Apprentissage continu en ligne de classificateurs un contre tous

Baptiste WAGNER, Denis PELLERIN, Serge OLYMPIEFF, Sylvain HUET

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
Baptiste.Wagner@gipsa-lab.grenoble-inp.fr, Denis.Pellerin@gipsa-lab.grenoble-inp.fr
Serge.Olympieff@gipsa-lab.grenoble-inp.fr, Sylvain.Huet@gipsa-lab.grenoble-inp.fr

Résumé – Dans le contexte d'apprentissage continu en ligne, les réseaux de neurones artificiels utilisés pour la classification d'images sont sujets au problème d'oubli catastrophique : la tendance à oublier des connaissances précédemment acquises lors de l'apprentissage sur de nouvelles informations. Afin d'entraîner de tels réseaux de neurones sur un flux de données d'images, une mémoire externe de taille fixe est généralement allouée pour stocker des images passées et les réintroduire dans l'entraînement. Or, avec cette méthode un biais est généralement observé vers les classes récentes dégradant ainsi les performances du modèle. Dans cet article, nous proposons d'utiliser des classificateurs *un contre tous* dans un contexte d'apprentissage continu. Le système entraîne uniquement le classificateur de la dernière classe observée sans mettre à jour ceux des classes précédentes, le rendant ainsi robuste à l'oubli catastrophique et au biais vers les classes récentes. Nous comparons notre modèle à l'état de l'art sur deux bases de données pour la classification d'images adaptées au contexte d'apprentissage continu en ligne.

Abstract – In the context of online continual learning, artificial neural networks for image classification tasks are prone to catastrophic forgetting: the tendency to forget previously acquired knowledge when learning new information. In order to alleviate this phenomenon, an external memory of fixed size is usually allocated to store past images and are fed back to the training loop. However, this method introduces a bias towards recent classes, which degrades the performance. In this paper, we propose to use *one vs all* classifiers in a continuous learning context. The system only trains the classifier of the last observed class without updating those of the previous classes, making it robust to catastrophic forgetting and to the bias towards recent classes. We compare our model to the state of the art on two datasets for image classification adapted to the online continual learning context.

1 Introduction

Les réseaux de neurones artificiels ont amené une grande avancée dans le domaine de la vision par ordinateur et la classification d'images. Cependant, lors de l'entraînement dans un contexte continu en ligne à partir d'un flux de données, ces réseaux ont tendance à se focaliser sur les données plus récentes et à oublier les connaissances précédemment acquises. Ce phénomène, couramment appelé *l'oubli catastrophique* est un problème majeur en apprentissage profond et a donné lieu au domaine d'étude sur l'apprentissage continu [1].

Dans ce contexte en ligne, les méthodes dites de rejeu de données ou *replay* se sont montrées particulièrement efficaces [2] [3] [4]. Ces approches allouent une mémoire externe de taille réduite et fixe afin de stocker des exemplaires des classes anciennes et de les introduire régulièrement dans la boucle d'entraînement afin de limiter l'effet d'oubli catastrophique.

Pour un problème de classification à C classes, les méthodes de l'état de l'art entraînent généralement une couche de classification avec C neurones de sortie de manière conjointe à l'aide d'une fonction d'activation *softmax* [2] [5]. Or, mettre à jour les poids du réseau correspondant aux anciennes classes avec les méthodes de *replay* peut amener à une dégradation des performances. En effet, on observe à la fin de l'entraînement que les neurones de sortie des classes récentes ont en moyenne des paramètres plus élevés que ceux des classes plus anciennes, le modèle a alors une chance plus grande de prédire une classe récente qu'une classe antérieure [6].

Afin de palier ce problème, nous proposons l'utilisation de classificateurs *un contre tous* (UCT) pour l'apprentissage continu en ligne. Un classificateur par classe est entraîné de manière indépendante, rendant ainsi le modèle robuste à l'oubli catastrophique ainsi qu'au biais vers les classes récentes observé généralement dans les méthodes de *replay* avec *softmax*.

Dans la suite, nous présentons notre modèle utilisant un schéma de classification *un contre tous* (UCT) dans le contexte de l'apprentissage continu en ligne. Puis, nous comparons les résultats obtenus à ceux de l'état de l'art grâce à différentes évaluations sur deux bases de données pour la classification d'images, MNIST et CIFAR-10, adaptées à l'apprentissage continu.

2 Méthode

Le problème de classification d'image supervisée à partir d'un flux de données de distribution inconnue $\mathcal{D} = \{(x_0, y_0), (x_1, y_1) \dots (x_{N-1}, y_{N-1})\}$ avec $(x_i, y_i) \in X \times Y$ et X l'ensemble des images et Y des étiquettes de classe est considéré dans la suite. Le but est d'entraîner continuellement un modèle $f_\psi : X \to \mathbb{R}^C$ sur le flux $\mathcal{D}$ pour un problème de classification à C classes avec comme paramètres ψ.

Ce scénario est plus difficile qu'un entraînement classique de

réseaux de neurones car le système n'a plus accès aux images qu'il a précédemment vues dans le flux et une unique passe sur les données est permise [6].

2.1 Classificateur *un contre tous* en apprentissage continu

On considère un réseau de neurones $f_\psi = h_\theta \circ g_\phi$ avec g_ϕ un extracteur de caractéristiques (e.g. un réseau de neurones convolutif) et h_θ une couche de classification composée de C classificateurs indépendants $\{o_0, o_1, \dots o_{C-1}\}$ de paramètres $\theta = \{\theta_0, \theta_1, \dots \theta_{C-1}\}$. Un classificateur o_y est défini comme une régression logistique sur les vecteurs caractéristiques extraits de g_ϕ :

$$o_y(x) = \sigma\left(w_y g_\phi(x) + b_y\right) \tag{1}$$

Nous notons x une image en entrée, $\theta_y = (w_y, b_y)$ les paramètres du classificateur o_y avec les poids w_y et le biais b_y pour la classe y et σ la fonction sigmoïde :

$$\sigma \quad : \quad u \in \mathbb{R} \quad \mapsto \quad \frac{1}{1+\exp^{-u}} \tag{2}$$

Le classificateur o_y prédit la probabilité que l'événement "l'image x correspond à la classe y" arrive (valeur de 1) ou non (valeur de 0) avec un seuil à 0,5. Les paramètres de chaque classificateur sont entraînés par descente du gradient stochastique à l'aide d'une fonction de perte d'entropie croisée binaire :

$$l_{xy} = -y \log\left(o_y(x)\right) - (1-y)\log\left(1 - o_y(x)\right) \tag{3}$$

Les classificateurs sont entraînés sur un schéma un contre tous (UCT) : pour le classificateur o_y, les images de la classe y sont étiquetées 1 et les autres 0. A l'inférence, la classe obtenant le plus grand score de classification est sélectionnée :

$$\hat{y} = \underset{y \in \{0,\dots C\}}{\arg\max}\ o_y(x) \tag{4}$$

2.2 Entraînement avec une mémoire externe

A l'instant t de l'entraînement, le modèle perçoit la paire (x_t, y_t) provenant du flux de données. L'entraînement est réalisé uniquement sur le classificateur o_{y_t} avec l'image x_t comme expliqué en section 2.1.

Afin que le classificateur o_{y_t} puisse s'entraîner sur des images négatives (i.e. de classe différente), une mémoire externe de taille fixe est allouée afin de stocker des exemplaires de chaque classe. Ainsi, à l'instant t, une image x_b de classe $y_b \neq y_t$ est tirée aléatoirement de la mémoire. Le classificateur o_{y_t} est alors entraîné à l'instant t sur une image positive et une image négative respectivement x_t et x_b.

Notre méthode s'inscrit dans les méthodes de *replay* qui ont montré de très bonnes performances dans le contexte d'apprentissage continu en ligne [5] [7]. Or, celles-ci entraînent généralement un réseau de neurones avec une fonction d'activation *softmax* : les classificateurs sont entraînés conjointement. Un biais vers les classes récentes est ainsi observé, notamment à cause du déséquilibre d'exemplaires entre les classes de la mémoire et les classes récentes provenant du flux. L'entraînement déséquilibré du modèle engendre des paramètres plus élevés pour les classes récentes et celles-ci sont favorisées lors de l'inférence finale entraînant une dégradation des performances [6].

En revanche, notre méthode ne présente pas ce problème généralement perçu dans les méthodes de *replay*. D'une part, les classificateurs sont mis à jours indépendamment des autres : si une classe n'apparaît plus dans le flux, le classificateur correspondant n'est plus modifié et l'oubli catastrophique est réduit. D'autre part, chaque classificateur est entraîné de manière binaire et équilibrée entre les classes du flux et de la mémoire. En effet, un classificateur perçoit au final un même nombre d'images positives que négatives grâce à l'implémentation de la mémoire externe.

Finalement, l'architecture un contre tous diffère des méthodes de l'état de l'art, notamment par les frontières de décisions apprises par la fonction sigmoïde plutôt que *softmax*.

2.3 Étape de consolidation hors-ligne

Dans le modèle UCT, les classificateurs $\{o_0, o_1, \dots o_{C-1}\}$ sont entraînés de manière indépendante avec les images du flux de données et les images de la mémoire.

Pour que les classificateurs soient également mis à jour sur des classes plus récentes qui n'étaient pas disponibles dans la mémoire au moment de leur entraînement, un apprentissage en 2 étapes est proposé dans le modèle $\text{UCT}_{\text{conso}}$: toutes les τ_{conso} images visitées dans le flux, une étape d'apprentissage hors-ligne est réalisée uniquement sur les images de la mémoire. Comme la mémoire présente des images de toutes les classes, les anciens classificateurs sont également entraînés sur des images de classes plus récentes lors de cette étape de consolidation hors-ligne.

3 Expérimentations

3.1 Bases de données

La méthode est évaluée sur deux bases de données **MNIST Split** et **CIFAR-10 Split** souvent utilisées dans la classification d'images continue [3] [5]. Ces deux bases utilisent les données de MNIST et CIFAR-10 composées de 10 classes chacune. Lors de l'entraînement sur une base, le modèle est entraîné de manière séquentielle sur 5 expériences composées de 2 classes disjointes chacune. Ce scénario permet de définir un flux délimité afin de mieux comparer les méthodes plutôt qu'avec un flux arbitraire.

3.2 Base de comparaison

Nous comparons notre approche aux modèles d'apprentissage continu suivants :

— **ER** [5] : une mémoire externe de taille fixe est allouée. La méthode de *reservoir sampling* [5] [7] est utilisée pour la gestion de la mémoire. La sélection d'images à

TABLE 2 – Résultats sur **CIFAR-10 Split**, la précision moyenne et l'oubli moyen sont présentées pour trois expériences avec différentes tailles de mémoire M définies en nombre d'images pouvant être stockées. Les meilleures performances sont en gras.

	Précision moyenne			Oubli moyen		
	M=200	M=500	M=1000	M=200	M=500	M=1000
iid en ligne	60.1 ±0.8	60.1 ±0.8	60.1 ±0.8	-	-	-
iid hors-ligne	63.7 ±0.8	63.7 ±0.8	63.7 ±0.8	-	-	-
GEM [2]	16.8 ±1.1	17.1 ±1.0	17.5 ±1.6	73.5 ±1.7	70.7 ±4.5	71.7 ±1.3
ER [5]	27.5 ±1.2	33.1 ±1.7	41.3 ±1.9	50.5 ±2.4	35.4 ±2.0	23.3 ±2.9
ER MIR [5]	29.8 ±1.1	40.0 ±1.1	47.6 ±1.1	50.2 ±2.0	30.2 ±2.3	17.4 ±2.1
ER $\text{MIR}_{\text{conso}}$	43.5 ±0.1	49.1 ±0.1	53.1 ±0.1	50.6 ±0.2	39.8 ±0.2	33.4 ±0.2
UCT	47.1 ±1.5	46.1±2.3	46.1 ±2.1	**11.2 ±1.9**	**8.9 ±1.8**	**9.1 ±2.2**
$\text{UCT}_{\text{conso}}$	**55.8 ±0.8**	**57.4 ±0.6**	**58.7 ±0.5**	32.0 ±3.8	19.8 ±2.2	22.6 ±3.4

stocker et le tirage pour l'entraînement se fait de manière aléatoire.

— **ER MIR** [5] : utilise une mémoire comme ER pour limiter l'oubli catastrophique. Les exemplaires choisis pour le *replay* sont sélectionnés en fonction de l'augmentation de la fonction de perte, compte tenu de la mise à jour des paramètres estimés avec le mini-batch entrant.

— **GEM** [2] : une autre méthode utilisant une mémoire externe qui contraint la mise à jour des paramètres avec les exemplaires des classes stockés. Cette méthode garantit que la fonction de perte des exemplaires n'augmente pas pendant l'entraînement.

Les deux scénarios suivants *non continu* sont utilisés pour définir une borne supérieure pour les modèles d'apprentissage continu où aucun oubli catastrophique n'est observé.

— **iid en ligne** : le modèle est entraîné de manière classique non-continue, c'est à dire que toutes les données sont accessibles au début de l'entraînement. Cependant, qu'une seule *epoch* sur ces données est réalisée pour une meilleure comparaison avec l'apprentissage en ligne qui est fait en une seule passe sur le flux.

— **iid hors ligne** : le modèle est également entraîné de manière classique non-continue, cependant on s'autorise 5 *epochs* pendant l'entraînement.

3.3 Résultats

Afin de permettre une meilleure comparaison aux différentes méthodes de l'état de l'art, nous avons repris les protocoles proposés dans [5] [7]. Chaque expérimentation est lancée 15 fois et nous indiquons la performance moyenne ainsi que l'écart type dans les tableaux 1 et 2 sur MNIST et CIFAR-10 respectivement. Deux métriques sont utilisées pour l'évaluation des performances de notre modèle : la précision moyenne et l'oubli moyen [6] [7]. Les métriques indiquées sont mesurées à la fin de l'entraînement sur le flux.

Sur MNIST Split, le modèle g_ϕ utilisé est un Perceptron de deux couches composées de 400 neurones chacune associé à un *learning rate* de 0.05. Nous utilisons un *batch size* de taille 20 composé de 10 images issues du flux et de 10 images tirées aléatoirement depuis la mémoire. La mémoire a une taille fixe de 500 images et la fréquence de consolidation utilisée est $\tau_{\text{conso}} = 250$. Les résultats sont donnés dans le tableau 1.

Sur CIFAR-10 Split, nous utilisons un ResNet-18 pour l'extracteur de caractéristiques g_ϕ. Cependant nous avons observé que notre méthode produisait de meilleurs résultats avec un CNN pré-entraîné sur ImageNet et figé tout au long de l'entraînement. Nous comparons ainsi à l'état de l'art deux modèles UCT et $\text{UCT}_{\text{conso}}$ définis dans la sous-section 2.3 sur 3 tailles de mémoire différentes (200, 500 et 1000). Un *batch size* de taille 20 est encore utilisé (10 images du flux, 10 de la mémoire). La période de consolidation est fixée à $\tau_{\text{conso}} = 2500$ pour les trois expériences sur $\text{UCT}_{\text{conso}}$. Une période τ_{conso} plus grande est possible dans cette expérience avec le CNN figé car seulement la couche de classification doit être consolidée. Pour une meilleure comparaison à l'état de l'art, nous indiquons également les performances de ER $\text{MIR}_{\text{conso}}$ correspondant au modèle de l'état de l'art entraîné dans les mêmes conditions que notre modèle, i.e. CNN pré-entraîné, figé et une étape de consolidation de même période. Les résultats sont donnés dans le tableau 2.

TABLE 1 – Résultats sur **MNIST Split**, une mémoire de taille 500 images est utilisée. Les meilleures performances sont indiquées en gras.

	Précision moyenne	Oubli moyen
iid en ligne	89.6 ±1.1	-
iid hors-ligne	95.0 ±0.2	-
GEM [2]	86.3 ±1.4	11.2 ±1.2
ER [5]	82.1 ±1.5	15.0 ±2.1
ER MIR [5]	**87.6 ±0.7**	**7.0 ±0.9**
UCT	**87.6 ±0.8**	7.3 ±1.9

4 Discussion

4.1 Comparaison avec l'état de l'art

Sur MNIST Split, notre méthode est performante en terme de précision et d'oubli moyen par rapport aux méthodes de l'état de l'art avec des performances similaires à ER+MIR [5]. De plus, contrairement à cette dernière méthode, le tirage d'images

depuis la mémoire est réalisé de manière complètement aléatoire et aucune méthode de sélection n'est implémentée.

Sur CIFAR-10 Split, nous évaluons nos deux modèles UCT et $\text{UCT}_{\text{conso}}$. Tout d'abord, notre méthode est particulièrement performante avec les mémoires de petites tailles.

La méthode sans consolidation UCT présente une précision moyenne constante quelque soit la taille de la mémoire dans les trois expériences et un oubli très faible par rapport à l'état de l'art : seule l'étape de consolidation profite d'une augmentation de la taille de la mémoire.

En outre, la méthode avec consolidation $\text{UCT}_{\text{conso}}$ permet un gain significatif en précision moyenne : pour une mémoire de taille M = 200, nous observons un gain de 12.3 points de précision moyenne par rapport à $\text{ER+MIR}_{\text{conso}}$. L'oubli moyen est plus important que la méthode sans consolidation, mais pour M = 1000, elle est à 1.4 points de précision en dessous de la borne supérieure définie par le modèle iid en ligne.

4.2 Compromis stabilité-plasticité

Le problème de l'oubli catastrophique est souvent présenté par l'intermédiaire du dilemme stabilité-plasticité [8] auquel sont sujets les réseaux de neurones. En effet, l'oubli catastrophique survient car ces réseaux sont généralement très plastiques : lors de l'entraînement sur de nouvelles données, ils ont tendance à réécrire sur ce qu'ils ont appris, effaçant alors toute connaissance acquise précédemment.

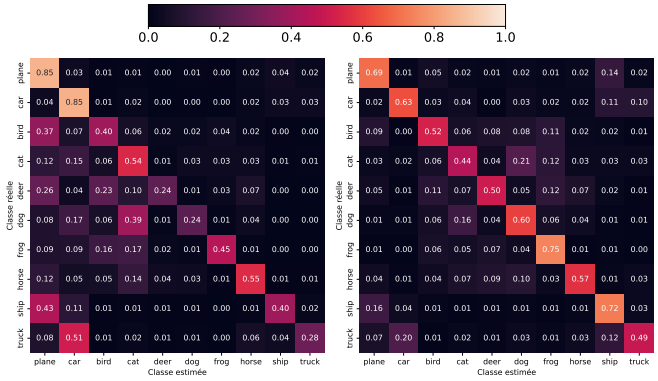
Notre modèle UCT montre une forte stabilité mais une faible plasticité. En effet, celui-ci préfère conserver les connaissances acquises auparavant au détriment de futures connaissances qu'il pourrait obtenir.

Ce phénomène est visible sur la matrice de confusion obtenue à la fin de l'entraînement sur CIFAR-10 Split (figure 1a). Les deux premières classes montrent une très bonne précision alors que celles apparues plus tard dans le flux présentent un grand nombre de faux positifs sur les premières classes. Ceci se traduit par un grand nombre de faux positifs dans la partie sous la diagonale de la matrice de confusion.

Cependant, dans la matrice de confusion du modèle avec consolidation $\text{UCT}_{\text{conso}}$ (figure 1b), la précision est mieux répartie sur l'ensemble des classes. Les premières classes présentent alors un oubli plus important au bénéfice des classes plus récentes qui présentent moins de faux négatifs.

La période τ_{conso} nous permet ainsi d'avoir un levier sur le compromis stabilité-plasticité. Une petite période accorde une forte plasticité en entraînant souvent les anciens classificateurs avec les nouvelles classes. Une grande période accorde plus de stabilité en limitant la mise à jour des anciens classificateurs.

En perspective, une méthode pour avoir un τ_{conso} adaptatif pourrait être considérée. Celle-ci permettrait d'obtenir le meilleur compromis entre stabilité et plasticité avec un nombre minimal de consolidations avec, par exemple, une plasticité plus forte en début d'apprentissage et une stabilité accrue lorsque beaucoup de connaissances ont été apprises.



(a) Modèle UCT (b) Modèle $\text{UCT}_{\text{conso}}$

FIGURE 1 – Matrices de confusion des deux modèles UCT et $\text{UCT}_{\text{conso}}$ sur CIFAR-10 Split avec une mémoire M=1000.

Références

[1] M. McCloskey and N. J. Cohen, "Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem," vol. 24 of *Psychology of Learning and Motivation*, pp. 109–165, Academic Press, 1989.

[2] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.

[3] T. Lesort, *Apprentissage continu : S'attaquer à l'oubli foudroyant des réseaux de neurones profonds grâce aux méthodes à rejeu de données*. Thèse, Institut Polytechnique de Paris, June 2020.

[4] P. Buzzega, M. Boschini, A. Porrello, and S. Calderara, "Rethinking Experience Replay : a Bag of Tricks for Continual Learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2180–2187.

[5] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia, "Online continual learning with maximal interfered retrieval," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification : An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, 2022.

[7] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "On tiny episodic memories in continual learning," *arXiv preprint arXiv :1902.10486*, 2019.

[8] S. Grossberg, "Studies of mind and brain : neural principles of learning, perception, development, cognition, and motor control," Boston studies in the philosophy of science 70. Reidel, Dordrecht, 1982.