbadger@guidehouse.com

ABSTRACT

Attention mechanisms that confer selective focus on a strict subset of input elements are nearly ubiquitous in language models today. We posit there to be downside to the use of attention: most information present in the input is necessarily lost. In support of this idea we observe poor input representation accuracy in transformers, but find more accurate representation in what we term masked mixers which replace self-attention with masked convolutions. Applied to TinyStories the masked mixer learns causal language tasks more efficiently than early transformer implementations and somewhat less efficiently than optimized, current implementations. The most efficient learning algorithm observed for this dataset is a transformer-masked mixer hybrid, suggesting that these models learn in an orthogonal manner. We hypothesized that the information loss exhibited by transformers would be much more detrimental to retrieval than generation, and to test this we introduce an efficient training approach for retrieval models based on existing generative model embeddings. With this method, embeddings from masked mixers are found to result in far better summary-to-story retrieval compared to embeddings from transformers.

Keywords Mixers · Transformers · Attention · Representation · Retrieval · Generation

1 Introduction

Since the introduction of the transformer [1], many deep learning application domains have witnessed a rapid increase in the state-of-the-art as a result of incorporation of this architecture. Originally designed for language processing tasks such as machine translation, the transformer has been found to be well suited to increases in model parameters such that ever-larger models may be trained on ever-larger datasets with continued increases in various metrics of goodness [2; 3; 4; 5].

In spite of this impressive success, it is less than clear how efficiently these models may be trained for various language tasks: many types of architectures experience increases in goodness with an increased parameter numbers and dataset sizes. One way to investigate this question indirectly is to observe how information passes from the input to various hidden layers of the model, which can be considered to be closely related to the question of how these hidden layers represent the model’s input. Current investigations on this topic often focus on the method by which transformers transfer information to and from different tokens (c.f. [6]), but we focus on a simpler question: *how much* rather than *how* information is passed from the input to various parts of the model. In particular, we focus on the information present in the model’s last hidden layer, as this is most commonly used for tasks such as causal modeling (before the language head is applied) and retrieval.

The goal of any attention operation is to focus on some input elements at the expense of others. Transformers used for language generation consist of many sequential self-attention modules (separated by MLPs) each with multiple attention heads per module, however, such that the restricted focus inherent in each attention transformation can no longer

*The author would like to thank Guidehouse for support during the research and writing of this paper. Code for this work may be found on <https://github.com/blbadger/maskedmixers>.

necessarily be ascribed to the model as a whole. In particular, deeper model layers no longer attend to embeddings of tokens but representations of attention of embeddings of tokens. In light of observations that trained language models tend to require high numerical precision from a tiny subset of all weights and activations which are often large in absolute value [7; 8] we hypothesized that the inherent information-limiting characteristic of attention does translate to the transformer model as a whole. Conversely, we hypothesized that a model with attention replaced via linear transformations would not exhibit these informational limitations.

In this work we observe that indeed transformer model hidden layers do not contain enough information to allow for accurate input reconstruction via gradient-based input representation algorithms, whereas the introduced ‘masked mixer’ models with self-attention substituted for masked convolutions do. We then refine the architecture of the convolutions using representation accuracy as a guide. We then compare causal language training efficiencies and autoregressive outputs of this model to that obtained for transformers, finding that the two architecture types are generally similarly efficient for causal language model training despite their very different input representation accuracies. We conclude by presenting evidence for the idea that masked mixers are more suitable than transformers for language retrieval.

1.1 Related Work

The body of work on deep learning representation is large. Much work towards understanding transformers has revolved around explanations for how attention transformations are capable of adding, removing, and transferring information from one token or patch to another (see for example [6] in the context of one-headed attention). For vision models, methods for applying gradient descent in order to observe representations in a model’s hidden layers were pioneered by such studies as [9]. Notably the method of representation employed in this work is in some ways much more straightforward than there: instead of performing smoothing and gradient orientation procedures to gain a comprehensible representation, we perform vanilla gradient descent followed by pseudoinversion to recover tokens.

The substitution of self-attention for convolutions in the transformer was introduced independently by [10] and [11] as the MLP-Mixer, a vision model designed primarily for image classification tasks. The authors of the former work noted that for a given fixed-sized image dataset the MLP-Mixer performed slightly worse than vision transformers, but do not directly compare their efficiencies for fixed compute training runs [10]. A language application of the MLP mixer based on Bloom-filtered inputs was introduced in [12] but was not designed for causal language modeling, and presents a number of architectural decisions that make this model unsuitable for autoregressive language modeling tasks that transformers are commonly trained for. To our knowledge there has not been any attempts to adapt the MLP-Mixer to causal language modeling, masked language modeling, or retrieval tasks.

The finding that information often passes linearly between tokens in simplified transformers [6] and the observation that purely linear deep models (ie with no nonlinearities between matrix multiplications) learn in very non-linear fashions even if they lack full representational power of nonlinear models [13] suggest that simple inter-token linear transformations (for example, 1-dimensional convolutions) may capture much of the dynamics of more complex inter-token transformations, such as those in self-attention modules.

2 Accurate self and non-self token representation in Masked Mixers

How much information does a model contain about its input? We use a gradient descent-based approach to address this question for language models. With respect to information that can be used to uniquely identify a language input, the answer for transformers is not very much. We introduce the masked mixer and show that this model accurately represents its inputs before and after training, and even accurately represent a limited number of non-self tokens.

2.1 Input Representation Background

In this work we measure the information present in a model’s hidden layer representation of the input by attempting to recover the input using gradient descent on an initially random input, minimizing a chosen metric on the activations of that hidden layer as described in [15]. The goal is to invert a model such that information in the hidden layer’s activations representation is used to identify the input that resulted in those activations. The development of this approach as it relates to language models is presented in a series of posts starting with [16].

Briefly, for some input a and some chosen layer l of model θ such that $O_l(a, \theta)$ is a vector space, gradient descent is performed on the norm of the difference between the hidden layer activations given some initially random input $a_0 = \mathcal{N}(a, \mu = 1/2, \sigma = 1/20)$ and its activations given a , where the only values that can be modified are the elements of the initially random input a_n . For example, this procedure using the L^1 metric is used in Equation (1).

$$a_{n+1} = a_n - \eta * \nabla_{a_n} ||O_l(a_n, \theta) - O_l(a, \theta)||_1 \quad (1)$$

We use a scheduled learning rate η that decreases linearly from η to $\eta/10$ as $n \rightarrow N$ (ie the iteration number n is increased until it reaches the final number of iterations N , which we set somewhat arbitrarily to $N = 500$) which empirically results in the fastest optimization with the fewest number of steps taken N .

Language models typically operate using discrete inputs, usually integer tokens. This means that one cannot immediately apply (1) but must instead either convert the input tokens to a differentiable vector space or else optimize some other value and convert that value to the discrete input a_N . Elsewhere it was found experimentally that the latter process tends to be more stable across a wide range of inputs [14], and that is the approach we use for this work.

To be specific, we optimize the embedding of the input rather than the input itself. The embedding may be computed by matrix-vector multiplication using the embedding weight matrix W such that $e = Wa$. With an initially random embedding $e_0 = \mathcal{N}(e, \mu = 1/2, \sigma = 1/20)$ we perform gradient descent on the embedding using an L^1 metric on the layer in question as denoted in (2).

$$e_{n+1} = e_n - \eta * \nabla_{e_n} ||O_l(e_n, \theta) - O_l(e, \theta)||_1 \quad (2)$$

In general the embedding weight matrix W is non-square and non-invertible. e_N is therefore converted back to the generated input a_N via left multiplication by the Moore-Penrose pseudo-inverse W^+ via (3), which is a form of generalized matrix inverse defined by Equation (4).

$$a_N = W^+ e_N \quad (3)$$

$$W^+ = \lim_{\alpha \rightarrow 0^+} (W^T W + \alpha I)^{-1} W^T \quad (4)$$

If positional encoding is applied to the embedding, it is subtracted before (3) is applied. This process is illustrated in Figure 1 for convenience.

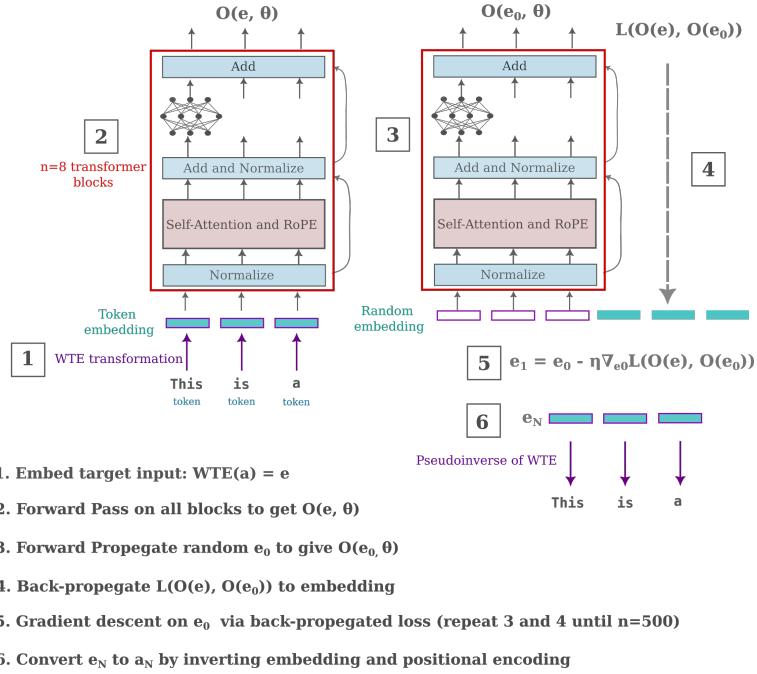


Figure 1: Indirect input representation method applied to a Llama-style transformer.

If the gradient is effectively backpropagated to the input e_n in Equation (2) then $||O_l(e_N, \theta) - O_l(e, \theta)||_1 < \epsilon$ for some small ϵ . We check this condition by adding a very small amount of Gaussian noise to the initial input embedding before observing the metric distance of the output of this noised input to a , ie $||O_l(e + N(e, \mu = 0, \sigma =$

$1/(20), \theta)) - O_l(e, \theta) \|_1 = \epsilon$. This allows us to confirm that iterations of (2) results in sufficient minimization of the metric on the output (in this case L^1) such that the corresponding a and a_N are nearly equivalent for the model. The particular amount of Gaussian noise to add in order to specify ϵ was determined in part by observing that the chosen value leads to no token changes upon pseudoinversion, meaning that typically $W^+e = W^+(e + \epsilon)$ although this depends somewhat on the transformation via the word-token embedding weights W .

2.2 Masked mixer architecture

Elsewhere it was observed that large and otherwise quite capable language models (Llama-2 7b and 70b, GPT2, etc.) exhibit relatively inaccurate input representation for all but the shallowest few hidden layers [14]. It was also observed that vision MLP-mixers ([11; 10]) have superior input representation accuracy for non-self tokens compared to vision transformers [17], leading us to wonder if a mixer adapted to causal language modeling could retain this accurate input representation. We hypothesized that adapting an MLP-mixer to the process of language generation would give us a model suitable for autoregressive language generation and other tasks, but with much more accurate input representation abilities.

We adapt the MLP-Mixer architecture for causal language modeling as follows: first all 1D convolutions (ie ‘MLPs’ on the sequence dimension) are reshaped and lower-triangular masked such that only inputs from tokens t_0, t_1, \dots, t_{n-1} have non-zero weights for the token t_n . As for causal language modeling (CLM) -style transformer models, output values are also shifted such that the loss compares $O(t_n)$ and t_{n+1} for all $n \in N$ such that the model may be trained on all tokens of an input with one forward and backward pass.

The details of the triangular masking process are as follows: first the 1D convolution weights are reshaped to place the model dimension and number of tokens in the last two tensor dimensions, a triangular mask is then applied to those weights, and finally the convolutional weight data is re-written with the masked weight data, reverted to its original shape. A succinct PyTorch [18] implementation of this convolutional weight masking (for any convolutional kernel size) using einops [19] and torch.tril lower triangular masking is as follows:

```
masked_conv = tril(rearrange(conv.weight, 'f d k -> k f d'))
conv.weight.data = rearrange(masked_conv, 'k f d -> f d k').contiguous()
```

where ‘f’ and ‘d’ are both equal to the number of tokens in the context window and ‘k’ is the 1-dimensional kernel size. A 1D convolution with a kernel size of 1 can also be implemented using ‘torch.linear’ on the sequence dimension in which the triangular weight masking is a more pithy ‘conv.weight.data = torch.tril(conv.weight)’ after reshaping, but experimentally this does not perform as well as the 1D convolutional kernel method above nor is it as flexible.

To see a diagrammatic interpretation of how these operations act to enforce causal language modeling, see Figure 2. The primary difference between these operations and the method used to CLM masking during training is that in that case a mask is usually applied to a self-attention module’s activations, whereas here the mask is applied to the weights directly.

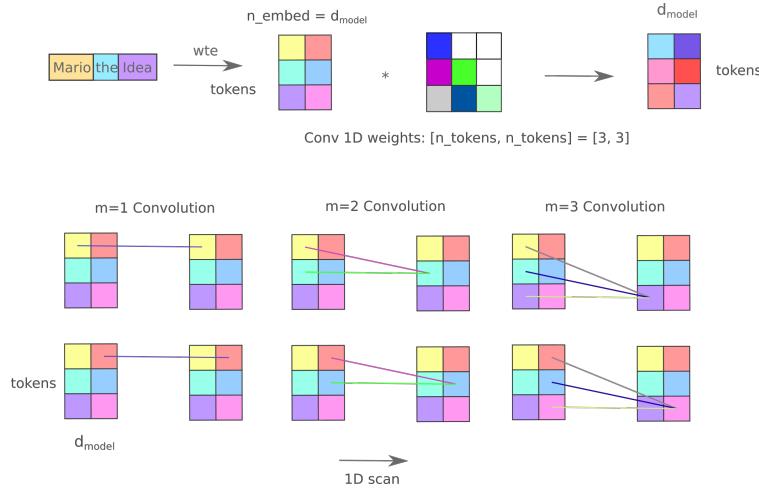


Figure 2: Causal language modeling via masking convolutional weights.

The forward pass is nearly identically to all-next-token training for transformers with some slight modification for reshaping. The use of the same loss function, tokenizer, and context window allows us to compare directly between transformer and masked mixer losses in order to assess training efficiency for these two models.

We call these models ‘Masked Mixers’ to emphasize 1) that the mask is an intrinsic part of the model for both training and inference and 2) these mixers in some cases no longer resemble MLPs, as the most performant versions use convolutions with non-unitary kernels or projected multi-headed kernels.

2.3 Masked mixers but not transformers exhibit accurate input representations

In the spirit of an information theoretic approach, we modify the normalized Hamming metric for comparing input representations to their corresponding inputs. Normalization and further modification is required because language inputs are not of fixed size, but masked mixers require fixed context windows as currently implemented.

We define this metric as follows: given input x and generated representation of that input y such that each element of x is an integer corresponding to an element in the tokenizer set $\{t_m\}$, which may be stated precisely as

$$x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n) \in \{0, 1, \dots, t_m\}^n \quad (5)$$

then the normalized Hamming metric is given in Equation (6). In words, this metric is the fraction of indices of the input where the generated representation’s token does not match the input’s token. The smaller the normalized Hamming distance, the more similar the input is to the model’s representation and the more information that representation carries. We measure before training as well as during and after 12 hours of RTX 3060 training on the first 2M examples of the TinyStories [20] dataset.

$$h(x, y) = \frac{1}{n} \text{Card}(\{x_i \neq y_i\}) : x_i \neq t_{pad} \quad (6)$$

As observed elsewhere for larger language models, small transformers designed for TinyStories modeling exhibit very poor input representation such that the information present in the last hidden layer is incapable of recovering very little of an input when (2) is applied (Figure 3), although this improves somewhat with training. Masked mixers exhibit near-perfect representation before training, and larger models retain this characteristic even after training (Figure 3).

Masked mixers also exhibit much more accurate input representation than transformers when gradient descent is performed using only the last hidden layer of the last token (Figure 3b), indicating that they pass more information between tokens than transformers. Neither transformers nor masked mixers exhibit accurate non-self token representation for larger context windows ($c = 512$ etc.) if only the last token’s last hidden layer is used for backpropagation, indicating that there are limits to the inter-token information bandwidth in mixers as well as transformers.

The original MLP-mixer architecture employs two sequential convolutional operations between each input token, with an expansion factor (usually set to two) for the layer between convolutions. We find that models with only one convolution operation between tokens (a ‘flat’ masked mixer) have superior inter-token information transfer (Figure S3), but that their self-token representation power is very similar.

Masked mixers of a sufficient d_m are in some sense biased towards accurate input representation, as untrained models nearly always exhibit perfect input representation regardless of their actual randomly initialized values. In this vein, transformers of a sufficient d_{model} are biased against accurate input representation. It is interesting therefore to observe that masked mixer and transformer input representation abilities converge somewhat during training (Figure 3d), which suggests that for the TinyStories dataset there is some range of optimal input representation accuracies that is associated with minimization of all-next-token prediction during causal language modeling. This convergence is very slow, however, especially for larger models such that the model type’s initial bias determines most behavior after training as well (Figure 3c). If we scale up the compute used for training these two models such that approximately ten times the amount of samples are shown to the model (using much larger batches) we find that indeed the gap in the Hamming metric between masked mixer and transformer is substantially reduced (Figure S4).

3 Masked Mixer and Transformer learning efficiencies

The ability of large transformer models to provide useful language modeling without accurate input representation suggests that accurate input representation is not necessary for many of the abilities of that model type. We reasoned that although input representational accuracy does not necessarily provide greater language modeling capabilities, it could perhaps lead to more efficient training.



Figure 3: Masked mixers exhibit more accurate input representation than transformers. All models are $n_l = 8$ and all transformers are llama models with $n_h = 32$. In d) the transformer is $d_m = 256$ and mixer $d_m = 512$.

To test the training efficiency of the Masked Mixer compared to the Transformer, we employ the same relatively small dataset as for representation studies: the first 2M training examples of TinyStories with all evaluation examples with a Llama-2 tokenizer of size 4096 trained on that dataset. Models are trained using a correspondingly small fixed compute budget: 12 hours on an Nvidia RTX 3060, or 2.25 hours on a 4x Nvidia V100 cluster for the training process. The Llama -style transformer models have no special architectural or hyperparameter settings, and as this model has the same loss function on the language modeling head as the Masked Mixer (cross-entropy loss, loss masks on padding tokens) with the same hyperparameters (a constant context window size of 512, AdamW optimizer [21], learning rates, cross-entropy loss function etc.) we can directly compare the loss achieved after the fixed compute has been applied.

3.1 Representation accuracy for Masked Mixers roughly correlates with training efficiency

A rough correlation is observed between representational accuracy and training efficiency for masked mixers: the original expanded-style architecture becomes more efficient as the hidden dimension increases, but the ‘flat’ masked mixers with one convolutional layer between tokens which have better self- and non-self token representation are more efficient learners still (Figure 4, Table S2, S3). Both the increased efficiency with increased d_{model} size as well as the increased efficiency of flat masked mixers relative to expanded ones correlate with increases in representation accuracy for these configurations (Figure 3). For Llama models the relationship between untrained or trained input accuracy and training efficiency is less clear (Table S1).

It should be noted that the expansion factors for expanded masked mixers are in some sense a superfluous hyperparameter when CLM maskind is applied. This is because any expansion factors not equal to one leads to rectangular inter-token convolutional weight matrices after reshaping, and triangular masking these matrices leads to unused rows or columns (see Figure S1 a graphical explanation of this phenomenon). This effectively makes expanded masked mixers with an expansion factor greater than one capable of viewing fewer samples per given time. But even when this effect is negated by considering losses at identical step number (rather than compute time) the expanded masked mixers are still slightly less efficient learners than ‘flat’ masked mixers. For expansion factors less than one, triangular causal language masking results in the loss of input token information (see Figure S2 for an example). Because of the less accurate

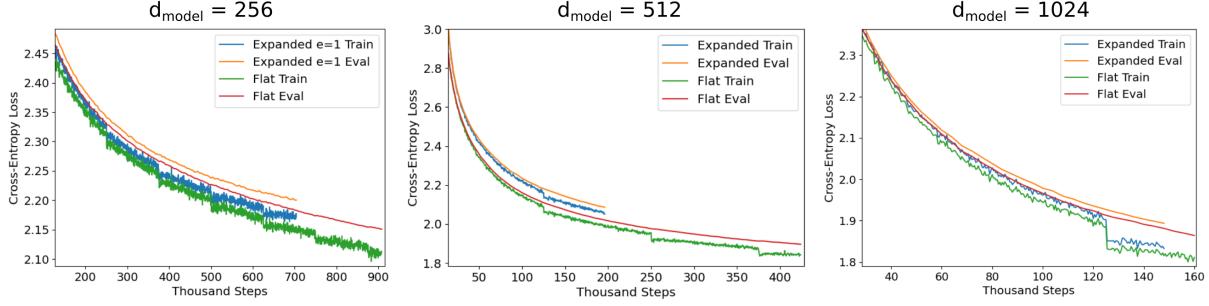


Figure 4: Flat mixers train more slightly more efficiently than expanded ones.

non-self token representation observed in the last section as well as the slightly worse training efficiency, we do not investigate expanded mixers further in this work.

3.2 Masked Mixers are more efficient learners than non-optimized modern transformers

Flat masked mixers turn out to not only be more efficient learners than expanded mixers but they are also more efficient than modern transformers (Llama-2 architecture) with default architectural hyperparameters, varying the d_{model} after previously optimizing other hyperparameters such as the number of layers per model and batch sizes (Figure (5, Table S1, Table S3). A comparison between flat masked mixer and transformer memory requirements for various context and d_m sizes may be found in Table S10 and Table S11.

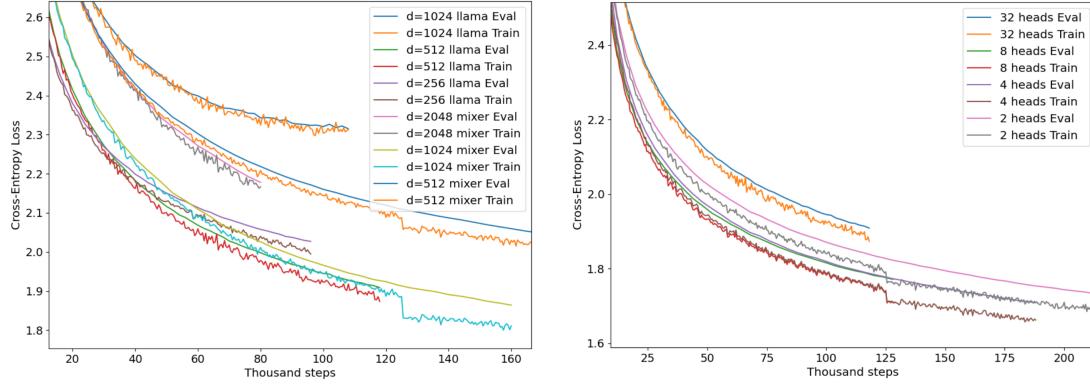


Figure 5: Left: Flat masked mixers are more efficient learners than modern transformers with a default attention head number, all with $b = 16$ batch size except for $d_m = 2048$ mixer and $d_m = 1024$ llama which have $b = 8$ to fit in memory. Right: Reducing the number of attention heads increases TinStories training efficiency substantially.

3.3 Optimized versions of modern transformers are somewhat more efficient learners than optimized masked mixers

Whilst optimizing the masked mixers and transformers in order to assess which model experiences a more efficient learning process, we focused on a number of hyperparameters: the hidden dimension d_{model} , the number of transformer or mixer block layers n_l , the AdamW learning rate, the number of attention heads and the batch size. For the Llama model we found that the most important single hyperparameter was the number of attention heads: decreasing this from 32 heads (the Llama 2 default) to 4 heads leads to a dramatic reduction in loss achieved per compute allocated, and somewhat surprisingly even on a per-step basis (Figure 5, Table S4). This fully optimized transformer learns more efficiently than d_m -optimized masked mixers using the 3060 compute node (Tables S3, S4) but less efficiently on the 4x V100 node (Tables S5, S6).

Amidst these investigations, one newly implemented optimization was found to result in a substantial boost in transformer training efficiency: the addition of Flash Attention 2 [22] led to optimized Llama models reaching training and test accuracies below those obtained for masked mixers (Table S7).

3.4 Masked mixer training efficiency with multiple heads or larger convolutional kernel sizes

Next we investigated whether increasing the number of inter-token parameters in the masked mixer would lead to more efficient training. As we had already found that flat mixers are more efficient TinyStories learners than expanded mixers, we started by testing the efficiency of masked mixers with parallel convolutions. We observed no benefits when using two parallel convolutions, however: the per-step loss is slightly lower but fewer optimizer steps are taken.

Increasing the convolutional kernel’s size leads to an increase in the number of inter-token parameters by a factor of the kernel size without increasing the depth of the inter-token transformations. A depiction of how a convolutional kernel of size 2 acts in the masked mixer is portrayed in Figure 6 for convenience. We found that masked mixers with a convolution size of 4 outperforms those with size 1 using the 3060 node (S3), but that there was little difference for the 4x V100 (Table S5).

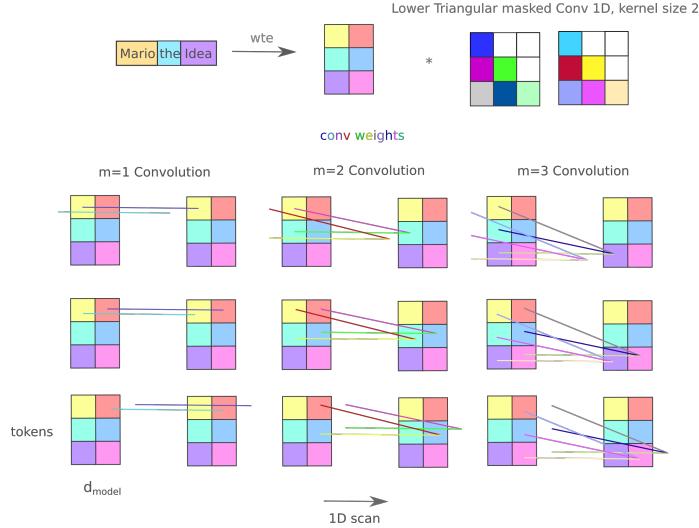


Figure 6: Convolutional kernel size 2 masked mixer with ‘same’ style padding

We also investigated whether or not the use of multiple masked convolutional heads would result in greater training efficiency. To do this, we project each input dimension into a certain number of heads and apply a 1D masked convolution on each head, before concatenating the outputs and projecting back to match the dimension of d_{model} before the feedforward layers are applied. The training efficiency of a two-headed mixer is very slightly superior to that of a single-headed mixer, and is on par with a mixer with convolutional kernel of size 4 (Table S6). Both for increases in kernel size as well as adding convolutional heads the per-step loss was lower than for the vanilla flat masked mixer, but as fewer steps were taken with fixed compute the overall losses remained similar.

It may be wondered whether or not an increase in focus in the masked mixer could lead to increased learning efficiency. One may do this by Softmax-transforming the 1D convolutional weights before re-masking. Adding the Softmax turns out to leads to much worse training efficiency than if these weights are not transformed regardless of the number of steps taken, however, indicating that increasing the ‘focus’ of the masked convolution via this transformation does not benefit training (Table S5).

3.5 Transformer-mixer hybrids learn most efficiently

Observations of the very different nature by which transformers and mixers represent their inputs suggests that these models might learn in fundamentally different but perhaps complementary ways. If the learned structures from these models were indeed complementary in some respect, combining these architectures could allow the resulting model to achieve greater training efficiency than either transformer or mixer.

We find this is indeed the case: inserting a masked 1D convolution into the output of the transformer layer and adding a residual (here termed the ‘Transfixer’) results in drops in cross-entropy loss even for optimized versions of Llama as

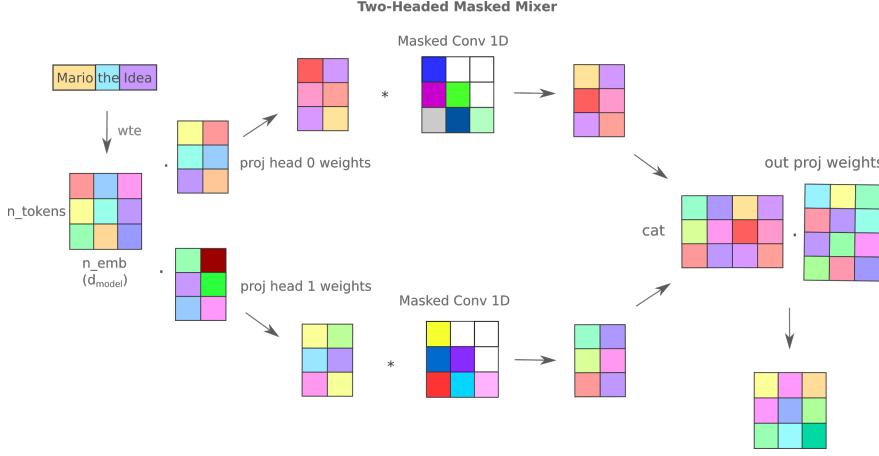


Figure 7: Multiheaded masked mixer architecture

shown in Figure 8. We observe a drop in both test and validation loss for the 3060 platform as well as the 4x V100 (the latter having Flash Attention 2 implemented) with the transformer-mixer hybrid relative to the transformer as shown in Figure 8 and Table S7.

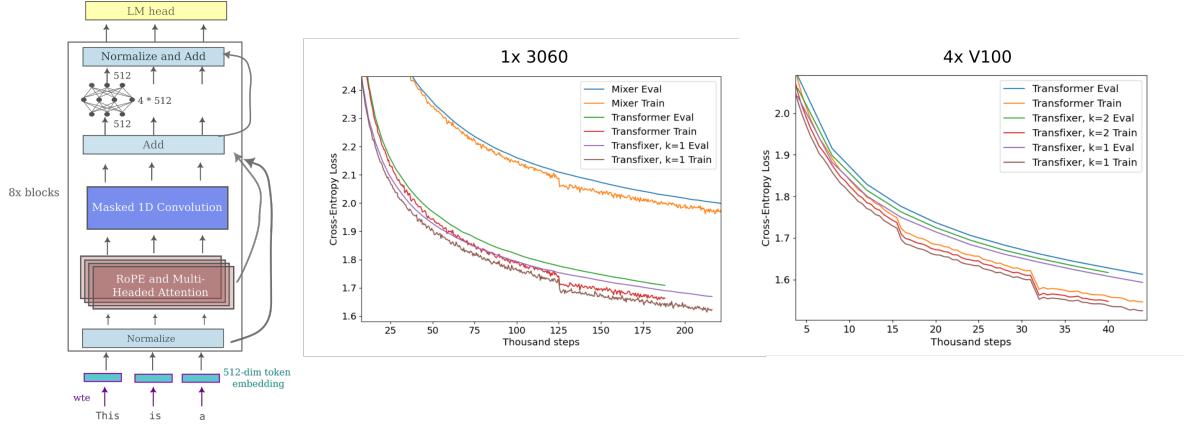


Figure 8: Transformer-mixer hybrid architecture and training efficiency

It is interesting to note that the placement of the convolutional layer in the transformer-mixer hybrid is important to this efficiency gain: if one instead places the masked convolution in parallel with the self-attention layer the resulting losses are identical to the transformer model without this convolution and if the masked convolution is placed before the self-attention layer the resulting training is somewhat less efficient than with no convolutional layer (Table S7).

3.6 Masked mixers are more efficient learners than early transformers

In some sense it is unsurprising that current transformer implementations are somewhat more efficient learners than masked mixers because transformers have seen a number of architectural improvements and compute optimizations since their introduction seven years ago. It is useful therefore to compare the learning efficiencies of early transformer implementations of all-next-token causal language modeling to masked mixers with the same training method, as these can be thought of as being similar in their developmental stage. We chose a model that was introduced before improvements such as Rotary Positional Encoding [23] and Flash Attention 2 to name a couple of notable improvements.

Specifically we tested the Hugging Face implementation of GPT [2], a CLM-style transformer model which was introduced shortly after the original transformer architecture. We find that GPT lags well behind the masked mixer's training and validation loss values both for default as well as learning rate, batch size, and head number-optimized versions of the transformer (Figure 9, Table S6). These results are consistent across compute scales, suggesting a general gap in learning efficiency between these model architectures for this dataset.

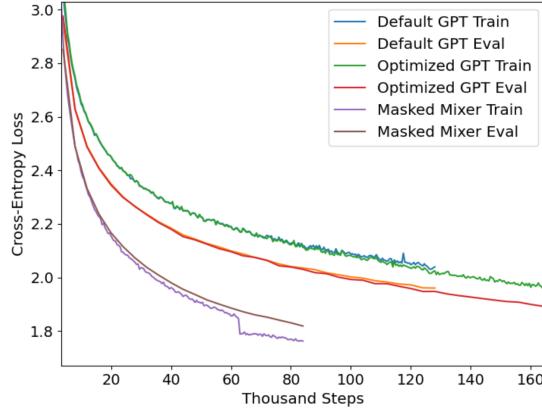


Figure 9: GPT versus mixer TinyStory training efficiency

3.7 Autoregressive inference for masked mixers

Cross-entropy loss for all-next-token training (using identical tokenizers) has allowed for a direct comparison of the abilities of masked mixers and transformers on the same dataset. The ultimate goal of causal language model training is often language generation, and now we compare the language generated by transformers and masked mixers.

Masked mixers cannot be inferred in a manner analogous to transformers because these models have a fixed feedforward context window size, at least without modification to the present implementation. This means that a full forward pass is required for each token generation such that the masked mixer inferences with time complexity $O(n^3 * d)$ where n is the number of tokens and $d = d_{model}$. In practice masked mixers inference only slightly slower than transformers that use key-value caching and are thus $O(n^2 * d)$ because of the much lower constant factors for the mixers.

This architecture requires a new inference method: rather than applying a trained model to a padded input to infer the last character, we instead apply a causal language mask the input just as is done during training and simply infer each next token at the start of the mask. Note that the causal language model mask must be retained during inference to prevent information from right-positioned tokens from influencing the next token prediction at the current token.

We introduce a simple inference method that makes use of all the information that the masked mixers learns per token position during training. Recall that the mixer has effectively a fixed positional encoding intrinsic in the model's convolutional weights, and that these weights are responsible for all of the inter-token information flow. One can make use of this positional encoding by iterating through the context indices while adding the token generated from the mixer's output at each index.

A Python implementation of this approach for is as follows:

```
for i in range(n_tokens_to_generate, 1, -1):
    loss, output = model(tokens)
    out_token = torch.topk(output, dim=1, k=1).indices.flatten()[-i]
    tokens[..., -i+1] = out_token
```

Using this inferencing method, we find that masked mixers generate similarly readable outputs to transformers at the same CLM cross-entropy loss value. For example, given the prompt

One day, a little boy named Tim went to play with his friend, Sam. They wanted to play a game with a ball. The game was to see who could get the best score. Tim threw the ball and Sam tried to catch it. Sam missed the ball and it went far away. Sam was angry. He said, "I don't like this game. I want to lie down and rest." Tim said, "Okay, let's lie down and rest. Then we can try again to get a good score." They lay down under a big tree and looked at the sky. They saw birds flying and clouds moving. After a while, they got up and

a 1024-dim flat masked mixer with validation cross-entropy loss of 1.81 yields a fairly coherent story completion:

played a game of catch. Tim threw the ball to Sam, and Sam caught it. They laughed and played until the sun went down. At the end of the day, Tim and Sam were tired but happy. They went home and took a nap. They dreamed of playing catch again tomorrow. And they did.

whereas a 256-dim transformer with a validation loss of 1.91 loss yields the the following completion,

started to play again. They played for a long time and had lots of fun. They laughed and played until the sun went down. Then, they went home and told their mom and dad about their game. Tim and Sam learned that it is important to be patient and wait for good things to happen. They had a fun day playing together.

with somewhat more deviation from the prompt's plot.

4 Masked mixers are more effective for retrieval than transformers

Along with generation one of the most frequently encountered language tasks today is retrieval, a task which usually involves finding one or more matches for a given language segment given a set of text embeddings. One of the most common uses for this is in retrieval-augmented generative search, whereby a search phrase is ‘read’ by a language model and the resulting embedding from that phrase is then used to match one or more segments of a corpus of text, which is in turn fed to a generative model along with the original search phrase [24].

As is the case for language generation, nearly every retrieval model used today is based on the transformer architecture. The results in this work lead to the hypothesis that attention is not well-suited for the task of retrieval because these transformations are biased towards non-invertibility, such that a mapping with this transformation becomes many-to-one (Figure 3). For predicting a single next token it appears that this mapping is not detrimental and may even be beneficial, as masked mixers learn to reduce the information passing to deep layers during causal language model training.

On the other hand, matching a sequences of tokens to another sequence of tokens may be thought of as an approximately bijective function which would not be expected to benefit from information reduction or non-invertibility. Masked mixers have been shown to be effectively invertible for small inputs and in that respect retain much more input information than mixers, suggesting that these models are better suited to retrieval than models with attention assuming that retrieval requires most input information. In this section we test this idea experimentally, and find that indeed last-hidden-layer last-token embeddings from masked mixers are far superior for retrieval tasks compared to the same embeddings from transformers.

4.1 Synthetic dataset generation and embedding

Testing retrieval ability using our existing models entails using the dataset these models were trained on, especially as the tokenizers used for these models will generally be unsuitable for other text corpora. To that end we generated a synthetic summary sentence for the first 200k stories in the TinyStories dataset using Llama-3 (8b) Instruct [5] by prompting that model to read each story and give a very short summary of that story. Due to the relatively simple nature of these stories, we found that this relatively small model was more than capable of accurately and concisely summarizing each story.

Each story and its corresponding summary were fed to either a Llama-style transformer or a masked mixer, and the last hidden layer activations were saved as the model’s embedding of its input. The embedding from the transformer’s last token was used, whereas the second-to-last token embedding was used for the masked mixer. This is because the ‘last’ masked mixer token is untrained due to the index shift that occurs during all-next-token training of these models. An example of a TinyStory and its one-sentence summary is found below:

Text: Once upon a time, there was a smelly old tree. By the tree, there was a big hole. In the hole, there was a shiny coin. A boy named Tim saw the coin and wanted it. Tim said to his friend, “I want that coin, but the tree is smelly and rot.” His friend said, “Be brave, Tim! You can get the coin.” So, Tim went to the smelly tree and tried to get the coin from the hole.\n\nAs Tim got closer to the hole, he saw a little mouse. The mouse said, “I will help you get the coin, but you must help me too.” Tim agreed, and they got the coin together. Tim was happy, and the mouse was happy too. And they both learned that even in smelly places, good things can happen.

Summary: A boy named Tim overcomes his fear of a smelly old tree to retrieve a shiny coin with the help of a friendly mouse.

4.2 Retrieval model architecture and training approach

A standard method for transformer model retrieval is to use the cosine distance on the output, defined as follows:

$$d(x, y) = 1 - \cos(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \quad (7)$$

which is usually implemented by first normalizing x, y such that $\|x\| = \|y\| = 1$ and then performing matrix multiplication on matrices X, Y of concatenated vectors of x, y . Here the

It has been found that cosine distance on an output is in some ways well-aligned with the dot product self-attention transformations (c.f. [14]) and thus this metric would not be expected to lead to as accurate retrieval for attentionless models like mixers. Indeed this is found to be the case, and preliminary experimentation revealed that L^1 and L^2 metrics are also unsuitable such that we instead decided to directly train for retrieval using fixed embeddings from previously-trained masked mixers and transformers.

Training models for the purpose of retrieval is not a new idea, with most state-of-the-art retrieval models receiving training to minimize the cosine distance for matching pairs of sequences while maximizing the distance for non-matching pairs. This training typically proceeds by modifying the parameters of the embedding model directly [25] which in some sense is very inefficient because one forward pass is required for each input for each batch of matching and non-matching sequences.

One can instead train far more efficiently by effectively compressing each input using a trained generative model, and then training a separate model on these embeddings such that a single forward pass is required for comparison of all inputs in each batch of embeddings of size c . This method has the added bonus of allowing one to examining the effectiveness of training one type of model on the inputs from various architectures without having to account for differences in these architectures.

For our retrieval model we chose to implement a mixer with bidirectional convolutions rather than masked ones, similar to MLP mixers but with only one layer rather than two between sequence elements. This was chosen because a comparison of all possible matches requires information from all input elements by definition, making transformer models unsuitable. This was experimentally verified, as transformers were found to be very poor retrieval models even for small comparison contexts (Table S8). Word -token embedding layers are removed (as the inputs are embeddings already) as are language heads, and instead each last hidden layer has a $d_{model} \rightarrow 1$ transformation where the unitary output corresponds to the likelihood (after Softmax transformation) of that input embedding being the correct match. Softmax transformation serves to stabilize the retrieval model’s loss, as increasing all logit values does not decrease the cross-entropy loss. We make use of the Pytorch Cross-Entropy loss function’s compatibility with class probabilities for this model.

Arbitrarily choosing the first ‘token’ to be the summary and all other tokens to be the potential matches, each input is then assembled by randomly sampling all story embeddings before replacing one input (at a random location) with the embedding of the matching story. A non-optimized version of this approach for generating inputs for the retrieval model is found in Algorithm 1.

The outputs after Softmax transformation are compared to the true distribution via standard cross-entropy loss, which is then back-propagated through the retrieval model but not further to the generative model or input embedding. Note that the logits are not Softmax-transformed during inference for speed, as this operation has no effect on topk choice without noise introduction. We use a relatively modest batch size compared to most retrieval model training approaches (512 total). See Figure 10 for a representation of the retrieval model.

4.3 Embeddings from masked mixers result in much more accurate retrieval compared to embeddings from transformers

In contrast to the rest of this work, the efficiency of an embedding for retrieval modeling is not measured in loss per compute amount or even loss per number of samples seen but instead minimum evaluation cross-entropy loss over many (200) epochs of training. Retrieval model training was found to have an unusual and stereotypically Sigmoid loss over time. In the worst case the retrieval model fails to break symmetry and predicts one identical output for all embedding inputs, and this corresponds to little decrease in loss and may coincide with gradient explosion and overflow after the first hundred or so epochs of training. More commonly, the retrieval model will train to near-perfect accuracy on training data and the generalization to the evaluation dataset may be examined (Figure 10).

Embeddings from masked mixers after causal language model training lead to far lower evaluation set cross-entropy loss minimization than embeddings from transformers with the same training approach (Figure 10, Table S9, and Table

Algorithm 1 Retrieval training dataset sampling

Require: $x = O_l(\text{summaries}, \theta_g)$ ▷ Summary embeddings
Require: $y = O_l(\text{stories}, \theta_g)$ ▷ Story embeddings
Require: $c \leftarrow \text{len}(\text{context})$ ▷ Number of samples trained per match
Require: vector a : $\text{len}(a) = c$ ▷ Vector of embeddings for model input
Require: vector q : $\text{len}(q) = c$, $q[i] = 0$ ▷ Label vector

- 1: $N \leftarrow |x| = |y|$
- 2: $n \leftarrow i : 0 <= i < N$
- 3: **while** $n \leftarrow 0, n++, n < N$ **do** ▷ Loop over summary/story pairs
- 4: $a[0] \leftarrow x[n]$ ▷ First input element is the summary to match
- 5: weights: $\text{weights}[0, 1, \dots, N] \leftarrow 1$
- 6: $\text{weights}[n] \leftarrow 0$
- 7: $r = \text{multinomial}(\text{weights}, c-1)$ ▷ Random sample of all input indices except matching
- 8: $a[1:c] \leftarrow y[r]$ ▷ Replace elements of a with lookup of embeddings
- 9: $m = \text{randint}(1, c)$ ▷ Random index to place matching story embedding
- 10: $a[m] \leftarrow y[n]$ ▷ Matching story embedding placed
- 11: $q[m] \leftarrow n$ ▷ One-hot label at position m
- 12: yield a, q
- 13: **end while**

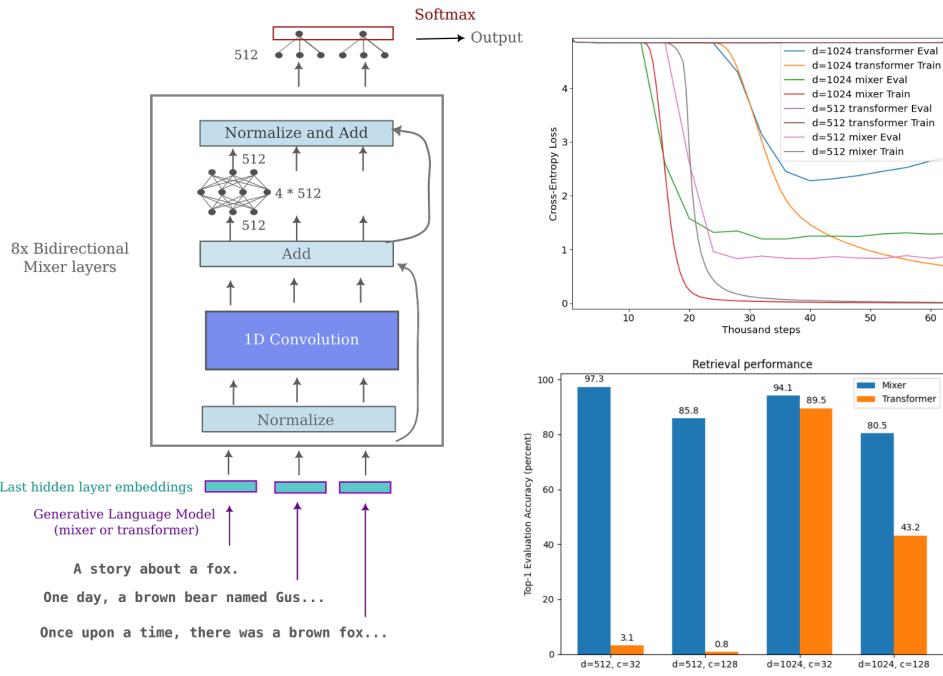


Figure 10: Left: Retrieval model architecture. Upper Right: Learning curves for the retrieval model with embeddings from CLM-trained masked mixers or transformers. Lower Right: Top-1 accuracy for evaluation samples (20k)

S10). This corresponds to far higher top-1 retrieval performance on the evaluation dataset 10. In particular, the best transformer embedding observed for 128-context retrieval has a cross-entropy loss of 2.28 and top-1 accuracy of 43.2%, compared to the best masked mixer embedding with a loss of 0.87 and top-1 accuracy of 85.8%.

Larger datasets predictably lead to increased accuracy for masked mixer embeddings: increasing the training dataset size by a factor of around 3 leads to a drop in 128-context comparison cross-entropy evaluation loss from 0.23 to 0.10 in the same validation dataset (for the 512-dimensional masked mixer CLM embeddings sampled with replacement). This corresponds to a top-1 accuracy of 97.0% for that context window size, an increase from 85.8% achieved with 200k samples.

4.4 Untrained or non-CLM trained models embeddings yield poor retrieval

We tested the ability of embeddings from untrained masked mixers to be used for effective retrieval modeling, as would be predicted by the more accurate input representation of untrained versus trained masked mixers (Figure 3). If it were indeed the case that one could forego causal language model training for the embedding model a great deal of compute could be saved particularly for larger datasets or instances where custom tokenizers are required. A masked mixer was initialized and then used produced a set of story and summary embeddings without the CLM-style training that normally precedes embedding generation. The standard bidirectional mixer retrieval model used previously trained very poorly when using these embeddings for matching batches of context size 128 or even 32, with practically no decrease in loss observed (Table S9). This suggests that training endows the embedding model with the ability to capture important parts of the language input, and that accurate input representation does not have a simple relationship to retrieval ability.

We also investigated whether the type of training is important to this retrieval process, and in particular whether or not all-next-token style causal language model training is necessary for effective retrieval. This can be done by swapping out the masked mixer generative model for one that learns to represent an input via a non-trivial autoencoding process. We implement this autoencoder using a encoder-decoder model where a standard masked mixer acts as the encoder, the last hidden state of the last token is taken as the embedding, and a decoder composed of a stack of masked mixer modules transforms this embedding (repeated for each input word) into the input sequence as shown in 11.

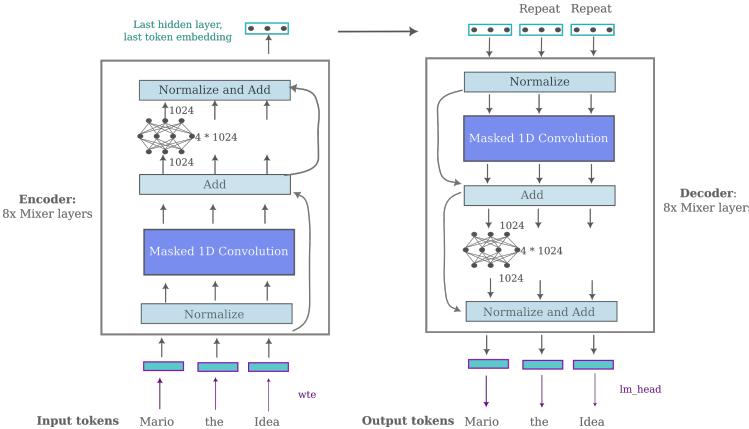


Figure 11: Bidirectional Mixer Autoencoder architecture for $d_m = 1024$

This mixer autoencoder reaches a cross-entropy loss on TinyStories comparable to non-optimized transformers and mixers (albeit with around twice the number of training steps required) and very little or no overfitting, indicating that the embedding created by the model’s encoder provides generalizable information on the input to the decoder, and in particular enough information for the decoder to reconstruct the entire input sequence approximately as accurately as trained all-next-token causal language models. It comes as some surprise, therefore, that the embeddings from this autoencoder are no better than embeddings from an untrained mixer when applied to retrieval training, even for the relatively small n=32 context sizes (Table S9).

It should be noted that this retrieval model approach is also efficient for inference as well as training: with batching, a retrieval model with a ‘context’ window of 128 samples with a batch size of 128 requires less than 6 GB vRAM, but is capable of matching a story to $128^2 = 16384$ samples in a two forward passes.

5 Limitations of this work

This study has been limited in scale: experiments were performed using a relatively small dataset and compute quantity. Although there are reasons to suspect that the results here would translate to much larger datasets and compute such as better scaling properties to larger context windows during training for masked mixers compared to transformers, this has yet to be explored. Another limitation inherent in this work is that compute efficiency is measured in a concrete hardware-specific fashion. We chose this route as it is immediately applicable to the primary goal of language model training (ie getting the lowest possible loss in the least amount of time), and one can recover the FLOPS per experiment using available data on each compute node.

A downside to our hardware-specific approach is that our GPUs (V100s and RTX 3060) are not typically used for language model training in industrial settings today, where instead H100 and A100s are favored. Newer GPUs are

specially tailored for transformer inference and training and thus our comparisons between transformer and mixer architectures would likely be somewhat different if performed on these newer GPUs. H100s in particular have dedicated transformer engines [26], which would be expected to provide speedups for that model relative to others.

It is also unclear how masked mixers would perform if given datasets with much more varied samples. TinyStories has a very large number of very similar data samples, which makes learning language generation relatively easy compared to more natural datasets used to train much larger models. On the other hand, the similarities in the inputs results in much more difficult for models to learn to match story summaries to stories: for a batch of 128 stories, there are typically dozens about a main character named ‘Tim’ who ends up having very similar exploits.

6 Discussion

6.1 Input representation Accuracy and model efficiency

This work has explored the use of input representation accuracy as a guide to improving deep learning model architectures for language generation and retrieval. Despite this perspective allowing for the improvement of model efficiency for both tasks, it has also become evident that there is no simple relationship between representation accuracy and either generation or retrieval or else very small masked mixers would outperform transformers, and untrained mixers would for effective embeddings for retrieval. Instead it is evident that other considerations (for example, the set of possible transformations a model is capable) are important in varying degrees.

Accurate input representation is evidently useful for creating language embeddings, but one might wonder if this property would also contribute to a model’s ability to memorize the inputs it was trained on. In other words, if a model is capable of retaining all the information in its input then is it prone to memorization? In this work we found that indeed the gap between training and test loss is slightly higher for masked mixers compared to transformers and in that sense the answer is yes. But we have yet to see any examples of severe overfitting of a language dataset of significant size for any masked mixer or transformer, which provides support for the ideas that memorization as it is currently observed is often a property of the dataset itself rather than a model’s architecture and that even very large models capable of memorizing large datasets are intrinsically biased against doing so when training using gradient descent due to a few characteristics of high-dimensional space [27].

6.2 Recurrent, semi-recurrent, and feedforward architectures

Transformers blend features of recurrent neural networks and pure feedforward models: during training they require only one forward and backward pass per input for training on all tokens of that input (like feedforward models) but during inference they may be used for language generation of an indeterminate number of tokens (assuming that the positional encoding method used is sufficiently flexible). The main downside is that these models are effectively $O(n^2d)$ in both time and space complexity during training because for each next token for an input of n tokens, the key and value projections must be re-computed. Masked mixers as implemented are pure feedforward models in the sense that these models have a fixed output size for both training and inference.

This lack of recurrent-like character for the masked mixer results in slightly longer evaluation or text generation times than for transformers of equivalent size. Language retrieval text corpora are typically chunked such that every input is approximately the same size, which means that transformers and mixers with the same number of operations would have very similar runtimes for the task of language embedding. That said, it is not difficult to imagine ways to reduce the memory and space requirements for masked mixers during inference: this may be done via lazy loading of convolutional weights only when they are required, or else by caching previously computed values.

6.3 Is attention useful for efficient language modeling?

The MLP mixer was originally introduced to test whether or not attention is all you ‘need’, and the answer was found to be no: in contrast, swapping feedforward MLP layers with attention in a vision transformer led to near-complete failures during training indicating that self-attention is not as important at least as the feedforward layers [11]. In this work we have found that attention is not necessary for training an effective generative language model as well, but primarily focus on the efficiency of the training process rather than its asymptotic characteristics.

Creating an efficient learning process is in some ways the fundamental goal nearly every deep learning architecture: as a one-hidden-layer feedforward network is universal for all computable functions [28], if efficiency was not a concern then that is the only model type that would be needed, provided it could be trained efficiently. There are indications that small feedforward networks applied to concatenated input elements (concatenated one-hot tokens) are similarly capable to small transformers for language-based sequence classification tasks, suggesting that a sufficiently large pure

feedforward model would effectively learn to generate natural language [29]. These models would be inefficient for longer sequences with bigger vocabularies, however, as they scale with $O(d^2tn)$ in parameter number with the length of the input n and token size t and hidden dimension d .

In this work we have seen evidence both for and against the idea that attention mechanisms are important for learning efficiency: the best attention-only models are slightly more efficient learners than convolution-only models, but on the other hand these attention models have been extensively optimized relative to the masked mixers introduced here. Given a similar effort to optimize the convolutional operations in the mixer, it is possible that this architecture would be more efficient than the most recent transformers.

From the results presented in this paper it would appear that properly optimized attention mechanisms do confer substantial benefits to the process of causal language model training, at least until more efficient training can be obtained using attentionless models such as masked mixers. On the other hand the evidence suggests that attention is generally not desirable for efficient language retrieval tasks, either in the generative model creating an embedding of each input or in a retrieval model trained to match embeddings.

6.4 How many parameters are necessary for language modeling?

The TinyStories dataset was introduced with the goal of testing how large a transformer model and in particular what hidden layer size is required in order to accurately model a limited subset of the English language [20]. In this work the most efficient transformer alternatives have generally benefited from a larger hidden layers than transformers, which leads to the question: model architecture aside, how many parameters are required for language modeling?

To start to answer this question, we can consider the language task of sentence completion. Suppose there were a huge number of valid English sentences, perhaps $m = 10^{570}$ as an upper bound. Without knowing how to model these sentences, we can view them as unique points in an arbitrarily high-dimension space and apply a result from the concentration of measure phenomenon to determine the number of dimensions required to accurately. The Johnson-Lindenstrauss lemma provides us with the result that the same m points may be represented with arbitrary precision in a space that is on the order of $8 \log m = 8 \ln 10^{570} \approx 1312$ dimensional. More precisely, this lemma states that for some small $\epsilon > 0$ for set X of m points in \mathbb{R}^N , for when

$$n > 8 \ln(m) / \epsilon^2 \quad (8)$$

there is a linear representation map $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$ such that for all $u, v \in X$ the following inequality holds:

$$(1 - \epsilon) \|u - v\|^2 \leq \|f(u) - f(v)\|_2 \leq (1 + \epsilon) \|u - v\|_2 \quad (9)$$

In words, (9) states that there exist a linear map f with dimensionality n that approximates X of dimension N arbitrarily well according to the L^2 distance between any two points in both sets, and the lower bound of the smaller dimension n is governed by (8).

For more concrete estimates based on datasets used today, consider that language models today are nearly universally trained to predict every ‘next’ token in a concatenated sequence of tokens. This means that for a training dataset of n tokens, the task is approximate n points in some presumably very high-dimensional space. The largest dataset any open-source model has been trained on is 15 trillion tokens (for Llama-3 [5]), and in this case we can represent this dataset arbitrarily well in a space of approximately $8 \ln(15 \times 10^{12}) \approx 8 * 30 = 240$ dimensional. The goal of language models is to generalize to larger datasets and thus we would hope the model would accurately predict the next tokens of a much larger dataset. But even supposing that this 15 trillion token dataset is only one millionth of the size of this generalized dataset, one would still only require a space of $8 \ln(15 \times 10^{18}) \approx 8 * 44 = 352$ dimensions.

How does this relate to the number of parameters necessary in a model? Assuming that the training algorithm is sufficiently powerful (which is not at all a safe assumption, more on this later) the number of parameters in a model could correspond to one of two things: either it is equivalent to the dimensionality of the model with respect to the points it can approximate, or else the model’s ‘width’ or hidden layer dimension is equivalent to the model’s dimensionality. Because the vector space of successive layers of practically every modern deep learning model are dependent (activations in layer $n + 1$ depend on layer n and perhaps layer $n - 1$ etc.) and dimensionality is defined on linear independence, it seems more likely that a model’s dimensionality best corresponds with its hidden layer dimension. If this is true then a model of no more than around 1300 hidden layer elements should be capable of completing any English language sentence, or a model with a width of 350 can accurately predict a next token for a massive dataset. If these hidden widths were used for default Llama architectures, the resulting models would be around 100 million (given a hidden width of $d_m = 350$) and 995 million (if $d_m = 1300$) parameters.

It cannot necessarily be assumed that a model with that number of parameters is actually trainable, however: it could be that training requires a large model that must then be converted into a small model. This is the approach used when performing pruning, where parameters are dropped depending on their importance for some output. Alternatively, instead of removing parameters one could reduce the memory required to store each parameters: this is the approach of quantization methods, which are perhaps the most effective methods currently available for shrinking the effective size of a model. The observation that weight quantization rather than pruning is the most effective method for reducing a transformer model’s effective size suggests that this particular architecture may indeed require nearly all the trained parameters in order to function effectively, although whether this is the case or not remains an open question.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf.
- [2] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [3] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.
- [4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [6] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- [7] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.
- [8] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization, 2022. URL <https://arxiv.org/abs/2110.02861>.
- [9] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.
- [10] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.
- [11] Luke Melas-Kyriazi. Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. *arXiv preprint arXiv:2105.02723*, 2021.
- [12] Francesco Fusco, Damian Pascual, Peter Staar, and Diego Antognini. pnlp-mixer: An efficient all-mlp architecture for language. *arXiv preprint arXiv:2202.04350*, 2022.
- [13] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, May 2019. ISSN 1091-6490. doi:10.1073/pnas.1820226116. URL <http://dx.doi.org/10.1073/pnas.1820226116>.
- [14] Benjamin L. Badger. Sentence representation with language models, 2023. URL <https://blbadger.github.io/language-representations-inputs.html>.
- [15] Benjamin L. Badger. Depth and representation in vision models, 2023.
- [16] Benjamin L. Badger. Language modeling and discrete encodings, 2023. URL <https://blbadger.github.io/language-discreteness.html>.
- [17] Benjamin L. Badger. Transformer and mixer features, 2023. URL <https://blbadger.github.io/transformer-features.html#mlp-mixer-feature-map>.

- [18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
- [19] Alex Rogozhnikov. Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=oapKSVM2bcj>.
- [20] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL <https://arxiv.org/abs/2305.07759>.
- [21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- [22] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.
- [23] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.
- [24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [25] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models, 2024. URL <https://arxiv.org/abs/2401.00368>.
- [26] Nvidia. Nvidia h100 tensor core gpu architecture, 2024. URL <https://resources.nvidia.com/en-us-tensor-core>.
- [27] Benjamin L. Badger. Why deep learning generalizes, 2022. URL <https://arxiv.org/abs/2211.09639>.
- [28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [29] Benjamin L. Badger. Small language models for tabular data, 2022. URL <https://arxiv.org/abs/2211.02941>.
- [30] Benjamin L. Badger. Information between tokens, 2024. URL <https://blbadger.github.io/smaller-lms.html>.
- [31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [32] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>, 2022.

7 Appendix

All code for this paper is available on Github (<https://github.com/blbadger/maskedmixers>). The motivations and background for this work are explained in more detail in a series of posts, see for example [30].

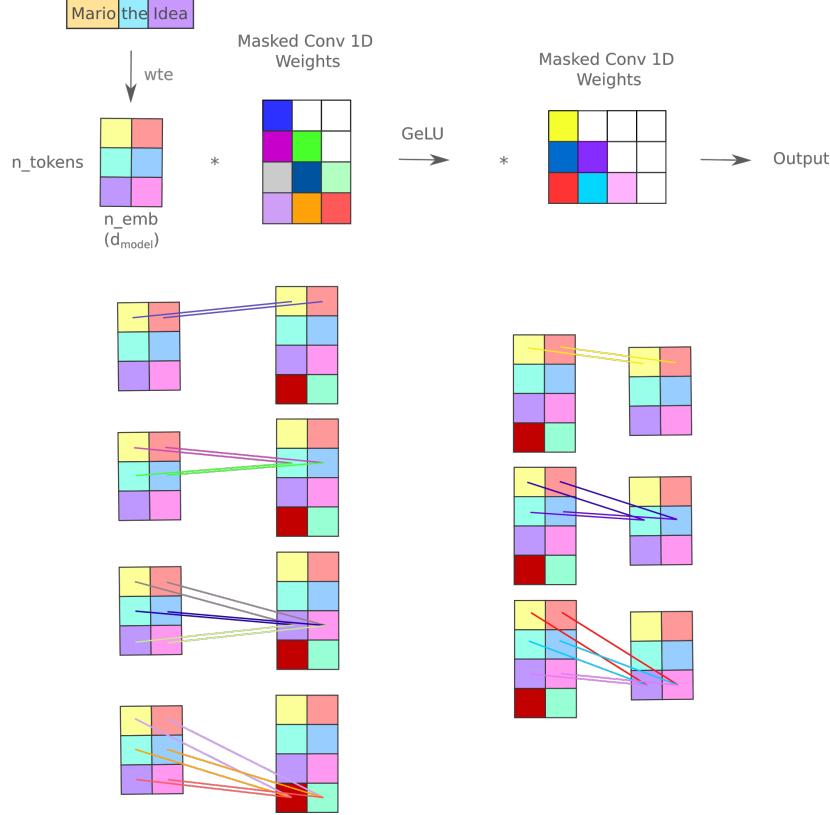


Figure S1: Mixers with expansion factors greater than one lead to unused parameters.

7.1 Masked Mixers with expansion factors $\neq 1$ lead to unused parameters or lost information

7.2 Compute

In many cases it is a trivial task to introduce a new architecture with better scores on a benchmark if one has access to variable amounts of compute: simply increasing the compute used for training compared to the SOTA, and ensure that the architecture has sufficient capacity to make use of an increase in compute. Because of this, we enforce constant compute in terms of clock time on fixed hardware (and software) on our model comparisons.

There are two units of compute used for CLM training in this paper: one is 12 hours of Nvidia RTX 3060 compute (with an i7-12700F CPU) and the other is 2.25 hours with 4x Nvidia V100s (a Gigabyte T180-G20 with 2x 2680 E5 V4 CPUs). Training on the 4x V100 cluster occurs via DDP such that first uses much smaller effective batch sizes compared to the latter, which provides a very preliminary estimate of the scaling properties of each architecture. It should be noted that the 4x V100 cluster has been limited on power consumption to reduce heat and cost considerations, and runs at 200W per GPU with 877 memory and 1005 application clock speeds. Empirically this decreases the cluster performance by a factor of around 15 percent.

Retrieval model training was not compute-limited, and was performed exclusively with the 4x V100 cluster using a 4x128=512 batch size, also using DDP.

7.3 Software infrastructure

The implementations of multi-headed attention have increased substantially in efficiency since the introduction of the transformer. Even in the duration of this particular study, new features such as Flash Attention 2 have been introduced.

Unless otherwise noted, the training runs in this study use either Hugging Face Transformers versions 4.36.0 (3060) or 4.41.1 (4x V100). All direct comparisons in this work are performed using identical versions of all major libraries (Pytorch, Transformers, Accelerate etc). For training runs where Flash Attention 2 is noted as included, we use Transformers version 4.42.3 for both nodes (FA2 was implemented in 4.42.0).

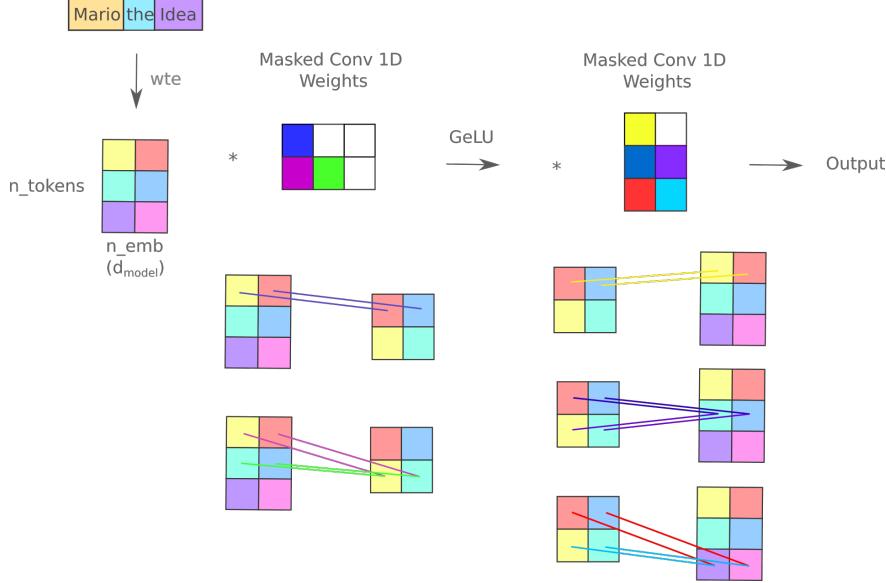


Figure S2: Mixers with expansion factors less than one lead to loss of information.

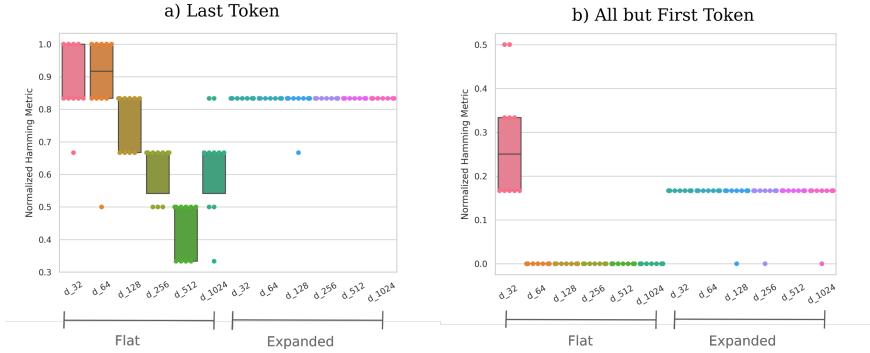


Figure S3: Flat masked mixers are more effective at passing information between tokens than expanded ($e=2$) masked mixers.

For both 3060 and 4x V100 nodes we use Pytorch 2.0 (2.3.1).

7.4 3060 compute fixed losses

The following tables detail minimum losses achieved after 12 hours of compute with a single Nvidia RTX 3060 (12GB). All models were trained using the ‘transformers.trainer’ utility [31] with evaluations every 4k steps and training statistics recorded every 500 steps.

	$d_{model} = 128$	$d_{model} = 256$	$d_{model} = 512$	$d_{model} = 1024, b=8$
Train	2.38	1.99	1.87	2.31
Eval	2.40	2.02	1.91	2.32

Table S1: Llama model training loss (n=8, h=32)

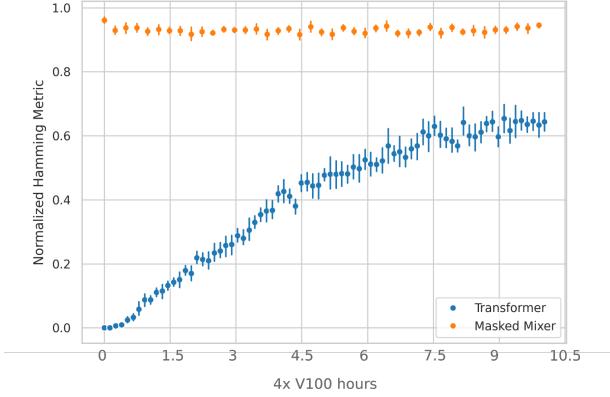


Figure S4: Normalized Hamming metric during extended training (10h on 4x V100s) for 256-dim Llama and 512-dim masked mixer models.

	$d_{model} = 256, e = 1$	$d_{model} = 512$	$d_{model} = 1024$
Train	2.17	2.05	1.83
Eval	2.20	2.08	1.89

Table S2: Expanded Mixer losses (e=2 unless otherwise noted) on 12h 3060

7.5 4x V100 fixed compute losses

The following details the fixed losses achieved after 2.25 hours on a 4x V100 (16GB) Gigabyte T180-G20 server node detailed above. The ‘transformers.trainer’ utility of the transformers library [31] was again used with 4k steps between evaluations and 500 steps between training statistic recordings. Distributed Data parallel training was performed using the Hugging Face accelerate ([32]) integrations in the trainer utility, which wraps the Pytorch-native DDP utilities ([18]).

7.6 Retrieval model losses

All retrieval models were trained on the 4x V100 cluster ($b = 128$ for each GPU for a total of $b = 512$).

7.7 Memory scaling properties

The efficiency of the masked mixer relative to a transformer of the same number of trainable parameters is apparent as follows: if

is compared to that for a transformer of the same width, we see that the masked mixer is between four and eight times as memory-efficient with increasing token length.

Note that the above tables do not capture the memory required for training with optimizers that must save multiple gradient values per trainable parameter, or for batched inputs multiple activation values. The smaller constant factors for memory complexity in masked mixers compared to transformers lead to even larger efficiency gains once these are accounted for.

	$d_{model} = 256$	$d_{model} = 512$	$d_{model} = 1024$	$d_{model} = 2048$	$d_{model} = 1024, k = 4$
Train	2.11	1.84	1.81	2.05	1.76
Eval	2.15	1.89	1.86	2.07	1.82

Table S3: Flat masked mixer losses on 12h 3060.

	$n_{heads} = 32$	$n_{heads} = 8$	$n_{heads} = 4$	$n_{heads} = 2$
Train	1.87	1.70	1.66	1.68
Eval	1.91	1.77	1.71	1.73

Table S4: Transformer Heads and loss for 12h 3060.

	1 head	2 heads	2 heads (softmax)	1 head (k=4)	1 head (k=4) and 2 wtes
Train	1.63	1.60	2.13	1.62	1.61
Eval	1.72	1.72	2.15	1.74	1.71

Table S5: Optimized Masked Mixer cross-entropy loss with 4x V100

	$b = 16, \eta = 0.02$	$b = 16$	$b = 32$	GPT, $b = 32$
Train	1.78	1.76	1.68	1.82
Eval	1.82	1.79	1.73	1.77

Table S6: 4x V100 Optimized transformer model loss (no Flash attention 2, Llama style $d_{model} = 512$, $n_l = 8$, $\eta = 0.005$, $n_h = 4$ unless otherwise noted)

	llama	attn -> conv (c2)	attn -> conv	attn + conv	conv -> attn
Train	1.55	1.55	1.53	1.55	1.68
Eval	1.61	1.62	1.59	1.61	1.75

Table S7: 4x V100 Transformer-mixer hybrids (all with Flash Attention 2 and kernel size 1 unless otherwise noted)

	2 Trans 32	Mixer 32	Mixer 128	Trans 32	Trans 128	Automixer 32	Automixer 128
Train	3.43	0.01	0.01	0.01	0.61	3.43	4.85
Eval	3.43	0.28	1.20	0.40	2.28	3.43	4.85

Table S8: Embedding model $d_{model} = 1024$ 200-epoch minimum losses ('2 Trans' denotes both embedding and retrieval models are transformers, 'Automixer' denotes the autoencoder embedding used for retrieval training)

	Mixer 128	Mixer 32	Trans 128	Trans 32	UMixer 32	UMixer 128
Train	0.01	0.01	4.85	3.42	0.62	4.85
Eval	0.87	0.10	4.85	3.42	3.36	4.85

Table S9: 200-epoch Embedding model $d_{model} = 512$ (UMixer denotes untrained mixer)

	$n_c=512$	$n_c=1024$	$n_c=2048$	$n_c=4096$	$n_c=8192$
$n_l=4$	2071	2341	2637	3573	6491
$n_l=8$	2431	2869	3425	5111	10527
$n_l=16$	2695	3159	3811	5879	OOM

Table S10: Flat Masked Mixer memory requirements ($b = 16$)

	$n_c=512$	$n_c=1024$	$n_c=2048$	$n_c=4096$	$n_c=8192$
$n_l=4$	2323	3275	6809	OOM	OOM
$n_l=8$	3176	4800	10126	OOM	OOM
$n_l=16$	4876	7750	OOM	OOM	OOM

Table S11: Transformer memory requirements ($n = 8$ layers, $h = 32$ attention heads, $b = 16$ batch size unless otherwise noted)

7.8 Retrieval model training

Aside from the sampling method detailed in 1, retrieval models are trained using the same techniques as generative models, albeit with much larger batch sizes (128x4 for an effective batch size of 512). All retrieval models were trained via the 4x V100 cluster, although in principle training on much smaller compute and memory (ie smaller batch sizes) is feasible too.