ANATOMY OF CATASTROPHIC FORGETTING: HIDDEN REPRESENTATIONS AND TASK SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Catastrophic forgetting is a recurring challenge to developing versatile deep learning models. Despite its ubiquity, there is limited understanding of its connections to neural network (hidden) representations and task semantics. In this paper, we address this important knowledge gap. Through quantitative analysis of neural representations, we find that deeper layers are disproportionately responsible for forgetting, with sequential training resulting in an erasure of earlier task representational subspaces. Methods to mitigate forgetting stabilize these deeper layers, but show diversity on precise effects, with some increasing feature reuse while others store task representations orthogonally, preventing interference. These insights also enable the development of an analytic argument and empirical picture relating forgetting to task semantic similarity, where we find that maximal forgetting occurs for task sequences with intermediate similarity.

1 INTRODUCTION

While the past few years have seen the development of increasingly versatile machine learning systems capable of learning complex tasks (Stokes et al., 2020; Raghu & Schmidt, 2020; Wu et al., 2019b), *catastrophic forgetting* remains a core capability challenge. Catastrophic forgetting is the ubiquitous phenomena where machine learning models trained on non-stationary data distributions suffer performance losses on older data instances. More specifically, if our machine learning model is trained on a sequence of tasks, accuracy on earlier tasks drops significantly. The catastrophic forgetting problem manifests in many sub-domains of machine learning including continual learning (Kirkpatrick et al., 2017), multi-task learning (Kudugunta et al., 2019), standard supervised learning through input distribution shift (Toneva et al., 2019; Snoek et al., 2019; Rabanser et al., 2019; Recht et al., 2019) and data augmentation (Gontijo-Lopes et al., 2020).

Mitigating catastrophic forgetting has been an important research focus (Goodfellow et al., 2013; Kirkpatrick et al., 2017; Lee et al., 2017; Li et al., 2019; Serrà et al., 2018; Ritter et al., 2018; Rolnick et al., 2019), but many methods are only effective in specific settings (Kemker et al., 2018), and progress is hindered by limited understanding of catastrophic forgetting‘s fundamental properties. How does catastrophic forgetting affect the hidden representations of neural networks? Are earlier tasks forgotten equally across all parameters? Are there underlying principles common across methods to mitigate forgetting? How is catastrophic forgetting affected by (semantic) similarities between sequential tasks? This paper takes steps to answering these questions, specifically:

1. With experiments on split CIFAR-10 and a novel distribution-shift CIFAR-100 variant, we analyze neural network layer representations, finding that higher layers are disproportionately responsible for catastrophic forgetting, the sequential training process erasing earlier task subspaces.
2. We investigate different methods for mitigating forgetting, finding that while all stabilize higher layer representations, some methods encourage greater feature reuse in higher layers, while others store task representations as orthogonal subspaces, preventing interference.
3. We study the connection between forgetting and task semantics, finding that semantic similarity between subsequent tasks consistently controls the degree of forgetting.
4. Informed by the representation results, we construct an analytic model that relates task similarity to representation interference and forgetting. This provides a quantitative empirical measure of task similarity, and together these show that forgetting is most severe for tasks with intermediate similarity.

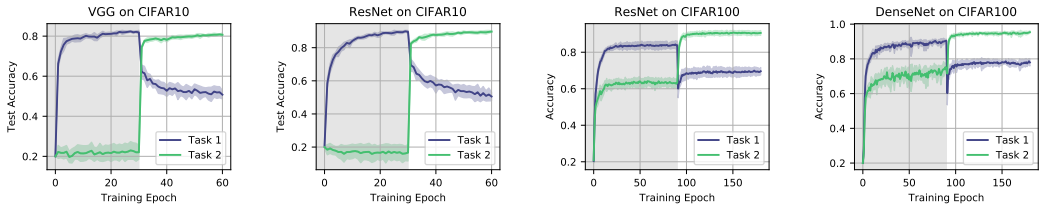

Figure 1: **Examples of catastrophic forgetting across different architectures and datasets.** We plot accuracy of Task 1 (purple) and Task 2 (green), on both the split CIFAR10 task, and the CIFAR100 distribution-shift task across multiple architectures. Catastrophic forgetting is seen as the significant drop of Task 1 accuracy when Task 2 training begins.

2 Related Work

Mitigation strategies Developing mitigation strategies for catastophic forgetting is an active area of research (Kemker et al., 2018). Popular approaches based on structural regularization include Elastic Weight Consolidation (Kirkpatrick et al., 2017) and Synaptic Intelligence (Zenke et al., 2017). An alternative, functional regularization approach is based on storing and replaying earlier data in a replay buffer (Schaul et al., 2015; Robins, 1995; Rolnick et al., 2019). We study how these mitigation methods affect the internal network representations to better understand their role in preventing catastrophic forgetting.

Understanding catastrophic forgetting Our work is similar in spirit to existing empirical studies aimed at better understanding the catastrophic forgetting phenomenon (Goodfellow et al., 2013; Toneva et al., 2019). We focus specifically on understanding how layerwise representations change, and on the relation between forgetting and task semantics, which have not previously been explored. This semantic aspect is related to recent work by Nguyen et al. (2019) examining the influence of task sequences on forgetting. It also builds on work showing that learning in both biological and artificial neural networks is affected by semantic properties of the training data (Saxe et al., 2019; Mandler & McDonough, 1993).

Fine-tuning and transfer learning While not studied in the context of catastrophic forgetting, layerwise learning dynamics have been investigated in settings other than continual learning. For example, Raghu et al. (2017) showed that layerwise network representations roughly converge *bottom-up*; (from input to output). Neyshabur et al. (2020) observed similar phenomena in a transfer-learning setting with images, as have others recently in transformers for NLP (Wu et al., 2020; Merchant et al., 2020). Furthermore, the observation that early layers in networks learn general features like edge detectors, while latter layers learn more task specific features is a well-known result in computer vision (Erhan et al., 2009), and here we quantitatively study its ramifications for sequential training and catastrophic forgetting.

3 Setup

Tasks: We use two tasks throughout this study. First, in the split CIFAR-10 task, we split the full ten-class dataset into m sequential tasks of n classes each. We generally uses $m = 2$ tasks, with each task comprising five classes. In this setting, our models consist of m classification heads, one for each task, along with a shared network *body*. Second, we introduce a CIFAR-100 based task to approximate *input distribution shift*. Here an individual task consists of distinguishing between n CIFAR-100 *superclasses*, which have training/test (input) data corresponding to a *subset* of their constituent classes. Different tasks have different subsets of constituent classes making up the input data, giving an input distribution shift across tasks. Full details are in Appendix A.3.

Models We perform experiments with three common neural network architectures used in image classification — VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2015) and DenseNet (Huang et al., 2016). Examples of catastrophic forgetting in these models are shown in Figure 1.

4 Catastrophic Forgetting and Hidden Layer Representations

We begin by investigating how catastrophic forgetting manifests in the hidden representations of the neural network. Do all parameters (and layers) forget equally? Or are specific components

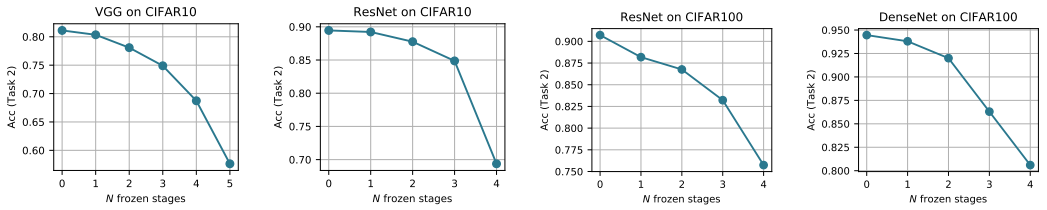

Figure 2: **Freezing lower layer representations after Task 1 training has little impact on Task 2 accuracy.** We freeze the parameters of a contiguous block of layers (starting from the lowest layer) after Task 1 training, and only train the remainder on Task 2, finding that freezing the lowest layers has little impact on Task 2 accuracy.

of the network particularly responsible for the drop in accuracy in sequential training? Through a combination of layer freezing experiments, and representational analysis, we find that *higher layers* (layers closest to output) are disproportionately responsible for catastrophic forgetting, with *lower layers* (layers closest to input) remaining representationally stable through sequential training. Further analysis of the representational subspaces of Task 1, Task 1 (post Task 2 training) and Task 2, provide insights on both feature reuse and interference.

4.1 Freezing Layer Representations

To study the effect of individual layers on forgetting, we measure the effect of freezing layer representations on Task 2 accuracy (Figure 2). Specifically, we freeze a contiguous block of layers (starting from the lowest layer) after training on Task 1, and only train the remaining layers on Task 2. Across different architectures and tasks, we observe that lower layers can be reliably frozen with very little impact on Task 2 accuracy. This suggests the possibility of lower layer features being reused between both tasks, with higher layers being the main contributor to catastrophic forgetting.

4.2 Representational Similarity through Sequential Training

The results of Figure 2 illustrate that lower layer representations on Task 1 can be reused for good performance on Task 2. To determine if this is what actually occurs during the training process, we turn to Centered Kernel Alignment (CKA) (Kornblith et al., 2019), a neural network representation similarity measure. CKA and other related algorithms (Raghu et al., 2017; Morcos et al., 2018) provide a scalar score (between 0 and 1) determining how similar a pair of (hidden) layer representations are, and have been used to study many properties of deep neural networks (Gotmare et al., 2018; Kudugunta et al., 2019; Wu et al., 2019a).

Specifically, letting $X \in \mathbb{R}^{nxp_1}$ and $Y \in \mathbb{R}^{nxp_2}$ be (centered) layer activation matrices of (the same) n datapoints and p_1, p_2 neurons, CKA computes

$$\frac{HSIC(XX^T, YY^T)}{\sqrt{HSIC(XX^T, XX^T)}\sqrt{HSIC(YY^T, YY^T)}}$$

for HSIC Hilbert-Schmidt Independence Criterion (Gretton et al., 2005). We use linear-kernel CKA.

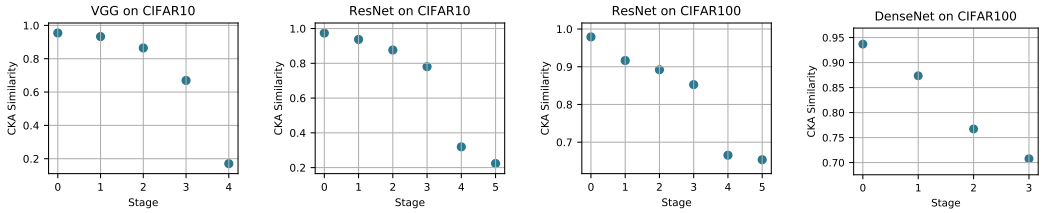

Figure 3: **CKA analysis of layer representations before/after Task 2 training shows lower layer feature reuse and higher layer representations changing significantly.** We compute CKA of the Task 1 layer representations before and after Task 2 training. We observe that later layers change more than early layers through second task training.

In Figure 3, we plot the results of computing CKA on Task 1 layer representations before and after Task 2 training. Across architectures and tasks, we observe that the lower layers have high representation similarity, suggesting that lower layer Task 1 features are reused in Task 2. The higher layers

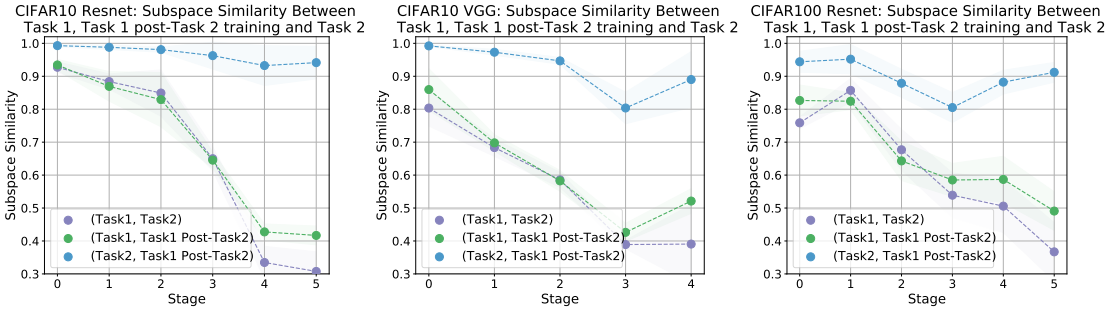

Figure 4: **Subspace similarity analysis shows feature reuse in lower layers and Task 1 subspace erasure in higher layers.** We compute subspace similarity between (i) (Task 1, Task 2) (purple line) (ii) (Task 1, Task 1 post-Task 2 training) (blue line) (iii) (Task 2, Task 1 post-Task 2 training) (blue line). All comparisons show high similarity in lower layers, indicative of feature reuse. Most striking is the comparison between (Task 1, Task 1 post-Task 2 training) and (Task 2, Task 1 post-Task 2 training), which show that after Task 2 training, higher layer representations for Task 1 are more similar to Task 2 than to Task 1! Specifically, the Task 1 subspace in higher layers is erased through Task 2 training.

however show significant decreases in representation similarity, suggesting they disproportionately contribute to catastrophic forgetting. These conclusions are further supported by additional layer reset experiments in Appendix Figure 11, which shows that rewinding higher layer parameters from post-Task 2 training to their pre-Task 2 training values significantly improves Task 1 performance.

4.3 Feature Reuse and Subspace Erasure

Further insights on how the representations of lower and higher layers evolve during sequential training is given through a subspace similarity analysis. Letting $X \in \mathbb{R}^{n \times p}$ be the (centered) layer activation matrix of n examples by p neurons, we compute the PCA decomposition of X, i.e. the eigenvectors $(v_1, v_2, ...)$ and eigenvalues $(\lambda_1, \lambda_2 ...)$ of $X^T X$. Letting V_k be the matrix formed from the top k principal directions, $v_1, ..., v_k$ as columns, and U_k the corresponding matrix for a different activation matrix Y, we compute

$$\text{SubspaceSim}_k(X, Y) = \frac{||V_k^T U_k||_2^2}{||V_k||_2 \cdot ||U_k||_2} .$$

This measures the overlap in the subspaces spanned by $(v_1, ..., v_k)$ and $(u_1, ..., u_k)$. Concretely, if X and Y correspond to layer activation matrices for two different tasks, $\text{SubspaceSim}_k(X, Y)$ measures how similarly the top k representations for those tasks are *stored* in the network.

Figure 4 shows the result of computing $\text{SubspaceSim}_k(X, Y)$ for X, Y being the layer activation matrices in (i) (Task 1, Task 2) (purple line) (ii) (Task 1, Task 1 post-Task 2 training) (blue line) (iii) (Task 2, Task 1 post-Task 2 training) (blue line). Supporting the CKA analysis, lower layers have high subspace similarity between Task 1 and Task 2 (feature reuse), but higher layers have low similarity (representations change significantly during Task 2 training.) Furthermore, (Task 1, Task 1 post-Task 2 training) also shows low subspace similarity in the higher layers, indicating that the representational change in later layers is not restricted to Task 2 data, but also alters the networks representation of the initial task. Strikingly, (Task 2, Task 1 post-Task 2 training) shows high similarity in all layers, despite being *different tasks*.

In summary, these results illustrate that during sequential training, effective feature reuse happens in the lower layers, but in the higher layers, after Task 2 training, Task 1 representations are mapped into the *same* subspace as Task 2. Specifically, Task 2 training causes *subspace erasure* of Task 1 in the higher layers. (Additional details and results are in Appendix Section A.5.)

5 Forgetting Mitigation Methods and Representations

Informed by this understanding of how catastrophic forgetting manifests in the hidden layers (feature reuse in lower layers and subspace erasure in higher layers), we next study methods to mitigate forgetting. Do popularly used mitigation methods act to stabilize higher layers — given possible alternate approaches such as weight orthogonalization (Appendix C)?

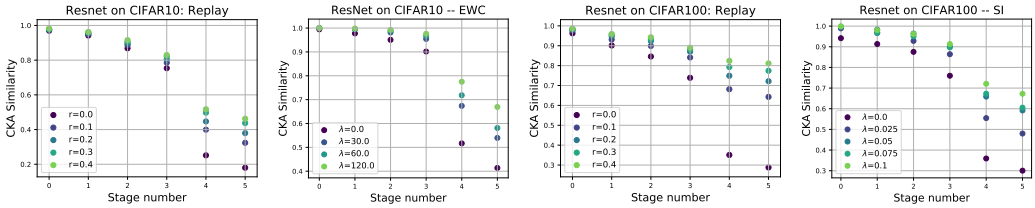

Figure 5: **CKA analysis of mitigation methods demonstrates that mitigation methods stabilize the higher layer representations.** We compute CKA between layer representations of Task 1 before and after Task 2 training, with varying amounts and types of mitigation. We find that across all mitigation methods, even a small amount of mitigation works to stabilize the higher layer representations. In Figure 6, we investigate whether this stabilization results in greater feature reuse or through finding independent subspaces.

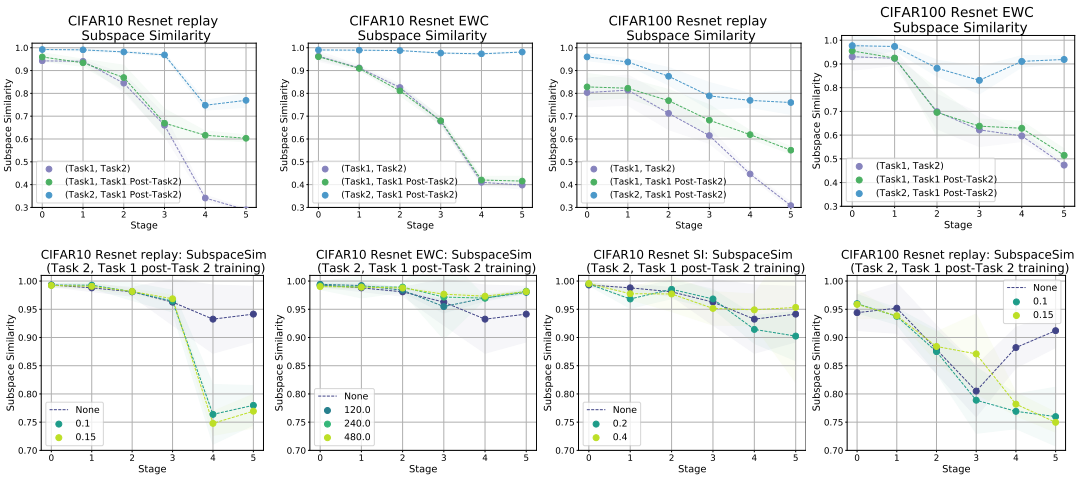

Figure 6: **Subspace similarity analysis reveals differences between different mitigation methods, with replay storing Task 1 and Task 2 representations in orthogonal subspaces, while EWC and SI promote feature reuse in the higher layers.** Top row: We plot subspace similarities like in Figure 4, for different mitigation methods. We observe that (Task 2, Task 1 post-Task 2 training) is much lower in replay compared to EWC/SI, as is (Task1, Task2) similarity. The bottom row plots (Task 2, Task 1 post-Task 2 training) for different amounts of mitigation. Again we see that replay has much lower (Task 2, Task 1 post-Task 2 training) similarity than no mitigation, while EWC/SI maintain the similar subspaces of Task 1 and Task 2 post Task 2 training.

We investigate these questions across different successful mitigation strategies (i) EWC (elastic weight consolidation) Kirkpatrick et al. (2017), a regularization-based approach (ii) replay buffer Schaul et al. (2015); Ratcliff (1990); Rolnick et al. (2019), a rehearsal-styled approach (iii) SI (Synaptic Intelligence) Zenke et al. (2017), another popular regularization-based approach. These mitigation methods have been specifically designed for the multi-head setting, so our experiments look at applying them to both split CIFAR-10 and split CIFAR-100.

In Figure 5, we plot the results of computing CKA on layer representations of Task 1 before/after training on Task 2, with different amounts of mitigation applied, and across different mitigation methods, tasks and architectures (full results in Appendix B.7.) Across all of these settings, we observe that even a small amount of mitigation results in a significant increase in similarity (stabilization) in the higher layers. However, there remains a central open question — is this increase in similarity due to *feature reuse in the higher layers* between Task 1 and Task 2, or does the network store Task 1 and Task 2 representations in non-interfering, *orthogonal subspaces*?

To answer this, we compute the Subspace Similarity measure introduced in Section 4.3, with results in Figure 6. Interestingly, the results show differences between the mitigation methods. When using replay buffer, the higher layer representations of Task 2 and Task 1 post-Task 2 training have much lower subspace similarity, indicating orthogonal network subspaces for the two tasks. In contrast, using EWC and SI shows high levels of subspace similarity in the higher layers, suggesting significant feature reuse between both tasks. Additional supporting experiments are in Appendix Section B.5.

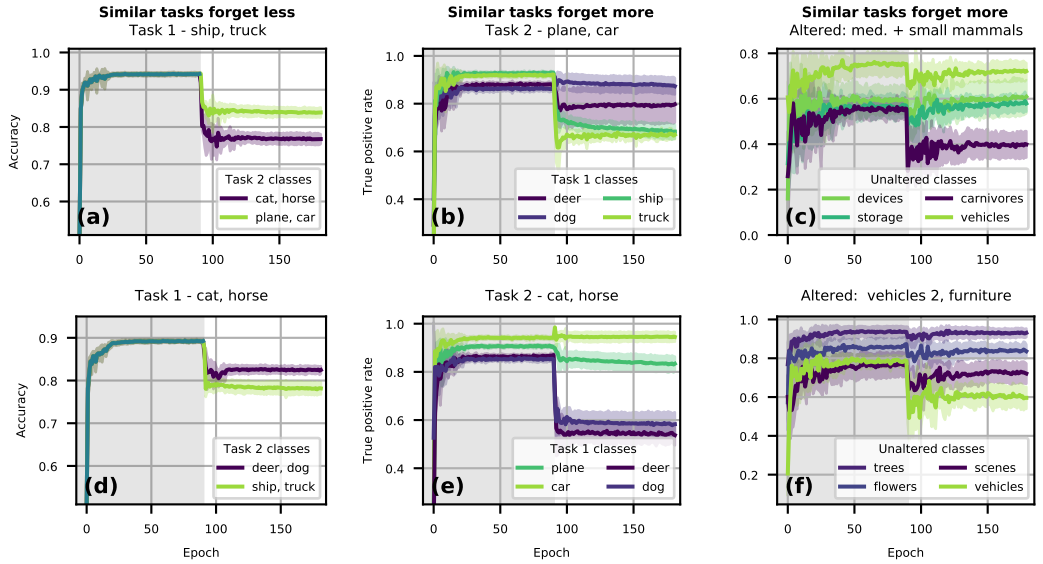

Figure 7: **Forgetting severity consistently varies along semantic divisions.** (a,d) For sequential binary classification tasks models trained on either objects or animals forgets less on a second object task of the same semantic category. (b,e) Models trained initially on four way classification of two object and two animal classes see more forgetting in semantically similar classes when trained on a second two class task of either objects or animals. (c,f) Models trained on sequential six-way classification with four unaltered categories (shown) and two altered categories show increased forgetting in the unaltered categories semantically most similar to altered categories. For all panes, performance is shown for ResNet, with additional results in Appendix B.9.

6 Semantics

With the insights on how catastrophic forgetting relates to the hidden representations, and the effects on network internals of different mitigation methods, we turn to investigating how the semantic similarity between the sequential tasks affects forgetting. Specifically the prior results on lower layer feature reuse suggest a key question: does similarity between tasks result in less forgetting? Surprisingly, the answer is quite nuanced, depending on how precisely the different tasks are represented in the network.

6.1 Forgetting, Task Semantics and a Similarity Puzzle

Our study of the relationship between task semantic similarity and forgetting reveals a surprising puzzle. In Figure 7, we observe that depending on the task setup, a semantically similar task may be forgotten less *or* more than a less similar task. This behavior is consistent across different settings and datasets. Specifically, in Figure 7a and7d, Task 1 and Task 2 consist of binary classifications between animals and objects from CIFAR-10, where we find that the similar categories forget less. But in Figure 7b and 7e, a four category classification between animals and objects followed by a binary classification task results in the similar categories being forgotten more. This effect of similar categories being forgotten more can also be observed in the CIFAR-100 distribution shift setting, where input distribution shift to the CIFAR-100 superclasses results in more similar categories being forgotten more, e.g. input shifts to natural superclasses (Figure 7c) or artifical superclasses (Figure 7f). Further details and supporting experiments can be found in Appendix B.9.

In the subsequent sections, we resolve this puzzle through (i) the development of an analytic model that formalizes the subtle connection between task similarity and forgetting (ii) an ensuing measure of task similarity, *trace overlap*, empirically quantifies task similarity (iii) showing maximal forgetting happens with intermediate task similarity.

6.2 An Analytic Model of Forgetting Semantics

We first formalize a model with a single classification head (like in distribution shift CIFAR-100), with the corresponding multi-head setup in Appendix C. We write the network output as $f(x) = \sum_{\mu=1}^{p} \theta_\mu g_\mu(w; x)$, where θ are the last layer weights and g are features. We consider training

sequentially on two tasks. For the first task, we train normally. For the second task, inspired by our empirical observation above that forgetting is driven by deeper layers, we freeze the the earlier layer features, $g_\mu(w;x) = g_\mu(\hat{w};x)$ when training on the second task. If we train with loss function $L(f,y)$, the network output SGD update step is

$$\Delta f_t(x) = -\eta \sum_{x',y' \in \mathcal{D}^{(2)}_{\text{train}}} \Theta(x,x') \frac{\partial L(f(x'),y')}{\partial f} \,. \tag{1}$$

Here $\mathcal{D}^{(2)}_{\text{train}}$ is the second-task training data and $\Theta(x,x') = \sum_{\mu=1}^{p} g_\mu(\hat{w};x) g_\mu(\hat{w};x')$ is the *overlap* (inner product) between the model last-layer representations on data point x and x'. If the overlap $\Theta(x,x')$ is small, then the change in the logits is small and forgetting is minimal, specifically:

Lemma 1 *Let f_t be the logits of a neural network trained via the frozen feature model. Let x be an element of the initial task test set. Let us further denote the vector of the feature overlap applied to the second task training set as $\vec{\Theta}(x) := \{\Theta(x,x') : x' \in X^{(2)}_{train}\}$ and the loss vector as $\vec{L} = \{L(f(x'),y') : x',y' \in \mathcal{D}^{(2)}_{train}\}$. With this, the change in the original task logits is bounded as*

$$|\Delta f_t(x)| \leq \eta \left|\left|\vec{\Theta}(x)\right|\right| \left|\left|\frac{\partial \vec{L}}{\partial f}\right|\right| \,. \tag{2}$$

Thus representational similarity is necessary for forgetting to occur, and in this model, sufficiently similar and sufficiently dissimilar tasks have minimal forgetting, while intermediate similar tasks have maximal forgetting.

6.3 Catastrophic Forgetting across varying Task Similarities

The analytical result enables us to make precise a measure of task similarity, through using the feature overlap matrix $\Theta(x',x)$. Concretely, letting Θ_{11} be the overlap matrix on Task 1, Θ_{22} on Task 2, and Θ_{12} on Task 1 and Task 2, we define:

$$\text{Trace overlap} = \frac{\text{Tr}\left(\Theta_{12}\Theta_{12}^T\right)}{\sqrt{(\text{Tr}\left(\Theta_{11}\Theta_{11}^T\right)\text{Tr}\left(\Theta_{22}\Theta_{22}^T\right))}} \,. \tag{3}$$

Trace overlap is directly analogous to CKA, where instead of comparing the similarity of different features on the same data, we are comparing the similarity of common features on different data.

Increasing task *dissimilarity* can help reduce forgetting In Lemma 1, we see that very dissimilar tasks, with small overlap matrix, exhibit minimal forgetting. In Figure 8, we test this in the context of sequential binary classification. We introduce an *other* category to the classification problem (of images from classes not included in the training tasks), which encourages the model to represent Task 1 and Task 2 dissimilarly (lower trace overlap in Figure8b), and reduces forgetting (Figure8a).

Maximal forgetting at intermediate similarity To empirically study the conclusion from the model that intermediate-similar tasks have maximal forgetting, we use mixup (Zhang et al., 2017), a standard data augmentation technique, to interpolate between the two tasks. Specifically, we first train a model on binary classification (with an other category to encourage dissimilarity) and for $\lambda \in [0,1]$, generate an mixup dataset $\mathcal{D}^{(\lambda)}_{\text{train}} = \{(\lambda x^{(2)} + (1-\lambda)x^{(1)}, \lambda y^{(2)} + (1-\lambda)y^{(1)})\}$. We compare model behavior across a range of mixup fractions, finding that the task similarity as measured through the trace overlap between task1 and task2 varies monotonically with mixup fraction (Figure 9b), quantifying that varying the mixing parameter, λ, tunes the degree of task similarity. Figure 9a shows that the most extreme forgetting happens for intermediate similarity (i.e. mixing fraction $\sim 0.1 - 0.2$).

Semantic similarity revisited Armed with these quantitative measures, we revisit the task semantics puzzle in Figure 7 where (i) for sequential binary classification, semantically more similar categories were forgotten less, while in (ii) four class classification followed by binary classification and input distribution shift, similar classes were forgotten more. Figure 9c (solid dots) shows that in (i) the net didn't represent objects and animals differently, and the slightly more similar task is forgotten less. But in (ii) Figure 9c (hollow dots), the model encodes animals and objects differently, minimizing forgetting on the dissimilar tasks, consistent with maximal forgetting for intermediate similarity.

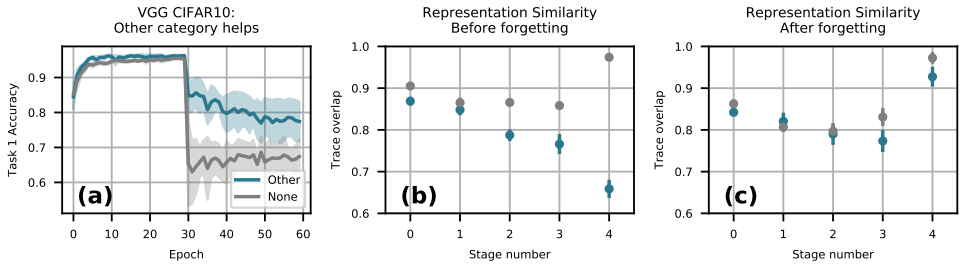

Figure 8: **Adding an *other* category reduces forgetting and increases orthogonalization**. (a) Models trained initially on classifying two objects: airplane/automobile, and subsequently on two animals: deer/dog exhibit diminished forgetting when we add an *other* category to task 1 training. The other category contains randomly sampled images from all classes not present in the initial or final tasks. (b,c) The presence of the other category reduces the similarity between the model's representations of each task. Additional plots are in Appendix B.8.

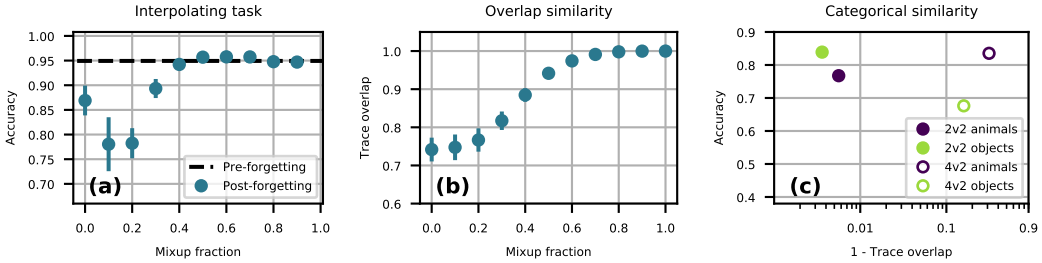

Figure 9: **Task similarity affects performance.** (a) We measure initial task performance for sequential binary classification as we dial the mixup fraction between 0.0, pure task 2 data, and 1.0, pure task 1 data. Maximum forgetting (minimal accuracy) occurs at an intermediate value of the mixup fraction, $\sim 0.1 - 0.2$. (b) We measure the similarity between task 1 and task 2 as we vary the mixup fraction, supporting the intuition that interpolating mixup fraction interpolates task similarity. (c) Task 1 performance after task 2 training is plotted against task 1 versus task 2 representational similarity for animal and object classes in the sequential binary classification tasks (solid) and four-way followed by binary classification tasks (hollow) of Figures 7a and Figures 7b. In the case of sequential binary tasks both object and animal second tasks are represented extremely similarly to the initial object task and the slightly less similar, object task, forgets more. For the four way classification task, animals are represented dissimilarly to the second object task and cause less forgetting.

7 Conclusion

In this paper, we have studied some of the fundamental properties of catastrophic forgetting, answering important open questions on how it arises and interacts with hidden representations, mitigation strategies and task semantics. By using representational analysis techniques, we demonstrated that forgetting is not evenly distributed throughout the deep learning model but concentrated at the higher layers, which change significantly and erase earlier task subspaces through sequential training. Consistent with this, we find mitigation methods all stabilize higher layer representations, but vary on whether they enforce more feature reuse, or store tasks in orthogonal subspaces. Informed by these insights, we formulate an analytic model to investigate connections between forgetting and task semantics. This gives a quantitative measure of task similarity and shows that intermediate similarity in sequential tasks leads to maximal forgetting. In summary, these results help provide a foundation for a deeper understanding of catastrophic forgetting and suggest new approaches for developing and measuring mitigation methods.

REFERENCES

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

Dar Gilboa and Guy Gur-Ari. Wider networks learn better features, 2019.

Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973*, 2020.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*, 2019.

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

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. *arXiv preprint arXiv:1904.00310*, 2019.

Jean M. Mandler and Laraine McDonough. Concept formation in infancy. *Cognitive Development*, 8(3):291 – 318, 1993. ISSN 0885-2014. doi: https://doi.org/10.1016/S0885-2014(93)80003-C. URL http://www.sciencedirect.com/science/article/pii/S088520149380003C.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to bert embeddings during fine-tuning?, 2020.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning?, 2020.

Cuong V Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: an empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, pp. 1394–1406, 2019.

Maithra Raghu and Eric Schmidt. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pp. 6076–6085, 2017.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97 2:285–308, 1990.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, pp. 3738–3748, 2018.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2): 123–146, 1995.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, pp. 348–358, 2019.

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, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1820226116. URL `https://www.pnas.org/content/116/23/11537`.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackerman, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. *ArXiv*, abs/1812.05159, 2019.

John M. Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Similarity analysis of contextual word representation models, 2020.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*, 2019a.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019b.

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. org, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.