An Empirical Investigation Towards Efficient Multi-Domain Language Model Pre-training

Kristjan Arumae, Qing Sun, & Parminder Bhatia
Amazon
Seattle, USA
{arumae, qinsun, parmib}@amazon.com

Abstract

Pre-training large language models has become a standard in the natural language processing community. Such models are pre-trained on generic data (e.g. BookCorpus and English Wikipedia) and often fine-tuned on tasks in the same domain. However, in order to achieve state-of-the-art performance on out of domain tasks such as clinical named entity recognition and relation extraction, additional in domain pre-training is required. In practice, staged multi-domain pre-training presents performance deterioration in the form of *catastrophic forgetting* (CF) when evaluated on a generic benchmark such as GLUE. In this paper we conduct an empirical investigation into known methods to mitigate CF. We find that elastic weight consolidation provides best overall scores yielding only a 0.33% drop in performance across seven generic tasks while remaining competitive in bio-medical tasks. Furthermore, we explore gradient and latent clustering based data selection techniques to improve coverage when using elastic weight consolidation and experience replay methods.

1 Introduction

Transformer (Vaswani et al., 2017) based language modeling has taken over many previous pre-training and initialization approaches (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Liu et al., 2019). Fine-tuning using these architectures yields state-of-the-art results in the order of a few hours. The caveat to these models is that the initial training can be on the scale of many days if not weeks, distributed across multiple GPUs (Strubell et al., 2019), a costly endeavour.

Pre-trained language models are adapted to perform strongly in more specific domains as well. For example, while the original BERT models (Devlin et al., 2019) were trained on English Wikipedia articles and BooksCorpus (Zhu et al., 2015), the same masked language modeling was continued on bio-medical data. BioBERT (Lee et al., 2019) was trained using Pubmed abstracts and full articles, meanwhile Clinical BERT (Alsentzer et al., 2019) was further refined using MIMIC-III clinical notes (Johnson et al., 2016). Evidence suggest that understanding the syntactic structure of scientific literature and clinical data from pre-training boosts performance in their respective downstream tasks (Peng et al., 2019). Pre-training is performed with the expectation of building robust, high capacity generalized language models which continue to absorb new domain knowledge.

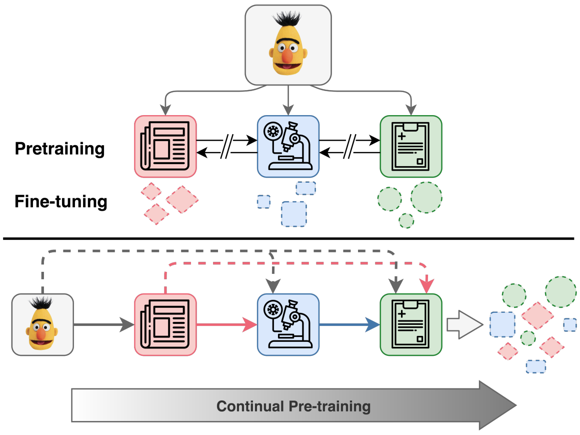


Figure 1: Traditional approaches (top) train independent domain specific language models (newswire, bio-medical, and clinical) which share no cross domain knowledge. They are further fine-tuned on their respective in-domain tasks. Our approach (bottom) shows how several domains are introduced in sequence, with knowledge retention using mitigation techniques across all domains. Here the final model has the capability to properly fine-tune on any domain specific task.

Unfortunately, continual learning (Ring, 1997) suffers from catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990) when incorporating domain data in a sequential manner. Parameters shift towards capturing the current task (or domain) and if previous data is no longer available the model will lose representation of it. For

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4854–4864,
November 16–20, 2020. ©2020 Association for Computational Linguistics

many tasks the straightforward solution is to combine datasets during training and approach this as a multi-task learning (MTL) (Ruder, 2017) problem. Mixing data has the desired effect of constraining parameters to find a space where both tasks reach close to optimal performance.

We argue that these expensive pre-trained models are an example where MTL is not feasible in practice for several reasons. Time and hardware accessibility are the largest constraints for developing such systems. Access to large scale training data is generally not possible (Radford et al., 2019; Devlin et al., 2019), and exact training configurations are equally difficult to gather with results being arduous to reproduce. Resource usage has recently been criticized from another perspective as well. Strubell et al. (2019) show that as deep neural architectures in the natural language community grow we increasingly trade results for carbon emissions.

Our work conducts an empirical investigation into suitable methods for multi-domain pre-training in a continual learning setting. We focus our efforts towards three methods: (i) elastic weight consolidation (EWC), (ii) learning rate control (LRC), and (iii) experience replay (ER). EWC (Kirkpatrick et al., 2017) is a parameter constraining method, an upgrade to vanilla regularization (e.g. L_2). LRC is borrowed from stage two of ULMFiT (Howard and Ruder, 2018) pre-training as a data independent method. Finally, as a scaled back version of MTL we investigate experience replay (ER), reintroducing data at a fixed scale from previous domains during pre-training. Furthermore we explore data selection approaches to improve efficiency for both ER, and EWC.

Our goal is to understand the trade-offs across these models in terms of resources and setup. To this end we conduct experiments across multiple domain shifts while pre-training. To evaluate the efficacy of the methods we use downstream fine-tuning tasks in the domains we study. To better understand how knowledge across domains is transferred, we perform layer-wise analysis and observe that outer layer are the most transferable.

Our contributions are as follows [1]:

- We provide empirical evidence of catastrophic forgetting mitigation with experience replay, learning rate control, and elastic weight consolidation, applied towards large scale language model pre-training. To this we add multiple domain shifts into bio-medical, and clinical data.
- We explore various data selection approaches for both elastic weight consolidation and replay based models.
- We investigate layer-wise understanding for continual pre-training across several domains to understand how best to mitigate forgetting and transfer knowledge understanding.

2 Continual Learning

We empirically study three forms of mitigation for catastrophic forgetting. Constraint based training in the form of EWC and learning rate control, and experience replay.

2.1 Elastic Weight Consolidation

EWC makes use of a simple Bayesian factorization of model representation (Kirkpatrick et al., 2017). This isolates the posterior of a learned task (A) while maintaining the objective of a current task (B). Due to the intractability of the true posterior, EWC makes use of a Fisher information (Frieden, 2004) matrix diagonal to approximate the effect of Task A on the parameters of a model. Intuitively speaking, if a parameter had a large effect on task A the Fisher value would be small yielding low variance to adapt to task B. This holds true inversely for when the Fisher value is large.

In practice, we initialize the Fisher matrix using gradients calculated with data sampled from Task A, which has already converged (Spall, 2005). This is demonstrated in Eq. 1 where i and j index parameters and data samples respectively.

$$F_{i,i} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\partial \mathcal{L}_A^{(j)}}{\partial \theta_i}\right)^2 \tag{1}$$

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \lambda F_{i,i}(\theta_i - \theta^*_{A,i})^2 \tag{2}$$

The full objective for task B is given in Eq. 2 where $\mathcal{L}_B(\theta)$ is the loss function of Task B, and EWC is represented as the second term regularizing model parameters. Specifically by weighting the shift of model parameters while training on Task B (here θ_i and $\theta^*_{A,i}$ being the currently updated and frozen Task A parameters at index i respectively). The EWC objective component is further adjusted by the hyperparameter λ.

[1] Our code is avaialble at https://github.com/aws-health-ai/multi_domain_lm

Model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI
$\text{BERT}_{\text{BASE}}$	57.82	92.09	86.74	88.13	87.49	84.01	90.79	64.98	53.52
BioBERT	37.78	89.68	88.44	87.40	86.96	83.19	89.79	60.29	28.17
Delta	20.04	2.41	-1.69	0.73	0.53	0.82	1.01	4.69	25.35

Table 1: Performance drop of BioBERT after further pre-training on Pubmed articles. The last row shows a positive value indicating the degree to which performance has dropped, and a negative value when it has increased.

2.2 Learning rate control

Our approach models the second stage of ULMFiT (Howard and Ruder, 2018), namely target task fine-tuning. We begin with a layer wise modifications by applying a decaying learning rate as a function of layer depth moving from the last layer towards model input.

$$\eta^{(l-1)} = \frac{\eta^{(l)}}{\rho} \tag{3}$$

Here η, l, and ρ denote learning rate, layer index and decay rate respectively. Depth plays a factor in our model since the network consists of 14 layers (i.e. 12 transformer layers, one layer for input, and one for the LM head).

2.3 Experience Replay

Given a replay buffer of a fixed, limited size we empirically investigate sample efficiency over a number of heuristic data selection methods. We focus our attention on how best to select data for this buffer, hypothesizing that domain coverage will increase performance. Recent work (de Masson d'Autume et al., 2019) has shown how this is crucial in strict lifelong learning when updating a fixed buffer size.

3 Catastrophic Forgetting in Language Modeling

We motivate our own experiments by first exploring off-the-shelf models to get a sense of the problem. To this end we fine tuned a $\text{BERT}_{\text{BASE}}$ architecture on all nine GLUE (Wang et al., 2018) tasks. These were compared directly against BioBERT, which has been further trained on full Pubmed articles. As reported in Table 1 an overall trend of performance deterioration is apparent with a relative increased error of 7.64% in the bio-medical model. Furthermore, we observed that on tasks which BERT struggles with, such as CoLA and WNLI, the performance decrease is amplified when switching pre-training domains.

4 Experimental Details

We first cover the data domains, fine-tuning tasks, and general modeling setup used in both our heuristic search as well as our main experiments in Section 6.2.2.

4.1 Pre-training Data

We processed publicly available bio-medical and non-bio-medical corpora for pre-training our models. For non-bio-medical data, we use BookCorpus and English Wikipedia data, CommonCrawl Stories (Trinh and Le, 2018), and OpenWebText (Gokaslan and Cohen, 2019). This combined corpus contains roughly 18B tokens. For bio-medical data, we use full Pubmed[2] articles which we processed to remove all tables, references, equations, and figures. This yields a dataset of over 4B tokens. For all datasets we retain training, validation, and test splits sampled at the document level with a respective ratio of 8:1:1.

4.2 Evaluation Data

We report the average accuracy across GLUE (Wang et al., 2018) tasks to track the performance of the model on generic natural language understanding. For measuring performance on GLUE, we further limit the selection of tasks to be the five most deteriorated (i.e. CoLA (Warstadt et al., 2018), SST-2 (Socher et al., 2013), MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016) and RTE (Giampiccolo et al., 2007)). Tasks such as QQP[3] and MRPC (Dolan and Brockett, 2005) are generally robust against domain change and perform well regardless of initialization. These five tasks reflect our findings from Table 1. Additionally we evaluate on CoNLL-03 (Tjong Kim Sang and De Meulder, 2003) named entity recognition (NER), and SQuAD 1.1 (Rajpurkar et al., 2016) question answering (QA). To demon-

[2] https://www.ncbi.nlm.nih.gov/pmc/

[3] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

strate domain shift we evaluate using BC5CDR (Li et al., 2016), Chemprot (Krallinger et al., 2017) and BioASQ (Nentidis et al., 2019) which are biomedical NER, relation extraction (RE), and QA tasks respectively. The first dataset is from the 2015 CDR challenge for identifying chemicals and diseases expertly annotated from Pubmed abstracts [4]. Chemprot contains annotations of chemical-protein reactions, also taken from Pubmed articles. Finally BioASQ appears in our paper using the same format and splits as described by Gu et al. (2020). Namely QA is treated as a binary classification of whether the answer to the query exists in the provided context.

4.3 Modeling

For modeling we use the RoBERTa architecture (Liu et al., 2019), and implement EWC, learning rate control, and experience replay changes directly into the model[5]. This extension of the original BERT removed next sentence prediction and is trained using only masked language modeling using very large batch sizes. We utilize all training hyperparameters as provided by Liu et al. (2019) unless otherwise noted, and use RoBERTa $_{\text{BASE}}$ as parameter initialization for all experiments. As a form of deterioration understanding, we continue to train a model using Pubmed articles (denoted as PMC) with no mitigation techniques.

5 Data Selection Methods

Data selection is an important component of both supervised, and unsupervised training. In our case, there is an abundance of data to build both the Fisher matrix, as well as the replay buffer. To do this efficiently for EWC and ER we need to severely restrict the number of datapoints we utilize. For example a mere 1.0% of generic pre-training data makes up over 400k segments. We require this subset to be comprehensively representative of the domain. Therefore, rather than randomly sampling data, we can use model generated features to induce better coverage of previous domains.

5.1 Gradient Analysis

We begin by treating the sum of squared gradients as a one-dimensional feature for data selection. The generic data is a skewed distribution with a mean at $1.04e^7$ and a standard deviation and max values of $4.89e^8$, and $1.82e^{11}$ respectively. The lower bound is, of course, 0 and arguably the samples closer towards that bound are more representative of the model in its generic state given this long tail.

	Sampling Type	GLUE	SQuAD	Avg.
	RoBERTa $_{\text{BASE}}$	87.56	90.20	88.00
	RoBERTa $_{\text{PMC}}$	83.00	88.73	83.95
ER	Random	84.23	89.43	85.10
	High	84.59	87.99	85.15
	Low	83.99	88.97	84.82
	Uniform	84.69	89.70	**85.53**
EWC	Random	86.93	90.32	87.50
	High	87.08	90.27	**87.61**
	Low	86.64	90.49	87.28
	Uniform	87.03	90.43	87.60

Table 2: Four sampling techniques used for pre-training and evaluated on GLUE and SQuAD 1.1. The results are compared against RoBERTa $_{\text{BASE}}$ and an unmitigated model trained on Pubmed articles (denoted using $_{\text{PMC}}$). The average column takes into account each of the individual GLUE tasks.

To be thorough we sampled data from this domain in four different ways: (i) randomly, (ii) low, (iii) high, and (iv) uniformly. For low and high sampling we order the samples according to this feature value and slice the list from the front or back. For uniform sampling we bin the data according to the gradient value, and sample from the bins uniformly, whereas random sampling is performed by treating all samples equally. For each of these experiments we sample 0.1% of the total corpus (roughly 42k segments). Furthermore in the same way that ER uses data to construct the replay buffer, EWC uses the samples to build the Fisher diagonal. We therefore test each sampling method across both mitigation techniques.

To test the effectiveness of our methods we pre-train RoBERTa $_{\text{BASE}}$ on one epoch of Pubmed data (with and without mitigation) and test retention performance by fine-tuning our models across GLUE and SQuAD 1.1. Looking at Table 2 we see that above all, using low gradients is the least useful signal. For ER, using uniform rather than low value selection has an average performance increase of 0.71 points. The other methods fall in line with uniform sampling indicating that including samples with larger gradients is helpful in representing of the source domain. EWC appears to be more robust

[4] We used a combined dataset: `https://github.com/cambridgeltl/bmip-2018`.

[5] `https://github.com/pytorch/fairseq/tree/master/examples/roberta`

	PCA	GMM	ER Avg.	EWC Avg.
<s>	50	5	85.04	87.46
	50	10	85.67	87.25
	100	5	85.46	**87.61**
	100	10	**85.74**	87.28
AVG. POOL	50	5	85.06	87.24
	50	10	85.04	87.20
	100	5	84.96	**87.83**
	100	10	**85.39**	87.24

Table 3: GLUE and SQuAD average performance for both ER and EWC when using two pooling techniques.

to data sampling with lower variance (1.8e−2 vs. 6.4e−2 for ER) across all models, with high and uniform selection improving most.

5.2 Sampling Latent Clusters

We further investigate more feature-rich representations in the form of sentence embeddings. Aharoni and Goldberg (2020) have demonstrated that transformer based LMs exhibit a keen ability to distinguish domains via clustering. The pre-training data for RoBERTa also comes from a variety of sources, with variation in prose, diction, and formality. We therefore cluster this data to see both how it is distributed and if uniformly sampling from these groups yields good performance for both EWC and ER.

Aharoni and Goldberg (2020) used average pooling across the last encoder layer to represent each segment, we test this method against using the vector representation of `<s>` (`[CLS]` in BERT) since it is frequently used in practice for sentence labeling. We then use PCA (Wold et al., 1987) to reduce the dimensionality to $d \in \{50, 100\}$ and apply a Gaussian Mixture Model (Reynolds, 2009) using $k \in \{5, 10\}$ as the number of clusters.

The resulting experiments for both ER and EWC can be seen in Table 3. Using PCA at 100 provides higher metrics for both ER and EWC, while the number of clusters for GMM does not give an interpretable signal across the experiments.

We note that from a practical perspective it is much faster to process data using clustering than gradients, largely due to the ability to batch data for clustering. Accumulating gradients for 1MM samples takes roughly five days using an NVIDIA V100, whereas acquiring latent representations from the same amount of data finishes in less than four hours (this does not account for PCA and clustering which takes an additional four to five hours).

6 Mitigation of Catastrophic Forgetting

We provide results for one and two stage domain shifts as given by fine-tuning tasks. Again, we apply mitigation only to pre-training and express our model performance by using them to fine-tune downstream tasks.

6.1 Setup

For a baseline and potential upper bound of performance we train a multi-domain learning (denoted as MDL) model which utilizes the full combined generic and bio-medical training sets as input data. For EWC (+EWC) we tune both λ [0.5, 1.0, 5.0, 10.0], and the size of the data used for fisher initialization [0.1%, 1.0%, 10.0%]; best values are underlined. For experience replay (+ER) we experiment with mixing non-bio-medical data (the same subset used for EWC init.) in each batch with a ratio proportional to their sizes. Additionally we showcase both a gradient based sampling (denoted with a subscript unif), and the GMM-PCA (subscript GMM) ($k = 5$, $d = 100$) for both ER and EWC. We tuned the decay rate, ρ in Eq. 3 [1.3, 1.7, 2.6] for LRC.

6.2 Results

Our experimental results are reported in Table 4. The first two rows contain the off-the-shelf RoBERTa as well as the PMC setting which received no catastrophic forgetting mitigation when further trained on bio-medical data. The lower section lists all mitigation based experimental settings as described above. For all models pre-trained using Pubmed data we fine-tune on tasks after a single epoch of pre-training.

We divide columns by task domain. The first three tasks (i.e. GLUE, SQuAD, and CoNLL) cover generic domain understanding. Just as in Section 5.1 we use the five worst GLUE tasks. For an overall understanding of forgetting we provide the average across all generic tasks. bio-medical tasks are displayed next followed by overall performance weighing the bio-medical and generic tasks equally [6]. NER and RE scores are reported using micro-F_1; all GLUE tasks we report accuracy on

[6] We take the mean of the generic and bio-medical average rather than treating each task equally since there are significantly more generic tasks.

	generic				bio-medical				
Model	GLUE	SQuAD	CoNLL	Avg.	BC5CDR	Chemprot	BioASQ	Avg.	Overall
RoBERTa $_{\text{BASE}}$	87.56	90.20	90.11	88.30	84.94	63.27	75.41	74.69	81.49
PMC	83.00	88.73	87.35	84.44	86.68	65.13	75.41	75.74	80.09
MDL	84.89	88.92	89.72	86.15	85.76	65.16	75.41	75.44	80.79
PMC +LRC	86.78	90.35	89.76	87.72	85.47	62.30	75.41	74.39	81.05
PMC +ER$_{\text{unif}}$	84.69	89.70	89.10	86.04	**87.20**	**67.40**	77.13	77.24	81.64
PMC +ER$_{\text{GMM}}$	84.25	88.50	89.78	85.65	86.83	63.70	**82.42**	**77.65**	81.65
PMC +EWC$_{\text{unif}}$	87.03	**90.43**	89.77	87.90	86.23	65.90	79.73	77.28	**82.59**
PMC +EWC$_{\text{GMM}}$	**87.08**	90.22	**90.46**	**88.01**	86.05	65.50	76.18	75.90	81.96

Table 4: Single stage domain adaptation. Other than RoBERTa $_{\text{BASE}}$, each model is pre-trained further on one epoch of bio-medical data. We average generic performance across five GLUE tasks, as well as QA (from SQuAD), and NER (CoNLL). The average across generic tasks considers all nine tasks equally. bio-medical performance is for BC5CDR (NER), Chemprot (RE), and BioASQ (QA) with the overall performance being the mean for bio-medical and generic averages.

the development set; SQuAD is evaluated using F_1; BioASQ uses accuracy.

6.2.1 Catastrophic Forgetting

Unsurprisingly among the first two rows RoBERTa $_{\text{BASE}}$ performs best overall on generic tasks with an average performance increase of 4.47% over the unmitigated (PMC) model. Conversely it underperforms on the bio-medical tasks, validating the need to further pre-train on domain specific data. When averaging across the three bio-medical tasks the PMC model has a 1.05 point F_1 edge. It should be noted here that four of the models achieved the same BioASQ F_1 score, this was not reported in error.

6.2.2 Mitigation Based Models

EWC and LRC both respond well during domain shifts, are our best candidates for combating catastrophic forgetting, and average only half a point in deterioration amongst the three of them when compared against RoBERTa $_{\text{BASE}}$. LRC has the benefit of tuning a single hyperparameter, the decay rate (ρ). Due to the depth of the models we found that a high value ($\rho = 2.6$) yields a model which has a negligible drop in performance for generic tasks (with an average of 88.28) but had a more difficult time with later domains.

We observed during hyper-parameter optimization that EWC was quite sensitive to λ values. With higher coefficients ($\lambda > 1.0$) EWC was able to halt deterioration nearly completely but performed quite poorly on bio-medical tasks. To better understand the importance of the Fisher values, we trained EWC with no Fisher (i.e removing $F_{i,i}$ from Eq. 2). We found that this resulted in less competitive bio-medical results (averaging 3.68% worse than the listed bio-medical EWC scores, and having overall the worst scores for both bio-medical tasks across all models), illustrating that giving equal weight to all the parameters results in poor generalization across source and target domains. MDL performed surprisingly average compared to the resource trade-off of the model. While it does produce better results than RoBERTa $_{\text{BASE}}$ in the bio-medical domain, the model struggles to retain generic knowledge. Experience replay grapples most with domain retention and produced the highest mitigated BC5CDR, Chemprot, and BioASQ results coupled with the lowest generic results.

When comparing sampling techniques across a larger number of fine-tuning experiments we echo results from Section 5. Experience replay is stronger when using gradient based sampling, while EWC functions better using clustered latent representations. Therefore, in practice, we would suggest latent representations for better efficiency.

6.2.3 Two Stage Domain Adaptation

To further evaluate mitigation methods we continue pre-training models using clinical data. We chose the clinical domain since although it may appear close to bio-medical text, health records have been shown to differ drastically in prose and diction even when the underlying information may be similar (Gu et al., 2020). We processed 659M tokens of de-identified clinical notes and continued training using the PMC +LRC, PMC +ER $_{\text{unif}}$, and PMC +EWC $_{\text{GMM}}$ from Table 4 (with this stage of model denoted with a subscript 2). RoBERTa $_{\text{BASE}}$ is the untouched model as presented in Table 4, and we continue to train (unmitigated) the PMC model from the same table (now denoted as PMC, clin.). We evaluate models on RE and NER from the i2b2

Model	Generic	bio-medical	i2b2 NER	i2b2 RE	ADE RE	Clin. Avg.	Overall
RoBERTa $_{\text{BASE}}$	88.30	74.69	81.12	77.16	87.82	82.03	81.67
PMC, clin.	82.98	76.53	85.96	79.44	88.96	84.79	81.43
LRC$_2$	**87.47**	74.33	85.03	77.93	86.84	83.26	81.69
ER$_2$	84.51	**75.85**	85.16	79.20	**88.23**	**84.20**	81.52
EWC$_2$	86.99	75.04	**85.43**	**79.59**	86.07	83.47	**81.91**

Table 5: Averaged performance for all generic, and bio-medical tasks (i.e. as seen in Table 4). Clinical average is across i2b2 NER and RE as well as n2c2 ADE RE are given as Micro-F_1

challenge after 5 epochs [7]. Additionally we use the n2c2 adverse drug reaction (ADE) (Henry et al., 2020) RE task.

Stage two results are reported in Table 5. The last column in this table indicates that average overall performance is about the same across models, however, when we take a closer look at the domain breakdown we see this is not the case. As expected the unmitigated model (PMC, clin.) suffers from performance deterioration in generic tasks, with GLUE dropping drastically (an error increase to 6.21% compared to RoBERTa $_{\text{BASE}}$). We find that LRC is still firmly holding onto generic representation, with the smallest drop in average generic performance of 0.83 points, when compared to stage one. Here we found that tuning ρ became more prevalent, with the range of average clinical scores for LRC being 1.49 points. ER, and EWC are the only mitigated models which achieve competitive numbers for clinical tasks, although they both show a drop in generic, and bio-medical results. Both of the latter models outperform the base model in average bio-medical and clinical metrics.

7 Analysis

To further understand learning and forgetting across different mitigation strategies, we conduct analyses to investigate how different layers of the model adapt to in-domain pre-training, whether the adaptation helps in transferring knowledge to downstream tasks, and how knowledge learned from in & out of domain data cooperates together.

7.1 Layer-wise analyses

7.1.1 Weight Similarity

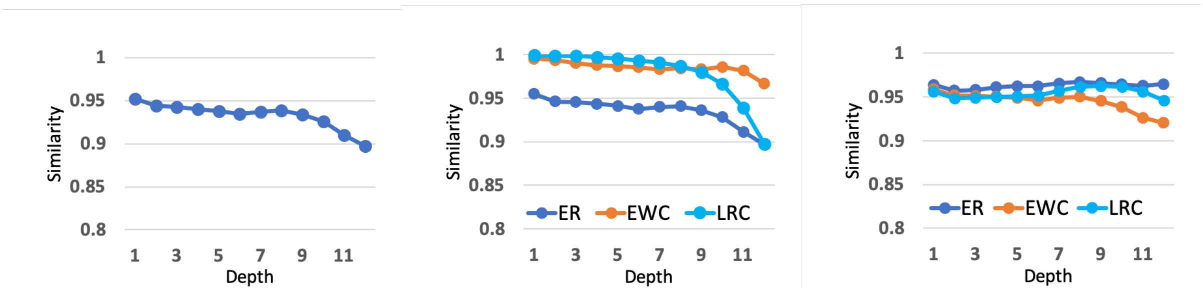
Figure 2 displays layer-wise weight (cosine-) similarity between models before and after pre-training on bio-medical data. We compare RoBERTa $_{\text{BASE}}$ (denoted as Generic) against the PMC model (row 2 in Table 4 and denoted as bio-medical in the Figure). In Figure 2a we discern similarity in layers closer towards the input. By comparing Figures 2b and 2c which illustrate how mitigated models behave compared to one another, we find that ER allows the model parameters to shift much closer towards the bio-medical data while EWC finds a shared space for parameters in both models. This is consistent with what we have observed in Section 6.2.2 where we find EWC is better at mitigating catastrophic forgetting compared to ER. It was important to see how LRC weights behave as well. Intuitively since the learning rate is close to 0 near the model input, these layers will change very little. This is indeed the case with only the last layer showing significant shift.

We investigate if constraining the weights to a shared space is enough to produce a good overall model. We observed that without the Fisher matrix, weight similarity between EWC and RoBERTa $_{\text{BASE}}$ is lower than 0.2, which is confirmed by the low F_1 scores noted in Section 6.2.2. This indicates that the Fisher diagonal plays an important role in fluctuating variance.

7.1.2 Transferability via Probing Tasks

To evaluate layer-wise transferability of pre-trained LMs, we use NER as a probing task and limit the capacity of task-specific layers to focus on what information has been learned by the model. We evaluate each layer of pre-trained LMs by extracting the model output as features and only fine-tuning task-specific layers. We observe in Figure 3 that (1) outer layers are most transferable to downstream tasks except for the last layer and (2) the performance of domain specific NER increases much faster than generic NER across layers, which indicates that grammatical understanding occurs in earlier layers, whereas segment level domain spe-

[7] To determine an appropriate stopping point we evaluated each epoch using the the clinical NER task until the Micro-F$_1$ plateaued.



(a) Generic vs. bio-medical (b) Mitigated Models vs. Generic (c) Mitigated Models vs. bio-medical

Figure 2: Weight distance vs. Depth across two domains. We compare RoBERTa $_{\text{BASE}}$ (trained on *generic* data) against PMC (denoted as bio-medical) and two mitigated models. Distance is given using cosine similarity.

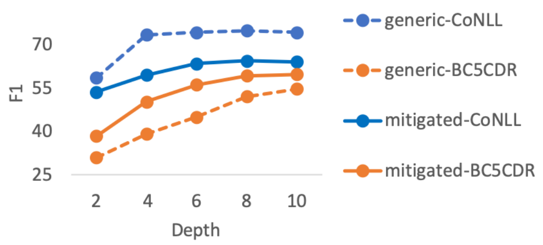


Figure 3: Transferability vs. Depth. Dashed curves denote generic models and solid curves denote mitigated models. After fine-tuning on bio-medical data, the performance of CoNLL drops while the performance of BC5CDR is boosted.

cific perception (i.e. semantics) appears in later layers. Both (1) and (2) are consistent with Figure 2a where weights change more in outer layers. This trend was also observed in previous works Belinkov et al. (2017); Jawahar et al. (2019).

Base on layer-wise analyses in this section, we empirically find that the adaptation in outer layers plays a key role in mitigation, which suggests that a decaying learning rate as a function of layer depth is worth being incorporated into different mitigation strategies.

7.2 Qualitative Examples

We observe that CF mitigation techniques are able to assist in generalization on rare words by composing knowledge from both generic and bio-medical domains. In Figure 4 (i) we observe that “Norilsk” occurs quite rarely in Newswire data, which is used for pre-training generic domain, however, it is frequent in Pubmed but size of pre-training data is small. Combining the two datasets in the form ER and EWC helps generalise the model understanding. We provide additional examples of this phenomenon in Figure 4 (ii) & (iii).

8 Related Work

Current work in catastrophic forgetting mitigation in NLP has been limited. Howard and Ruder (2018) introduced a multi stage training scheme for fine tuning LSTM based universal language models (ULMFiT). The authors proposed that current methods, rather than data, are ineffective and focused on learning rate control across layers, as well as modifying learning rate scheduling. A larger category of work deals with constraining model parameters to a latent space where they continue to capture previous tasks. Initial work focused on model regularization and varying activations (Goodfellow et al., 2013). Kirkpatrick et al. (2017) provided a more sophisticated solution constraining weights individually termed elastic weight consolidation (EWC). We make use of both EWC and ULMFiT and provide further technical detail in this paper. The final approach is focused on experience replay. Using small samples of data from previous tasks coupled with local adaptation de Masson d’Autume et al. (2019) demonstrate improvement in a lifelong learning training scheme. Chaudhry et al. (2019) also explore lifelong learning by experimenting with updating the memory bank for experience replay. Our work focuses on both of these techniques with the major difference being problem scale. Many existing works apply these solutions on small networks whereas we experiment on architectures having several orders of magnitude more parameters.

There has been a recent focus on more effective pre-training which focuses on narrowing the pre-training domain as we move closer towards fine-tuning. STILTs (Phang et al., 2018) and TandA (Garg et al., 2019) use intermediate tasks (in a data rich domain) training to lower variance during target task fine-tuning. This intuition was also covered

	Text	Model	Label	conf.
(i):	Entire social infrastructures in the icy Far North where Norilsk is based depend on the company, and government has said that expenditure could far outstrip Norilsk 's debts. **[Norilsk]** officials declined to comment.	Ground Truth	S-ORG	–
		RoBERTa $_{\text{BASE}}$	S-MISC	0.609
		PMC	S-MISC	0.983
		PMC+ER	S-ORG	1.000
(ii):	President Arafat's position is clear that such a meeting should come after successful negotiations so that the meeting would have positive results. Especially since the **[Hebron]** issue has not been agreed yet and the crucial disputed issues have not been resolved.	Ground Truth	S-LOC	–
		RoBERTa $_{\text{BASE}}$	S-PER	0.998
		PMC	O	1.000
		PMC+ER	S-LOC	0.994
(iii):	The committee said the Italian club had violated regulations by failing to inform Feyenoord, with whom the player was under contract. Blinker was fined 75,000 Swiss francs ($57,600) for failing to inform the English club of his previous commitment to **[Udinese]**.	Ground Truth	S-ORG	–
		RoBERTa $_{\text{BASE}}$	S-LOC	0.815
		PMC	S-LOC	1.000
		PMC+ER	S-ORG	1.000

Figure 4: Multi-task effect: generalization of a model on rare words using shared knowledge of pre-training on Newswire and Pubmed data. Example spans (taken from the CoNLL test split) are passed through an NER system initialized with various pre-trained encoders. We provide the labels and confidences for each.

in the visio-linguistic domain by Singh et al. (2020). Finally Gururangan et al. (2020) work on MLM pre-training and provide conclusive evidence at scale of the works listed above. This last body of work, although dealing with pre-training is different from our work in that we study mitigation of domain forgetting, rather than reducing variance by adding intermediate domains or tasks to pre-training.

9 Conclusion

In this work, we empirically investigated the existence of catastrophic forgetting in large language model pre-training. We further explored constraint and replay based mitigation techniques to close the performance gap between general and domain specific natural language tasks. We find that training a single model across multiple domains is possible. Due to practical considerations, we would suggest using latent representation for data selection when working with a data dependent model such as ER or EWC. When no previous data is available LRC provides a simple yet powerful solution for retaining prior domain knowledge. In the future work wish to explore more data independent methods such as LRC, for both speed and lack of data dependency, as well as manipulation of the decay w.r.t. what we have discovered from our layer-wise analysis.

Acknowledgments

We would like to thank Byron Wallace, Kevin Small, Ramesh Nallapati, and members of Amazon Comprehend Medical for their help in shaping this work over the last year, as well as the conference reviewers for providing thoughtful feedback.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

B Roy Frieden. 2004. *Science from Fisher information: a unification*. Cambridge University Press.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2019. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext (gokaslan's distribution, 2019), gpt-2 tokenized.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

S. Henry, Kevin P. Buchan, Michele Filannino, A. Stubbs, and Özlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association : JAMIA*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

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. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13143–13152. Curran Associates, Inc.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553–568. Springer.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Douglas Reynolds. 2009. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA.

Mark B. Ring. 1997. Child: A first step towards continual learning. In *Machine Learning*, pages 77–104.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

James C Spall. 2005. Monte carlo computation of the fisher information matrix in nonstandard settings. *Journal of Computational and Graphical Statistics*, 14(4):889–909.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.