A COMPREHENSIVE, APPLICATION-ORIENTED STUDY OF CATASTROPHIC FORGETTING IN DNNs

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a large-scale empirical study of catastrophic forgetting (CF) in modern Deep Neural Network (DNN) models that perform sequential (or: incremental) learning. A new experimental protocol is proposed that takes into account typical constraints encountered in application scenarios. As the investigation is empirical, we evaluate CF behavior on the hitherto largest number of visual classification datasets, from each of which we construct a representative number of Sequential Learning Tasks (SLTs) in close alignment to previous works on CF. Our results clearly indicate that there is no model that avoids CF for all investigated datasets and SLTs under application conditions. We conclude with a discussion of potential solutions and workarounds to CF, notably for the EWC and IMM models.

1 INTRODUCTION

This article is in the context of sequential or incremental learning in Deep Neural Networks (DNNs). Essentially, this means that a DNN is not trained once, on a single task $D$, but successively on two or more sub-tasks $D_1, \ldots, D_n$, one after another. Learning tasks of this type, which we term Sequential Learning Tasks (SLTs) (see Fig. 1a), are potentially very common in real-world applications. They occur wherever DNNs need to update their capabilities on-site and over time: gesture recognition, network traffic analysis, or face and object recognition in mobile robots. In such scenarios, neural networks have long been known to suffer from a problem termed “catastrophic forgetting” (CF) (e.g., French (1999)) which denotes the abrupt and near-complete loss of knowledge from previous sub-tasks $D_1, \ldots, D_{k-1}$ after only a few training iterations on the current sub-task $D_k$ (see Fig. 1b compared to Fig. 1c). We focus on SLTs from the visual domain with two sub-tasks each, as DNNs show pronounced CF behavior even when only two sub-tasks are involved.

![Figure 1: Scheme of incremental training experiments conducted in this article (a) and representative outcomes with (b) and without CF (c). The sequential learning tasks used in this study only have two sub-tasks: $D_1$ and $D_2$. During training (white background) and re-training (gray background), test accuracy is measured on $D_1$ (blue, △), $D_2$ (green, □) and $D_1 \cup D_2$ (red, ○). The blue curve allows to determine the presence of CF by simple visual inspection: if there is significant degradation w.r.t. the red curve, then CF has occurred.](image-url)
1.1 RELATED WORK ON AVOIDING CF

The field of incremental learning is large, e.g., Parisi et al. (2018) and Gepperth & Hammer (2016). Principal recent approaches include ensemble methods (Ren et al., 2017; Fernando et al., 2017), dual-memory systems (Shin et al., 2017; Kemker & Kanan, 2017; Rebuffi et al., 2017; Gepperth & Karaoguz, 2015) and single-DNN regularization approaches which are in focus here. Whereas Goodfellow et al. (2013) suggest Dropout for alleviating CF, the EWC method (Kirkpatrick et al., 2016) proposes to add a term to the energy function that protects weights that are important for the previous sub-task(s). Importance is determined by approximating and analyzing the Fisher information matrix of the DNN. A somewhat related approach is pursued with the Incremental Moment Matching technique (IMM) (see Lee et al. (2017)), where weights from DNNs trained on a current and a past sub-tasks are “merged” using the Fisher information matrix. Other regularization-oriented approaches are proposed in Aljundi et al. (2018); Srivastava et al. (2013) and Kim et al. (2018) which focus on enforcing sparsity of neural activities by lateral interactions within a layer. In the following, we will review recent regularization approaches w.r.t. their feasibility in application scenarios.

Number of tested datasets In general, most methods referenced here are evaluated only on a few datasets, usually on MNIST (LeCun et al., 1998) and various derivations thereof (permutation, rotation, class separation). Some studies make limited use of CIFAR10, SVHN, the Amazon sentiment analysis problem, and non-visual problems such as data from Q-learning of Atari games. A large-scale evaluation on a huge number of qualitatively different datasets is still missing

Model selection and prescience Model selection (i.e., selecting DNN topology and hyper-parameters) is addressed in some approaches (Goodfellow et al., 2013) but on the basis of a “pre-scient” evaluation where the best model is selected after all tasks have been processed, an approach which is replicated in Kirkpatrick et al. (2016). This amounts to a knowledge of future sub-tasks which is problematic in applications. Most approaches ignore model selection (Lee et al., 2017; Srivastava et al., 2013; Aljundi et al., 2018; Kim et al., 2018), and thus implicitly violate causality.

Storage of data from previous sub-tasks From a technical point of view, DNNs can be retrained without storing training data from previous sub-tasks, which is done in Goodfellow et al. (2013) and Srivastava et al. (2013). For regularization approaches, however, there are regularization parameters that control the retention of previous knowledge, and thus must be chosen with care. In Kirkpatrick et al. (2016), this is $\lambda$, whereas two such quantities occur in Lee et al. (2017): the “balancing” parameter $\alpha$ and the regularization parameter $\lambda$ for L2-transfer. The only study where regularization parameters are obtained through cross-validation (which is avoided in other studies) is Aljundi et al. (2018) (for $\lambda_{SN1}$ and $\lambda_1$) but this requires to store all previous training data.

This review shows that enormous progress has been made, but that there are shortcomings tied to applied scenarios which need to be addressed. We will formalize this in Sec. 1.2 and propose an evaluation strategy that takes these formal constraints into account when testing CF in DNNs.

1.2 INCREMENTAL LEARNING IN APPLIED SCENARIOS

When training a DNN model on SLTs, first of all the model must be able to be retrained at any time by new classes (class-incremental learning). Secondly, it must exhibit retention, or at least graceful decay, of performance on previously trained classes. Some forgetting is probably unavoidable, but it should be gradual and not immediate, i.e., catastrophic. However, if a DNN is operating in, e.g., embedded devices or autonomous robots, additional conditions may be applicable:

Low memory footprint Data from past sub-tasks cannot be stored and used for re-training, or else to determine when to stop re-training.

Causality Data from future sub-tasks, which are often known in academic studies but not in applications, must not be utilized in any way, especially not for DNN model selection. This point might seem trivial, but a number of studies such as Kirkpatrick et al. (2016); Goodfellow et al. (2013) and Srivastava et al. (2013) perform model selection in hindsight, after having processed all sub-tasks.

Constant update complexity Re-training complexity (time and memory) must not depend on the number of previous sub-tasks, thus more or less excluding replay-based schemes such as Shin et al. (2017). Clearly, even if update complexity is constant w.r.t. the number of previous sub-tasks, it should not be too high in absolute terms either.

\footnote{Although the comparisons performed in Aljundi et al. (2018) include many datasets, the experimental protocol is unclear, so it is uncertain how to interpret these results.}
1.3 CONTRIBUTION AND PRINCIPAL CONCLUSIONS

The original contributions of our work can be summarized as follows:

- We propose a novel training and model selection paradigm for incremental learning in DNNs that incorporates typical application constraints, see Sec. 1.2.
- We investigate the incremental learning capacity of various DNN approaches (Dropout, LWTA, EWC and IMM) using the largest number of qualitatively different classification datasets so far described. We find that all investigated models are afflicted by catastrophic forgetting, or else in violation of application constraints and discuss potential workarounds.
- We establish that the “permuted” type of SLTs (e.g., “permuted MNIST”) are not recommended when testing for CF as they can be solved by almost any model for any dataset.

1.4 APPROACH OF THIS ARTICLE

We collect a large number of visual classification datasets, from each of which we construct SLTs according to a common scheme, and compare several recent DNN models using these SLTs. The experimental protocol is such that application constraints, see Sec. 1.2, are enforced.

2 METHODS

In the following section, the used models, hyper-parameters, the sequential learning tasks, the tested datasets and the general procedure for evaluation of the experiments are described.

2.1 USED DNN MODELS

For all DNN models described here, we use a TensorFlow (v1.7) implementation under Python (v3.4 and later). The source code for all processed models, the experiment-generator and evaluation routine can be found on our public available repository (dummy_link_for_double_blind_review_test_with_same_length.org).

**FC** A normal, fully-connected (FC) feed-forward DNN with a variable number and size of hidden layers, each followed by ReLU, and a softmax readout layer minimizing cross-entropy.

**CONV** A convolutional neural network (CNN) based on the work of Cirean et al. (2011). It is optimized to perform well on image classification problems like MNIST. We use a fixed topology: two conv-layers with 32 and 64 filters of size $5 \times 5$ plus ReLU and $2 \times 2$ max-pooling, followed by an fc-layer with 1024 neurons and softmax readout layer minimizing a cross-entropy energy function.

**EWC** The Elastic Weight Consolidation (EWC) model presented by Kirkpatrick et al. (2016).

**LWTA** A fully-connected DNN with a variable number and size of hidden layers, each followed by a Local Winner Takes All (LWTA) transfer function as proposed in Srivastava et al. (2013).

**IMM** The Incremental Moment Matching model as presented by Lee et al. (2017). We examine the weight-transfer and L2-transfer techniques in our experiments, using the provided implementation.

**D-FC** and **D-CONV** Motivated by Goodfellow et al. (2013) we combine the FC and CONV models with Dropout as an approach to solve the CF problem. Only FC and CONV are eligible for this, as EWC and IMM include dropout by default, and LWTA is incompatible with Dropout.

2.2 HYPER-PARAMETERS AND MODEL SELECTION

We perform model selection in all our experiments by a combinatorial hyper-parameter optimization, whose limits are imposed by the computational resources available for this study. In particular, we vary the number of hidden layers $L \in \{2, 3\}$ and their size $S \in \{200, 400, 800\}$ (CNNs excluded), the learning rate $\epsilon_1 \in \{0.01, 0.001\}$ for sub-task $D_1$, and the re-training learning rate $\epsilon_2 \in \{0.001, 0.0001, 0.00001\}$ for sub-task $D_2$. The batch size (batchsize) is fixed to 100 for all experiments, and is used for both training and testing. As in other studies, we do not use a fixed number of training iterations, but specify the number of training epochs (i.e., passes through the whole dataset) as $E = 10$ for each processed dataset (see Sec. 2.3), which allows an approximate comparison of different datasets. The number of training/testing batches per epoch, $B$, can be calculated from the batch size and the currently used dataset size. The set of all hyper-parameters for
a certain model, denoted $P$, is formed as a Cartesian product from the allowed values of the hyperparameters $L$, $S$, $\epsilon_1$, $\epsilon_2$ and complemented by hyper-parameters that remain fixed ($E$, batch size) or are particular to a certain model. For all models that use dropout, the dropout rate for the input layer is fixed to 0.2, and to 0.5 for all hidden layers. For CNNs, the dropout rate is set to 0.5 for both input and hidden layers. All other hyper-parameters for CNNs are fixed, e.g., number and size of layers, the max-pooling and filter sizes and the strides ($2 \times 2$) for each channel. These decisions were made based on the work of Goodfellow et al. (2013). The LWTA block size is fixed to 2, based on the work of Srivastava et al. (2013). The model parameter $\lambda$ for EWC is set to $\lambda_1/\epsilon_2$ (set but not described in the source code of Kirkpatrick et al. (2016)). For all models except IMM, the momentum parameter for the optimizer is set to $\mu = 0.99$ (Sutskever et al., 2013). For the IMM models, the SGD optimizer is used, and the regularizer value for the L2-regularization is set to 0.01 for L2-transfer and to 0.0 for weight transfer.

2.3 Datasets

We select the following datasets (see Tab. 1). In order to construct SLTs uniformly across datasets, we choose the 10 best-represented classes (or random classes if balanced) if more are present.

- **MNIST** (LeCun et al., 1998) is the common benchmark for computer vision systems and classification problems. It consists of gray scale images of handwritten digits (0-9).
- **EMNIST** (Cohen et al., 2017) is an extended version of MNIST with additional classes of handwritten letters. There are different variations of this dataset: we extract the ten best-represented classes from the By Class variation containing 62 classes.
- **Fruits 360** (Murean & Oltean, 2017) is a dataset comprising fruit color images from different rotation angles spread over 75 classes, from which we extract the ten best-represented ones.
- **Devanagari** (Acharya, 2015) contains gray scale images of Devanagari handwritten letters. From the 46 character classes (1.700 images per class) we extract 10 random classes.
- **FashionMNIST** (Xiao et al., 2017) consists of images of clothes in 10 classes and is structured like the MNIST dataset. We use this dataset for our investigations because it is a “more challenging classification task than the simple MNIST digits data” (Xiao et al., 2017).
- **SVHN** (Netzer & Wang, 2011) is a 10-class dataset based on photos of house numbers (0-9). We use the cropped digit format, where the number is centered in the color image.
- **CIFAR10** (Krizhevsky, 2009) contains color images of real-world objects e.g., dogs, airplanes etc.
- **NotMNIST** (Bulatov Yaroslav) contains grayscale images of the 10 letter classes from “A” to “J”, taken from different publicly available fonts.
- **MADBase** (Abdelazeem Sherif & El-Sherif Ezzat) is a modified version of the “Arabic Digits dataBase”, containing grayscale images of handwritten digits written by 700 different persons.

Table 1: Overview of each dataset’s detailed properties. Image dimensions are given as width × height × channels. Concerning data imbalance, the largest percentual difference in sample count between any two classes is given for training and test data, a value of 0 indicating a perfectly balanced dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Properties</th>
<th>number of elements</th>
<th>class balance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>image size</td>
<td>train</td>
<td>test</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>32 × 32 × 3</td>
<td>50,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Devanagari</td>
<td>32 × 32 × 1</td>
<td>18,000</td>
<td>2,000</td>
</tr>
<tr>
<td>EMNIST</td>
<td>28 × 28 × 1</td>
<td>345,035</td>
<td>57,918</td>
</tr>
<tr>
<td>FashionMNIST</td>
<td>28 × 28 × 1</td>
<td>60,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Fruits 360</td>
<td>100 × 100 × 3</td>
<td>6,148</td>
<td>2,052</td>
</tr>
<tr>
<td>MADBase</td>
<td>28 × 28 × 1</td>
<td>60,000</td>
<td>10,000</td>
</tr>
<tr>
<td>MNIST</td>
<td>28 × 28 × 1</td>
<td>55,000</td>
<td>10,000</td>
</tr>
<tr>
<td>NotMNIST</td>
<td>28 × 28 × 1</td>
<td>529,114</td>
<td>18,724</td>
</tr>
<tr>
<td>SVHN</td>
<td>32 × 32 × 3</td>
<td>73,257</td>
<td>26,032</td>
</tr>
</tbody>
</table>

2.4 Sequential Learning Tasks (SLTs)

As described in Sec. 1, each SLT consists of two sub-tasks $D_1$ and $D_2$. For each dataset (see Sec. 2.3), these are defined by either subdividing classes into disjunct groups, or by applying differ-
ent spatial permutations to all image data (see Tab. 2). SLTs resulting from a subdivision of classes are denoted, e.g., D9-1a. The first and second digits represent the number of classes assigned to D1 and D2, respectively. The last letter is used to distinguish between different choices of classes for both sub-tasks. From each dataset, we create three SLTs of the type D9-1, eight SLTs of the type D5-5 and one SLT termed DP10-10, resulting from permutation.

Table 2: Overview of all SLTs. The assignment of classes to sub-tasks (STs) are disjunct, except for DP10-10 where two different random image permutations are applied for D1 and D2.

<table>
<thead>
<tr>
<th>SLT</th>
<th>D5-5</th>
<th>D9-1</th>
<th>DP10-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>a b c d e f g h</td>
<td>a b c d e f g h</td>
<td>a b c</td>
</tr>
<tr>
<td>D1</td>
<td>0-4 02468 34689 02567 01345 03489 05678 02368</td>
<td>0-8 1-9 0,2-9</td>
<td>0-9</td>
</tr>
<tr>
<td>D2</td>
<td>5-9 13579 01257 13489 26789 12567 12349 14579</td>
<td>9 0 1</td>
<td>0-9</td>
</tr>
</tbody>
</table>

3 EXPERIMENTS

This study presents just one, albeit very large, experiment, whose experimental protocol implements the constraints from Sec. 1.2. Every DNN model from Sec. 2.1 is applied to each SLT as defined in Sec. 2.4 while taking into account model selection, see Sec. 2.2.

A precise definition of our application-oriented experimental protocol is given in Alg. 1. For a given model m and an SLT (D1 and D2), the first step is to determine the best hyper-parameter vector p∗ for sub-task D1 only (see lines 1-4), which determines the model m∗ used for re-training.

In a second step, m∗ (from line 5) is used for re-training on D2, with a different learning rate ϵ2 which is varied separately. We introduce two criteria for determining the (ϵ2-dependent) quality of a re-training phase (lines 6-10): “best”, defined by the highest test accuracy on D1 ∪ D2, and “last”, defined by the test accuracy on D1 ∪ D2 at the end of re-training. Although the “best” criterion violates the application constraints of Sec. 1.2 (requires D1), we include it for comparison purposes.

Finally, the result is computed as the highest ϵ2-dependent quality (line 11).

Independently of the second step, another training of m∗ is conducted using D1 ∪ D2, resulting in what we term baseline accuracy.

Evaluation for IMM differs slightly: in line 5 a copy of m∗ is kept, termed m1∗, and the weights of m∗ are re-initialized. After selecting the best re-trained model m2∗ as a function of ϵ2, final performance q∗ is obtained by “merging” the models m1∗ and m2∗ and testing the result.

Algorithm 1: The application-oriented evaluation strategy used in this study.

Data: model m, SLT with sub-tasks D1, D2, hyper-parameter value set P

Result: incremental learning quality for model with hyper-parameters p∗: q∗

1. for all p ∈ P do
2. for t ← 0 to E · B do
3. train(m∗, D1train, ϵ1)
4. q∗,t ← test(m∗, D1test, t)
5. m∗ ← model m∗,t with maximum q∗,t // find best model with max. accuracy on D1
6. for all ϵ2 do
7. m∗ ← m∗
8. for t ← 0 to E · B do
9. train(m∗, ϵ2, D2train, ϵ2)
10. q∗,t,ϵ2 ← test(m∗, ϵ2, D2test, t)
11. q∗ ← max best/last, q∗,t,ϵ2 // find parameter set with the best accuracy on D2

4 FINDINGS

The results of the experiment described in Sec. 3 are summarized in Tab. 4 and in Tab. 4 for IMM. They lead us to the following principal conclusions:
Permutation-based SLTs should not be used when investigating CF. We find that DP10-10, the SLT based on permutation, does not show CF for any model and dataset, which is exemplary visualized for the FC model in Fig. 2 which fails completely for the other SLTs.

All examined models exhibit CF. While this is not surprising for FC and CONV, D-FC as proposed in Goodfellow et al. (2013) performs poorly (see Fig. 3), as does LWTA (Srivastava et al., 2013). For EWC and IMM, the story is slightly more complex and will be discussed below.

EWC is mildly effective against CF for simple SLTs. Our experiments show that EWC is effective against CF for D9-1, at least when the “best” evaluation criterion is used, which makes use of $D_1$. This, in turn, violates the application requirements of Sec. 1.2. For the “last” criterion not making use of $D_1$, EWC performance, though still significant, is much less impressive. We can see the origins of this difference illustrated in Fig. 4.

EWC is ineffective against CF for more complex problems. Tab. 3 shows that EWC cannot prevent CF for D5-5 type SLTs, see Fig. 5 for a visualization.
Apparently, the EWC mechanism cannot protect all the weights relevant for $D_1$ here, which is likely to be connected to the fact that the number of samples in both sub-tasks is similar. This is not the case for D9-1 type tasks where EWC does better and where $D_2$ has about 10% of the samples in $D_1$.

Figure 5: Best EWC experiments for SLT D5-5d constructed from all datasets, to be read as Fig. 2. We observe that CF happens for all datasets.

Table 3: Summary of incremental learning quality $q_{\ell^p}$, see Alg. 1, over SLTs of type D9-1, D5-5 and DP10-10. For aggregating results over SLTs of the same type, the minimal value of $q_{\ell^p}$ is taken. For DP10-10 and D5-5 type tasks, CF (failure to retain $D_1$) is indicated by qualities $< 0.5 = \theta$, failure to learn $D_2$ by qualities of $\approx \theta$. The corresponding value for D9-1 type tasks is $\theta = 0.1$. Only when $\theta$ is exceeded, re-training can be said to be successful. Each cell contains two qualities evaluated according to the “best” and “last” criteria, see Alg. 1.
**IMM is effective for all SLTs but unfeasible in practice.** As we can see from Tab. 3, wtIMM (we performed the same experiments for l2IMM and obtained slightly better but qualitatively similar results) clearly outperforms all other models compared in Tab. 3. Especially for the D5-5 type SLTs, a modest incremental learning quality is attained, which is however quite far away from the baseline accuracy, even for MNIST-derived SLTs. This is in contrast to the results reported in [Lee et al. (2017)] for MNIST: we attribute this discrepancy to the application-oriented model selection procedure using only $D_1$ that we perform. In contrast, in [Lee et al. (2017)], a model with 800/800/800 neurons, for which good results on MNIST are well established, is chosen beforehand, thus arguably making implicit use of $D_2$. A significant problem of IMM is the determination of the balancing parameter $\alpha$, exemplarily illustrated in Fig. 6. Our results show that the optimal value cannot simply be guessed from the relative sizes of $D_1$ and $D_2$, as it is done in [Lee et al. (2017)], but must be determined by cross-validation, thereby requiring knowledge of $D_1$ (violates constraints). Apart from these conceptual issues, we find that the repeated calculation of the Fisher matrices is quite time and memory-consuming (>4h and >8GB), to the point that the treatment of SLTs from certain datasets becomes impossible even on high-end machine/GPU combinations when using complex models. This is why we can only evaluate IMM for a few datasets. It is possible that this is an artifact of the TensorFlow implementation, but in the present state IMM nevertheless violates not one but two application constraints from Sec. 1.2. Fig. 7 and Fig. 8 give a visual impression of training an IMM model on D9-1 and D5-5 type SLTs, again illustrating basic feasibility, but also the variability of the “tuning curves” we use to determine the optimal balancing parameter $\alpha$.

![Figure 6](image6.png)

**Figure 6:** Accuracy measurements of best IMM model on SLT D5-5f for Devanagari dataset. On the left-hand side, the blue curve ($\triangle$) measures the accuracy of the first DNN trained on $D_1$, the green curve ($\circ$) the accuracy of the second DNN trained on $D_2$. Additionally, the baseline (dashed line) is delineated. The right-hand side shows the tested accuracy on $D_1 \cup D_2$ of the merged DNN as a function of $\alpha$, both for mean-IMM (red $\bullet$) and the mode-IMM (orange $\triangledown$) variants.

![Figure 7](image7.png)

**Figure 7:** Best wtIMM experiments for SLT D5-5b constructed from datasets we were able to test. The blue surfaces (epochs 0-10) represent the test accuracy during training on $D_1$, the green surfaces the test accuracy on $D_2$ during training on $D_2$ (epochs 10-20). The white bars in the middle represent baseline accuracy, whereas the right part shows accuracies on $D_1 \cup D_2$ for different $\alpha$ values, computed for mean-IMM (orange surfaces) and mode-IMM (red surfaces).

5 CONCLUSIONS

The primary conclusion from the results in Sec. 4 is that CF still represents a major problem when training DNNs. This is particularly true if DNN training happens under application constraints as...
Table 4: Summary of incremental learning quality $q^*$, see Alg. 1, for the IMM model, evaluated on SLTs of type D9-1, D5-5 (DP10-10 is omitted because near-perfect performance was always attained). For aggregating results over SLTs of the same type, the minimal value of $q^*$ (the best) is taken, as the presence of CF is indicated by a single occurrence of it in any SLT of the same type. To be interpreted as Tab. 3.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR10 D5-5</th>
<th>Devanagari D5-5</th>
<th>F_MNIST D5-5</th>
<th>MADBase D5-5</th>
<th>MNIST D5-5</th>
<th>SVHN D5-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode</td>
<td>.31 .73 .70</td>
<td>.43 .85 .78</td>
<td>.70</td>
<td>.91 .91</td>
<td>.84 .87</td>
<td>.56 .60</td>
</tr>
<tr>
<td>mean</td>
<td>.30 .67 .62</td>
<td>.43 .85 .78</td>
<td>.70</td>
<td>.91 .92</td>
<td>.82 .88</td>
<td>.50 .59</td>
</tr>
</tbody>
</table>

outlined in Sec. 1.2. Some of these constraints may be relaxed depending on the concrete application: if some prior knowledge about future sub-task exists, it can be used to simplify model selection and improve results. If sufficient resources are available, a subset of previously seen data may be kept in memory and thus allow a “best” type evaluation/stop criterion for re-training, see Alg. 1.

In general application scenarios without prior knowledge or extra resources, however, an essential conclusion we draw from Sec. 4 is that model selection must form an integral part of training a DNN on SLTs. Thus, a wrong choice of hyper-parameters based on $D_1$ can be disastrous for the remaining sub-tasks, which is why application scenarios require DNN variants that do not have extreme dependencies on hyper-parameters such as layer number and layer sizes.

Furthermore, from the comparison of “last” and “best” evaluation criterion in Fig. 4, the importance of finding a good stopping criterion for DNN re-training can be observed. Such a criterion would unambiguously indicate when to stop re-training before too much knowledge about $D_1$ is lost, ideally without making use of $D_1$. This is possible, e.g., in prototype-based classification methods such as LVQ and variants, see Nova & Estévez (2014), where the change of internal parameters due to re-training can provide a measure of damage to knowledge about $D_1$.

Lastly, our findings indicate workarounds that would make EWC or IMM practicable in at least some application scenarios. If model selection is addressed, a small subset of $D_1$ may be kept in memory for both methods; to determine optimal values of $\alpha$ for IMM and to determine when to stop re-training for EWC. Fig. 4 shows that small changes to $\alpha$ do not dramatically impact final accuracy for IMM, and Fig. 4 indicates that accuracy loss as a function of re-training time is gradual in most cases for EWC. The inaccuracies introduced by using only a subset of $D_1$ would therefore not be very large for both algorithms.

To conclude, this study shows that the consideration of applied scenarios significantly changes the procedures to determine CF behavior, as well as the conclusions as to its presence in latest-generation DNN models. We propose and implement such a procedure, and as a consequence claim that CF is still very much of a problem for DNNs. More research, either on generic solutions, or on workarounds for specific situations, needs to be conducted before the CF problem can be said to be solved. A minor but important conclusion is that the use of “permuted” SLTs should be avoided in future studies on CF.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.
REFERENCES

Abdelazeem Sherif and El-Sherif Ezzat. AHDBase. URL http://datacenter.aucegypt.edu/shazeem/.

Shailesh Acharya. Deep Learning Based Large Scale Handwritten Devanagari Character Recognition. 2015.


