GMML IS ALL YOU NEED

Sara Atito\textsuperscript{a,b} Muhammed Awais\textsuperscript{a,b} Srinivasa Nandam\textsuperscript{b} Josef Kittler\textsuperscript{a}

\textsuperscript{a} Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey
\textsuperscript{b} Surrey Institute for People-Centred AI, University of Surrey, Guildford, GU2 7XH, UK

ABSTRACT

Vision transformers (ViTs) have generated significant interest in the computer vision community because of their flexibility in exploring contextual information, whether it is sharply confined local, or long range global. However, they are known to be data hungry and therefore often pretrained on large-scale datasets, e.g. JFT-300M or ImageNet. An ideal learning method would perform best regardless of the size of the dataset, a property lacked by current learning methods, with merely a few existing works studying ViTs with limited data. We propose Group Masked Model Learning (GMML), a self-supervised learning (SSL) method that is able to train ViTs and achieve state-of-the-art (SOTA) performance when pre-trained with limited data. The GMML uses the information conveyed by all concepts in the image. This is achieved by manipulating randomly groups of connected tokens, successively covering different meaningful parts of the image content, and then recovering the hidden information from the visible part of the concept. Unlike most of the existing SSL approaches, GMML does not require momentum encoder, nor relies on careful implementation details such as large batches and gradient stopping. Pretraining, finetuning, and evaluation codes are available under: https://github.com/GMML.

Index Terms— Self-supervised Learning, Vision Transformers, Group Masked Model Learning, Deep Learning.

1. INTRODUCTION

Vision transformers (ViTs) \cite{dosovitskiy2020image} have shown tremendous potential thanks to their self-attention mechanism which is able to model global context. Similar to Natural Language Processing (NLP) \cite{vaswani2017attention}, the ViTs treat an image as a 1D sequence of visual tokens. The modelling of visual structure by a ViT induces no intrinsic inductive bias. While the lack of bias is advantageous from the modelling point of view, the training of ViTs becomes more challenging, and requires orders of magnitude more data \cite{dosovitskiy2020image}. Recently, ViTs have been shown to perform well on ImageNet-1K \cite{deng2009imagenet} without external data \cite{caron2021emerging}. However, they need distillation approaches from CNNs. A tremendous progress in SSL for visual data has been marked by recent methods \cite{caron2021emerging, chen2021improved, jiang2021vision}. A common theme of these methods is the learning of invariant representations for different views/augmentations of the visual data by maximising the similarity between these different views. However, without careful implementation tricks such as the use of large batches, gradient stopping, momentum encoding, asymmetric projection head, etc., these approaches commonly suffer from collapse, i.e. trivial solutions.

In contrast to existing unsupervised learning approaches, GMML exploits information redundancy and complementarity in the image data by learning to reconstruct local content by linking it to context. In spirit, this principle is similar to the masked language modelling (MLM) used in BERT \cite{devlin2018bert}, which recovers masked words from context. In computer vision, we take the inspiration from the principle of a denoising autoencoder \cite{vincent2008extracting} and from the idea of context encoder \cite{useda2018unsupervised}, which has been studied for unsupervised learning using CNNs. The main aim of this study is to draw on the principles of MLM, denoising autoencoders, and context encoders to create an effective self-supervised learning procedure for vision transformers. The GMML addresses the issues of data-efficiency of ViT, clearly demonstrating how vision transformers can be trained from scratch, using limited data, by means of self-supervised pretraining, without using any external data. The proposed methodology of transformer pretraining by self-supervision is expected to have a significant impact on the advancement of science by enabling the wider research community starved of resources to contribute to deep learning. The main contributions and findings of this study are summarised as follows:

- We introduce GMML, a simple method for SSL by transformers, which has been inspired by MLM of BERT, denoising autoencoders and context encoders.
- We show that the amount of labelled data required to learn a downstream task by finetuning is two orders of magnitude lower than that needed for with supervised pretraining.
- GMML outperforms SOTA supervised and self-supervised methods in small, medium and even large datasets with large margins reaching +35\% improvement.
- Among concurrent SSL methods, GMML is almost the only contender which neither suffers from trivial solutions nor requires careful implementation details, others being Barlowtwins \cite{zbontar2021barlow} and VICReg \cite{chacon2021vicreg}.
2. METHODOLOGY

There are several considerations when designing a masked image modelling (MIM) alternative for the image domain. These considerations are discussed in this section. The system diagram of GMML is shown in Figure 1.

**Journey from MLM to GMML/MIM:** Data-tokens in NLP, i.e. words, often represent semantic concepts. Thus, randomly masking a small percentage of tokens and recovering them from context in NLP can induce semantic understanding of the missing data. In contrast, individual data-tokens in an image, i.e. small visual patches, often do not represent a semantic concept. Therefore, randomly masking a small percentage of isolated tokens is not as fruitful as in NLP.

Instead, we propose to randomly mask groups of connected tokens. Such masked tokens are more likely to represent meaningful parts of different semantic concepts in an image. Hence, recovering these meaningful parts from the local and global contextual information can induce learning of concepts in the vision transformers. We note that the groups of randomly masked tokens will fall on different semantic concepts present in the image.

The key hypothesis is that, if the ViTs are able to model missing information from groups of masked tokens on different objects, then they will implicitly learn the semantic representations of these objects in the image. This forms the basis of the thesis that GMML is able to learn information from all the concepts. We refer to this mechanism of modelling missing information from groups of masked tokens as GMML.

The intuition is that, by modelling all semantic concepts, a GMML-trained transformer will generalise better for unseen objects, whether they are related to an object, a distributed object, or to the whole visual signal.

**Realisation of GMML via autoencoder:** The next question is how to promote the learning of the transformer weights via some self-supervised loss function. As a vehicle for realising the idea of GMML in practice, we proposed the use of transformer based masked autoencoders. Specifically, we use $\ell_1$-loss between the reconstructed image from GMML manipulated images and the original image.

There are key difference between the proposed masked autoencoder and vanilla autoencoders which are used commonly in CV. The existing autoencoder are commonly constituted by convolutional encoders with non-linearity and pooling operations for downsampling a bottleneck representation and a decoder which consists of transposed convolutions or upsampling and convolutions. These decoders are usually expensive in terms of parameters as well as storage of the feature maps. Due to the isotropic architecture design of vision transformers and their ability to exploit local or global contextual information, we employ a very light decoder. Our decoder consisting of two point-wise convolution layers (aka MLP layers in transformers) with ReLU non-linearity and a transposed convolution layer to return back to image space. Since the GMML architecture, including both the transformer blocks as well as point-wise convolution, is isotropic, some of the transformer blocks may act as a decoder.

**Choice of GMML based Image Manipulation:** There can be several variations to manipulate the images using GMML. Some are basic, others introduce the notion of alien concepts. The variation of alien concepts have a different impact on performance. These choices are discussed below.

- **Masking with zeros:** The most straightforward approach to introduce an alien concept is to mask the groups of connected tokens with zeros. We found empirically that it works well, however, it is not the most effective way as it is less difficult to model the regions masked with zeros.

- **Masking with noise:** A slightly more complex alien concept is to mask the tokens with random noise. Empirically, it works marginally better than masking with zeros.

- **Replace with visually plausible alien concept:** By introducing alien concepts from another image in the batch. This manipulation challenges the transformers as the network has to first model the visually plausible alien concepts which are injected at random locations via GMML. After modelling the visually plausible alien concepts in initial blocks the transformers use a few blocks gradually to diffuse the information from the native concepts into the regions distorted by alien concepts by drawing on the local and global context of the image, as shown in Figure 2.

Empirically, we found that the best choice is to combine the three strategies, where each image in the batch is manipulated using a random selection of the GMML based manipulation strategies described above.

Fig. 1: Overall architecture of Group Mask Model Learning (GMML).
The objective of the image reconstruction is to restore the original image \( \hat{x} \) from the GMML manipulated image \( \bar{x} \). For this task, we use the \( \ell_1 \)-loss between the original and the reconstructed image as shown in Equation 1.

\[
L(W) = \sum_k \left( \sum_i \sum_j M_{i,j}^k \|x_{i,j}^k - \bar{x}_{i,j}^k\|_1 \right) 
\]

where \( W \) denotes the learnable parameters, \( N \) is the batch size, \( M \) is a binary mask with 1 indicating the manipulated pixels, and \( \bar{x} \) is the reconstructed image.

To improve the performance of the autoencoder, we introduce skip connections from several intermediate transformer blocks to the decoder. These additional connections can directly send the feature maps from the earlier layers of the transformers to the decoder, which helps to use fine-grained details learned in the early layers to construct the image. Besides, skip connections in general make the loss landscape smoother which is leading to a faster convergence.

The reconstructed image \( \hat{x} \) is obtained by averaging the output features from intermediate blocks from the transformer encoder \( E(.) \) and feeding the output to a light decoder \( D(.) \) as follows: \( \hat{x} = D(\sum_{b \in B} E_b(\bar{x})) \), where \( E_b(.) \) are the output features from block \( b \) and \( B \) is a pre-defined index set of the transformer blocks that are included in the decoding process. In this work, we set \( B \) to \( \{6, 8, 10, 12\} \).

## 3. EXPERIMENTAL RESULTS

The common evaluation to demonstrate the generalisation of the learnt features by self-supervised methods is to pretrain and fine-tune on another dataset. As shown in Figure 3, this confirms that GMML greatly benefits the model in unsupervised fashion, followed by fine-tuning on a downstream task like image classification, object detection, segmentation, etc. In this work, we conduct several experiments on 6 well-known datasets to show the effectiveness of our proposed self-supervised vision transformer.

### 3.1. Classification

It is well known that ViTs are data-hungry which makes them hard to train, mostly, due to the lack of the typical convolutional inductive bias. Consequently, the common protocol for SSL with transformers is to pretrain the model on a large scale dataset, such as ImageNet or even larger datasets. The compute and data demand of the vision transformers limit their adoption, particularly by AI researchers with smaller resource budget. Therefore, in the first set of the experiments we investigate the possibility of training ViTs from scratch with limited data. In particular, we compare our proposed GMML approach with the SOTA SSL methods when the pretraining and fine-tuning are performed only on the target dataset. Table 1 shows that our method outperforms the state-of-the-art with a large margin with an improvement of +1.3%, +17.4%, +23.2%, +35.2%, and +23.7% on Flowers, Pets, CUB, Aircraft, and Cars datasets, respectively.

Moreover, we show that longer pretraining achieve better performance rates with an improvement of +9.2%, +11.9%, +4.9%, +6.7%, and +2.6% on the aforementioned datasets, as shown in Figure 3. This confirms that GMML greatly benefits from longer pretraining, where the performance is steadily improving even after 3000 epochs of pretraining.

Additionally, in order to study the effectiveness of GMML on bigger models, we pretrain GMML employing a ViT-
Small for 3000 epochs. As shown in Table 1, we find that using a bigger transformer for self-supervised pretraining using GMML further improves the accuracy with an improvement of +4.1%, +2.1%, +6.2%, +0.4%, and +0.4% on the aforementioned datasets, respectively, compared to pretraining on the ViT-T variant of transformers. Further, GMML significantly outperforms MAE [18] with a large margin on small datasets with an improvement of +7.6%, +15.0%, +18.0%, +15.5%, and +2.1% on the aforementioned datasets, respectively. Note that, for a fair comparison with MAE, we pretrained MAE for twice the number of epochs as compared to GMML, i.e. 6000 epochs. We attribute the poor performance of MAE on small datasets to the lack of information in the encoder part, as the modelling of the spatial visual structures in MAE is only considered in the decoder part, which is excluded during the finetuning phase.

### 3.1.1. Transfer Learning

After demonstrating the applicability of training transformers from scratch with limited data, we study the transfer ability of the representations learnt using GMML. In Table 2, we report the top-1 accuracy obtained in cross domain experiments employing ViT-T. In particular, the off-diagonal cells indicate the performance when the models are pretrained and finetuned on the same dataset and the on-diagonal cells evaluate transfer performance across datasets.

We observe that the proposed approach generalises well across different datasets even if the pretrained dataset and the target dataset are not from the same domain, e.g. CUB and Cars. This is attributed to the fact that the GMML approach leverages unlabelled data in a task-agnostic way during the pretraining stage, hence the representations are not directly tailored to a specific classification task.

The second observation is that the size of the pretraining dataset matters. The more data the model sees during the pretraining, the better the accuracy, except for the MNIST dataset. MNIST is a toy dataset which has only 10 concepts, i.e. the digits, without any sort of variations in the background. In fact, it was expected that the pretrained model on MNIST dataset would not transfer well to other datasets. Yet, we note that the performance of the pretrained model on MNIST is much better than the performance when the model is trained from scratch with an improvement of 16.7%, 36.1%, 28.5%, 42.6%, and 57.9% on Flowers, Pets, CUB, Aircraft, and Cars datasets, respectively. These results demonstrate that pretraining the model with GMML enables the ViT to learn local image structures even without any inductive bias.

Further, we show the benefits of transfer learning from large-scale dataset like ImageNet-1K. As shown in Table 3, pre-training the model in the self-supervised fashion using GMML on ImageNet-1K outperforms supervised pre-training with a large margin, with an improvement of +0.6% +0.6% +5.1% +7.5% +2.9% +4.2% on Flowers, Pets, CUB, Aircraft, Cars, and datasets, respectively. An important characteristic of GMML is the ability of training transformers from scratch on the tiny datasets. This is reflected in Table 3. Note the reduction in performance between the GMML pretraining on ImageNet-1K and the GMML pretraining on the tiny dataset itself without any external information. In particular, the gap in the case of Aircraft and Cars reduces to around 1% between GMML based ImageNet-1K pretraining and the GMML pretraining on the dataset itself without any external knowledge.

### 4. Conclusion

In this work we presented a self-supervised vision transformer, trained with unlabelled data to perform pre-text tasks. During pre-training, the transformer, with innovative architectural features, is used as an autoencoder. The image reconstruction capability enables the transformer to be trained using the GMML strategy, which is instrumental in modelling contextual information present in all the concepts in the image. The key impact of the proposed GMML is that it makes it possible for transformers to train on small and medium size datasets. It is not only data efficient, but its outstanding information extraction ability enables it to outperform state-of-the-art supervised and self-supervised methods with large margins. The additional advantages include the simplicity and elegance of training, without the need to use large batches, momentum encoders, gradient stopping and other tricks to avoid a solution collapse. GMML is an outstanding mechanism to extract information from a given dataset and instil this information into transformer’s weights.

**Acknowledgements:** This work was supported in part by the EPSRC grants MVSE (EP/V002856/1), JADE2 (EP/T022205/1) and EP-SRC/dstl/MURI project 0184EP/R56/1.

---

**Table 2:** Domain Transfer. The results of fine-tuning self-supervised pretrained models employing ViT-T.

| Pretrain | Fine-tuning |
|----------|-------------|
| | MNIST | Flowers | Pets | CUB | Aircraft | Cars |
| **random init** | 58.1 | 31.8 | 23.8 | 14.6 | 12.3 |
| **Transfer learning from toy dataset** | 99.6 | 74.8 | 67.9 | 52.3 | 57.2 | 70.2 |

| **Transfer learning from small datasets** |
|----------------|
| MNIST | 99.6 | 74.8 | 67.9 | 52.3 | 57.2 |
| Flowers | 99.6 | 90.0 | 78.7 | 61.8 | 67.4 | 80.2 |
| Pets | 99.5 | 88.8 | 86.0 | 61.7 | 69.1 | 82.7 |
| CUB | 99.5 | 89.1 | 84.8 | 71.2 | 77.9 | 88.7 |
| Aircraft | 99.5 | 89.2 | 84.4 | 66.7 | 85.1 | 89.7 |
| Cars | 99.6 | 89.2 | 85.7 | 69.4 | 81.1 | 92.7 |

**Table 3:** Transfer Learning from ImageNet-1K employing ViT-T.* is reported by IDMM [13].

| Pretraining | Fine-tuning |
|-------------|-------------|
| | Flowers | Pets | CUB | Aircraft | Cars | ImageNet-1K |
| **Training using only the given dataset** |
| Random Init* | 58.1 | 31.8 | 23.8 | 14.6 | 12.3 | – |
| GMML | 90.4 | 86.0 | 71.2 | 84.1 | 92.7 | 76.4 |

| **Transfer learning from ImageNet-1K** |
|----------------|
| DEIT [4] (SL) | 97.3* | 88.6* | 76.8* | 78.7* | 90.3* | 72.2 |
| GMML (SSL) | 97.9 | 89.2 | 81.9 | 86.2 | 93.2 | 76.4 |
5. REFERENCES

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR, 2021.

[2] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, IEEE, 2009.

[4] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, pages 10347–10357, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning, pages 1597–1607, 2020.

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9640–9649, 2021.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, L. Bottou, K. Kavukcuoglu, and E. Sutskever, editors, Advances in Neural Information Processing Systems, volume 33, pages 21271–21284, 2020.

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the 38th International Conference on Machine Learning, volume 139, pages 12310–12320, 18–24 Jul 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, June 2019.

[10] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, pages 1096–1103, 2008.

[11] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2536–2544, 2016.

[12] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In ICLR, 2022.

[13] Hao Yu Yun-Hao Cao and Jianxin Wu. Training vision transformers with only 2040 images. The European Conference on Computer Vision, 2022.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in Neural Information Processing Systems, 33:18661–18673, 2020.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9650–9660, 2021.

[17] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Computer Vision–ECCV: 17th European Conference, pages 456–473. Springer, 2022.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15979–15988, 2022.