Forget Me Not: Reducing Catastrophic Forgetting for Domain Adaptation in Reading Comprehension

1st Ying Xu
IBM Research Australia
Melbourne, Australia
ying.jo.xu@au1.ibm.com

2nd Xu Zhong
IBM Research Australia
Melbourne, Australia
peter.zhong@au1.ibm.com

3rd Antonio Jose Jimeno Yepes
IBM Research Australia
Melbourne, Australia
antonio.jimeno@au1.ibm.com

4th Jey Han Lau
University of Melbourne
Melbourne, Australia
jeyhan.lau@gmail.com

***Abstract*—The creation of large-scale open domain reading comprehension data sets in recent years has enabled the development of end-to-end neural comprehension models with promising results. To use these models for domains with limited training data, one of the most effective approach is to first pre-train them on large out-of-domain *source* data and then fine-tune them with the limited *target* data. The caveat of this is that after fine-tuning the comprehension models tend to perform poorly in the source domain, a phenomenon known as catastrophic forgetting. In this paper, we explore methods that reduce catastrophic forgetting during fine-tuning without assuming access to data from the source domain. We introduce new auxiliary penalty terms and observe the best performance when a combination of auxiliary penalty terms is used to regularise the fine-tuning process for adapting comprehension models. To test our methods, we develop and release 6 narrow domain data sets that can potentially be used as reading comprehension benchmarks.**

***Index Terms*—Domain adaptation, Reading comprehension, Catastrophic forgetting, Question Answering**

I. Introduction

Reading comprehension (RC) is the task of answering a question given a context passage. Related to Question-Answering (QA), RC is seen as a module in the full QA pipeline, where it assumes a related context passage has been extracted and the goal is to produce an answer based on the context. In recent years, the creation of large-scale open domain comprehension data sets [1]–[6] has spurred the development of a host of end-to-end neural comprehension systems with promising results.

In spite of these successes, it is difficult to train these modern comprehension systems on narrow domain data (e.g. biomedical), as these models often have a large number of parameters. A better approach is to transfer knowledge via fine-tuning, i.e. by first pre-training the model using data from a large *source domain* and continously training it with examples from the small *target domain*. It is an effective strategy, although a fine-tuned model often performs poorly when it is re-applied to the source domain, a phenomenon known as catastrophic forgetting [7]–[10]. This is generally not an issue if the goal is to optimise purely for the target domain, but in real-word applications where model robustness is an important quality, over-optimising for a development set often leads to unexpected poor performance when applied to test cases in the wild.

Methods for reducing catastrophic forgetting are categorised into three types based on three assumptions, i.e. with *full*, *partial* or *no* access to the source domain data. When *full* access is available, the most straightforward way to reduce catastrophic forgetting is to perform multitask learning, where the model is learned to perform well for both source and target domains [11], [12]. For *partial* access assumption, the gradient episodic memory (GEM) [13] was proposed to store a piece of data from the source domain, which is used to regularise the fine tuning process. However, in real-world applications, data became inaccessible for a number of reasons, such as an expired data sharing agreement, physical damage to data storage, accidental deletion of data, and the introduction of new data privacy laws. Here we focus on exploring methods to reduce catastrophic forgetting assuming *no* access to the source domain data.

In this paper, we explore strategies to reduce forgetting for comprehension systems during domain adaptation. Our goal is to preserve the source domain's performance as much as possible, while keeping target domain's performance optimal and assuming no access to the source data. We experiment with a number of auxiliary penalty terms to regularise the fine-tuning process for three modern RC models: `QANet` [14], `decaNLP` [15] and `BERT` [16]. We observe that combining different auxiliary penalty terms results in the best performance, outperforming benchmark methods that require source data. Technically speaking, the application of the methods we propose are not limited to domain transfer for reading comprehension. We also show that the methodology can be used for transferring to entirely different natural language processing tasks. With that said, we focus on comprehension here because it is a practical problem in real world applications, where the target domain often has a small number of QA pairs and over-fitting occurs easily when we fine-tune based on a small development set. In this scenario, it is as important to develop a robust model as achieving optimal development performance.

To demonstrate the applicability of our approach, we apply topic modelling to MSMARCO [2] — a comprehension data set based on internet search queries — and collect examples that belong to a number of salient topics, producing 6 small to medium sized RC data sets for the following domains: *biomedical*, *computing*, *film*, *finance*, *law* and *music*. We focus on extractive RC, where the answer is a continuous sub-span in

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

the context passage.[1] Scripts to generate the data sets are available at: https://github.com/ibm-aur-nlp/domain-specific-QA.

Summarising our contributions: (1) we experiment with a number of penalty terms to regularise the fine-tuning process for adapting comprehension systems, and found that a combination of them produces the most robust model that both preserves performance in the source domain and achieves optimal performance in the target domain; and (2) we develop and release six narrow-domain extractive comprehension data sets, facilitating research on domain adaptation for reading comprehension.

II. Related Work

Most large comprehension data sets are open-domain because non-experts can be readily recruited via crowdsourcing platforms to collect annotations. Development of domain-specific RC data sets, on the other hand, is costly due to the need of subject matter experts and as such the size of these data sets is typically limited. Examples include BIOASQ [18] in the biomedical domain, which has less than 3k QA pairs — orders of magnitude smaller compared to most large-scale open-domain data sets [2]–[4], [6].

Wiese et al. [8] explore supervised domain adaptation for reading comprehension, by pre-training their model first on large open-domain comprehension data and fine-tuning it further on biomedical data. This approach improves the biomedical domain's performance substantially compared to training the model from scratch. At the same time, its performance on source domain decreases dramatically due to catastrophic forgetting [7], [19], [20].

This issue of catastrophic forgetting is less of a problem when data from multiple domains or tasks are present during training. For example in [15], their model `decaNLP` is trained on 10 tasks simultaneously — all cast as a QA problem — and forgetting is minimal. For multi-domain adaptation, Daumé III, H. [11] and Kim et al. [12] propose using a K+1 model to capture domain-general pattern that is shared by K domains, resulting in a more robust model. Using multi-task learning to tackle catastrophic forgetting is effective and generates robust models. The drawback, however, is that when training for each new domain/task, data from the previous domains/tasks has to be available.

Several studies present methods to reduce forgetting with limited or no access to previous data [9], [10], [13], [21], [22]. Inspired by synaptic consolidation, Kirkpatrick et al. [9] propose to selectively penalise parameter change during fine-tuning. Significant updates to parameters which are deemed important to the source task incur a large penalty. Lopez-Paz et al. [13] introduce a gradient episodic memory (`gem`) to allow beneficial transfer of knowledge from previous tasks. More specifically, a subset of data from previous tasks are stored in an episodic memory, against which reference gradient vectors are calculated and the angles with the gradient vectors for the current task is constrained to be between $-90°$ and $90°$. Riemer et al. [10] suggest combining `gem` with optimisation based meta-learning to overcome forgetting. Among these three methods, only that of [9] assumes zero access to previous data. In comparison, the latter two rely on access to a memory storing data from previous tasks, which is not always feasible in real-world applications (, as is mentioned in the previous section).

[1] Although RC with free-form answers is arguably a more challenging and interesting task, evaluation is generally more difficult [17].

Partition	Domain	#Examples	#Unique Q	Mean C Length	Mean Q Length	Mean A Length
Train	MS-BM	22,134	21,902	70.9	6.4	13.7
	MS-CP	3,021	3,011	67.2	5.5	18.9
	MS-FM	3,522	3,481	65.8	6.4	6.5
	MS-FN	6,790	6,720	71.9	6.4	14.0
	MS-LW	3,105	3,078	64.7	6.2	18.5
	MS-MS	2,517	2,480	68.6	6.4	6.6
	BIOASQ	3,083	387	35.4	11.0	2.4
Dev	MS-BM	4,743	4,730	71.2	6.4	13.7
	MS-CP	647	646	65.4	5.3	19.6
	MS-FM	755	751	65.9	6.6	5.9
	MS-FN	1,455	1,453	71.6	6.5	14.4
	MS-LW	665	664	65.8	6.2	20.0
	MS-MS	539	536	69.2	6.4	6.1
	BIOASQ	674	83	39.7	11.1	2.4
Test	MS-BM	4,743	4,728	70.5	6.4	13.5
	MS-CP	648	645	66.6	5.6	18.3
	MS-FM	755	755	66.7	6.3	6.2
	MS-FN	1,455	1,452	70.8	6.5	13.6
	MS-LW	666	663	65.1	6.2	18.9
	MS-MS	540	540	67.4	6.6	7.0
	BIOASQ	631	84	34.9	13.2	2.9

TABLE I: Statistics of our seven target domain data sets (Q: Question; C: Context; and A: Answer).

III. Data Set

We use SQUAD v1.1 [3] as the source domain data for pre-training the comprehension model. It contains over 100K extractive (context, question, answer) triples with only answerable questions.

To create the target domain data, we leverage MSMARCO [2], a large RC data set where questions are sampled from Bing™ search queries and answers are manually generated by users based on passages in web documents. We apply LDA topic model [23] to passages in MSMARCO and learn 100 topics.[2] Given the topics, we label them and select 6 salient domains: *biomedical* (MS-BM), *computing* (MS-CP), *film* (MS-FM), *finance* (MS-FN), *law* (MS-LW) and *music* (MS-MS). A QA pair is categorised into one of these domains if its passage's top-topic belongs to them. We create multiple (context, question, answer) training examples if a QA pair has multiple contexts,[3] and filter them to keep only extractive examples.[4]

In addition to the MSMARCO data sets, we also experiment with a real biomedical comprehension data set: BIOASQ [26].

[2] When collecting the passages, we include only those being selected as useful for answering a query (i.e. `is_selected` $= 1$). We tokenise the passages with Stanford CoreNLP [24] and use MALLET [25] for topic modelling.

[3] We only consider context passages that are marked as being useful by annotators in the original data (i.e. `is_selected` $= 1$).

[4] A (context, question, answer) triple is defined to be extractive if the answer has a case-insensitive match to the context.

Each question in BIOASQ is associated with a set of snippets as context, and the snippets are single sentences extracted from a scientific publication's abstract/title in PubMed Central™. There are four types of questions: factoid, list, yes/no, and summary. As our focus is on extractive RC, we use only the extractive factoid questions from BIOASQ. As before, we create multiple training examples for QA pairs with multiple contexts.

For each target domain, we split the examples into 70%/15%/15% training/development/test partitions.[5] We present some statistics for the data sets in Table I.

IV. METHODOLOGY

We first pre-train a general domain RC model on SQUAD, our source domain. Given the pre-trained model, we then perform fine-tuning (`finetune`) on the MSMARCO and BIOASQ data sets: 7 target domains in total. By fine-tuning we mean taking the pre-trained model parameters as initial parameters and update them accordingly based on data from the new domain. To reduce forgetting on the source domain (SQUAD), we experiment with incorporating auxiliary penalty terms (e.g. L2 between new and old parameters) to the standard cross entropy loss to regularise the fine-tuning process.

We explore 3 modern RC models in our experiments: `QANet` [14]; `decaNLP` [15]; and `BERT` [16]. `QANet` is a Transformer-based [27] comprehension model, where the encoder consists of stacked convolution and self-attention layers. The objective of the model is to predict the position of the starting and ending indices of the answer words in the context. `decaNLP` is a recurrent network-based comprehension model trained on ten NLP tasks simultaneously, all casted as a question-answer problem. Much of `decaNLP`'s flexibility is due to its pointer-generator network, which allows it to generate words by extracting them from the question or context passages, or by drawing them from a vocabulary. `BERT` is a deep bi-directional encoder model based on Transformers. It is pre-trained on a large corpus in an unsupervised fashion using a masked language model and next-sentence prediction objective. To apply `BERT` to a specific task, the standard practice is to add additional output layers on top of the pre-trained `BERT` and fine-tune the whole model for the task. In our case for RC, 2 output layers are added: one for predicting the start index and another the end index. Devlin et al. [16] demonstrates that this transfer learning strategy produces state-of-the-art performance on a range of NLP tasks. For RC specifically, `BERT` (`BERT-Large`) achieved an F1 score of 93.2 on SQUAD, outperforming human performance by 2 points.

Note that `BERT` and `QANet` RC models are extractive models (goal is to predict 2 indices), while `decaNLP` is a generative model (goal is to generate the correct word sequence). Also, unlike `QANet` and `decaNLP`, `BERT` is not designed specifically for RC. It represents a growing trend in the literature where large models are pre-trained on big corpora and further adapted to downstream tasks.

To reduce the forgetting of source domain knowledge, we introduce auxiliary penalty terms to regularise the fine-tuning process. We favour this approach as it does not require storing data samples from the source domain. In general, there are two types of penalty: selective and non-selective. The former penalises the model when certain parameters diverge significantly from the source model, while the latter uses a pre-defined distance function to measure the change of all parameters.

For selective penalty, we use elastic weight consolidation (EWC: [9]), which weighs the importance of a parameter based on its gradient when training the source model. For non-selective penalty, we explore L2 [8] and cosine distance. We detail the methods below.

Given a source and target domain, we pre-train the model first on the source domain and fine-tune it further on the target domain. We denote the optimised parameters of the source model as θ^* and that of the target model as θ. For vanilla fine-tuning (`finetune`), the loss function is:

$$\mathcal{L}_{ft} = \mathcal{L}_{ce}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss.

For non-selective penalty, we measure the change of parameters based on a distance function (treating all parameters as equally important), and add it as a loss term in addition to the cross-entropy loss. One distance function we test is the L2 distance:

$$\mathcal{L}_{+l2} = \mathcal{L}_{ce} + \lambda_{l2}\text{L2}(\theta, \theta^*)$$

where λ_{l2} is a scaling hyper-parameter to weigh the contribution of the penalty. Henceforth all scaling hyper-parameters are denoted using λ.

We also experiment with cosine distance, based on the idea that we want to encourage the parameters to be in the same direction after fine-tuning. In this case, we group parameters by the variables they are defined in, and measure the cosine distance between variables:

$$\mathcal{L}_{+cd} = \mathcal{L}_{ce} + \lambda_{cd}\frac{1}{|V|}\sum_{v}\text{CD}(\theta_v, \theta_v^*)$$

where θ_v denotes the vector of parameters belonging to variable v.

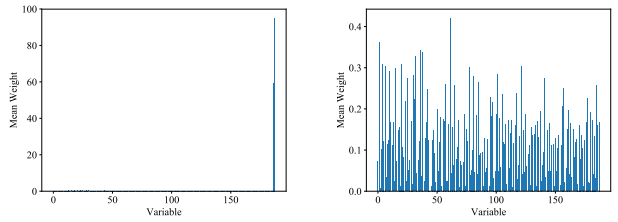
For selective penalty, EWC uses the Fisher matrix F to measure the importance of parameter i in the source domain. Unlike non-selective penalty where all parameters are considered equally important, EWC provides a mechanism to weigh the update of individual parameters:

$$\mathcal{L}_{+ewc} = \mathcal{L}_{ce} + \lambda_{ewc}\sum_{i}(F_i \cdot (\theta_i - \theta_i^*))$$

$$F = E[(\frac{\partial \mathcal{L}_{ce}(f_{\theta^*}, (x,y))}{\partial \theta^*})^2 | \theta^*]$$

where $\frac{\partial \mathcal{L}_{ce}(f_{\theta^*}, (x,y))}{\partial \theta^*}$ is the gradient of parameter update in the source domain, with f_{θ^*} representing the model and x/y the data/label from the source domain.

[5] Partitioning is done at the question level to ensure the same question does not appear in more than one partition.



(a) EWC (b) Normalised EWC

Fig. 1: Mean Fisher Matrix values for variables in QANet on SQUAD.

In preliminary experiments, we notice that EWC tends to assign most of the weights to a small subset of parameters. We present Figure 1a, a plot of mean Fisher values for all variables [6] in QANet after it was trained on SQUAD, the source domain. We see that only the last two variables have some significant weights (and a tiny amount for the rest of the variables). Due to the considerably different contribution of variables in different layers , using the Fisher matrix above directly may leads to ignoring the difference in the contribution of different parameters within the same variable. We therefore propose a new variation of EWC, *normalised EWC*, by normalising the weights within each variable via min-max normalisation, which brings up the weights for parameters in other variables (Figure 1b):

$$F_i^* = \frac{F_i - min(\{F\}^{v_i})}{max(\{F\}^{v_i}) - min(\{F\}^{v_i})}$$
$$\mathcal{L}_{+ewcn} = \mathcal{L}_{ce} + \lambda_{ewcn} \sum_i (F_i^* \cdot (\theta_i - \theta_i^*))$$

where $\{F\}^{v_i}$ denotes the set of parameters for variable v where parameter i belongs.

Among the four auxiliary penalty terms, L2 and EWC are proposed in previous work while cosine distance and normalised EWC are novel penalty terms. Observing that EWC and normalised EWC are essentially weighted $l1$ distances, L2 is based on $l2$ distance and cosine distance focuses on the vector angle between variables (and ignores the magnitude), we propose combining them altogether as these different distance metrics may complement each other in regularising the fine-tuning process:

$$\begin{aligned}\mathcal{L}_{+all} = \mathcal{L}_{ce} &+ \lambda_{l2}\text{L2}(\theta, \theta^*) \\ &+ \lambda_{cd}\frac{1}{|V|}\sum_v \text{CD}(\theta_v, \theta_v^*) \\ &+ \lambda_{ewcn}\sum_i (F_i^* \cdot (\theta_i - \theta_i^*))\end{aligned}$$

Since EWC and normalised EWC are both weighted $l1$ distances and our results demonstrate that our proposed normalised EWC perform better than the vanilla EWC, we only incorporate normalised EWC in the combined loss $\mathcal{L}_{+all}$.

[6]By variable we intend a parameter matrix in the attention or feed-forward layers, e.g. the kernel matrix of the 6th layer defined in a BERT model.

V. EXPERIMENTS

We test 3 comprehension models: QANet, decaNLP and BERT. To pre-process the data, we use the the models' original tokenisation methods.[7] For BERT, we use the smaller pre-trained model with 110M parameters (BERT-Base).

A. Fine-Tuning with Auxiliary Penalty

We first pre-train QANet and decaNLP on SQUAD, tuning their hyper-parameters based on its development partition.[8] For BERT, we fine-tune the released pre-trained model on SQUAD by adding 2 additional output layers to predict the start/end indices (we made no changes to the hyper-parameters). We initialise word vectors of QANet and decaNLP with pre-trained GloVe embeddings [28] and keep them fixed during training. We also freeze the input embeddings for BERT.[9] To measure performance, we use the standard macro-averaged F1 as the evaluation metric, which measures the average overlap of word tokens between prediction and ground truth answer.[10] Our pre-trained QANet, decaNLP and BERT achieve an F1 score of 80.47, 75.50 and 87.62 respectively on the development partition of SQUAD. Note that the test partition of SQUAD is not released publicly, and so all reported SQUAD performance in the paper is on the development set.

Given the pre-trained SQUAD models, we fine-tune them on the MSMARCO and BIOASQ domains. We test vanilla fine-tuning (finetune) and 5 variants of fine-tuning with auxiliary penalty terms: (1) EWC (+ewc); normalised EWC (+ewcn); cosine distance (+cd); L2 (+l2); and combined normalised EWC, cosine distance and L2 (+all). As a benchmark, we also perform fine-tuning with gradient episodic memory (gem), noting that this approach uses the first m examples from SQUAD ($m = 256$ in our experiments).

To find the best hyper-parameter configuration, we tune it based on the development partition for each target domain (and report the performance on the their test partitions). For a given domain, finetune and its variants (+ewc, +ewcn, +cd, +l2 and +all) all share the same hyper-parameter configuration.[11] Detailed hyper-parameter settings are given in the supplementary material if requested.

As a baseline, we train QANet, decaNLP and BERT from scratch (scratch) using the target domain data. As before, we tune their hyper-parameters based on development performance. We present the full results in Table II.

For each target domain, we display two F1 scores: the source SQUAD development performance ("SQUAD (dev)");

[7]We use spaCy (https://spacy.io/), revtok (https://github.com/jekbradbury/revtok), and WordPiece for QANet, decaNLP and BERT, respectively.

[8]We tune for dropout, batch size, learning rate and number of training iterations, keeping other hyper-parameters in their default configuration.

[9]The input embeddings of BERT is a sum of token, segment and position embeddings; we freeze only the token embeddings.

[10]If there are multiple ground truths, the maximum F1 is taken.

[11]The only exception are the scaling hyper-parameters (λ_{ewc}, λ_{ewcn}, λ_{cd} and λ_{l2}), where we tune them separately for each model.

Model	Partition	Domain	scratch	finetune	+ewc	+ewcn	+cd	+l2	+all	gem
QANet	SQUAD (dev)	MS-BM	—	62.92	63.35	63.93	63.49	64.93	**65.54**	63.22
		MS-CP	—	39.13	41.62	43.43	41.19	41.61	**51.84**	43.49
		MS-FM	—	56.32	58.23	58.46	57.01	58.48	**60.79**	57.53
		MS-FN	—	65.08	65.45	67.03	65.36	66.27	**68.14**	66.53
		MS-LW	—	68.29	68.64	68.63	68.75	68.38	**69.39**	69.04
		MS-MS	—	69.60	69.96	70.11	69.72	69.74	**71.13**	70.63
		BIOASQ	—	59.85	62.87	63.57	62.83	62.50	**66.11**	62.52
		Avg.	—	60.17	61.45	62.17	61.19	61.70	**64.71**	61.85
	Target (test)	MS-BM	62.75	68.45	67.96	67.85	67.80	68.05	67.33	68.31
		MS-CP	60.67	68.86	69.26	69.86	70.27	69.42	70.42	69.17
		MS-FM	59.57	73.84	72.70	74.13	73.94	73.50	73.47	72.00
		MS-FN	63.62	70.96	70.70	70.60	70.49	70.15	70.27	69.18
		MS-LW	61.66	71.29	71.27	71.39	71.25	71.28	71.41	71.49
		MS-MS	58.36	69.58	69.94	69.89	69.62	69.92	70.67	71.47
		BIOASQ	29.83	65.81	67.17	65.93	67.26	65.57	66.82	66.42
		Avg.	56.64	69.83	69.86	69.95	70.09	69.70	70.06	69.72
decaNLP	SQUAD (dev)	MS-BM	—	62.99	63.00	63.26	63.27	62.43	63.82	**64.95**
		MS-CP	—	56.48	58.19	59.44	61.96	60.73	62.61	**63.37**
		MS-FM	—	58.69	59.21	59.18	62.66	58.32	**64.04**	63.36
		MS-FN	—	58.21	61.63	63.43	59.25	58.80	**66.55**	62.47
		MS-LW	—	57.86	58.14	59.73	58.17	56.89	60.75	**61.76**
		MS-MS	—	59.75	**64.92**	62.01	62.00	60.06	63.62	63.89
		BIOASQ	—	67.42	67.19	67.21	67.44	67.46	67.49	**68.94**
		Avg.	—	60.20	61.75	62.04	62.11	60.67	**64.13**	64.11
	Target (test)	MS-BM	62.01	66.90	67.39	67.52	67.61	67.19	67.41	67.02
		MS-CP	63.7	66.67	67.11	68.15	66.37	67.82	67.55	67.90
		MS-FM	63.28	70.45	70.47	70.83	69.08	70.36	68.04	69.73
		MS-FN	64.41	64.59	64.57	64.35	64.32	64.87	64.32	64.88
		MS-LW	66.36	73.43	73.28	73.34	73.42	74.13	73.04	72.89
		MS-MS	64.65	68.67	67.12	67.93	67.34	69.40	66.51	68.28
		BIOASQ	43.25	63.80	63.89	63.89	63.96	63.96	64.70	66.36
		Avg.	61.09	67.79	67.69	68.00	67.44	68.25	67.37	68.15
BERT	SQUAD (dev)	MS-BM	—	72.55	74.24	76.51	72.36	74.14	**77.32**	74.14
		MS-CP	—	68.41	69.63	75.65	76.92	75.98	**77.86**	73.37
		MS-FM	—	73.82	75.175	79.75	75.28	74.71	**81.42**	76.89
		MS-FN	—	72.59	74.27	75.52	73.22	74.84	**78.18**	76.16
		MS-LW	—	71.93	81.11	81.05	78.77	77.97	**83.11**	75.90
		MS-MS	—	72.59	78.06	**83.56**	75.67	74.29	83.54	76.99
		BIOASQ	—	75.04	85.28	85.62	85.76	84.23	**86.88**	75.89
		Avg.	—	72.42	76.82	79.67	76.85	76.59	**81.19**	75.62
	Target (test)	MS-BM	66.83	68.30	68.20	68.00	68.04	68.24	67.87	68.02
		MS-CP	65.99	70.57	71.21	71.41	69.33	69.57	69.49	70.40
		MS-FM	72.59	74.73	74.75	74.36	73.73	74.85	75.78	74.63
		MS-FN	66.70	69.13	70.42	70.60	69.07	70.05	69.15	69.54
		MS-LW	67.38	69.99	70.73	71.59	70.57	70.91	68.59	68.87
		MS-MS	70.45	73.56	73.19	73.07	72.97	73.43	72.50	72.73
		BIOASQ	54.09	71.62	75.84	78.50	79.47	78.86	76.93	68.87
		Avg.	66.29	71.13	72.05	72.50	71.88	72.27	71.47	70.44

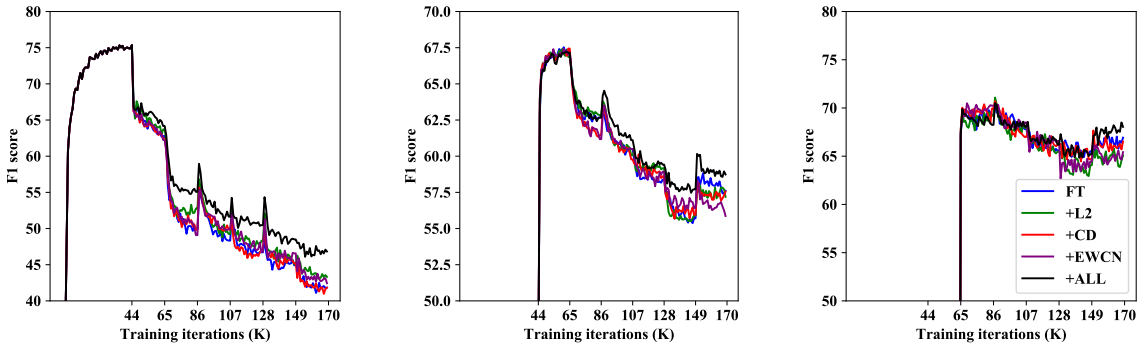
TABLE II: RC results over all domains. Pre-trained QANet/decaNLP/BERT performance on SQUAD (dev) = 80.47/75.50/87.62. Boldface indicates optimal performance for SQUAD (dev) and Underline indicates best performance for target domains (test).

and the target domain's test performance ("Target (test)")[12]. We first compare the performance between scratch and finetune. Across all domains for QANet, decaNLP and BERT, finetune substantially improves the target domain's performance compared to scratch. The largest improvement is seen in BIOASQ for QANet, where its F1 improves two-fold (from 29.83 to 65.81). Among the three RC models, BERT has the best performance for both scratch and finetune in most target domains (with a few exceptions such as MS-FN and MS-LW). Between QANet and decaNLP, we see that decaNLP tends to have better scratch performance but the pattern is reversed in finetune, where QANet produces higher F1 than decaNLP in all domains except for MS-LW.

[12]Since we are evaluating different auxiliary penalty terms in terms of their effectiveness in reducing catastrophic forgetting, we intentionally tuned the hyper-parameters for different auxiliary penalty terms to generate similar performance on target domains. Therefore, we are more interested in comparing the performance on the source domain, e.g. SQUAD (dev).

In terms of SQUAD performance, we see that finetune degrades it considerably compared to its pre-trained performance. The average drop across all domains compared to their pre-trained performance is 20.30, 15.30 and 15.07 points for QANet, decaNLP and BERT, respectively. For most domains, F1 scores drop by 10-20 points, while for MS-CP the performance is much worse for QANet, with a drop of 41.34. Interestingly, we see BERT suffers from catastrophic forgetting just as much as the other models, even though it is a larger model with orders of magnitude more parameters.

We now turn to the fine-tuning results with auxiliary penalties (+ewc, +ewcn, +cd and +l2). Between +ewc and +ewcn, the normalised versions consistently produces better recovery for the source domain (one exception is MS-MS for decaNLP), demonstrating that normalisation helps. Between +ewcn, +cd and +l2, performance among the three models vary depending on the domain and there's no clear winner.



(a) SQUAD (b) MS-BM (c) MS-CP

Fig. 2: decaNLP's F1 performance during continuous learning.

Combining all of these losses (+all) however, produces the best SQUAD performance for all models across most domains. The average recovery (+all- finetune) of SQUAD performance is 4.54, 3.93 and 8.77 F1 points for QANet, decaNLP and BERT respectively, implying that BERT benefits from these auxiliary penalties more than decaNLP and QANet.

When compared to gem, +all preserves SQUAD performance substantially better, on average 2.86 points more for QANet and 5.57 points more BERT. For decaNLP, the improvement is minute (0.02). Also, as gem requires partial access to the training data from the source domain (SQUAD training examples in this case), the auxiliary penalty techniques are more favourable for real world applications.

Does adding these penalty terms harm target performance? Looking at the "Test" performance between finetune and +all, we see that they are generally comparable. We found that the average performance difference (+all-finetune) is 0.23, −0.42 and 0.34 for QANet, decaNLP and BERT respectively, implying that it does not (in fact, it has a small positive net impact for QANet and BERT). In some cases it improves target performance substantially, e.g. in BIOASQ for BERT, the target performance is improved from 71.62 to 76.93, when +all is applied.

Based on these observations, we see benefits for incorporating these penalties when adapting comprehension models, as it produces a more robust model that preserves its source performance (to a certain extent) without trading off its target performance. In some cases, it can even improve the target performance.

B. Continuous Learning

In previous experiments, we fine-tune a pre-trained model to each domain independently. With continuous learning, we seek to investigate the performance of finetune and its four variants (+l2, +cd, +ewcn and +all) when they are applied to a series of fine-tuning on multiple domains. For the remainder of this paper, we experiment only with decaNLP.

We have one model for each of the five fine-tuning methods, e.g. finetune, +l2, +cd, +ewcn, and +all. Including the pre-training on SQUAD, all five models are trained for a total of 170K iterations: SQUAD (0–44K), MS-BM (45K–65K), MS-CP (66K–86K), MS-FN (87K–107K), MS-MS (108K–128K), MS-FM (129K–149K) and MS-LW (150K–170K). When computing the penalties, we consider the trained model for the previous domain as the source model.[13] Figure 2 (a)-(c) demonstrate the performance of the five models on the development set of SQUAD and test sets of MS-BM and MS-CP, respectively, when they are adapted to MS-BM, MS-CP, MS-FN, MS-MS, MS-FM and MS-LW in sequence.[14] We exclude plots for the latter domains as they are similar to that of MS-CP.

We first look at the recovery for SQUAD in Figure 2a. +all (black line; legend in Figure 2c) trails well above all other models after a series of fine-tuning, followed by +ewcn and +cd, while finetune produces the most forgetting. At the end of the continuous learning, +all recovers more than 5 F1 points compared to finetune. We see a similar trend for MS-BM (Figure 2b), although the difference is less pronounced. The largest gap between finetune and +all occurs when we fine-tune for MS-FM (iteration 129K–149K). Note that we are not trading off target performance when we first tune for MS-BM (iteration 45K–65K), where finetune and +all produces comparable F1.

For MS-CP (Figure 2c), we first notice that there is considerably less forgetting overall (MS-CP performance ranges from 65–75 F1, while SQUAD performance in Figure 2a ranges from 45–75 F1). This is perhaps unsurprising, as the model is already generally well-tuned (e.g. it takes less iterations to reach optimal performance for MS-CP compared to MS-BM and SQUAD). Most models perform similarly here. +all produces stronger recovery when fine-tuning on MS-FM (129K–149K) and MS-LW (150K–170K). At the end of the continuous learning, the gap between all models is around 2 F1 points.

C. Task Transfer

In decaNLP, curriculum learning was used to train models for different NLP tasks. More specifically, decaNLP was

[13]The implication is that we have to re-compute the Fisher matrix for the last domain before we fine-tune the model on a new domain.

[14]In terms of hyper-parameters, we choose a configuration that is generally good for most domains.

Partition	Task	finetune	+ewc	+ewcn	+cd	+l2	+all
SQUAD (dev)	SUM	8.60	11.65	12.48	11.28	9.34	**14.00**
	SRL	50.51	51.30	56.99	55.40	51.56	**57.64**
	SP	6.95	9.69	10.20	10.61	19.39	**28.36**
	MT	3.55	4.03	**4.29**	3.48	3.15	3.59
	SA	1.74	2.69	2.38	3.63	2.51	**6.43**
Target (test)	SUM	20.06	19.79	19.99	20.01	20.38	20.12
	SRL	71.69	71.80	71.74	72.12	71.90	72.56
	SP	92.52	92.77	92.70	92.62	92.59	91.11
	MT	24.99	25.10	25.04	25.00	24.90	24.90
	SA	84.79	86.38	84.84	85.06	86.27	85.89

TABLE III: decaNLP's SQUAD and target performance for several tasks.

first pre-trained on SQUAD and then fine-tuned on 10 tasks (including SQUAD) jointly. During the training process, each minibatch consists of examples from a particular task, and they are sampled in an alternating fashion among different tasks.

In situations where we do not have access to training data from previous tasks, catastrophic forgetting occurs when we adapt the model for a new task. In this section, we test our methods for task transfer (as opposed to domain transfer in previous sections). To this end, we experiment with decaNLP and monitor its SQUAD performance when we fine-tune it for other tasks, including semantic role labelling (SRL), summarisation (SUM), semantic parsing (SP), machine translation (MT), and sentiment analysis (SA)[15]. Note that we are not doing joint or continuous learning here: we are taking the pre-trained model (on SQUAD) and adapting it to new tasks independently. Description of these tasks are detailed in [15].

A core novelty of decaNLP is that its design allows it to generate words by extracting them from the question, context or its vocabulary, and this decision is made by the pointer-generator network. Based on the pointer-generator analysis in [15], we know that the pointer-generator network favours generating words using: (1) context for SRL, SUM, and SP; (2) question for SA; and (3) vocabulary for MT.

As before, finetune serves as our baseline, and we have 5 variants with auxiliary penalty terms. Table III displays the F1 performance on SQUAD and the target task; the table shares the same format as Table II. In terms of target task performance ("Test"), we see similar performances for all five models. This is a similar observation we saw in previously, and it shows that the incorporation of the auxiliary penalty terms does not harm target task or domain performance. For the source task SQUAD, +all produces substantial recovery for SUM, SRL, SP and SA, but not for MT. We hypothesise that this is due to the difference in nature between the target task and the source task: i.e. for SUM, SRL and SP, the output is generated by selecting words from context, which is similar to SQUAD; MT, on the other hand, generate using words from the vocabulary and question, and so it is likely to be difficult to find an optimal model that performs well for both tasks.

[15]We did not test task transfer on QANet or BERT because they are designed only for reading comprehension.

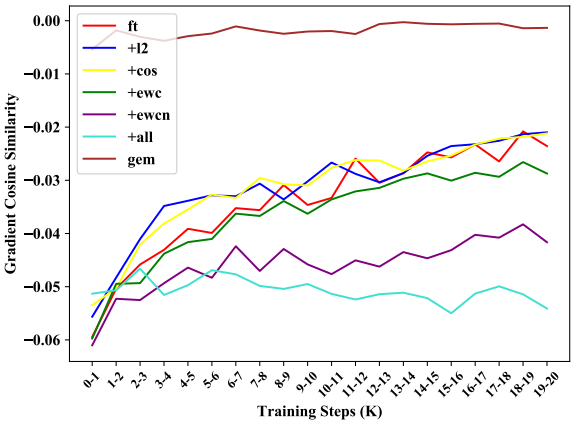


Fig. 3: Averaged gradient cosine similarities on MS-FN.

VI. Discussion

A. Gradient angle analysis

The improved performance achieved by the combined auxiliary loss leads us to question that why it performed better than individual penalty losses. An intuitive explanation is that different distance penalties, e.g. $l1$-distance, $l2$-distance and vector angles, form different constraints to prevent the model from being tuned too far away from the pre-trained model. Combining these distance penalties result in a more strict constraints that helps further reducing the catastrophic forgetting. In addition to this, we hypothesise that the combined loss may be better at minimising the angle between the gradient vector w.r.t the target domain data and the gradient vector w.r.t the source domain data, during the process of fine-tuning, which is essentially the rationale behind the gem method [13].

To validate this hypothesis, we conduct gradient analysis for all five auxiliary penalty terms. During fine-tuning, at each step t, we calculate the gradient cosine similarity $sim(g_t, g'_t)$,

$$g_t = \frac{\partial \mathcal{L}(f_{\theta_t}, M)}{\partial \theta_t}$$
$$g'_t = \frac{\partial \mathcal{L}(f_{\theta_t}, (x, y))}{\partial \theta_t}$$

where M is a memory containing SQUAD examples, and x/y is training data/label from the current domain. According to [13], when the gradient cosine similarity is $\geq$0, the fine-tuning process is unlikely to harm the performance of the model on original domain. We smooth the scores by averaging over every 1K steps, resulting in 20 cosine similarity values for 20K steps. Figure 3 plots the gradient cosine similarity scores for the five models in MS-FN.

Curiously, our best performing model +all produces the lowest cosine similarity at most steps (the only exception is between 0-1K steps). finetune, on the other hand, maintains relatively high similarity throughout. Similar trends are found for other domains. These observations imply that the approach gem adopted — i.e. constraining a positive dot

product between g_t and g'_t — is sufficient but not necessary for reducing catastrophic forgetting.

B. Decaying auxiliary loss scale

Conventionally, the λ scaling hyper-parameter for controlling the contribution of the penalty terms has a static value. In preliminary experiments, we notice that this loss starts at a very low value (close to zero), as initially there is little change to the model parameters. As such in the early iterations of fine-tuning, the model tends to focus on optimising for the target domain/task, and that results in a sharp drop of performance for the source domain/task.

In light of that, we explore using a dynamic λ scale that starts at a larger value that decays over time. With just simple linear decay, we found substantial improvement in `+ewc` for recovering SQUAD's performance, although the results are mixed for other penalties (particularly for `+ewcn`). We therefore only report results that are based on static λ values in this paper. With that said, we contend that this might be an interesting avenue for further research, e.g. by exploring more complex decay functions.

VII. Conclusion

To reduce catastrophic forgetting when adapting comprehension models, we explore several auxiliary penalty terms to regularise the fine-tuning process. We experiment with selective and non-selective penalties, and found that a combination of them consistently produces the best recovery for the source domain without harming its performance in the target domain. We also found similar observations when we apply our approach for adaptation to other tasks, demonstrating its general applicability. To test our approach, we develop and release six narrow domain reading comprehension data sets for the research community.

References

[1] Y. Yang, W.-t. Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2013–2018.

[2] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *CoRR*, vol. abs/1611.09268, 2016.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 2016, pp. 2383–2392.

[4] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1601–1611.

[5] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789.

[6] T. Kocisky, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette, "The narrativeqa reading comprehension challenge," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 317–328, 2018.

[7] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[8] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Vancouver, Canada, 2017, pp. 281–289.

[9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 3521–3526, 2017.

[10] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," *arXiv preprint arXiv:1810.11910*, 2018.

[11] H. Daume III, "Frustratingly easy domain adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 256–263.

[12] Y.-B. Kim, K. Stratos, and R. Sarikaya, "Frustratingly easy neural domain adaptation," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 387–396.

[13] D. Lopez-Paz *et al.*, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[14] A. W. Yu, D. Dohan, M. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *CoRR*, vol. abs/1804.09541, 2018.

[15] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *CoRR*, vol. abs/1806.08730, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019.

[18] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos *et al.*, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC bioinformatics*, vol. 16, no. 1, p. 138, 2015.

[19] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

[20] M. Riemer, E. Khabiri, and R. Goodwin, "Representation stability as a regularizer for improved text analytics transfer learning," *arXiv preprint arXiv:1704.03617*, 2017.

[21] M. Riemer, T. Klinger, M. Franceschini, and D. Bouneffouf, "Scalable recollections for continual lifelong learning," *arXiv preprint arXiv:1711.06761*, 2017.

[22] J. Serra, D. Surís, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," *arXiv preprint arXiv:1801.01423*, 2018.

[23] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[24] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[25] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[26] G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, E. Gaussier, P. Gallinari, T. Artieres, M. R. Alvers, M. Zschunke *et al.*, "Bioasq: A challenge on large-scale biomedical semantic indexing and question answering," in *2012 AAAI Fall Symposium Series*, 2012.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.