De l'amélioration de l'accord des caractéristique en apprentissage continu grâce aux explications cohérentes

Antoine MONTMAUR[1] Ngoc-Son VU[1]

[1]Laboratoire ETIS, 6 avenue du poànceau, 95000 Cergy , France

Résumé – Nous présentons une approche novatrice pour l'apprentissage continu, intégrant l'apprentissage auto-supervisé et la distillation de connaissances afin d'améliorer la robustesse des caractéristiques face aux changements de distribution. En utilisant des techniques d'explication comme Grad-CAM, notre méthode renforce la cohérence des réseaux à travers les tâches. Nos expérimentations sur plusieurs ensembles de données et méthodes auto-supervisées montrent que notre approche surpasse les solutions existantes, améliorant la stabilité de l'apprentissage à long terme sans recours aux mémoires de rejeu.

Abstract – We introduce a novel approach to continual learning that integrates self-supervised learning and knowledge distillation to enhance feature robustness against distribution shifts. Using explanation-based techniques like Grad-CAM, our method improves network consistency across tasks. Additionally, we propose a gradient-free feature-matching strategy, reducing computational overhead while maintaining effective network alignment. Our experiments across multiple datasets and self-supervised methods demonstrate that our approach outperforms existing solutions, significantly improving long-term learning stability without relying on replay buffers.

1 Introduction

L'apprentissage continu (CL) est apparu comme une solution prometteuse, permettant aux réseaux neuronaux de s'entraîner sur un flux continu d'expériences sans avoir accès aux exemples précédents. En aidant les modèles à s'adapter aux évolutions des distributions de données tout en préservant les connaissances passées, le CL atténue l'oubli catastrophique (CF), un phénomène où de nouvelles informations interfèrent avec et remplacent celles déjà acquises [12].

Un facteur clé du CF est le dilemme stabilité-plasticité, qui traduit le compromis entre la rétention des connaissances et l'intégration efficace de nouvelles informations. Le CF est souvent lié à la confusion des représentations : des représentations distribuées favorisent la généralisation mais augmentent l'interférence, tandis que des représentations localisées limitent l'oubli mais réduisent l'adaptabilité [15]. Les stratégies de CL doivent équilibrer ces facteurs pour garantir un apprentissage robuste et durable, ce qui a récemment été fait via la régularisation de fonctions.

Nous proposons ECD (Explanation Consistent Distillation), un cadre qui étend la régularisation de fonction en intégrant de nouveaux schémas de distillation de connaissances (KD) basés sur l'appariement des caractéristiques. Contrairement à la KD classique, qui se concentre sur la cohérence des sorties, ECD impose une cohérence au niveau des caractéristiques à différentes échelles. Il aligne ainsi les représentations du modèle étudiant sur celles du modèle enseignant en garantissant la similitude de leurs caractéristiques explicatives (comme Grad-CAM ou Grad-CAM++), facilitant l'utilisation de caractéristiques analogues lors des prédictions.

Nos principales contributions se résument ainsi :

- **Nous proposons une nouvelle technique de régularisation pour l'apprentissage continu auto-supervisé (CSSL)**, qui aligne directement les caractéristiques des anciens et nouveaux modèles. En exploitant l'explicabilité, notre approche compare leurs explications via une similarité cosinus multi-échelle, assurant la préservation des caractéristiques essentielles à travers les tâches (**stratégie basée sur l'explication**).
- **Notre stratégie permet des améliorations significatives sans recours au rejeu**, répondant ainsi aux préoccupations liées à la confidentialité. Nos résultats expérimentaux montrent que notre méthode sans relecture surpasse les approches CSSL de pointe dans des conditions exigeantes.

2 Contexte

2.1 Appentissage Continu

En raison de son intérêt majeur, l'apprentissage continu a suscité beaucoup d'attention, et de nombreuses méthodes ont été proposées ces dernières années. Les axes de recherches se sont principalement concentrés dans deux directions : les méthodes reposant sur une optimisation particulière et les méthodes utilisant la régularisation.

Les approches basées sur **l'optimisation** modifient directement la dynamique d'apprentissage en ajustant le processus d'optimisation. Une technique courante dans ce cadre est la projection de gradient [10], qui empêche l'interférence entre les tâches passées et nouvelles, avec plusieurs variantes proposées pour améliorer son efficacité.

Les techniques basées sur la **régularisation** atténuent l'oubli en ajoutant des contraintes explicites lors de l'entraînement pour équilibrer la rétention des connaissances passées et l'intégration de nouvelles informations. Une stratégie courante est la régularisation des poids [9], qui met à jour ou préserve les paramètres du modèle en fonction de la matrice d'information de Fisher. De plus, les méthodes de régularisation fonctionnelle [7] ajustent directement les sorties du réseau pour assurer

la cohérence à travers les phases d'apprentissage.

2.2 Distillation de connaissances

Dans le contexte de l'apprentissage continu (CL), la distillation de connaissances (KD) est largement utilisée dans la régularisation de fonction, où le modèle actuel agit comme l'étudiant et les modèles précédents comme les enseignants. Étant donné que tous les anciens échantillons d'entraînement sont indisponibles en CL, des méthodes alternatives de distillation ont été proposées, comme l'utilisation de nouveaux échantillons d'entraînement, une portion d'anciens échantillons, ou des données externes ou générées non étiquetées.

Des travaux pionniers comme Learning without Forgetting [9] permettent d'apprendre de nouveaux échantillons tout en préservant les connaissances des tâches passées en calculant les coûts de distillation à partir des prédictions des modèles précédents. LwM améliore la distillation en utilisant des cartes d'attention des nouveaux échantillons d'entraînement, tandis que EBLL utilise des autoencodeurs spécifiques à la tâche pour maintenir l'intégrité de la reconstruction des caractéristiques, empêchant ainsi les modifications indésirables des représentations apprises. GD étend la KD en utilisant des données non étiquetées à grande échelle provenant de sources externes pour améliorer la généralisation du modèle dans le cadre de l'apprentissage continu auto-supervisé (CSSL). Lorsque les anciens échantillons d'entraînement sont correctement reconstruits, l'efficacité de la régularisation de fonction est considérablement améliorée.

2.3 Explicabilité

Les algorithmes d'explicabilité se sont avérés être un outil puissant pour analyser comment les modèles d'apprentissage profond font des prédictions précises. Depuis les travaux fondamentaux de [20, 19], un nombre important de recherches s'est concentré sur la modification des cartes de caractéristiques pour mettre en évidence les éléments les plus pertinents pour les prédictions du modèle. Ces approches utilisent diverses méthodes de mise à l'échelle, comme la multiplication du gradient pixel par pixel dans [19], tandis que d'autres utilisent des outils mathématiques pour analyser directement les caractéristiques.

Remarque. Bien que le concept de distillation des explications partage des similitudes avec les travaux précédents [2, 14], notre algorithme se distingue de ces approches en termes d'application et de méthodologie. À notre connaissance, il s'agit du premier travail à explorer la distillation des explications pour l'apprentissage continu (CL).

3 Méthode

Des études récentes [14] ont montré que la cohérence explicative améliore de manière constante les résultats de la distillation des connaissances (KD) dans divers contextes. L'orientation du modèle permet aux réseaux étudiants de se concentrer sur les caractéristiques pertinentes tout en ignorant les autres [18, 16, 14]. Étant donné que le CSSL implique l'apprentissage de représentations génériques, il est crucial de distinguer les caractéristiques spécifiques à la tâche des caractéristiques plus générales. Nous définissons ici la **cohérence explicative** comme la similarité entre les cartes d'explication générées par deux versions successives d'un modèle (enseignant/étudiant), à partir des mêmes entrées, pour des couches correspondantes.

3.1 Coût d'explication cohérente $\mathcal{L}_{CEL}$

Nous introduisons la fonction de coût $\mathcal{L}^l_{CEL}$ pour mesurer la similarité entre les cartes Grad-CAM (ou variantes) extraites d'une même couche l de deux modèles : f_{t-1} (enseignant, gelé) et f_t (étudiant, en cours d'apprentissage). La mesure utilisée est une distance cosinus :

$$\mathcal{L}^l_{CEL} = 1 - \left\langle L_l^{f_{t-1}}(x_t), L_l^{f_t}(x_t) \right\rangle \tag{1}$$

Cette procédure est illustrée à la Figure 3.1.

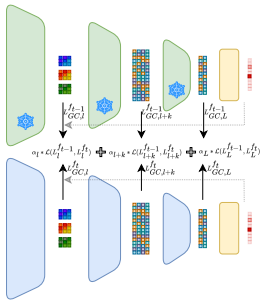

FIGURE 1 : Détails de l'ECD avec $\mathcal{L}_{CEL}$ où $L^{f_{t-1}}_{GC,l}$ figure la représentation Grad-CAM du réseau f_{t-1} à la couche l. α_l représente un coefficient d'échelle pour chaque composant. Les lignes pointillées indiquent le flux de gradient qui retourne pour créer les cartes de gradient (pas de flux de gradient en b). f_{t-1} dans la branche supérieure est gelé.

Pour obtenir une régularisation multi-échelle, ces coûts sont combinés sur un sous-ensemble $P(l)$ des couches du réseau :

$$\mathcal{L}_{CEL} = \sum_{l \in P(l)} \alpha_l \cdot \mathcal{L}^l_{CEL} \tag{2}$$

avec la contrainte : $\sum_{l \in P(l)} \alpha_l = 1$.

Clarifications supplémentaires :

- $P(l)$ est un sous-ensemble de couches sélectionnées régulièrement (par exemple, une couche par bloc résiduel), fixé manuellement avant l'entraînement.
- Les coefficients α_l sont fixés à des valeurs égales par défaut (uniformes), mais peuvent être affinés par validation croisée.

Remarque : les cartes d'explication utilisées sont basées sur Grad-CAM [19] ou Grad-CAM++ [5], produites à partir des activations des couches et des gradients associés à une prédiction cible. Ces cartes sont normalisées pour garantir la compatibilité des distances cosinus.

3.2 Fonction de coût totale

La fonction de perte finale combine le terme standard d'auto-apprentissage SSL et notre régularisation basée sur les explications :

$$\mathcal{L}_{total} = \mathcal{L}_{CSSL} + \mathcal{L}_{CEL} \tag{3}$$

où $\mathcal{L}_{CSSL}$ comprend :

- $\mathcal{L}_{plast}$: apprentissage des représentations à partir de vues augmentées d'un échantillon
- $\mathcal{L}_{stab}$: stabilité inter-tâches via la comparaison entre représentations projetées de f_t et f_{t-1}

$$\mathcal{L}_{CSSL} = \mathcal{L}_{SSL}(z^{f_t}, z_+^{f_t}) + \mathcal{L}_{SSL}(p(z_+^{f_t}), z_+^{f_{t-1}}) \quad (4)$$

Où :

- z^{f_t} et $z_+^{f_t}$ sont les représentations de deux vues augmentées de x_t par le modèle f_t
- $p(\cdot)$ est un projecteur utilisé avant la comparaison

Remarque : La formulation est compatible avec plusieurs coûts SSL (SimCLR, BYOL, Barlow Twins), utilisés sans modification.

4 Expériences

Configuration d'apprentissage incrémental de classes (C-IL). Pour appuyer notre analyse, nous menons des expériences approfondies dans un cadre d'apprentissage incrémental de classes (C-IL), un scénario d'apprentissage continu où les frontières entre les tâches partitionnent les classes et ne sont pas fournies au modèle ni pendant l'entraînement ni lors de l'inférence. Chaque tâche consiste à apprendre un sous-ensemble disjoint de classes.

Architecture du modèle. Toutes les expériences sont menées à l'aide d'un modèle ResNet-18, largement utilisé pour sa stabilité en apprentissage SSL. Chaque version du modèle est suivie d'un projecteur MLP (Multi-Layer Perceptron) composé de deux couches avec ReLU intermédiaire. Le projecteur est utilisé avant le calcul des pertes $\mathcal{L}_{SSL}$ et $\mathcal{L}_{stab}$.

Couches utilisées pour $\mathcal{L}_{CEL}$. Nous sélectionnons une couche par bloc résiduel dans le ResNet-18 (soit 4 couches au total). Ces couches sont utilisées pour extraire les cartes Grad-CAM. Tous les α_l sont fixés à 0.25.

Méthodes comparées. Nous comparons notre approche à des méthodes CSSL récentes (CaSSLe, Sy-Con, LUMP) et à des approches supervisées adaptées (EWC, ER, DER). Certaines méthodes utilisent un rejeu mémoire (CaSSLe+, CroMo-Mixup). Pour toutes les méthodes, les hyperparamètres sont sélectionnés selon les configurations par défaut ou recommandées dans les articles d'origine.

4.1 Métriques

Nous considérons la précision moyenne A sur toutes les tâches passées et présentes, ainsi que le transfert avant (FT), mesurant dans quelle mesure les représentations acquises facilitent l'apprentissage de tâches futures. Les formules exactes sont données dans l'équation (6) et (7) du document original.

4.2 Résultats

La Table 1 résume les performances sur CIFAR100-Split5 avec trois variantes SSL : SimCLR, Barlow Twins et BYOL. Notre approche ECD dépasse toutes les méthodes de l'état de l'art, y compris celles utilisant des tampons mémoire.

Notamment, avec BYOL, ECD surpasse CaSSLe+ de plus de 5% en précision moyenne, malgré l'absence de rejeu.

TABLE 1 : Comparaison avec les méthodes CSSL SOTA sur CIFAR100-Split5 dans le cadre du paramètre C-IL. Le meilleur résultat est en gras.

Strategy	**SimCLR**		**Barlow Twins**		**BYOL**	
	A (↑)	FT (↑)	A (↑)	FT (↑)	A (↑)	FT (↑)
Fine-tuning	48.9	33.5	54.3	39.2	52.7	35.9
EWC [1]	53.6	33.3	56.7	39.1	56.4	39.9
ER [17]	50.3	32.7	53.3	40.3	54.2	38.7
Less-Forget [8]	52.5	35.8	58.0	41.0	58.6	41.1
POD [6]	51.3	33.8	55.9	40.3	57.9	41.1
DER [3]	50.7	33.2	55.5	41.5	57.0	41.1
LUMP [11]	52.3	34.5	57.8	40.1	58.6	41.1
CaSSLe [7]	57.6	-	60.6	-	56.9	-
Sy-Con [4]	58.9	-	60.4	-	57.3	-
Ours (with $\mathcal{L}_{CEL}$)	**60.9**	**38.2**	**62.4**	**43.1**	**62.3**	**43.7**
CaSSLe+	59.5	-	61.3	-	57.4	-
CroMo-Mixup [13]	62.7	-	65.5	-	60.6	-
Offline	65.1	-	70.0	-	66.7	-

5 Conclusion

Dans ce travail, nous proposons une nouvelle méthode appelée ECD pour le CSSL, qui améliore l'accord des caractéristiques entre les apprenants précédents et actuels. Notre conclusion principale met en évidence le potentiel de l'accord des caractéristiques multi-échelles comme un outil efficace pour atténuer l'oubli catastrophique (CF) dans l'apprentissage en continu. L'efficacité de cette approche découle de la transférabilité des représentations génériques entre l'enseignant et l'étudiant, qui partagent la même architecture sous-jacente. De plus, nous démontrons qu'une méthode bien conçue peut exploiter cette transférabilité pour améliorer les performances en utilisant des explications basées sur le gradient afin de sélectionner les caractéristiques pertinentes qui renforcent l'accord entre les apprenants passés et actuels.

Une piste intéressante pour les travaux futurs serait d'explorer l'influence des paramètres α_l et β_l sur l'évolution de chaque coût. D'après nos résultats, il semble que l'accord multi-échelles soit bénéfique dans plusieurs scénarios d'apprentissage en continu. Toutefois, des recherches supplémentaires sont nécessaires pour déterminer si toutes les échelles, ou couches, utilisées dans le processus sont également efficaces.

Références

[1] Abhishek AICH : Elastic weight consolidation (EWC) : nuts and bolts. *CoRR*, 2021.

[2] Pedro R. A. S. BASSI, Andrea CAVALLI et Sergio DECHERCHI : Explanation is all you need in distillation : Mitigating bias and shortcut learning, 2024.

[3] Pietro BUZZEGA, Matteo BOSCHINI, Angelo PORRELLO, Davide ABATI et Simone CALDERARA : Dark experience for general continual learning : a strong, simple baseline. *In NeurIPS*, 2020.

[4] Sungmin CHA, Kyunghyun CHO et Taesup MOON : Regularizing with pseudo-negatives for continual self-supervised learning. *In ICML*, 2024.

[5] Aditya CHATTOPADHAY, Anirban SARKAR, Prantik HOWLADER et Vineeth N BALASUBRAMANIAN : Grad-cam++ : Generalized gradient-based visual explanations for deep convolutional networks. *In WACV*, 2018.

[6] Arthur DOUILLARD, Matthieu CORD, Charles OLLION, Thomas ROBERT et Eduardo VALLE : Podnet : Pooled outputs distillation for small-tasks incremental learning. *In ECCV*, 2020.

[7] Enrico FINI, Victor G. Turrisi da COSTA, Xavier ALAMEDA-PINEDA, Elisa RICCI, Karteek ALAHARI et Julien MAIRAL : Self-supervised models are continual learners. *In CVPR*, 2022.

[8] Saihui HOU, Xinyu PAN, Chen Change LOY, Zilei WANG et Dahua LIN : Learning a unified classifier incrementally via rebalancing. *In CVPR*, 2019.

[9] Zhizhong LI et Derek HOIEM : Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[10] David LOPEZ-PAZ et Marc'Aurelio RANZATO : Gradient episodic memory for continual learning. *In NeurIPS*, 2022.

[11] Divyam MADAAN, Jaehong YOON, Yuanchun LI, Yunxin LIU et Sung HWANG : Rethinking the representational continuity : Towards unsupervised continual learning. *In ICLR*, 2021.

[12] Michael MCCLOSKEY et Neal J. COHEN : Catastrophic interference in connectionist networks : The sequential learning problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165, janvier 1989.

[13] Erum MUSHTAQ, Yasin BAKMAN, Xialei LIU, Chao TAO et Dimitris DIMITRIADIS : Cromo-mixup : Augmenting cross-model representations for continual self-supervised learning. *In ECCV*, 2024.

[14] Amin PARCHAMI-ARAGHI, Moritz BÖHLE, Sukrut RAO et Bernt SCHIELE : Good teachers explain : Explanation-enhanced knowledge distillation. *In ECCV*, 2024.

[15] Julien POURCEL, Ngoc-Son VU et Robert M. FRENCH : Online task-free continual learning with dynamic sparse distributed memory. *In ECCV*, 2022.

[16] Sukrut RAO, Moritz BÖHLE, Amin PARCHAMI-ARAGHI et Bernt SCHIELE : Studying how to efficiently and effectively guide models with explanations. *In ICCV*, 2023.

[17] Anthony ROBINS : Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 1995.

[18] Andrew Slavin ROSS, Michael C. HUGHES et Finale DOSHI-VELEZ : Right for the right reasons : Training differentiable models by constraining their explanations. *In IJCAI*, 2017.

[19] Ramprasaath R. SELVARAJU, Michael COGSWELL, Abhishek DAS, Ramakrishna VEDANTAM, Devi PARIKH et Dhruv BATRA : Grad-cam : Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 2019.

[20] Bolei ZHOU, Aditya KHOSLA, Agata LAPEDRIZA, Aude OLIVA et Antonio TORRALBA : Learning deep features for discriminative localization. *In CVPR*, 2015.