Mitigating Catastrophic Forgetting for Zero-Shot Cross Lingual Transfer

Husun Bano
Computer and Info. Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
bano4003847@cloud.neduet.edu.pk

Fatima Azim
Computer and Info. Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
azim4002323@cloud.neduet.edu.pk

Safiya Mehmood
Computer and Info. Systems Engineering
NED University of Engineering & Technology
Karachi, Pakistan
mehmood4030214@cloud.neduet.edu.pk

Abstract*— *Pre-trained language models such as BERT [1] are being widely used in the cross-lingual setting. For optimal performance, fine-tuning of specific task labeled data is needed which is abundant for high resource languages. However, these models are prone to catastrophic forgetting during the fine-tuning stage as the parameters tend to overwrite crucial features which were originally learned. When used in a zero-shot environment, it performs inadequately on low resource languages and hence compromises the cross-linguality of the original language modeling (LM) task [2]. To mitigate this issue, we implement state of the art techniques: MAS [3] and Recall and Learn [4]. Our methods achieve better results than the baselines.

Keywords—Continual Learning, Zero-shot learning, catastrophic forgetting, pretrained language models, cross-language understanding, cross-lingual transfer

I. INTRODUCTION

The effectiveness of pre-trained language models (LMs), such as mBERT, has caused a stir in the NLP community. These models have been trained on huge corpora of multilingual data, and can be fine-tuned on task specific labeled data to perform well on these tasks (e.g. sentiment analysis, name entity tagging etc.). mBERT, for example, has been trained on 100+ languages, which means that the model understands the general use of all these languages.

However, pre-trained models [90] have shown to be prone to catastrophic forgetting [31]. When fine-tuning a pre-trained multilingual model using labeled data in a high resource locale (such as English), important multilingual features learned during the original language modeling task may be overwritten by the high-resource-locale based downstream task. This would make the model more over fitted to the source locale and affect its transferability to low-resource languages in both zero and few shot settings.

In NLP research, an abundant collection of annotated data is present but for very few languages, like English, French, Mandarin (Chinese) and Arabic [68]. That leaves all the other languages with little or no linguistic resources. Consequently, natural language understanding tasks become impossible to perform in these low-resource languages.

One solution to this problem is to collect annotated data for all languages, which is rather infeasible. We, instead, aim to train a model for particular tasks in a particular high-resource language and apply it to another low-resource language.

For this purpose, we implement MAS [3], and Recall and Learn [4], to mitigate catastrophic forgetting during fine-tuning and thus, preserving the cross-linguality of the model, when tested on languages like Urdu and Swahili. We perform regularized fine-tuning on popular pre-trained multilingual language model, mBERT-base, on three tasks: paraphrase detection [63], natural language inference [65], and intent classification and slot labeling [39].

Our contributions are as follows:

- We review and summarize some classic techniques used to alleviate catastrophic forgetting.
- We perform experiments to show that LMs do catastrophically forget.
- We apply MAS and RecAdam fine-tuning to a pre-trained LM, mBERT.
- We test our models on zero-shot cross-lingual scenarios.
- We perform ablation studies to investigate performance changes.
- We visualize weight changes in mBERT.

We believe our strategy will allow one to use the large amount of training data available for high-resource-languages for the benefit of low-resource-languages.

The rest of the paper is organized as follows. In section II we do a short literature of some relevant recent papers. In section III we describe the two techniques which we have chosen. The procedures of our experiments, as well as the results, have been detailed in section IV. Section V dives deeper into the experiments by testing performance after removing some features of the system. In section VI we study the inner workings of BERT by visualizing weight changes during training. In section VII we identify some beneficiaries of our work. In section VII we discuss some future work.

II. Related Works

A. Catastrophic Forgetting

Forgetting old lessons is part of the natural cognitive process. In humans, this happens gradually over a course of many years. Computers, on the other hand, forget dramatically and abruptly, a phenomenon commonly referred to in the AI community as 'catastrophic forgetting'. Catastrophic forgetting or catastrophic interference [27, 31, 32] is the tendency of a model to 'forget' previously learned knowledge when it learns new knowledge [5, 25, 26, 27]. Until recently, catastrophic forgetting was believed to be an inevitable feature of connectionist networks [5]. However, in recent years, much research has taken place in this domain, with various techniques successfully proving to mitigate catastrophic forgetting[5,7,9,8,10,11,12,13,17,24]. Techniques designed to alleviate catastrophic forgetting fall under one of the three categories [28, 29]:

- Regularization methods, which add constraints to weight updates so that important parameters from previous tasks are not updated when learning new tasks. These methods can be divided into node-based, which focus on node-level importance [25, 33, 34, 35, 12], weight-based, which compute importance of each weight in a given neural network [3, 36], and Lasso regularization methods [37], which have also been used for the continual learning problem [38].
- Architectural methods, in which particular layers, architectures, activation functions, or weight-freezing strategies are used to reduce the effects of catastrophic forgetting.
- Replay / Rehearsal methods, in which previously acquired knowledge is periodically mixed with newly, learned knowledge to strengthen older memory connections.

Elastic Weight Consolidation [5] is a regularization method that extenuates catastrophic forgetting by regularizing the loss i.e. slows down the parameter learning process important for prior tasks. It is implemented as a soft, quadratic constraint whereby each weight is retreated towards its old values by an amount proportional to its importance for performance on previously-learnt tasks. The weight update rule is constrained to ensure that the information learnt previously is retained when finding parameters to solve a new task. After a network is trained on one task, the "importance" of parameters of this network is calculated using the Fisher Information Matrix [6]. The Fisher matrix efficiently describes the posterior probability i.e. the probability that the weight was important given a past task.

Learn To Grow [8] is a continual structure learning framework to overcome catastrophic forgetting comprising two components: Neural Structure Optimization and Parameter Learning Component. In the first component, Neural Architecture Search (NAS) is employed wherein the goal is to choose the optimal choice for the current layer given a super network consisting of the training data for the present task and shared weights of layers. Three choices are: REUSE (using the existing weights fixed during learning), ADAPTATION (weight of the original kernel is fixed while parameters of the adapter are learned only), NEW (weights are randomly initialized and trained). The latter component either adds a regularization term such as L2 or EWC [5] or keeps the weights unchanged.

GEM [10] is a replay-based method in which the main component is an 'episodic memory', which stores samples from previous tasks. The GEM framework uses these samples to constrain gradients in a way which stops the loss on previous tasks from increasing. GEM minimizes backward transfer, which is the influence learning a new task has on previous tasks. In other words, it minimizes catastrophic forgetting. However, no significant positive forward transfer (zero-shot learning) is obtained.

A-GEM [11] is an improved version of GEM. Its performance matches that of GEM, while being as computationally efficient as EWC [5]. It provides constraints on the optimization problem by averaging gradients from past class exemplars, with the goal to stop the average episodic memory loss over previous tasks from increasing. According to A. Chaudhry et al., A-GEM is about 100 times faster than GEM and its memory requirements are 10 times less.

Hard Attention to Task [12] framework, works by penalizing modifications made to important layer activations, by learning hard attention masks with each task. HAT is able to retain previous task information, while learning new tasks, by using attention vectors of previous tasks to define almost binary masks and then providing constraints on the updated network weights on current tasks. This technique has been shown to reduce catastrophic forgetting by 45% to 80%.

REMIND [13] or replay using memory indexing is a brain-inspired approach which uses the hippocampal indexing theory [14]. Hippocampus is a complex brain structure which stores long-term memories and the hippocampal index represents neocortical regions activated by particular events. Hence, according to this theory, any cue which reactivates the stored hippocampal index would also reactivate the associated neocortical region, resulting in memory replay [15]. REMIND relies on the Product Quantization (PQ) [16] algorithm to store, and later replay, hidden intermediate representations (like CNN feature maps).

LAMOL [60] is an architecture-based approach where the model simultaneously learns the source task (generating answers given context) and generates pseudo-samples (generating the question, the answer and the context). These pseudo-samples are then jointly trained with new samples for the target task. A more efficient version of LAMOL, MFK-LAMOL [61], generates only those pseudo-samples that have been forgotten most.

Synaptic Intelligence [62] computes the importance of parameters using two quantities: i) the change in loss over the entire training trajectory (calculated during training), ii) how much does the newly learned parameter deviate from the previous parameter. Then, during the training phase, the model is penalized for deviating from parameters deemed important.

B. Continual Learning in NLP

Various techniques have been proposed to address the problem of continual learning [46] in NLP. [53] propose a strategy called "Mix-review" which involves replaying pre-train data during fine-tuning. [47] explore techniques to mitigate catastrophic forgetting during domain adaptation and test on reading comprehension benchmarks. Their findings conclude that using a combination of auxiliary penalty terms(normalized EWC, cosine distance and L2 [48]) during fine-tuning produces the best results. For the domain adaptation setup, forgetting also occurs in neural machine translation (NMT) [54]. [55] and [56] use EWC to mitigate forgetting in NMT. [49] introduce ARPER (Adaptively Regularized Prioritized Exemplar Replay), a method which replays prioritized past exemplars, together with EWC. In spoken language understanding (SLU), [58] introduce 'ProgModel', a novel progressive slot filling model for the semantic slot filling task. [59] study the continual learning scenario for sequence-to sequence tasks in NLP.

C. Zero-Shot Cross-Lingual Transfer

Zero-shot learning (ZSL) [86, 87, 88, 89] involves recognizing new classes with just a high level description of them. In the context of NLP, the model is trained using no labelled data or parallel text from the target language [30]. Existing work in this domain mostly deals with few-shot and one-shot scenarios [69, 70, 71]. ZSL with the help of meta-learning has also been studied [72, 67, 73, 74]. ZSL has been widely used to solve machine translation problems for under-resourced languages [75, 76, 77, 78] and for spoken language understanding tasks [79, 80]. Additionally, other problems in NLP have been addressed through ZSL, such as, fine-grained named entity typing [81], entity extraction from Web pages [82] and semantic utterance classification [83].

Cross-lingual transfer, in particular, has come under the spotlight in recent years. Researchers are increasingly realizing the value of using high-resource languages to solve tasks in languages for which resources are limited. This would have a variety of practical applications, such as, in goal-oriented dialogue systems [57]. [39] evaluates various cross-lingual transfer methods using MultiAtis++, a multilingual natural language understanding (NLU) corpus. Cross-lingual transfer learning has also been studied for named entity recognition tasks [40, 41, 42] and sequence tagging [43, 44, 45].

D. Zero-Shot Cross-Lingual Transfer

Very little research has been done for continual zero-shot learning [84, 85]. [84] were the first to address the problem of life-long ZSL using Variational Auto-Encoders, by integrating Knowledge Distillation, unified semantic embedding and selective retraining. This method, however, only works with multi-head settings. For the single-head setting, [5] builds on A-GEM.

A major reason for carrying out this research is the lack of linguistic resources for most languages, which stands in the way of any advanced language technology being developed in these low-resource languages.

III. METHODOLOGY

In this section, we describe the working of MAS and RecAdam. We also discuss the deviations we have made from the original MAS implementation in order to adapt to our setting. We chose this technique because it does not require retraining of mBERT and has proved to be the most effective regularization technique in continual learning [92, 93].

A. Memory Aware Synapses

Fine-tuning of the pre-trained language model is divided into two phases: i) computing importance and ii) regularized fine-tuning. For a learned function F, parameterized by θ, importance of each parameter can be estimated by measuring how sensitive function F is to changes δ in its parameters for each data point x_k:

$$F(x_k\,;\,\theta + \delta) - F(x_k : \theta) \qquad (1)$$

This value can be approximated by taking the gradient of the network output with respect to its parameters, as shown in (2) (In contrast, the gradients of loss with respect to parameters are calculated in classic deep learning). In our setting, network output refers to outputs of off-the-shelf mBERT.

$$g_{i_j}(x_k) = \frac{\partial[l_2^2 F(x_k;\theta)]}{\partial \theta_{i_j}} \qquad (2)$$

Finally, we accumulate the gradients over the entire dataset of size N to obtain importance weights omega:

$$\Omega_{i_j} = \frac{1}{N}\sum_{k=1}^{N} ||g_{i_j}(x_k)|| \qquad (3)$$

In order to compute network outputs, we do not require labels. Thus, unlabeled data can be used in this phase. We use the English dataset to compute importance.

In the next phase, that is, regularized fine-tuning, we fine-tune mBERT on the English corpus using the loss function in (4):

$$L(\theta) = L_n(\theta) + \frac{\lambda}{2}\sum_{ij}\Omega_{i_j}\left(\theta_{i_j} - \theta^*_{ij}\right)^2 \quad (4)$$

where θ^* refers to off-the-shelf mBERT parameters and theta refers to the parameters of mBERT with classifier on top

i.e. the model that is being fine-tuned. $L_n(\theta)$ is the loss function of the fine-tuning task. λ is a hyper parameter that controls the degree of forgetting. During fine-tuning, the model is penalized for making drastic updates to parameters that were deemed important in the earlier phase. Note that in this training phase, labeled data is required. We then evaluate and report the model's accuracy on low-resource locales.

B. Recall and Learn

Based on Adam optimizer and EWC (a classic catastrophic forgetting regularization technique); the approach uses the concept of multi-task learning i.e. learns pretraining and downstream tasks simultaneously instead of learning both tasks using Sequential Transfer Learning. We chose this technique because it does not require retraining of BERT and has shown to do well on GLUE tasks.

The method is divided into two phases i) Pre-training Simulation, and ii) Objective Shifting. Pretraining Simulation helps the model to recall information about previous tasks using only pretrained parameters. During this phase, the loss of the source task is evaluated as the quadratic penalty between the model and pretrained parameters so the model parameters are approximately close to the pretrained parameters, and thus multi-task learning is achieved.

$$Loss_S = \frac{1}{2}\gamma \sum_i (\theta_i - \theta_i^*)^2 \qquad (5)$$

where γ is the coefficient of the quadratic penalty.

Objective Shifting gradually learns downstream tasks by slowly shifting the objective function with the annealing coefficient. Initially, the training model learns knowledge from the pretrained tasks and then step by step acquires knowledge of target tasks and computes the final optimization objective $Loss_T$. The loss function incorporates the annealing coefficient with multi-task learning:

$$Loss = \lambda(t) Loss_T + (1 - \lambda(t)) Loss_S \qquad (6)$$

where $\lambda(t)$ is the sigmoid annealing function that updates timesteps t during fine-tuning, $Loss_T$ is the fine-tuning loss and $Loss_S$ is the pretraining loss.

Finally, the RecAdam optimizer integrates the quadratic penalty with the annealing coefficient by decoupling their respective gradients. The optimization steps result in the adaptation of only the gradient of the downstream task ∇f(θ) and all weights of the training model are penalized with the same rate $(1 - \lambda(t))\gamma$.

IV. Experiments

A. Model used

Multilingual BERT: BERT [1] stands for Bidirectional Encoder Representations from Transformers. mBERT is an unsupervised cross lingual language model, which has been pretrained on monolingual texts from 104 languages. BERT's model architecture is a multi-layer bidirectional Transformer encoder based on [90]. We fine-tune the top layer (i.e., Transformer block) on our three NLP tasks. The model size we specifically use is bert-base-multilingual-cased which has 12 layers, 768 hidden states, 12 self-attention heads and a total of 110M parameters.

B. Datasets

We evaluate our model on 3 datasets: PAWS-X, XNLI and MultiAtis++, commonly used in cross lingual language understanding (XLU) research.

PAWS-X [63], an extension of PAWS (Paraphrase Adversaries from Word Scrambling)[64], contains sentence pairs along with a paraphrase judgment, stating whether the sentences are paraphrases of each other or not. This means that the task on which we fine-tune mBERT is paraphrase detection. PAWS-X contains 23,659 human translated example pairs, and 49,401 machine translated pairs, each into the six languages: Spanish, Korean, Chinese, German, Japanese and French.

XNLI [65] is an extension of the test and dev sets of the Multi-Genre Natural Language Inference Corpus (MultiNLI), and performs the task of natural language inference. Simply put, it determines, given a 'premise', whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral). XNLI contains 7500 human-annotated dev and test set examples, in 15 languages, making it a total of 112500. These languages consist of high as well as low resource languages, namely, English, German, Spanish, French, Russian, Turkish, Greek, Arabic, Chinese, Bulgarian, Thai, Vietnamese, Hindi, Urdu and Swahili.

MultiAtis++[39] is a natural language understanding(NLU) corpus, which we will use to fine-tune mBERT on the tasks, intent classification and slot labeling. As suggested by the name, intent classification deals with predicting the intent of the speaker's utterance, while in slot labeling, the model extracts semantic entities in a sentence, which are related to the predicted intent. MultiAtis++ adds six new languages to the original Multilingual ATIS [66] corpus, making it a total of 9 languages: English, Spanish, Portuguese, German, French, Chinese, Japanese, Hindi and Turkish. It consists of 37,084 training examples and 7,859 test examples.

C. Baselines

We compare our methods to two baselines. (i) **Vanilla Finetune**: finetune both mBERT and XLNet on the three NLP tasks (ii) **Off-the-shelf model** : freezing all layers of the pre-trained model and training only the classifier. This is the lower bound. In all the above experiments we train the model for 3-5 epochs with a learning rate of 1e-5 and a batch size of 32. We use the Adam optimizer for our baselines.

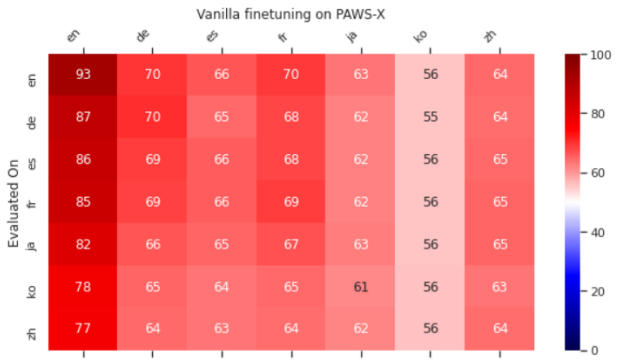

Fig. 1. Heatmap showing the average accuracy after training and testing on all languages in the PAWS-X dataset, without applying any special fine-tuning technique to fight catastrophic forgetting.

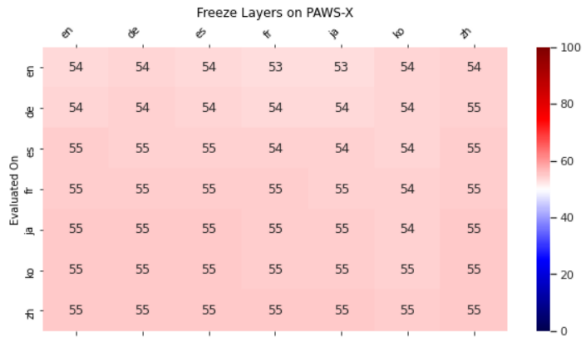

Fig. 2. Heatmap showing how the average accuracy declines when we freeze all layers in a model, and only train the classifier. The model has been trained and tested on all 7 languages. The PAWS-X dataset has been used in this example.

D. Setup and Implementation

We utilize the multilingual BERT base model from the GluonNLP[1] library for MAS experiments, and from the HuggingFace[2] library for RecAdam experiments. We train the two catastrophic forgetting fighting methods, MAS and RecAdam, on each of the three tasks and then compare the results to baselines.

E. Results

We use classification accuracy as our metric for evaluation of PAWS-X and XNLI, and intent classification and slot labeling for MultiAtis++.

PAWS-X: Table 1 compares the results of baselines with the zero-shot results of the RecAdam fine-tuning method, and MAS fine-tuning method, on the PAWS-X dataset. The models have been trained on English data, and then directly evaluated on all other languages. **Off-the-shelf model** and **Vanilla fine-tuning** are our two baselines. **MAS fine-tuning** and **RecAdam fine-tuning** consistently outperforms vanilla fine-tuning on all languages, with an overall improvement of **6%** and **3.5%** respectively.

[1] https://nlp.gluon.ai/model_zoo/bert/index.html
[2] https://huggingface.co/transformers/model_doc/bert.html

TABLE I. ZERO-SHOT RESULTS ON PAWS-X

	en	de	fr	ja	ko	zh	Avg
Off-the-shelf model	54	55	55	55	55	55	54.83
Vanilla*	82.2	86.0	85.2	70.5	71.7	75.8	78.56
MAS	88.48	87.82	87.24	84.6	79.32	79.18	**84.44**
RecAdam	84.45	86.85	87.9	76.75	76.75	79.75	**82.08**

***Table 1**: Zero-shot results of training on English dataset with all layers frozen, vanilla fine-tuning, MAS and RecAdam fine-tuning on the PAWS-X dataset. *Vanilla fine-tuning results taken from the official PAWS-X paper (zero-shot) [63].*

Figure 3: Zero-shot results using GluonNLP's BERTAdam optimizer on PAWS-X

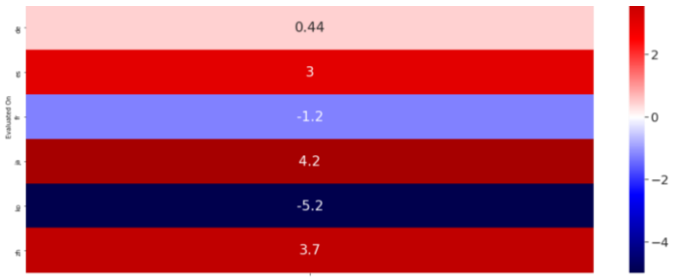

Figure 4: We used 500 of human translated data of all non-English languages to compute MAS-based parameter importance. These are zero-shot results obtained on PAWS-X.

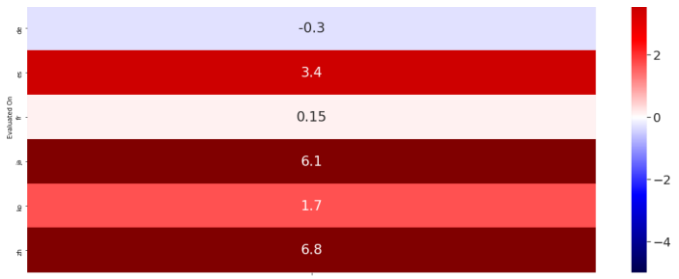

Figure 5: We used 10% of the machine translated training data of all non-English languages to compute MAS-based parameter importance. These are zero-shot results obtained on PAWS-X.

Figures 3-5 illustrate the language-wise percentage improvements from the vanilla results in multiple scenarios.

XNLI: Table 2 and 3 show that MAS fine-tuning significantly improves the zero-shot performance on all

languages, with an average improvement of **5%**. Interestingly, the performance declines with RecAdam, with an average deterioration of 5%. The results of 14 languages in the XNLI dataset have been distributed between table 2 and 3.

MultiAtis++ results on Naive MAS and MAS: MultiAtis++ is a relatively small dataset due to which we were able to train and zero-shot test MAS fine-tuning on every language. As shown in Figures 6(a) and (b), these results on Naïve MAS are rather unstable, showing improvement as high as 17% on intent classification and 35% on slot labeling, and as low as -27% on intent classification and -13% on slot labeling. Table 4 shows an average improvement of **14.4%** with **MAS fine-tuning.**

TABLE II. ZERO-SHOT RESULTS ON XNLI

	ar	bg	de	el	es	fr	hi
Off-the-shelf model	35.6	36.2	36.1	36.6	35.7	35.2	33.2
Vanilla*	63.3	66.8	70.0	64.8	73.6	72.9	58.4
MAS	65.8	68.0	72.2	69.1	76.3	78.7	63.1
RecAdam	61.0	66.8	66.5	57.8	72.2	70.6	48.5

***Table 2**: Zero-shot results of training on English dataset with all layers frozen, vanilla fine-tuning, MAS and RecAdam fine-tuning on the XNLI dataset for Arabic, Bulgarian, German, Greek, Spanish, French and Hindi. *Vanilla fine-tuning results taken from [30].*

TABLE III. ZERO-SHOT RESULTS ON XNLI

	ru	sw	th	tr	ur	vi	zh	AVG
Off-the-shelf model	34.8	36.2	35.2	35.7	35.0	35.2	36.0	35.47
Vanilla*	67.3	47	51	60	56	69	68	**63.44**
MAS	70.7	55.8	57.0	67.8	61.3	73.0	79.2	**68.43**
RecAdam	66.1	47.5	43.1	53.5	55.2	40.5	67.8	**58.36**

***Table 3**: Zero-shot results of training on English dataset with all layers frozen, vanilla fine-tuning, MAS and RecAdam fine-tuning on the XNLI dataset for Russian, Swahili, Thai , Turkish, Urdu, Vietnamese and Chinese. *Vanilla fine-tuning results taken from [30].*

TABLE IV. ZERO-SHOT RESULTS ON MULTIATIS++

	es	de	fr	ja	pt	hi	zh	AVG
Off-the-shelf model	66.1	65.3	66.2	40.6	63.6	56.5	51.7	56.4
Vanilla*	74.9	82.6	75.7	35.7	74.0	31.2	62.3	57.5
Naïve MAS	74.9	82.1	75.7	35.7	74.0	31.0	62.2	**57.4**
MAS	84.9	87.1	84.5	65.8	84.7	48.5	80.0	**71.9**

***Table 4**: Zero-shot results of training on English dataset with all layers frozen, vanilla fine-tuning, Naïve MAS and MAS fine-tuning on the MULTIATIS++ dataset for Spanish, German, French , Japanese, Portuguese, Hindi and Chinese. *Vanilla fine-tuning results taken from [39].*

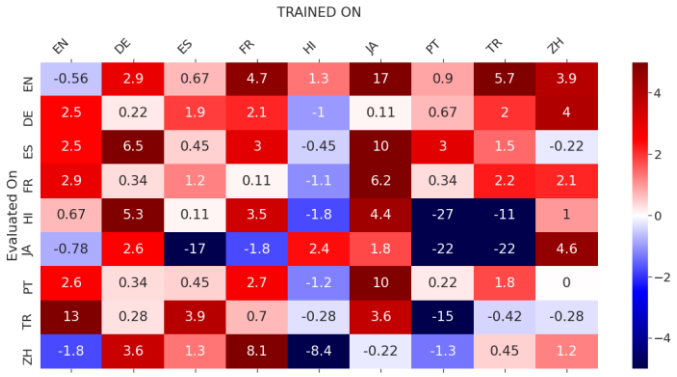

Figure 6(a): MultiAtis++ on Naive MAS (Intent classification)

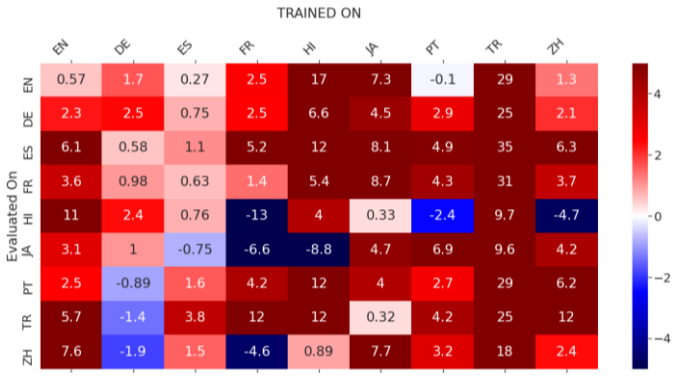

Figure 6(b): MultiAtis++ on Naive MAS (Slot Labeling)

V. ABLATION STUDIES

During the computing importance phase, we removed the classifier layer, layernorm, all bias layers, embeddings and pooler layers, that is, we do not regularize these layers. We found that this strategy reduces performance. Removing the classifier and embedding layer gives best results. We also found that increasing the number of epochs during phase I does not improve results nor does it hurt performance.

When computing Ω, we do update the weights. We found that updating weights during an Ω update gives better

performance than not updating weights. After updating Ω, we use an off-the-shelf mBERT for fine-tuning in phase II.

Some results of our ablation studies have been presented in Appendix A.

VI. BERTology

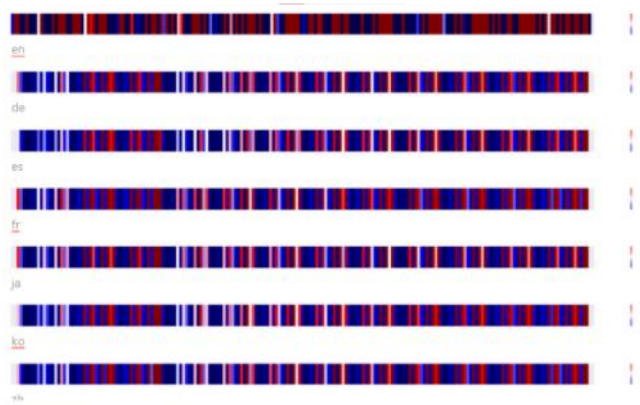

Figure 7(a) Parameter Importance Mean per Layer

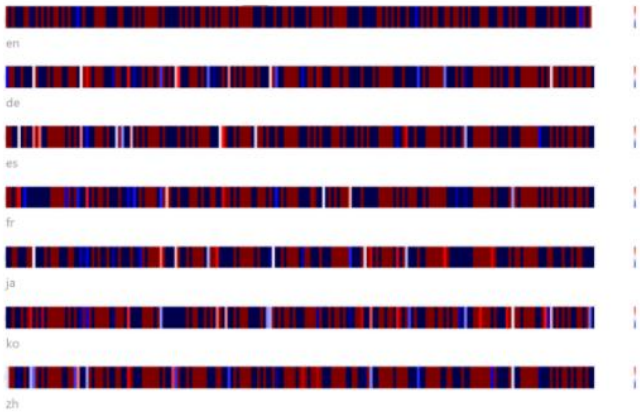

Figure 7(b) Weight difference mean per layer

Figure 7(a) shows the mean of parameter importance (Ω) per layer. The classifier and embedding layers are omitted since these layers are not regularized. It can be seen that non-English languages have almost the same parameter importance pattern and the important layers for English are different. The heatmap of the English's parameter importance and weight difference is almost identical which shows that Omega was able to encapsulate parameter importance rather closely. The weight differences in different languages are different, meaning that different parameters do correspond to different languages.

In English's omega, layernorm beta consistently has high importance. Pooler holds high importance. The attention cells, especially key and query layers, have the least importance, which means that they are least sensitive to the output and that they do not change significantly during training.

VII. Target Market

Our approach towards zero-shot cross-lingual transfer can help accelerate progress in the following research areas:

- Goal-Oriented Dialogue Systems: To carry out effective conversation, with the user goal in focus, dialogue systems need to be able to understand users from all backgrounds. This means that they should be able to deal with non-English speakers just as well as with English speakers. Cross-lingual transfer greatly improves system performance in this case.
- Multi-lingual systems: Contemporary NLP systems require huge collections of labeled data to learn tasks. Since most data is almost only available in English, how do we make sure that advanced language technology is made available in all languages? Collecting data in every language is an impractical solution. This is where zero-shot cross lingual transfer comes in. Systems used in major international products may encounter inputs from various languages; hence efficient cross-lingual transfer is necessary.
- Plagiarism Detection: Paraphrase detection (plagiarism) is widely used to remove duplicate content from the internet. As more and more information becomes available on the internet in languages other than English, the need to perform this task in low resource languages also increases.

VIII. Conclusion and Future Work

In this paper we compare two techniques to tackle the problem of catastrophic forgetting in pre-trained LMs. The techniques we have chosen do not require retraining of LMs and hence, are computationally less expensive. Our experiments show that mBERT indeed does catastrophically forget and that continual learning techniques can be used to alleviate this phenomenon in language models.

We hope that our strategy will help reduce the dependency on labeled training data which is absent for most languages, and bridge the digital divide between high and low resource languages.

Our work is primarily focused on exploring MAS and RecAdam to mitigate catastrophic forgetting. We leave the study of various other techniques, like A-GEM in a cross-lingual setting, for future work. Techniques which require retraining of LMs can also be explored, given enough resources. Ablation studies and hyper parameter tuning can be performed to achieve better results. In addition, visualization techniques can be used to better understand how weights change during fine-tuning. A performance comparison can be done between other language models such as ALBERT, RoBERTa, and XLM-R as an extension of this work.

Acknowledgments

We thank Dr. Ali Ismail and Batool Haider for their guidance and support throughout the research. This work was partially funded by the National Grassroots Research Initiative (NGIRI).

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.

[2] D. Yogatama et al., "Learning and Evaluating General Linguistic Intelligence." Accessed: Apr. 22, 2021. [Online]. Available: https://rajpurkar.github.io/SQuAD-explorer/.

[3] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory Aware Synapses: Learning What (not) to Forget," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11207 LNCS, pp. 144–161, 2018, doi: 10.1007/978-3-030-01219-9_9.

[4] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," arXiv, pp. 7870–7881, 2020, doi: 10.18653/v1/2020.emnlp-main.634.

[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei ARusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2016. Overcoming catastrophic forgetting in neural networks.Proceedings of the National Academy of Sciences of the United States of America, 114(13):3521

[6] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks.arXiv preprintarXiv:1301.3584, 2013

[7] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network.InNIPS Deep Learning and Representation Learning Workshop.

[8] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to Grow: A Continual Structure Learning Framework for Overcoming Catastrophic Forgetting."

[9] F. Zenke, B. Poole, and S. Ganguli, "Continual Learning Through Synaptic Intelligence," 2017.

[10] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in Advances in Neural Information Processing Systems, 2017, vol. 2017-December.

[11] A. Chaudhry, R. Marc'Aurelio, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2019.

[12] J. Serra, D. Suris, M. Mirón, and A. Karatzoglou, "Overcoming Catastrophic forgetting with hard attention to the task," 35th Int. Conf. Mach. Learn. ICML 2018, vol. 10, pp. 7225–7234, 2018.

[13] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "REMIND Your Neural Network to Prevent Catastrophic Forgetting," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12353 LNCS, pp. 466–483, 2020, doi: 10.1007/978-3-030-58598-3_28.

[14] T. Teyler and J. Rudy, "The hippocampal indexing theory and episodic memory: Updating the index", Hippocampus, vol. 17, no. 12, pp. 1158-1169, 2007. Available: 10.1002/hipo.20350.

[15] J. O'Neill, B. Pleydell-Bouverie, D. Dupret and J. Csicsvari, "Play it again: reactivation of waking experience and memory", Trends in Neurosciences, vol. 33, no. 5, pp. 220-229, 2010. Available: 10.1016/j.tins.2010.01.006.

[16] H. Jégou, M. Douze and C. Schmid, "Product Quantization for Nearest Neighbor Search", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, 2011. Available: 10.1109/tpami.2010.57.

[17] C. Lee, K. Cho, and W. Kang, "Mixout: Effective regularization to finetune large-scale pretrained language models," arXiv. 2019.

[18] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," CoNLL 2017 - 21st Conf. Comput. Nat. Lang. Learn. Proc., vol. 2, pp. 281–289, 2017, doi: 10.18653/v1/k17-1029.

[19] B. Romera-Paredes and P. H. S. Torr, "An embarrassingly simple approach to zero-shot learning.

[20] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," Accessed: Jul. 11, 2021. [Online]. Available: https://github.com/.

[21] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," Accessed: Jul. 11, 2021. [Online]. Available: https://github.com/facebookresearch/cc.

[22] Y. Liu et al., "Multilingual Denoising Pre-training for Neural Machine Translation," doi: 10.1162/tacl.

[23] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," Accessed: Jul. 11, 2021. [Online]. Available: https://pypi.org/project/langdetect/.

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. In 14th European Conference on Computer Vision (ECCV 2016), volume 9908 LNCS, pages 614–629, Amsterdam, Netherlands, 2016.

[25] S. Jung, H. Ahn, S. Cha, and T. Moon, "Continual Learning with Node-Importance based Adaptive Group Sparse Regularization," no. NeurIPS, 2020, [Online]. Available: http://arxiv.org/abs/2003.13726.

[26] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., 2014.

[27] R. Ratcliff, "Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions," Psychol. Rev., vol. 97, no. 2, pp. 285–308, 1990, doi: 10.1037/0033-295X.97.2.285.

[28] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," 32nd AAAI Conf. Artif. Intell. AAAI 2018, pp. 3390–3398, 2018.

[29] P. Keung, Y. Lu, J. Salazar, and V. Bhardwaj, "Don't Use English Dev: On the Zero-Shot

[30] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios", Neural Networks, vol. 116, pp. 56-73, 2019. Available: 10.1016/j.neunet.2019.03.010.Cross-Lingual Evaluation of Contextual Embeddings," Accessed: Jul. 17, 2021. [Online]. Available: https://github.com/facebookresearch/MLDoc.

[31] M. McCloskey and N. J. Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem," Psychol. Learn. Motiv. - Adv. Res. Theory, vol. 24, no. C, pp. 109–165, Jan. 1989, doi: 10.1016/S0079-7421(08)60536-8.

[32] R. Ratcliff, "Psychological Review Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions," Psychol. Assoc. Inc, vol. 97, no. 2, pp. 285–308, 1990.

[33] H. Ahn, S. Cha, D. Lee, and T. Moon, "Uncertainty-based continual learning with adaptive regularization," Adv. Neural Inf. Process. Syst., vol. 32, no. NeurIPS, 2019.

[34] R. Aljundi, T. Tuytelaars, and M. Rohrbach, "Selfless sequential learning," 7th Int. Conf. Learn. Represent. ICLR 2019, pp. 1–17, 2019.

[35] S. Golkar, M. Kagan, and K. Cho, "Continual Learning via Neural Pruning," 2019, [Online]. Available: http://arxiv.org/abs/1903.04476.

[36] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr, "Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11215 LNCS, no. 1, pp. 556–572, 2018, doi: 10.1007/978-3-030-01252-6_33.

[37] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 68, no. 1, pp. 49–67, 2006, doi: 10.1111/j.1467-9868.2005.00532.x.

[38] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., pp. 1–11, 2018.

[39] W. Xu, B. Haider, and S. Mansour, "End-to-End Slot Alignment and Recognition for Cross-Lingual NLU," pp. 5052–5063, 2020, doi: 10.18653/v1/2020.emnlp-main.410.

[40] A. Zirikly and M. Hagiwara, "Cross-lingual transfer of named entity recognizers without parallel corpora," ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., vol. 2, pp. 390–396, 2015, doi: 10.3115/v1/p15-2064.

[41] C. T. Tsai, S. Mayhew, and D. Roth, "Cross-lingual named entity recognition via wikification," CoNLL 2016 - 20th SIGNLL Conf. Comput. Nat. Lang. Learn. Proc., pp. 219–228, 2016, doi: 10.18653/v1/k16-1022.

[42] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell, “Neural cross-lingual named entity recognition with minimal resources,” Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 369–379, 2020, doi: 10.18653/v1/d18-1034.

[43] D. Yarowsky, G. Ngai, and R. Wicentowski, “Inducing multilingual text analysis tools via robust projection across aligned corpora,” pp. 1–8, 2001, doi: 10.3115/1072133.1072187.

[44] O. Täckström, D. Das, S. Petrov, R. McDonald, and J. Nivre, “Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging,” Trans. Assoc. Comput. Linguist., vol. 1, pp. 1–12, 2013, doi: 10.1162/tacl_a_00205..

[45] B. Plank and Ž. Agić, “Distant supervision from disparate sources for low-resource part-of-speech tagging,” Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 614–620, 2020, doi: 10.18653/v1/d18-1061..

[46] Ring, M., 2021. Continual Learning in Reinforcement Environments. [online] Cs.utexas.edu. Available at: <https://www.cs.utexas.edu/~ring/Ring-dissertation.pdf>.

[47] Y. Xu, X. Zhong, A. J. J. Yepes, and J. H. Lau, “Forget Me Not: Reducing Catastrophic Forgetting for Domain Adaptation in Reading Comprehension,” Proc. Int. Jt. Conf. Neural Networks, 2020, doi: 10.1109/IJCNN48605.2020.9206891.

[48] G. Wiese, D. Weissenborn, and M. Neves, “Neural domain adaptation for biomedical question answering,” CoNLL 2017 - 21st Conf. Comput. Nat. Lang. Learn. Proc., vol. 2, no. CoNLL, pp. 281–289, 2017, doi: 10.18653/v1/k17-1029.

[49] F. Mi, L. Chen, M. Zhao, M. Huang, and B. Faltings, “Continual Learning for Natural Language Generation in Task-oriented Dialog Systems,” pp. 3461–3474, 2020, doi: 10.18653/v1/2020.findings-emnlp.310.

[50] A. A. Rusu et al., “Progressive Neural Networks,” 2016, [Online]. Available: http://arxiv.org/abs/1606.04671.

[51] L. Mou et al., “How transferable are neural networks in NLP applications?,” EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 479–489, 2016, doi: 10.18653/v1/d16-1046.

[52] A. Chronopoulou, C. Baziotis, and A. Potamianos, “An embarrassingly simple approach for transfer learning from pretrained language models,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 2089–2095, 2019, doi: 10.18653/v1/n19-1213..

[53] T. He et al., “Analyzing the Forgetting Problem in the Pretrain-Finetuning of Dialogue Response Models,” 2019, [Online]. Available: http://arxiv.org/abs/1910.07117.

[54] [54] D. Saunders, F. Stahlberg, A. de Gispert, and B. Byrne, “Domain adaptive inference for neural machine translation,” ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 222–228, 2020, doi: 10.18653/v1/p19-1022.

[55] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, “Overcoming catastrophic forgetting during domain adaptation of neural machine translation,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. 2018, pp. 2062–2068, 2019, doi: 10.18653/v1/n19-1209.

[56] D. Variš and O. Bojar, “Unsupervised pretraining for neural machine translation using elasticweight consolidation,” ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Student Res. Work., pp. 130–135, 2019, doi: 10.18653/v1/p19-2017..

[57] S. Lee, “Toward Continual Learning for Conversational Agents,” 2017, [Online]. Available: http://arxiv.org/abs/1712.09943.

[58] Y. Shen, X. Zeng, and H. Jin, “A progressive model to enable continual learning for semantic slot filling,” EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 1279–1284, 2020, doi: 10.18653/v1/d19-1126.

[59] Y. Li, L. Zhao, K. Church, and M. Elhoseiny, “Compositional Language Continual Learning,” Iclr, pp. 1–15, 2020.

[60] F.-K. Sun, C.-H. Ho, and H.-Y. Lee, “LAMOL: LAnguage MOdeling for Lifelong Language Learning,” ICLR, pp. 1–15, 2019, [Online]. Available: http://arxiv.org/abs/1909.03329.

[61] H. Choi and P. Kang, "Lifelong Language Learning With the Most Forgotten Knowledge," in IEEE Access, vol. 9, pp. 57941-57948, 2021, doi: 10.1109/ACCESS.2021.3071787.

[62] F. Zenke, B. Poole, and S. Ganguli, “Continual learning through synaptic intelligence,” 34th Int. Conf. Mach. Learn. ICML 2017, vol. 8, pp. 6072–6082, 2017.

[63] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, “PAWS-X: A cross-lingual adversarial dataset for paraphrase identification,” EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 3687–3692, 2020, doi: 10.18653/v1/d19-1382.

[64] Y. Zhang, J. Baldridge, and L. He, “PAWS: Paraphrase adversaries from word scrambling,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. 2, pp. 1298–1308, 2019.

[65] A. Conneau et al., “XNLI: Evaluating Cross-lingual Sentence Representations,” Proc. 2018 Conf. Empir. Methods Nat. Lang. Process., pp. 2475–2485, 2018, [Online]. Available: http://arxiv.org/abs/1809.05053.

[66] S. Upadhyay, M. Faruqui, G. Tür, H. Dilek and L. Heck, "(Almost) Zero-Shot Cross-Lingual Spoken Language Understanding," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6034-6038, doi: 10.1109/ICASSP.2018.8461905..

[67] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, “Zero-Shot Cross-Lingual Transfer with Meta Learning,” pp. 4547–4562, 2020, doi: 10.18653/v1/2020.emnlp-main.368.

[68] "The #BenderRule: On Naming the Languages We Study and Why It Matters", The Gradient, 2021. [Online]. Available: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/. [Accessed: 09- Sep- 2021].

[69] X. Han et al., “Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation,” Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp. 4803–4809, 2020, doi: 10.18653/v1/d18-1514.

[70] M. Yu et al., “Diverse few-shot text classification with multiple metrics,” NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 1206–1215, 2018, doi: 10.18653/v1/n18-1109..

[71] N. Rethmeier and I. Augenstein, “Data-Efficient Pretraining via Contrastive Self-Supervision,” 2020, [Online]. Available: http://arxiv.org/abs/2010.01061..

[72] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. K. Li, “Meta-Learning for Low-Resource Neural Machine Translation,” pp. 3622–3631, 2018..

[73] Z. Wang, Q. Wang, and D. W. Wang, “Bayesian network based business information retrieval model,” Knowl. Inf. Syst., vol. 20, no. 1, pp. 63–79, 2009, doi: 10.1007/s10115-008-0151-5.

[74] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” 34th Int. Conf. Mach. Learn. ICML 2017, vol. 3, pp. 1856–1868, 2017.

[75] O. Firat, “Zero-Resource Neural Machine Translation with,” pp. 268–277, 2016.

[76] M. Johnson et al., “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation,” Trans. Assoc. Comput. Linguist., vol. 5, pp. 339–351, 2017, doi: 10.1162/tacl_a_00065.

[77] H. Nakayama and N. Nishida, “Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot,” Mach. Transl., vol. 31, no. 1–2, pp. 49–64, 2017, doi: 10.1007/s10590-017-9197-z.

[78] H. Zheng, Y. Cheng, and Y. Liu, “Maximum expected likelihood estimation for zero-resource neural machine translation,” IJCAI Int. Jt. Conf. Artif. Intell., vol. 0, pp. 4251–4257, 2017, doi: 10.24963/ijcai.2017/594.

[79] Ferreira, E., Bassam Jabaian and F. Lefèvre. “Zero-shot semantic parser for spoken language understanding.” INTERSPEECH (2015).

[80] M. Yazdani and J. Henderson, “A model of zero-shot learning of spoken language understanding,” Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process., no. September, pp. 244–249, 2015, doi: 10.18653/v1/d15-1027.

[81] Y. Ma, E. Cambria, and S. Gao, “Label embedding for zero-shot fine-grained named entity typing,” COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap., pp. 171–180, 2016.

[82] P. Pasupat and P. Liang, “Zero-shot entity extraction from web pages,” 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf., vol. 1, pp. 391–401, 2014, doi: 10.3115/v1/p14-1037.

[83] Y. N. Dauphin, G. Tur, D. Hakkani-Tür, and L. Heck, “Zero-shot learning for semantic utterance classification,” 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc., pp. 1–9, 2014.

[84] K. Wei, C. Deng, and X. Yang, “Lifelong zero-shot learning,” IJCAI Int. Jt. Conf. Artif. Intell., vol. 2021-January, pp. 551–557, 2020, doi: 10.24963/ijcai.2020/77.

[85] I. Skorokhodov and M. Elhoseiny, “Class Normalization for (Continual)? Generalized Zero-Shot Learning,” no. 1, pp. 1–31, 2020, [Online]. Available: http://arxiv.org/abs/2006.11328.

[86] Y. Cai, Z. Ding, B. Yang, Z. Peng, and W. Wang, “Zero-Shot Learning Through Cross-Modal Transfer,” Phys. A Stat. Mech. its Appl., vol. 514, pp. 729–740, 2019.

[87] L. Zhang, T. Xiang, and S. Gong, “Learning a deep embedding model for zero-shot learning,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-January, pp. 3010–3019, 2017, doi: 10.1109/CVPR.2017.321.

[88] B. Zhao, X. Sun, Y. Fu, Y. Yao, and Y. Wang, “MSplit LBI: Realizing feature selection and dense estimation simultaneously in few-shot and zero-shot learning,” 35th Int. Conf. Mach. Learn. ICML 2018, vol. 13, pp. 9421–9432, 2018.

[89] L. Chen, H. Zhang, J. Xiao, W. Liu and S. Chang, "Zero-Shot Visual Recognition Using Semantics-Preserving Adversarial Embedding Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1043-1052, doi: 10.1109/CVPR.2018.00115.

[90] A. Vaswani et al., “Attention is all you need,” Adv. Neural Inf. Process. Syst., vol. 2017-December, no. Nips, pp. 5999–6009, 2017.

[91] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” Adv. Neural Inf. Process. Syst., vol. 32, no. NeurIPS, pp. 1–11, 2019.

[92] M. Elhoseiny, F. Babiloni, R. Aljundi, M. Rohrbach, M. Paluri, and T. Tuytelaars, “Exploring the Challenges Towards Lifelong Fact Learning,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11366 LNCS, no. Lll, pp. 66–84, 2019, doi: 10.1007/978-3-030-20876-9_5.

[93] M. Delange et al., “A continual learning survey: Defying forgetting in classification tasks,” IEEE Trans. Pattern Anal. Mach. Intell., no. c, pp. 1–29, 2021, doi: 10.1109/TPAMI.2021.3057446.

Appendix A

	en
en	93.6
de	89.05
es	88.3
fr	87.675
ja	85.03
ko	79.7583
zh	79.85

Figure A.1: *Results on PAWS-X by omitting omega of classifier only*

	en
en	92.35
de	87.175
es	86.5167
fr	86.05
ja	83.45
ko	78.1167
zh	77.9286

Figure A.2: *Results on PAWS-X by omitting omega of pooler, bias, embeddings and classifier layers*

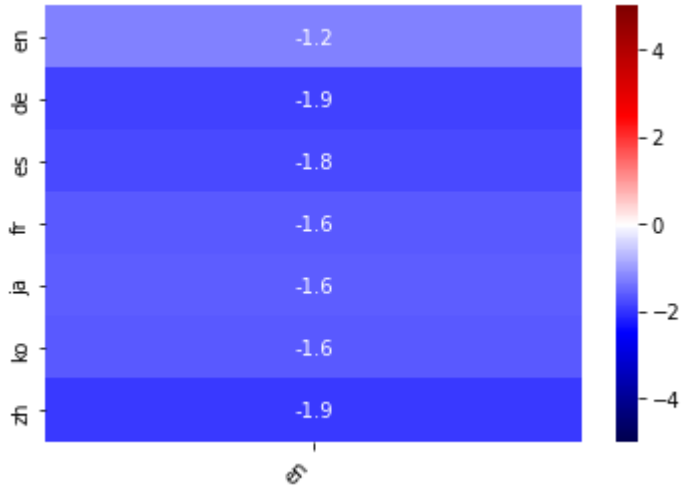

Figure A.3: The difference of accuracies before and after removing selected layers from Omega