Supplementary Material for ”IL2M: Class Incremental Learning With Dual Memory”

Eden Belouadah
CEA, LIST,
F-91191 Gif-sur-Yvette, France
eden.belouadah@cea.fr

Adrian Popescu
CEA, LIST,
F-91191 Gif-sur-Yvette, France
adrian.popescu@cea.fr

1. Introduction

In this supplementary material, we provide:

- a detailed description of the used datasets,
- the top-5 results for all algorithms,
- the top-1 error analysis of past and new classes over incremental batches,
- detailed plots of top-1 obtained results for all datasets with $Z = 10$ and $K = \{20000, 5000\}$,
- algorithm implementation details.

2. Description of evaluation datasets

The datasets used in the evaluation are designed for three visual classification tasks: object, face and tourist landmark recognition. To facilitate reproducibility, we chose to perform the evaluation with publicly available datasets whose training set main statistics are provided in Table 1.

2.1. ILSVRC

ILSVRC [9] is the well known subset of ImageNet used in the ILSVRC competitions and is reused here. The statistics from Table 1 show that the training set is well balanced, with an average of 1231.2 images per class and a 70.2 standard deviation. The dataset is available for download from http://image-net.org/download.

2.2. VGGFace2

VGGFace2 [2] is a recent dataset focused on face recognition. It includes over 9000 unique identities. We selected the 1000 identities which have the largest number of associated images for the evaluation in order to have a dataset similar in size to ILSVRC. VGGFace2 is well balanced and includes a mean of 491.7 images per class, with 49.4 standard deviation. The dataset includes loosely cropped face images and, following the usual face recognition pipeline, we extracted tighter crops before training and testing. Face detection was done using the publicly available MTCNN [10] framework. The dataset is available for download from http://www.robots.ox.ac.uk/~vgg/data/vgg_face2/.

	ILSVRC	VGGFace2	Landmarks
Train images mean	1231.2	491.7	374.4
Train images std	70.2	49.4	103.8

Table 1: Main statistics for evaluation datasets. The two lines provide: (1) the mean number of train images per class and (2) the standard deviation of the number of train images per class.

2.3. Google Landmarks

Google Landmarks [6] (Landmarks below) is a dataset built for tourist landmark recognition. It includes over 2 million images for over 30000 landmarks across the world. Again, we selected the 1000 landmarks which have the largest number of associated images for the evaluation. The selected train subset is more imbalanced than ILSVRC and VGGFace2, with a mean number of 374.4 images per class and 103.8 standard deviation. The dataset is available for download from https://www.kaggle.com/google/google-landmarks-dataset

3. Top-5 accuracy results

In addition to the top-1 results from the paper, we provide top-5 results obtained by all methods to facilitate comparability with earlier works [3, 5, 8]. Overall, the results follow the same trend as top-1. It is noteworthy that the differences between the FT baseline and the methods built on top of it are globally lower than top-1 results. This is particularly true for the VGGFace2 and Landmarks, the easier datasets tested here, where the imbalance inherent to incremental learning matters less than in the case of ILSVRC. The smaller performance differences are explained by the fact that top-5 accuracy has a smoothing effect on results. $IL2M$ is still the best method in a majority of tested con-

States	$Z=10$												$K=5000$					
Dataset	ILSVRC				VGGFace2				Landmarks				ILSVRC		VGGFace2		Landmarks	
K	$20k$	$10k$	$5k$	$0k$	$20k$	$10k$	$5k$	$0k$	$20k$	$10k$	$5k$	$0k$	Z=5	Z=20	Z=5	Z=20	Z=5	Z=20
$iCaRL$	62.5	61.4	60.9	43.8	84.5	83.9	83.6	48.3	84.4	83.6	83.0	46.3	61.0	56.3	89.4	71.6	89.0	71.2
$DeeSIL$	74.5	74.3	**74.2**	**73.9**	92.6	92.6	92.5	**92.3**	94.2	94.1	94.0	93.6	**79.2**	69.0	**96.4**	87.2	**96.4**	90.3
FT	77.0	70.1	60.0	20.5	97.1	96.0	94.1	21.3	**97.6**	96.5	94.4	21.3	61.9	64.5	95.6	94.4	94.6	93.8
FT^{NEM}	**79.4**	74.5	69.6	20.5	96.7	95.7	94.1	21.3	96.8	95.8	93.9	21.3	71.2	**71.4**	95.4	**94.6**	93.2	93.6
FT^{BAL}	77.5	73.4	65.0	20.5	**97.2**	**96.2**	94.3	21.3	97.5	96.5	94.6	21.3	70.1	67.8	96.1	94.5	95.4	**94.0**
$IL2M$	78.3	**75.2**	71.2	20.5	**97.2**	**96.2**	**94.9**	21.3	**97.6**	**96.6**	**94.7**	21.3	75.6	66.1	**96.4**	94.5	95.3	93.6
$Full$	92.3				99.2				99.1				92.3		99.2		99.1	

Table 2: Top-5 average accuracy (%) for the different methods tested. The available memory K and the number of states Z are varied to test their effect on the performance of the tested methods. Following [3], accuracy is averaged only for incremental states (i.e. excluding the initial, non-incremental state). Best results are in bold. $Full$ is the non-incremental upper-bound performance obtained with all data available for all classes.

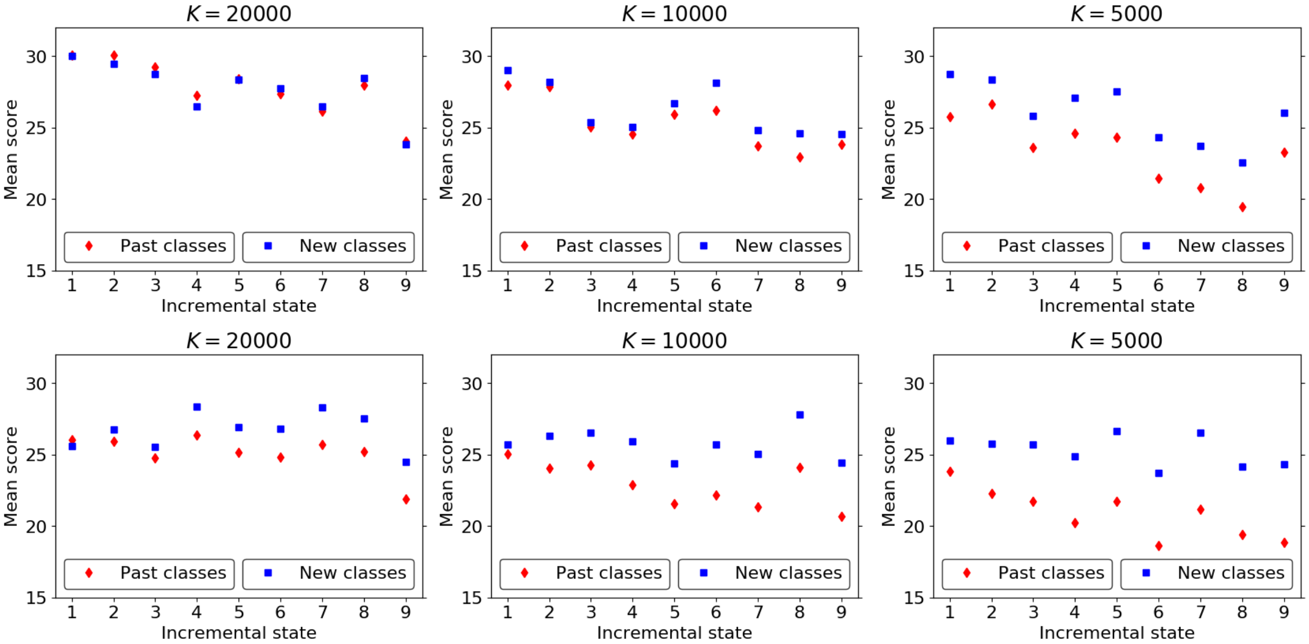

Figure 1: Prediction scores for Landmarks [6] (up) and VGGFace2 [2] (down) datasets with memory $K = \{20000, 10000, 5000\}$ exemplars and $Z = 10$ states. We select the scores of the true class for train images and then average them for past and new classes. Incremental states from 1 to 9 are represented. The initial state (0) does not include past classes and is not represented. (*Best viewed in color.*)

figurations. A first notable difference is that FT^{NEM} gives slightly better results for three configurations instead of one for top-1. A second difference is that $DeeSIL$ has best performance for all datasets with $K = 5000$ and $Z = 5$. This is due to the fact that the initial representation is stronger when it includes a higher number of classes. $DeeSIL$ has the best top-5 performance for ILSVRC with $K = 5000$ and $IL2M$ comes second in this case.

Compared to $Full$, the non-incremental training, the best class IL algorithms with $Z = 10$ and memory $K = 20000$ loses 12.9, 2 and 1.5 top-5 points for ILSVRC, VGGFace2 and Landmarks respectively . This gap is rather small for VGGFace2 and Landmarks, but more work is still needed for difficult tasks like ILSVRC. Naturally, the gap increases when the memory is reduced and the number of states increases. As expected, it becomes very important without memory. In this last case, which is not in focus here, the $DeeSIL$ baseline performs best for all three datasets.

4. Effect of data imbalance on predicted scores

In Figure 1, we provide scores plots for past and new classes for VGGFace2 and Landmarks. This figure is a

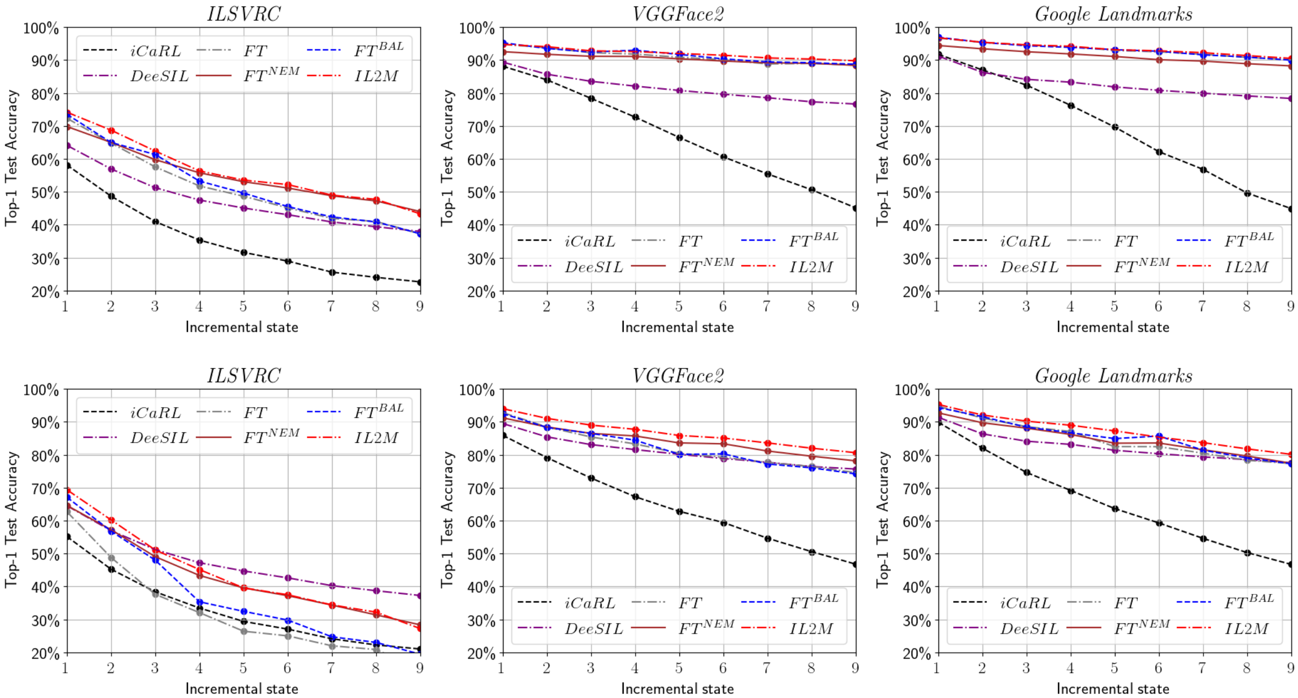

Figure 2: Top-1 accuracy for object, face and landmark recognition with $Z = 10$ states and memory $K = 20000$ (up) and $K = 5000$ (down). To be aligned with the results from paper in Table 2, only the incremental states are represented. (*Best viewed in color.*)

complement to Figure 2 of the paper, where similar analysis was provided for ILSVRC. The difference of mean scores between past and new classes for VGGFace2 and Landmarks grows as memory is reduced from left to right of the figure. This trend is natural since imbalance increases and it was also observed for ILSVRC in Figure 2 of the paper. Compared to ILSVRC, the differences between predicted scores of past and new classes are much smaller for VGGFace2 and negligible for Landmarks when $K = 20000$. This explains the very small contribution of $IL2M$ score rectification in this configuration.

5. Error analysis

The analysis from the previous two sections shows that data imbalance inherent to class IL with memory produces a classification bias toward new classes. In Table 3, we enrich the analysis by providing an analysis of error types before (FT) and after ($IL2M$) score rectification with memory $K = 10000$ and $Z = 10$ states.

Before rectification, the largest number of errors is of type $e(p, n)$, that is test images of past classes mistaken for images of new classes. We will look closely at the incremental state 9 of ILSVRC, which includes 45000 and 5000 test images for past and new classes respectively. 30740/45000 (68%) of test images of past classes were predicted as new and only 8746/45000 images were correctly predicted. 4267/5000 (85.34%) of test images of new classes are predicted correctly and only 66/5000 of them are assigned to past classes. These statistics further confirm the bias in favor of new classes and the need for score rectification.

After rectification with $IL2M$, the distributions of correct predictions and of errors changes quite significantly. For ILSVRC, there are significantly more correct predictions for past classes, accompanied by a lower performance for new classes. In state 9 of ILSVRC, correct predictions of past test images increase from 19.43% with FT to 32.86% with $IL2M$. The corresponding performance for new classes drops from 85.34% to 70.2%. $IL2M$ ensures a better performance balance between past and new classes. The errors of type $e(p, p)$, where images of a past class are mistaken for images of another past class are increasingly frequent toward later incremental states. This covers a majority of cases for states from 5 to 9. The number of images of past classes predicted as new decreases significantly and these errors cover only 21.32% of test images for past classes in state 9 of ILSVRC.

6. Implementation details

iCaRL [8] was run with SGD optimizer and binary cross entropy loss for classification (+ distillation term) following the same parameterization given by authors in their

			Incremental states								
Dataset			1	2	3	4	5	6	7	8	9
$ILSVRC$	FT	$c(p)$	2621	4327	5730	6702	7600	7980	8576	9169	8746
		$e(p,p)$	194	690	1360	2203	3035	4016	4462	6100	5514
		$e(p,n)$	2185	4983	7910	11095	14365	18004	21962	24731	30740
		$c(n)$	4139	4314	4145	4155	4251	4319	4236	4376	4267
		$e(n,n)$	779	608	771	762	692	619	694	560	667
		$e(n,p)$	82	78	84	83	57	62	70	64	66
	$IL2M$	$c(p)$	3223	5913	7744	9279	11233	11899	13115	13563	14791
		$e(p,p)$	433	2010	3374	5324	9177	11239	13984	16780	20614
		$e(p,n)$	1344	2077	3882	5397	4590	6862	7901	9657	9595
		$c(n)$	3940	3791	3815	3816	3484	3774	3552	3900	3510
		$e(n,n)$	666	409	582	553	352	361	398	347	341
		$e(n,p)$	394	800	603	631	1164	865	1050	753	1149
$VGGFace2$	FT	$c(p)$	4619	8887	13114	17234	21279	25163	29084	32617	36893
		$e(p,p)$	62	275	580	898	1270	1638	2051	2649	3145
		$e(p,n)$	319	838	1306	1868	2451	3199	3865	4734	4962
		$c(n)$	4789	4814	4847	4868	4873	4879	4878	4868	4884
		$e(n,n)$	167	129	115	87	90	88	86	92	88
		$e(n,p)$	44	57	38	45	37	33	36	40	28
	$IL2M$	$c(p)$	4657	9122	13436	17780	22031	26232	30353	34024	38506
		$e(p,p)$	78	378	813	1382	1885	2601	3287	4039	4781
		$e(p,n)$	265	500	751	838	1084	1167	1360	1937	1713
		$c(n)$	4776	4762	4814	4810	4806	4802	4798	4802	4784
		$e(n,n)$	161	112	94	63	70	55	56	72	57
		$e(n,p)$	63	126	92	127	124	143	146	126	159
$Landmarks$	FT	$c(p)$	1894	3649	5423	7170	8847	10414	12070	13570	15093
		$e(p,p)$	31	85	174	329	516	643	858	1128	1437
		$e(p,n)$	75	266	403	501	637	943	1072	1302	1470
		$c(n)$	1937	1952	1957	1954	1969	1960	1963	1965	1960
		$e(n,n)$	49	32	32	37	18	22	27	24	29
		$e(n,p)$	14	16	11	9	13	18	10	11	11
	$IL2M$	$c(p)$	1907	3718	5493	7230	8951	10599	12245	13826	15358
		$e(p,p)$	45	107	218	384	587	834	1067	1462	1711
		$e(p,n)$	48	175	289	386	462	567	688	712	931
		$c(n)$	1934	1896	1935	1949	1944	1947	1955	1940	1922
		$e(n,n)$	42	30	29	33	16	19	21	18	28
		$e(n,p)$	24	74	36	18	40	34	24	42	50

Table 3: Analysis of top-1 errors for (FT) and ($IL2M$) methods with memory $K = 10000$ and $Z = 10$ states. p and n stand for past and new classes; c and e stand for correct and erroneous predictions. For instance $e(p, n)$ designates the number of wrong predictions of past classes as new ones.

Tensorflow implementation[1]. The rest of baselines were implemented using Pytorch [7] with SGD optimizer and multi-label cross entropy loss. A ResNet-18 [4] architecture was used in all experiments. $Full$ as well as the first non-incremental model of FT and FT^{BAL} are run for 100 epochs with initial learning rate 0.1 and divided by 10 when the error plateaus for 10 consecutive epochs. For the subsequent batches, FT and FT^{BAL} are run with initial $lr = \frac{0.1}{z}$, where z is the incremental state count ranging between 2 and Z. The learning rate is divided by 10 when the error plateaus for 5 epochs. The weights decay is 0.0001

[1] https://github.com/srebuffi/iCaRL

and the momentum is 0.9. FT was run for 25 epochs while FT^{BAL} was run for 25 epochs for the imbalanced step and 15 epochs for the balanced one continuing with the same learning rate from the imbalanced step.

For the SVM training in $DeeSIL$ [1], we split the training set of the initial batch using a $\frac{90}{10}$ training/validation division. The validation set is used to optimize the SVMs. The optimal regularizer for all configurations was $C = 1$. We frozen it for all the subsequent batches.

For fine tuning based approaches (FT and FT^{BAL}), training images are randomly cropped then resized (224 × 224). After this, they are randomly horizontally flipped and finally normalized.

References

[1] Eden Belouadah and Adrian Popescu. Deesil: Deep-shallow incremental learning. *TaskCV Workshop @ ECCV 2018.*, 2018. 5

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74, 2018. 1, 2

[3] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, pages 241–257, 2018. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2016. 4

[5] Khurram Javed and Faisal Shafait. Revisiting distillation and incremental classifier learning. In *ACCV*, 2018. 1

[6] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, pages 3476–3485. IEEE Computer Society, 2017. 1, 2

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshops*, NIPS-W, 2017. 4

[8] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2017. 1, 3

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[10] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016. 1