EFFECT OF SCALE ON CATASTROPHIC FORGETTING IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Catastrophic forgetting presents a challenge in developing deep learning models capable of continual learning, i.e. learning tasks sequentially. Recently, both computer vision and natural-language processing have witnessed great progress through the use of large-scale pretrained models. In this work, we present an empirical study of catastrophic forgetting in this pretraining paradigm. Our experiments indicate that large, pretrained ResNets and Transformers are significantly more resistant to forgetting than randomly-initialized, trained-from-scratch models; this robustness systematically improves with scale of both model and pretraining dataset size. We take initial steps towards characterizing what aspect of model representations allows them to perform continual learning so well, finding that in the pretrained models, distinct class representations grow more orthogonal with scale. Our results suggest that, when possible, scale and a diverse pretraining dataset can be useful ingredients in mitigating catastrophic forgetting.

Continual learning is a current challenge in designing machine learning systems: when trained on multiple tasks in sequence, models tend to suffer performance losses on earlier tasks—this is known as *catastrophic forgetting* (McCloskey & Cohen, 1989). Catastrophic forgetting occurs in many settings, from supervised learning for classification (Zenke et al., 2017) to reinforcement learning for games (Kirkpatrick et al., 2017) and downstream applications such as translation (Bapna & Firat, 2019) or medical diagnoses (Gupta et al., 2021). While many methods have been proposed to help mitigate forgetting, none of these have fully solved the problem; see Kemker et al. (2017) for a review.

Increasing model and dataset size have been key to the success of deep learning (Kolesnikov et al., 2020; BIG-bench collaboration, 2021; Radford et al., 2019; Zhai et al., 2021). A recent trend has been the use of increasingly large models pretrained on diverse datasets, which are subsequently fine-tuned or evaluated without fine-tuning on downstream tasks of interest; see Bommasani et al. (2021) and references therein. A particularly striking realization of this has been large pretrained language models, which show consistent improvement across scales (Kaplan et al., 2020) and can perform well on diverse tasks with no finetuning (Brown et al., 2020). From a continual learning perspective, this behavior is especially dramatic as, without finetuning, any training signal for these tasks appears mixed into a drastically larger pretraining dataset, and thus more than likely appeared many gradient steps before evaluation. This observation in large part inspired our work.

Here, we empirically study continual learning in large pretrained models. Our results suggest that pretrained models are significantly more resistant to catastrophic forgetting than their randomly-initialized counterparts, and that this resistance improves with the scale of both the model and pretraining data. Most of our experiments concern vision—specifically image classification; but we also conduct a preliminary investigation of language models.

Our key observations are as follows:

- Resistance to catastrophic forgetting improves with model and pretraining dataset scale (Figure 1).
- Large, pretrained image-classification models are significantly more resistant to forgetting than models trained from random initialization (Figure 4).
- Both supervised and unsupervised, SimCLR (Chen et al., 2020), pretraining improve forgetting (Figure 7).

- We perform initial experiments aimed at exploring why scale and pretraining reduce catastrophic forgetting, showing that distinct class representations in pretrained models are more orthogonal than in trained-from-scratch models (Figure 8).
- We observe similar phenomena in language models (Figure 9).

1 Related work

Catastrophic forgetting Catastrophic forgetting was identified as a problem for neural networks as early as the late 1980s, as it occurred in the single-hidden-layer networks that were studied at the time (see Delange et al. (2021) for a review). Within the catastrophic forgetting literature, a number of potential solutions to the problem have been devised; some involve replaying previous-task examples during training on later tasks (Rebuffi et al., 2016; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018; Shin et al., 2017; Riemer et al., 2018), while others like elastic weight consolidation (Kirkpatrick et al., 2017) limit the extent to which parameters can change once learned (Zenke et al., 2017; Lee et al., 2017; Farajtabar et al., 2019). More recently, meta-learning algorithms (Beaulieu et al., 2020; Caccia et al., 2021) have been used to directly learn methods to continually learn task sequences.

While some works have investigated the effect of architectural choices on forgetting (Goodfellow et al., 2015; Ramasesh et al., 2021), a systematic study of scaling—especially using modern pretraining methods—has not been conducted. This is the gap our work seeks to fill.

Large-scale models The observation—initially in language models (Kaplan et al., 2020)—that model performance scales predictably with model and dataset size spurred an increased interest in training and deploying increasingly large pretrained models. Large-scale models emerged in natural-language processing with models such as GPT (Radford et al., 2019) and BERT (Devlin et al., 2019), which are pretrained in an unsupervised manner and thus able to make use of the large corpora of natural text available on the internet. More recently, image-classification models have also benefited from pretraining on large datasets (Kolesnikov et al., 2020), both supervised and unsupervised; a key example is the Vision Transformer architecture of Dosovitskiy et al. (2020), which—despite not using the inductive bias provided by convolutions—attains downstream performance comparable to convolutional networks after pretraining on ImageNet-21k (Deng et al., 2009) or JFT-300M (Sun et al., 2017). On many current NLP and computer vision benchmarks, large pretrained models achieve top performance.

2 Experimental setup

Tasks: Most of the experiments in this paper are conducted in the standard, task split setting (Kirkpatrick et al., 2017; Zenke et al., 2017). We consider *split CIFAR-10* task sequences – the ten class dataset is split and sequentially trained on two tasks of 5 classes each. We also consider sequential 10 and 50 class splits of CIFAR-100.

We also consider the more recent *input-distribution-shift CIFAR-100* setup (Ramasesh et al., 2021). In this case each task has the same target types made up a fixed subset of the 20 CIFAR-100 superclasses (e.g. aquatic mammals, fish, ...). The sequence of tasks consists of modifying which subclasses are sampled from. For example Task A may have aquatic mammals drawn from beaver and dolphin, while Task B may have otter and whale. This setup allows one to probe forgetting without changing the model head, in a setting where task identity is not known to the model.

For our language model experiments, we consider next token prediction generative tasks and use the IMDb Reviews (Maas et al., 2011) and english Wikipedia (Foundation) datasets.

Vision models: In most of our experiments we use the Vision Transformer (ViT) and ResNet architectures. For the Vision Transformers, we use the ViT-S and ViT-B configurations from Dosovitskiy et al. (2020) and a smaller configuration we call ViT-xS – ranging from 5.7-86.7 million parameters. For ResNets, we use five models: ResNet26, ResNet50, ResNet101, ResNet152, and ResNet200 – ranging from 14.0-62.6 million parameters; as in Kolesnikov et al. (2020), these models use group normalization (Wu & He, 2018) and weight standarization (Qiao et al., 2019) in place of batch normalization. Details can be found in Tables 1 and 2 in Appendix A.1.2.

Pretraining: Unless otherwise specified, we pretrain our image models in a supervised manner on the ImageNet21k dataset, which contains approximately 14 million images in about 26000 distinct categories using the Adam optimizer (Kingma & Ba, 2017). Further details are in the Appendix.

Finetuning: When fine-tuning on multiple tasks in sequence, we use the SGD optimizer with momentum $\beta = 0.9$. The first task training employs a cosine decay schedule with a linear warmup; later tasks in the sequence are trained with fixed learning rates. For the split-CIFAR-10 task sequences, we use a multi-head setup in which each head is initialized to zeros.

The architectures and finetuning setup for language modeling are described in Section 4.

3 Results

3.1 Forgetting improves with model scale

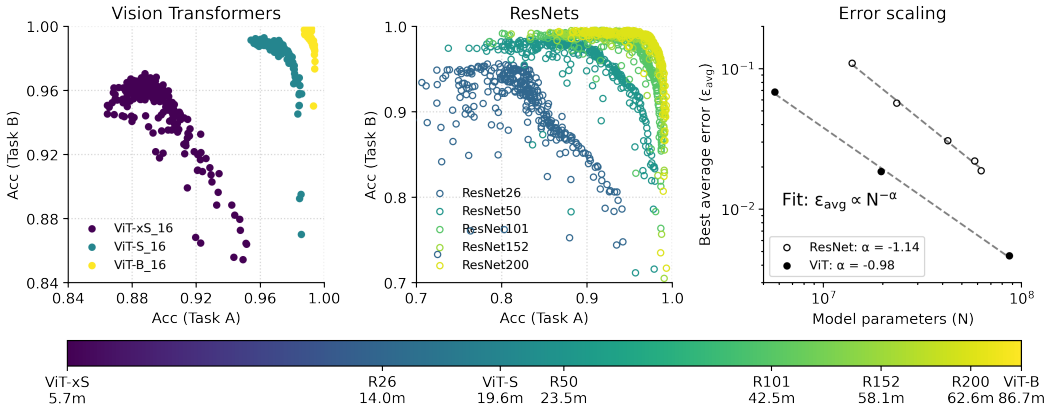

Figure 1: **Forgetting frontier across model scales**. Task A versus Task B performance for both pretrained vision transformers (left) and ResNets (center). Each point represents a different choice of learning rate or finetuning step for Task B. All models were pretrained on ImageNet21k for 90 epochs. (right) the best average Task A/B error improves systematically with model size.

Our main finding is that for both pretrained Vision Transformers and pretrained ResNets, catastrophic forgetting is mitigated to a large extent by scale: that is, larger models suffer less from forgetting. This is shown in Figure 1 for the two-task split CIFAR-10 sequence: Task A {airplane, automobile, bird, cat, deer} $\rightarrow$ Task B {dog, frog, horse, ship, truck}. In order to make a fair comparison, we plot the test accuracies of the model on both tasks at different steps of Task B fine-tuning for a variety of learning rates. This allows us to see, for each model, the maximum performance achievable on the new task for a given performance on the initial task. We call this the forgetting frontier.

Figure 1 shows that the forgetting frontier for both ViT and ResNet models improves significantly with scale. While a 5.7M-parameter ViT-xS model loses about 6.5% in Task A accuracy over the course of learning to perform Task B, an 86.7M-parameter ViT-B model loses less than half a percent. To put this in context, in a previous studies of split-CIFAR-10 task sequences, a ResNet18 model lost close to 40% accuracy with no mitigation methods, 5% when using a replay buffer, and 20% when using elastic weight consolidation (see Figure 1 of Ramasesh et al. (2021) for an example).

It is also important to note that this effect of scale and pretraining on improving forgetting frontiers does not seem to be architecture-dependent. Both vision transformers—based on self-attention—and ResNets—based on convolutions—exhibit improvements in their forgetting frontiers as they grow larger, as long as they are pretrained.

We see this improvement in forgetting performance beyond the simple two-task split-CIFAR-10 scenario. Across a 10-task sequence of 10-class split-CIFAR-100 tasks (Figure 2 left), sequential 50-class splits of CIFAR-100 (Figure 2 right), and the CIFAR-100 single-head input-distribution-

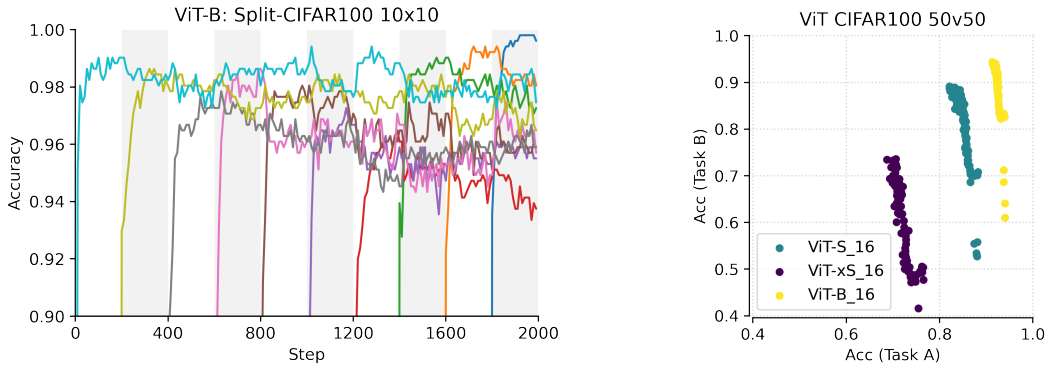

Figure 2: **Split CIFAR-100 tasks**. (left) Sequential training of a pretrained ViT-B model on a 10-task sequence, where each task consists of classifying between 10 categories in the CIFAR-100 dataset. In this continual learning setup, the model is able to learn all 10 tasks without much forgetting: on average, the model loses 1.8% accuracy on each task; at most, the model loses 2.9%. (right) Forgetting frontiers for vision transformers on a 2-task, 50-class split CIFAR-100, showing the same increase with scale as on the split-CIFAR-10 version.

shift task (Figure 3), large-scale pretrained models exhibit resistance to forgetting, and continual learning ability appears to improve systematically with scale.

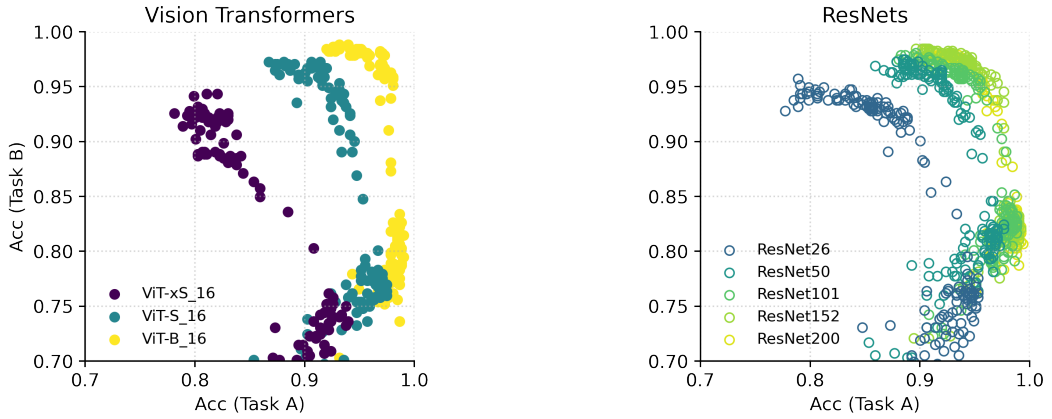

Figure 3: **Single-head models on CIFAR-100 distribution shift task**. As with the multi-head setup, the forgetting in Vision Transformers (left) and ResNets (right) is largely mitigated by pretraining and scale. The full task details are given in Appendix A.1.3

3.2 Forgetting in pretrained models vs. models trained from scratch

In the previous section, we saw that the amount of catastrophic forgetting in models pretrained on large datasets decreased with increasing model size, to the point that the largest models (ViT-B, with 86.7M parameters and ResNet200 with 62.6M parameters) suffer very little performance loss from second-task training. In this section, we investigate whether this improvement in forgetting is simply a consequence of the improvement in performance on Task A. That is, do pretraining and scale have benefits with respect to forgetting beyond simply improving the fine-tuned performance on downstream tasks?

To check this, we compare the performance and amount of forgetting in pretrained ResNet models to models trained from scratch. We handicap the pretrained ResNets so that their maximum accuracy on on Task A matches that of the corresponding trained-from-scratch ResNet; this is done simply by reducing the number of steps for which we fine-tune. Figure 4 shows the results of this experiment, with additional results and details in Appendix B.1. While the trained-from-scratch ResNets achieve the same performance as their pretrained counterparts on Task A, their Task A performance degrades

significantly more than the pretrained ResNets when trained on Task B; pretrained ResNets forget less, and this improves with model scale.

This result suggests that pretraining and model scale confer benefits upon models which go beyond simply improving performance. From Figure 4 it is also clear (see also Figure 11 in the appendix, which plots the same data in a different way) that pretraining is required to see the benefits of scale. Simply scaling up the model size in trained-from-scratch ResNets does not yield clear improvements in their forgetting performance.

We focused in this section exclusively on ResNets (rather than Vision Transformers) because obtaining decent performance on image classification tasks with Vision Transformers seems to require pretraining. In our experiments, for example, a trained-from-scratch ViT-S model was unable to achieve better than 75% test accuracy on five-class CIFAR-10, while the same model, pretrained, achieved better than 98% accuracy.

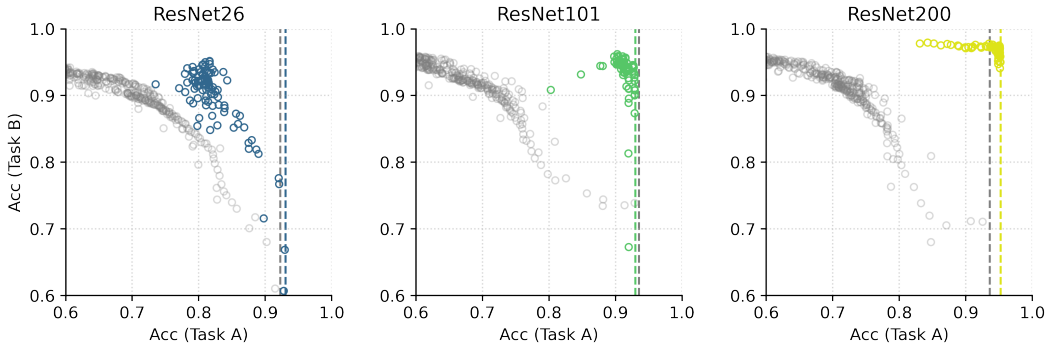

Figure 4: **Pretrained versus trained-from-scratch models**. ResNet models trained from scratch (gray) exhibit inferior forgetting performance than pretrained models (colored), even for the pretrained models shown here, which were handicapped during fine-tuning, in order to match Task A performance. Futhermore, the trained-from-scratch models are unable to take advantage of model scale – the forgetting frontier is largely independent of model size (see also Figure 11 in the Appendices). Dashed vertical lines show the maximum accuracy on Task A.

3.3 Dependence on pretraining time, pretraining dataset size, and finetuning time

Having seen in the previous sections that pretraining on large datasets improves models' resilience to catastrophic forgetting, in this section we explore the dependence of this phenomenon on the duration of pretraining, the size of the pretraining dataset, and the number of steps for which we finetune on Task A.

First, we study the influence of pretraining time on downstream forgetting. This is motivated by previous studies (e.g. Hernandez et al. (2021)) showing that downstream performance does not always monotonically improve with pretraining time and performance. By contrast, in the case of forgetting we find (see Figure 5 for ViT-S results) that as the pretraining process continues, both downstream performance and the forgetting frontier improve. This was the case for all models (ViTs and ResNets) and downstream task sequences (CIFAR-10, CIFAR-100, etc.) we studied. We found no evidence of forgetting frontiers getting worse with increased pretraining.

Along similar lines, we show the influence of pretraining data scale on downstream forgetting in Figure 6, pretraining a model on varying fractions of ImageNet21k. While the full dataset contains over 14 million images, we find that training on even one-sixteenth of the dataset (roughly the size of the standard ImageNet ILSVRC 2012 dataset) yields a model which only loses about 3% performance in forgetting on split-CIFAR-10. This suggests that one does not need to pretrain models on unreasonably large datasets in order to reap benefits for continual learning.

Finally, we study the forgetting frontier of pretrained models as we vary the amount of first-task fine-tuning (number of training steps). This is partly inspired by a recent study (Andreassen et al., 2021) of robustness in pretrained models, which found that during fine-tuning, pretrained models

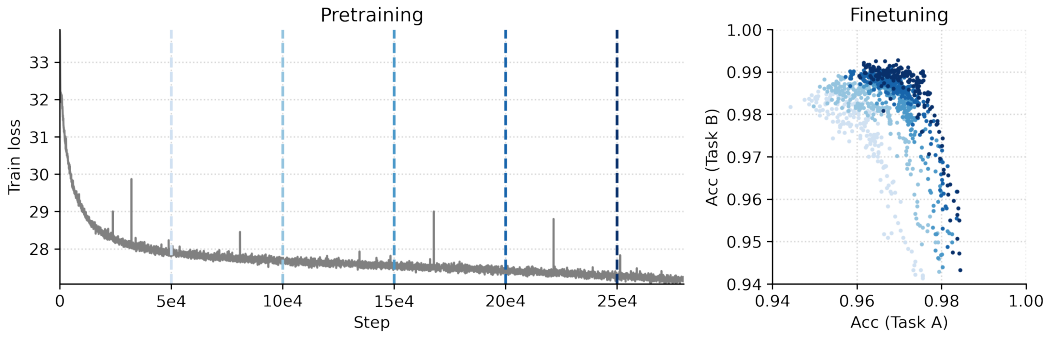

Figure 5: **Varying pretraining time**. We take pretrained checkpoints at 50k, 100k, 150k, 200k, and 250k steps (shown in colored ticks at the left) for a ViT-S model pretrained on ImageNet21k. (right) The forgetting frontiers improve with increased pretraining time.

evolved some robustness to distribution shift which vanished at the end of fine-tuning. Figure 6 shows forgetting frontiers for pretrained ViT-B and ResNet101 models, fine-tuned for a number of steps between 100 and 5000. The results in the figure suggest that in our setting, increasing the fine-tuning steps to maximize performance on Task A does not appear to cause increased forgetting. On the contrary, even training for about ten times more steps than it takes to achieve saturating performance does not impact forgetting.

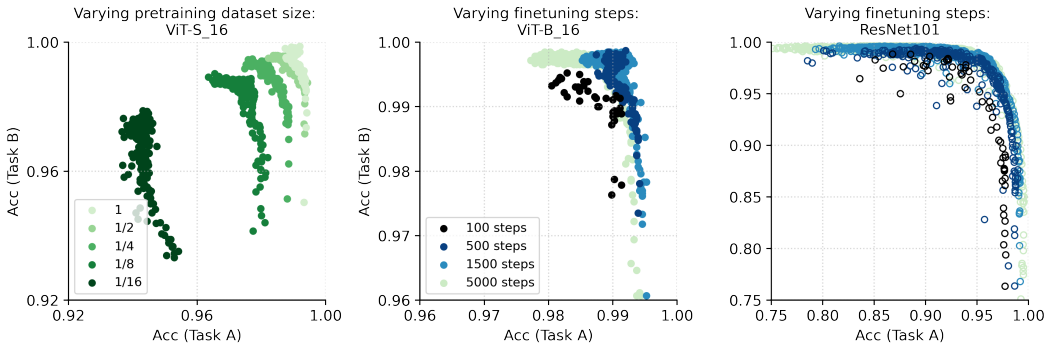

Figure 6: **Varying dataset size and finetuning time**. (left) The forgetting frontier for a vision transformer (ViT-S16) finetuned on a two-task CIFAR-10 sequence, pretrained on varying fractions of the ImageNet21k dataset shows improvement with pretraining dataset size. (center, right): The forgetting frontiers for models pretrained on ImageNet21k, with varying amounts of fine-tuning – Task A performance saturates around 500 steps; even training for ten times as many steps does not affect the forgetting frontier.

3.4 Forgetting in models with unsupervised pretraining

In previous sections, the models we studied were pretrained in a supervised manner on ImageNet21k. Part of the conceptual motivation for this work came from the properties of language models trained in an unsupervised fashion. Furthermore, recently, unsupervised pretraining methods have yielded vision models which achieve good downstream performance; among these are SimCLR, MoCo, DINO, and BYOL. Given the increased prevalence of such unsupervised pretraining methods, we investigate in this section whether unsupervised pretraining, combined with scale, provides resistance to forgetting similarly to supervised pretraining. We take publicly-available ResNet models which were pretrained using SimCLR (Chen et al., 2020) on the ImageNet ILSVRC-2012 dataset (Russakovsky et al., 2015), and finetune them on the two-task split CIFAR-10 sequence. The resulting forgetting frontiers, shown in Figure 7, show that as in supervised pretraining, SimCLR pretraining yields models which are resilient to forgetting, especially as scale is increased. We leave a thorough exploration of the continual-learning properties of other unsupervised pretraining methods for future work.

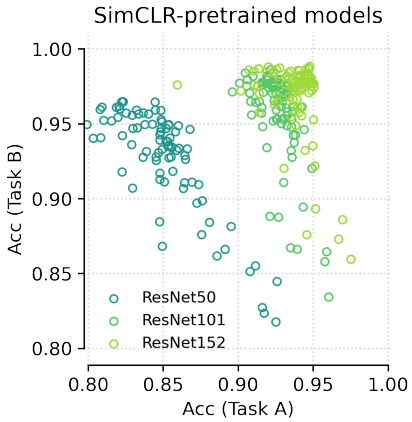

Figure 7: **Unsupervised pretraining**. ResNets pretrained using unsupervised SimCLR on ImageNet1k exhibit a similar trend to that observed in the supervised setting of forgetting robustness improving with model scale.

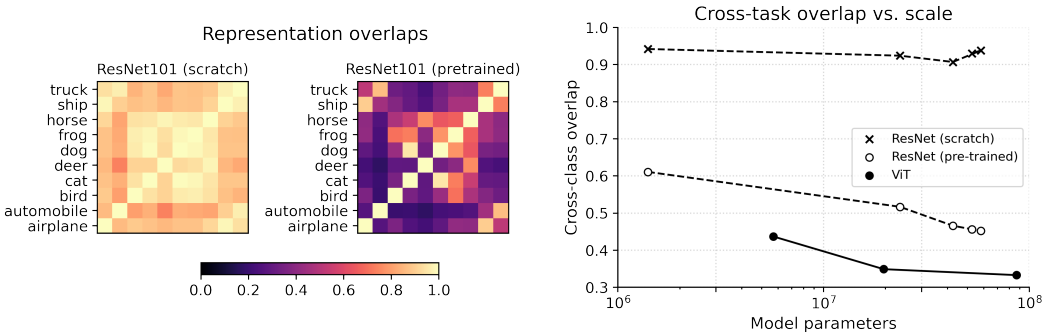

Figure 8: **Trace overlap of CIFAR-10 class representations**. (left) The cross-class representation overlap matrices (Equation 1) are significantly lower for a pretrained ResNet101 than for one trained from scratch. (right) For both Vision Transformers and ResNets, the cross-class overlap decreases as the model size increases. See Appendices for additional models and measurements.

3.5 Representation overlap and forgetting

In this section we take steps toward understanding why large pretrained models seem to be resistant to forgetting. Ramasesh et al. (2021) introduce the *trace overlap* of model representations. They show empirically that forgetting is maximal for intermediate values of the overlap, and derive this property analytically in a simplified model. In particular Ramasesh et al. (2021) propose that a minimal trace overlap (corresponding to orthogonality of representations of different tasks) will lead to minimal forgetting. Inspired by this, we measure representation overlaps in models finetuned on split CIFAR-10, taking representations (pre-logit activations) at the end of Task A training.

In detail, if we denote the P model features evaluated on datasets $\mathcal{D}_A$ and $\mathcal{D}_B$ by $f_{A,B} \in \mathbb{R}^{|\mathcal{D}_{A,B}| \times P}$, and define $\Theta_{AB} := f_A f_B^T$, then the trace overlap, S_{AB}, is given by

$$S_{AB} = \frac{\mathrm{Tr}(\Theta_{\mathrm{AB}}\Theta_{\mathrm{AB}}^{\mathrm{T}})}{\sqrt{\left(\mathrm{Tr}(\Theta_{\mathrm{AA}}\Theta_{\mathrm{AA}}^{\mathrm{T}})\mathrm{Tr}(\Theta_{\mathrm{BB}}\Theta_{\mathrm{BB}}^{\mathrm{T}})\right)}}. \tag{1}$$

We consider the 10x10 matrix, S, formed by taking the indices A and B to run over the different classes in CIFAR-10.

Using the above definition, we return to the original split-CIFAR-10 setting of §3.1 and §3.2. For each of the models used in those experiments (see Tables 1 and 2), we compute the representation overlap matrix. We focus on two comparisons: (i) how does the representation overlap matrix

differ between pretrained and trained-from-scratch ResNets? and (ii) how does the orthogonality of representations vary as we scale the model size, for both ResNets and Vision Transformers?

Pretrained vs. trained-from-scratch representations: We show representative trace overlap matrices for both pretrained and trained-from-scratch ResNet101 models in Figure 8(left and center) (see Figure 12 for additional models). Each element of this matrix corresponds to the representation overlap between two batches, each drawn entirely from a single class. The representation overlap between disjoint classes (off-diagonal class overlap matrix elements) is much smaller for the pretrained ResNets than those trained from scratch. Averaging cross-class overlaps over class pairs from different tasks quantifies this difference: for trained-from-scratch ResNet101, the average overlap is 0.907, while for the pretrained model this value is only 0.466. For all sizes of ResNet, the pretrained models exhibit representations which have much smaller overlap than their trained-from-scratch counterparts.

Representation overlap across scales—Figure 8(right) shows the average cross-task overlap for both ResNets and Vision Transformers plotted against model size (number of parameters). For pretrained Vision Transformers and ResNets, this overlap decreases with model size, while for trained-from-scratch ResNets, there is not a clear decreasing trend. Interestingly, Vision Transformers appear to form more orthogonal representations than ResNets do, a fact which can also be seen in the overlap matrices presented in supplementary Figure 13, and perhaps explains their improved forgetting performance.

These observations suggest that pretrained models store representations of different classes with much less overlap than trained-from-scratch models, and that this orthogonality increases with scale. As this mirrors the the observed trend in forgetting in these models, it is suggestive that the increased orthogonality of representations is a partial explanation for the forgetting resilience seen in large pretrained models.

4 Forgetting in Language Models

Transfer learning from unsupervised pretraining is commonplace in Natural Language Processing. It has become standard to pretrain models on data crawled from the internet and then finetune on smaller, downstream tasks (Devlin et al., 2019; Radford et al., 2019; Raffel et al., 2020; Brown et al., 2020). This bears similarities with both the unsupervised pretraining and the experiments in Figure 3 where a single head was used to train on different distributions (subclasses). Indeed, as mentioned above, part of the conceptual motivation for this work came from the observation that large language models with diverse pretraining perform well on tasks with no finetuning – ostensibly retaining information for many gradient steps. In this subsection and in the Supplement, we include some preliminary empirical results probing catestrophic forgetting directly in large language models.

We consider decoder only transformer models pretrained on data scraped from the web, including C4 (Raffel et al., 2020). These models range in size from 57 million to 8.6 billion non-embedding parameters and were originally trained as part of the BIG-bench collaboration (2021). In Figure 9, we consider sequentially finetuning the models on Task A: IMDb Reviews for 1.5k steps followed by Task B: english Wikipedia for 10k steps. The longer finetuning time for Task B is intended to highlight forgetting, as these models are relatively robust.

We find that, as in the case of pretrained vision models, the best joint Task A, Task B performance improves with scale.

5 Discussion

The central observation we have reported in this paper is that scaling up models pretrained on large datasets is effective at mitigating catastrophic forgetting in the settings studied. While it was known that large pretrained models can achieve impressive performances in downstream tasks, our experiments show that pretraining and scale can confer benefits beyond simply the achievable downstream loss or accuracy values. This is similar in spirit to Andreassen et al. (2021), which found that pretrained models exhibited high out-of-distribution robustness; and to the observations of zero-shot or few-shot task learning in, e.g., GPT models (Radford et al., 2019). We would be excited to see

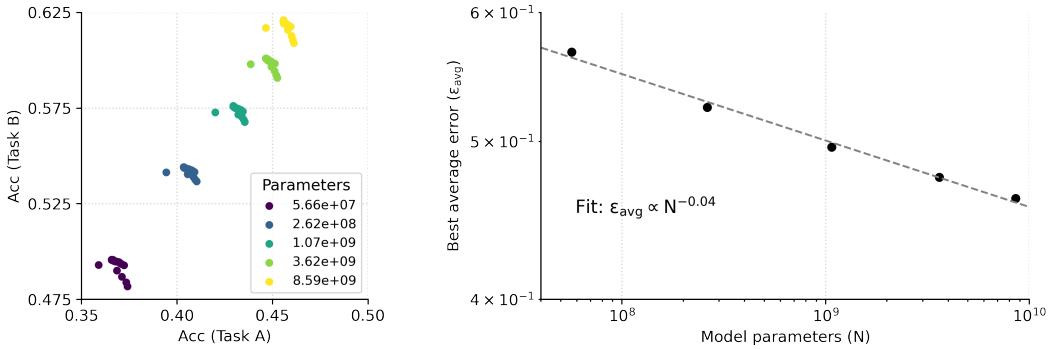

Figure 9: **Forgetting in sequential language-modeling**. Token accuracy for decoder-only transformer models show systematic improvement in joint Task A, Task B performance across scales.

future work exploring other potential benefits of large-scale pretrained models which are not simply reflected in downstream performance.

A natural question is what relationship between pretraining and finetuning tasks is required to confer benefit. In the vision setting, we considered ImageNet pretraining followed by CIFAR classification, so both pretraining and finetuning distributions consisted of natural images (albeit with different resolutions and collection techniques). In the language setting, finetining and pretraining tasks both consisted of English language modeling. It would be interesting to explore to what degree distributional overlap of pretraining and finetuning data underlies the improved resistence to forgetting.

Interestingly, our results provide slight evidence against a common intuition expounded in, e.g., §6.6 of Hernandez et al. (2021)—namely that if one has infinite data, one achieves the best model by omitting pretraining and instead training entirely on-distribution. It is possible, given our results, that if one is concerned with more than simply on-distribution performance—i.e., if one is concerned with performance after learning tasks in sequence, models trained only on-distribution (in this case trained only on data from Task A) would be more susceptible to catastrophic forgetting than models pre-trained on a different distribution. To truly probe this experimentally, one would need a vast quantity of finetuning (Task A) data, so at this point our results are simply suggestive, not conclusive.

Our work suggests natural follow-up investigations to further understand continual learning in large pretrained models. First, while we have taken initial steps towards identifying the reason for the resistance to catastrophic forgetting we observed – noting that increased orthogonality of representations is a plausible mechanism – we have not performed extensive experiments showing that this is causally linked to the forgetting behavior. Second, we explored unsupervised pretraining only via the SimCLR method; confirming that our results hold in other label-free methods such as DINO, BYOL, MoCo, and others would also be a valuable study.

Another natural direction for future work is understanding what underlies the systematic improvement with model scale (e.g. Figures 1 and 9 right). The rate at which performance improves with model size in Figure 1, with power-law exponent close to -1, is reminiscent of the *variance limited* behavior discussed in Bahri et al. (2021). Namely when the number of parameters is much larger than the training dataset size (as is the case for split-CIFAR fintetuning) performance is predicted to scale as N^{-1}. This behavior is noticeably not seen in Figure 9 for language modeling – this is reminiscent of the disparate scaling behaviors observed in Sharma & Kaplan (2020) between language modeling and classification tasks. More generally, understanding what underlies and predicting the rate of improvement to forgetting with model size is an important direction for future research.

REFERENCES

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws, 2021.

Ankur Bapna and Orhan Firat. Non-parametric adaptation for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1921–1931, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1191. URL https://aclanthology.org/N19-1191.

Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O. Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn, 2020.

BIG-bench collaboration. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*, 2021. URL https://github.com/google/BIG-bench/.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Massimo Caccia, Pau Rodriguez, Oleksiy Ostapenko, Fabrice Normandin, Min Lin, Lucas Caccia, Issam Laradji, Irina Rish, Alexandre Lacoste, David Vazquez, and Laurent Charlin. Online fast adaptation and knowledge accumulation: a new approach to continual learning, 2021.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *ArXiv*, abs/1812.00420, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. ISSN 1939-3539. doi: 10.1109/tpami.2021.3057446. URL http://dx.doi.org/10.1109/TPAMI.2021.3057446.

Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pp. 248–255, 06 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning, 2019.

Wikimedia Foundation. Wikimedia downloads. URL `https://dumps.wikimedia.org`.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015.

Sharut Gupta, Praveer Singh, Ken Chang, Liangqiong Qu, Mehak Aggarwal, Nishanth Thumbavanam Arun, Ashwin Vaswani, Shruti Raghavan, Vibha Agarwal, Mishka Gidwani, Katharina Hoebel, Jay B. Patel, Charles Lu, Christopher P. Bridge, Daniel L. Rubin, and Jayashree Kalpathy-Cramer. Addressing catastrophic forgetting for medical domain expansion. *CoRR*, abs/2103.13511, 2021. URL `https://arxiv.org/abs/2103.13511`.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1611835114. URL `https://www.pnas.org/content/114/13/3521`.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *Neural information processing systems*, pp. 4652–4662, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. 1989.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan L. Yuille. Weight standardization. *CoRR*, abs/1903.10520, 2019. URL `http://arxiv.org/abs/1903.10520`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LhY8QdUGSuw.

Sylvestre-Alvise Rebuffi, Alexander I Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2016.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *CoRR*, abs/2004.10802, 2020. URL https://arxiv.org/abs/2004.10802.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pp. 2990–2999, 2017.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017.

Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018. URL http://arxiv.org/abs/1803.08494.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3987–3995. JMLR, 2017.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2021.

A Experimental details

A.1 Vision

A.1.1 Supervised pretraining

Here we detail the settings used to in supervised pretraining, resulting in the models used in the experiments of sections 3.1, 3.2, 3.3, and 3.4. We pretrain seven models: (ViT-B, ViT-S, ViT-xS, R26x1, R50x1, R101x1, R152x1, R200x1) on ImageNet21k for 90 epochs. Pretraining is done using the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$); for all models, we use a batch size of 4096. Our learning-rate schedule includes a warmup of 10k steps to a maximum learning rate $\eta = 10^{-3}$, followed by a linear decay to 10^{-5}. For the Vision Transformer models, we use a weight decay penalty of 0.03 and a dropout rate of 0.1; for ResNet models, we do not use weight decay or dropout. These settings are largely identical to those found in Dosovitskiy et al. (2020). During pretraining, we perform augmentation, applying Inception-style cropping and random horizontal flipping.

A.1.2 VISION TRANSFORMER MODEL DETAILS

For full description of the Vision Transformer architecture, see the original paper by Dosovitskiy et al. (2020). Abbreviated details are given here. We process 2D images of height H, width W, and C channels, $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ by first reshaping them into a 1D sequence of flattened patches–in this work, we exclusively use patch size 16. Each patch is mapped (via a learned linear transformation) to an embedding vector of dimension D (for our models, $D_{xS} = 256$, $D_S = 516$, and $D_B = 768$, as in Table 1). As both our pretraining and finetuning is supervised, we prepend a (learnable) '[class]' token embedding to this sequence of patches, add a learnable 1D position embedding, and pass the sequence through the Transformer encoder. The Transformer encoder alternates between multi-headed self-attention layers and MLP (plus residual connection) layers, with layernorm applied before both the attention and MLP layers.

For classification, we use a classification head which acts on only the transformed '[class]' token representation. During pretraining, this classification head is an single-hidden-layer MLP, as in the original Vision Transformer work; during finetuning, the classification head is a linear layer (in the split-CIFAR-10 setting, we use a multi-head setup with a different classification layer for each task in the sequence). We initialize all classification layers with zeros.

Table 1: Configurations of ViT models

Model	Layers	Width	Dim.	Heads	Params
ViT-xS	4	256	256	4	5.7M
ViT-S	6	516	516	6	19.6M
ViT-B	12	768	768	12	86.7M

Table 2: Configurations of ResNet models

Model	Resblocks	Params
R26	[2, 2, 2, 2]	14.0M
R50	[3, 4, 6, 3]	23.5M
R101	[3, 4, 23, 3]	42.5M
R152	[3, 8, 36, 3]	58.1M
R200	[3, 24,36,3]	62.6M

A.1.3 CIFAR-100 DISTRIBUTION SHIFT TASK

Here we provide more information about the CIFAR-100 distribution-shift task, which was introduced in Ramasesh et al. (2021) and forms the basis of our experiments in section A.1.3 of the main text. The idea behind this task is to mimic a way in which catastrophic forgetting might occur in practice, in image-classification settings where the class labels remain constant but the distribution of the images changes. A key features of this type of setting is that the identity of the task is not known to the model at inference time, so task-specific components (such as multiple readout layers) are not allowed. For example, in a medical setting, one might seek to classify images as displaying signs of a disease or not, and the system might be trained sequentially on images taken at various hospitals. While the CIFAR-100 dataset features images divided into 100 distinct classes, these classes are grouped into 20 *superclasses*. The CIFAR-100 distribution shift task sequence involves classifying an image by its superclass; individual tasks in the sequence draw images from a single subclass for each superclass. As a concrete example, in the experiments in the main text we use the five superclasses *aquatic mammals*, *fruits and vegetables*, *household electrical devices*, *trees*, and *vehicles-1*[1]. For the first task in the sequence, Task A, we uses the classes dolphin (for aquatic mammals), apple (for fruits and vegetables), lamp (for household electrical devices), maple tree (for trees), and bicycle (for vehicles-1). For the second task, Task B, we use the classes whale (for aquatic mammals), orange (for fruits and vegetables), television (for household electrical devices), willow (for trees), and motorcycle (for vehicles-2).

A.2 LANGUAGE

Models are decoder-only transformers which were pretrained using Adafactor and learning rate 1.0, context length 1024 and batch size 256. We finetuned with batch size of 8 and set a constant learning rate equal to the learning rate at the end of pretraining rescaled by the relative batch size, $8/256$.

[1]The CIFAR-100 dataset features two vehicle superclasses, denoted *vehicles-1* and *vehicles-2*

B ADDITIONAL RESULTS

B.1 COMPARING PRETRAINED AND TRAINED-FROM-SCRATCH RESNETS

In the main text, we showed forgetting frontiers on the split-CIFAR-10 task sequence for selected ResNet models: ResNet26, ResNet101, and ResNet200. Here, we show the forgetting frontiers for all ResNets (10), and we also plot the data in a slightly different way (11) to make it apparent that while the forgetting frontiers of the pretrained ResNet models clearly improve with model size, the forgetting frontiers of trained-from-scratch ResNet models do not.

B.2 REPRESENTATION OVERLAP MATRICES FOR ALL RESNET AND VIT MODELS

In the main text, we described the computation of the classwise representation overlap matrices for our models at the end of Task A training, showing an example of such a matrix for both a pretrained and a trained-from-scratch ResNet101 model. In Figure 12, we show representation overlap matrices for all of the ResNet models, which bears out our statement in the main text that pretrained ResNet representations were much closer to orthogonal than their trained-from-scratch counterparts. In Figure 13, we show representation overlap matrices for the (pretrained) Vision Transformer models. As with the pretrained ResNet models, the cross-class representation overlaps become closer to orthogonal as we increase the model size.

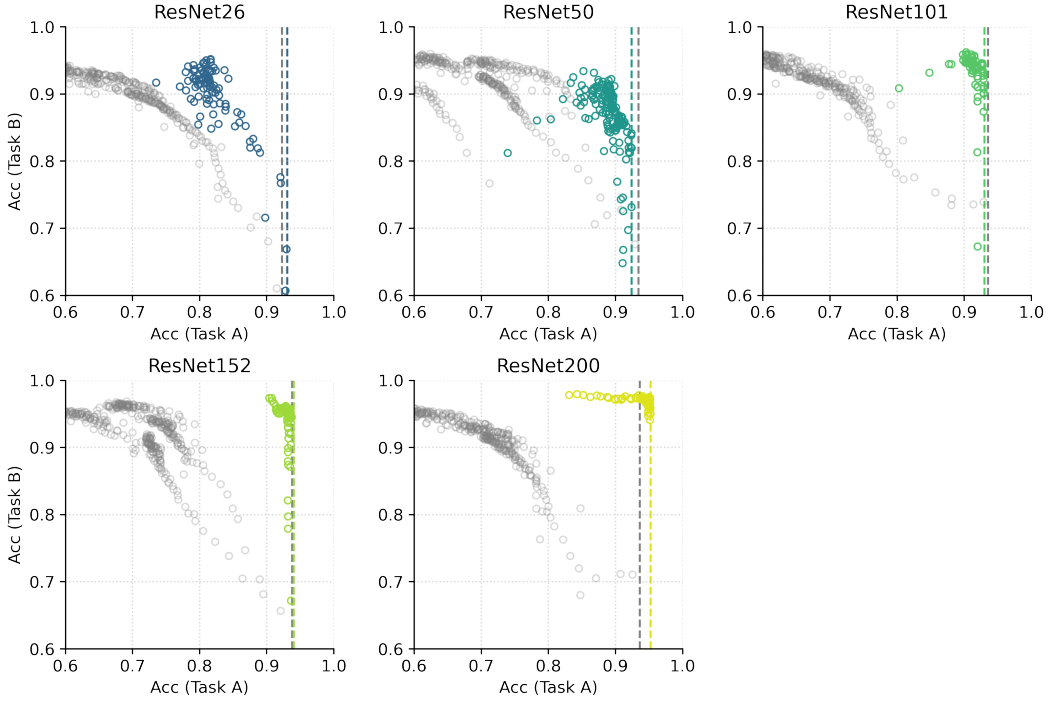

Figure 10: Forgetting frontier for ResNets trained from scratch, vs. pretrained ResNets (on ImageNet21k). The pretrained ResNets were handicapped during fine-tuning, in order to match the Task A performance of the from-scratch model. This figure is identical to Figure 4 in the main text, with the addition of the ResNet50 and ResNet152 models.

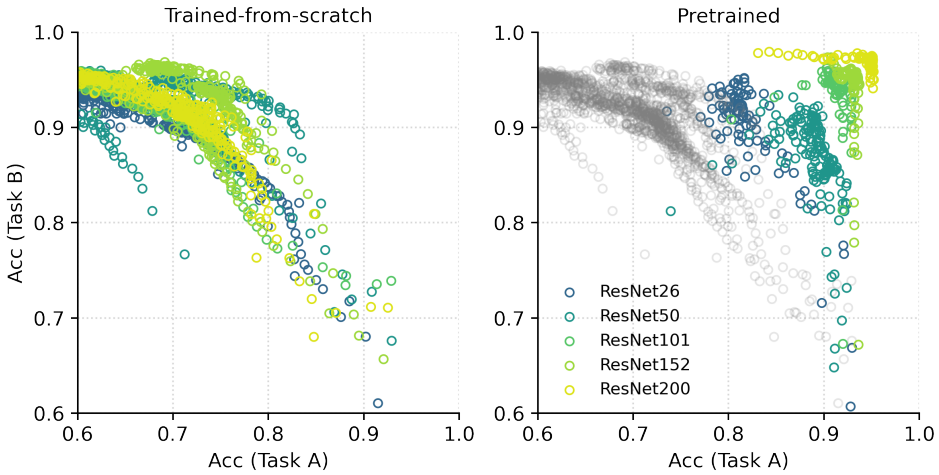

Figure 11: Forgetting frontier for ResNets trained from scratch, vs. pretrained ResNets (on ImageNet21k). The pretrained ResNets were handicapped during fine-tuning, in order to match the Task A performance of the from-scratch model. The data plotted here is the same as that plotted in Figure 4 in the main text; here, the intention is to make clear that the performance of the trained-from-scratch models do not exhibit any apparent benefit with scale, while the pretrained models do.

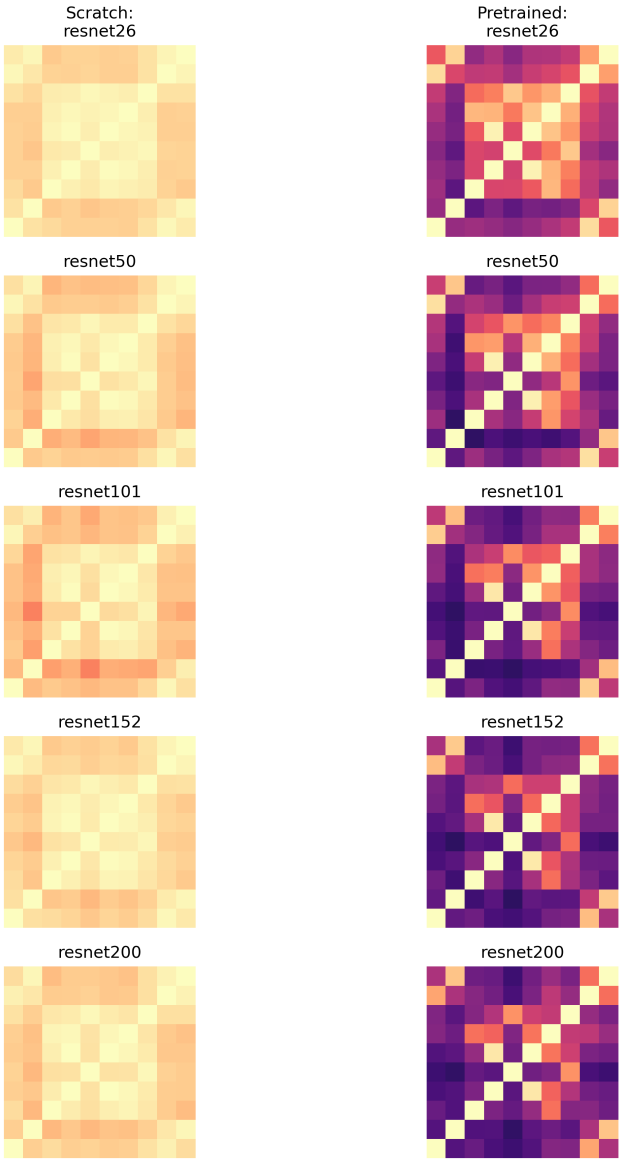

Figure 12: Classwise representation matrices for ResNet models, taken after the end of training on the first task (Task A) of a two-task split CIFAR-10 sequence. Classes are ordered alphabetically, left to right and down to up, i.e. the order is *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship*, *truck*.

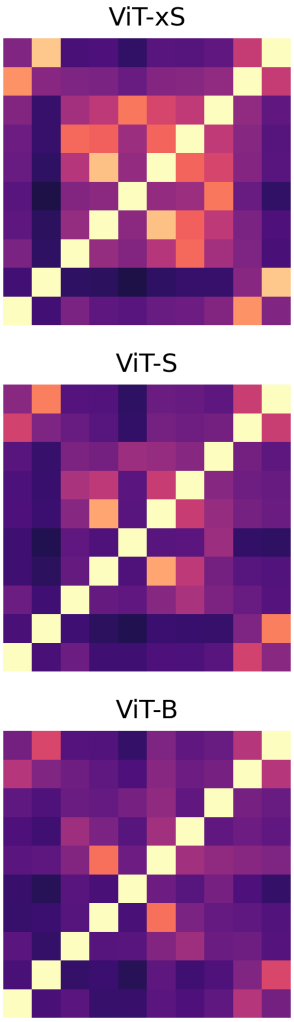

Figure 13: Classwise representation matrices for Vision Transformer models, taken after the end of training on the first task (Task A) of a two-task split CIFAR-10 sequence. Classes are ordered alphabetically, left to right and down to up, i.e. the order is *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship*, *truck*.

B.3 More language results

For completeness, Figure 14 displays the token accuracy on task A as a function of the number of steps.

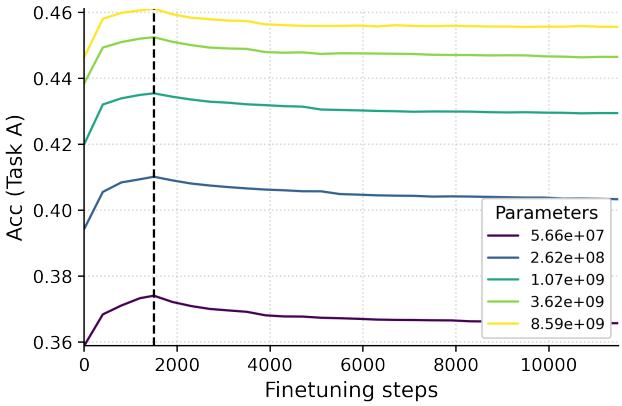

Figure 14: **Forgetting in sequential language-modeling**. Token accuracy as a function of steps for the different models.