Échantillonnage basé sur l'entropie pour l'apprentissage en continu avec l'approche Move-to-Data sur la vidéo

Meghna P AYYAR[1] Jenny BENOIS-PINEAU[1] Akka ZEMMARI[1]
[1]Univ. de Bordeaux, CNRS, LaBRI, UMR 5800, F-33400, Talence, France

Résumé – L'entraînement des réseaux de neurones profonds repose sur des corpus volumineux de données annotées. Cependant, les données peuvent changer au fil du temps. L'apprentissage continu est un type d'apprentissage incrémental où les réseaux profonds apprennent séquentiellement au fil de l'eau. Dans cet article, nous proposons des critères de sélection de données basés sur l'incertitude pour améliorer notre précédente méthode d'apprentissage en continu : la méthode Move-to-Data (MTD). La nouvelle approche appelée EMTD est basée sur le calcul d'entropie. Nous utilisons un réseau de type *Transformer* pour analyser MTD, EMTD et leurs variantes. Nous comparons les performances d'EMTD avec MTD et une méthode populaire d'apprentissage continu. EMTD est capable de surpasser la méthode de base MTD, et EMTD avec re-ajustement obtient des résultats proches de ceux de la méthode de l'état de l'art tout en étant environ 1.2 fois plus rapide.

Abstract – The training of deep neural networks relies on large annotated datasets. However, the real-world is not static and data can change over time. Streaming learning is a type of incremental learning where networks learn sequentially and on the fly. We propose an uncertainty based data selection criteria to improve our previous fast streaming learning method Move-to-Data (MTD), called Entropy-based MTD (EMTD). We use a transformer network to analyse MTD, EMTD and a popular method from the State-of-the-art. EMTD is able to outperform baseline MTD, and EMTD with an adjustment achieves close results to the SOA method and is $\sim$1.2 times faster.

1 Introduction

Dans un contexte temps réel, les données évoluent constamment avec le temps et les réseaux d'apprentissage profond doivent s'adapter aux nouvelles données (plasticité) sans perdre en performance sur les anciennes données (stabilité). Cet équilibre entre stabilité et plasticité est l'objectif des approches d'apprentissage incrémental/continu, qui gagnent en popularité pour une série d'applications[3].

L'apprentissage continu (ou *streaming learning* en anglais) est un type d'apprentissage incrémental dans lequel les paramètres du réseau doivent être mis à jour séquentiellement en fonction des données qui arrivent [3]. La contrainte temps réel se traduit par la disponibilité immédiate, mais courte, des données et la difficulté de les stocker sur une longue période [5]. Une solution "naïve" consiste à continuer l'entraînement du réseau avec les données qui arrivent (avec leurs étiquettes) [9]. Toutefois, cette approche peut se révéler longue, car elle nécessite une optimisation par descente de gradient et donc un calcul systématique du gradient de la fonction de perte sur les données.

Le "Move-to-Data" (MTD) [8] est un apprentissage continu qui *bouge* les paramètres du réseau dans une direction sous-optimale sans calculer le gradient. Ceci permet de réduire le temps de calcul. Cependant, l'approche peut induire une dérive du modèle. Pour réduire cette dérive, nous avons proposé un déplacement périodique des paramètres dans la direction optimale opposée au gradient recalculé pour recibler le réseau[3].

Dans cet article, nous introduisons un critère de sélection pour choisir "les meilleures données" pour l'ajustement des paramètres. Nous nous plaçons dans un scénario où les données n'ont pas de nouvelles classes, mais où le domaine des classes précédentes évolue avec le temps. Bien que la méthode soit universelle, nous l'appliquons au corpus spécifique de données vidéo enregistré pour détecter les situations à risque chez les personnes fragiles [7]. Ainsi, dans ce travail, nous proposons un critère basé sur l'entropie maximale pour sélectionner des échantillons informatifs dans le flux afin d'améliorer le MTD.

2 Travaux connexes

Dans cette section, nous discutons le phénomène de "l'oubli catastrophique" que peuvent rencontrer les méthodes d'apprentissage continu. Ensuite, nous présentons le scénario d'apprentissage considéré dans cet article et enfin la méthode de sélection de données pour un entraînement "à la volée".

2.1 L'oubli catastrophique : est-ce crucial ?

Dans un scénario d'apprentissage continu, les réseaux de neurones s'adaptent aux nouvelles données qui arrivent, mais peuvent perdre de manière drastique la capacité de reconnaître les anciens données. Ce phénomène porte le nom de "l'oubli catastrophique". Plusieurs méthodes ont été proposées pour circonvenir ce phénomène, comme [6]. Cependant, dans plusieurs scénarios où les données évoluent avec le temps, il n'y a pas de nécessité de reconnaître des anciennes données. Tout évolue, le contexte, les données, comme dans l'observation longitudinale des personnes, pour la détection des situations à risque [7].

2.2 Scénario d'apprentissage continu

L'apprentissage continu à partir de flux de données est un cas d'apprentissage incrémental dans lequel le réseau ingère un nouvel échantillon à la fois. Dans l'apprentissage "par lots" (Incremental Batch Learning (IBL)[3]), le réseau reçoit des lots

de données labélisées et peut ainsi itérer les phases d'apprentissage de menière "classique". A contrario, l'apprentissage continu comporte une première phase, Phase-0, au cours de laquelle un modèle de base est entraîné sur l'ensemble des données disponibles. Cette phase est suivie d'une Phase-1 au cours de laquelle l'extracteur de caractéristiques entraîné lors de la Phase-0 est gelé et la partie classifieur du réseau est entraînée sur les échantillons du flux de données.

ExStream [5] et REMIND [6] sont des méthodes récentes qui utilisent un tampon de données afin d'éviter un oubli catastrophique. La méthode Move-to-Data (MTD) avec reciblage proposée par [3] adapte l'algorithme MTD [8] pour atténuer la dérive du modèle par le recalcul du gradient de la fonction objective à intervalles fixes (reciblage). Néanmoins, pour limiter la dérive du modèle, un choix spécifique des données pour sa mise à jour ainsi que pour son reciblage peuvent être proposés.

2.3 Sélection de données basée sur l'entropie

La sélection de "meilleurs" données a été fortement étudiée en apprentissage actif [4]. La solution proposée se base sur l'entropie, donc sur la quantité d'information moyenne apportée par l'échantillon par rapport à la taxonomie des classes du problème de classification. Dans [10], l'entropie des données a été également utilisée comme mesure pour effectuer une sélection d'échantillons dans le cadre d'une stratégie de répétition pour l'apprentissage continu. Dans un scénario réaliste où de nouvelles données étiquetées sont disponibles par portions, il est possible de mettre les données en mémoire-tampon afin de sélectionner les plus informatives. Nous ajoutons donc une mémoire-tampon à notre méthode MTD de base et utilisons l'entropie des échantillons pour i) la sélection d'échantillon d'apprentissage et ii) l'ajustement du modèle (reciblage).

3 Méthodologie

3.1 Move-to-Data (MTD) et le reciblage

La méthode Move-to-Data (MTD) proposée par [8] effectue la mise à jour sur un seul neurone de la dernière couche classificatrice du réseau en déplaçant le vecteur de ses poids synaptiques dans la direction du vecteur de caractéristiques de l'échantillon de données. Pour un réseau à M couches, la mise à jour est effectuée sur le vecteur de poids entre la $(M-1)^{\text{ème}}$ couche et le $j^{\text{ème}}$ neurone de la couche M, où j est le *vraie étiquette de classe* du nouvel échantillon. L'ajustement est contrôlé par ϵ, cf. l'Eq. 1.

$$W_j^{'[M]} = W_j^{[M]} + \epsilon \cdot \left(\left\| W_j^{[M]} \right\| * \frac{f^{[M-1]}}{\left\| f^{[M-1]} \right\|} - W_j^{[M]} \right) \quad (1)$$

où $f^{[M-1]}$ est le vecteur de caractéristiques de la couche $M-1$ du nouvel échantillon, les poids du neurone j sont $W_j^{[M]}$ et $1 > \epsilon > 0$. Comme la direction de descente de la MTD n'est pas opposée au gradient, elle est sous-optimale et une dérive du modèle peut se produire. Le MTD avec reciblage a été proposé par [3] pour atténuer cette dérive. La descente est "reciblée" dans la direction opposée au gradient pour la correction du modèle lorsque la performance du réseau tombe en dessous d'un seuil. La précision de validation a été utilisée pour mesurer la baisse de performance qui déclenche le reciblage, comme indiqué dans l'équation 2. Ici, Acc^* est la meilleure précision de validation pour les étapes d'entraînement 1 à $i-1$ et Acc_i est la précision de validation pour l'étape actuelle i. Si la performance tombe en dessous du seuil, une étape descente du gradient est effectuée uniquement pour la dernière couche (avec les échantillons observés depuis le reciblage précédent).

$$\frac{(Acc^* - Acc_i)}{Acc^*} > \alpha \quad (2)$$

3.2 MTD basée sur l'entropie (EMTD)

Le critère d'entropie peut être utilisé pour calculer l'incertitude et donc le niveau informatif [1] des échantillons pour un meilleur apprentissage.

***EMTD* :** Nous décomposons d'abord le réseau en deux fonctions : $F(.)$ - les couches de l'extracteur de caractéristiques et $G(.)$ - la couche du classifieur, de sorte que $\hat{y}^t = G(F(I^t))$ où t représente le moment où I^t l'échantillon de données d'entrée était disponible dans le flux de données et $\hat{y}^t$ est la sortie du réseau. En outre, $F(I^t) = X^t$ où X^t représente les caractéristiques de l'entrée I^t. Nous proposons d'utiliser un tampon d'apprentissage $B_X(I^t)$ de taille n (petit) qui cumule les caractéristiques des nouveaux échantillons :

$$B_X(I^t) = \left\{ X_1^t, X_2^t, ..., X_n^t \right\} \quad (3)$$

La dimension des caractéristiques X^t de l'entrée est généralement inférieure à la dimension de l'entrée I^t elle-même. Par exemple, la taille de la dernière couche de l'extracteur de caractéristiques pour le Transformer [7] est de 1×768 comparé à la taille de l'entrée des images vidéo $8 \times 3 \times 224 \times 224$. L'entropie de chacune de ces caractéristiques est calculée en utilisant la sortie *softmax* de la dernière couche pour les classes J à l'aide de l'équation 4. Ici p_j est la sortie *softmax* pour la classe j et X_i^t est le vecteur de caractéristiques de l'échantillon i pris dans le tampon de caractéristiques $B_X(I^t)$. Les valeurs d'entropie de tous les échantillons sont ensuite enregistrées dans le mémoire-tampon d'entropie $B_e(I^t)$.

$$\begin{aligned} e(X_i^t) &= -\sum_{j=1}^{J} p_j \log(p_j) \\ B_e(I^t) &= \{e(X_1^t), e(X_2^t), ..., e(X_n^t)\} \end{aligned} \quad (4)$$

L'entropie d'un échantillon est maximale lorsque les probabilités de classes sont égales pour cet échantillon. Nous effectuons une sélection de l'entropie maximale à partir de $B_e(I^t)$ pour choisir un échantillon informatif correspondant X^{t*} afin de mettre à jour notre modèle avec MTD :5.

$$e^* = \text{argmax}(B_e(I^t)),\ X^{t*} = B_X(I^t)[e^*] \quad (5)$$

Le tampon de caractéristiques dans l'équation 3 est similaire au tampon ExStream [5]. Mais, contrairement à ExStream, nous n'utilisons qu'un seul tampon pour toutes les classes. Alors que l'échantillon ayant l'entropie la plus élevée est utilisé pour la mise à jour du MTD, l'échantillon ayant l'entropie la plus faible est remplacé par la nouvelle instance (voir Eq. 6). Nous appelons cette méthode MTD basée sur l'entropie (EMTD).

$$idx = \text{argmin}\, B_e(I^t),\ B_X(I^t)[idx] = X^t \quad (6)$$

***EMTD avec reciblage* :** Après quelques itérations de l'EMTD, nous devons nous déplacer dans la direction opposée

au gradient de la fonction objective, comme nous l'avons fait dans la MTD de base. Contrairement au reciblage du MTD de base, nous n'utilisons pas toutes les données entre deux reciblages pour la descente du gradient. Au lieu de cela, nous maintenons une courte mémoire tampon de reciblage $B_r(I^t)$ de taille fixe m qui comporte les m échantillons précédemment sélectionnés utilisés pour la mise à jour de l'EMTD. Ainsi, nous ne reciblons qu'avec les échantillons ayant l'entropie la plus élevée. Le reciblage est effectué sur les valeurs des paramètres W^{t-1}*avant* l'étape actuelle t, car les valeurs actuelles W^t sont considérées comme dérivées.

4 Résultats et discussion

4.1 Description du corpus de données

Nous utilisons le corpus BIRDS [7]. Il contient 19 500 vidéos de 6 situations à risque dans la taxonomie des classes à détecter et une classe de rejet "non-risque". Nous supposons qu'à la Phase-0, les données ont été acquises de telle sorte que toutes les situations de risque ont été observées.

Comme les situations vécues par des personnes peuvent varier considérablement dans le temps pour les mêmes classes, nous créons un nouvel ensemble de validation en étiquetant continuellement les données de la Phase-1. Nos travaux antérieurs [2] montrent qu'un tel enrichissement des données annotées est réaliste. Nous avons divisé l'ensemble de données BIRDS en Phases 0 et 1 conformément au scénario décrit ci-dessus. En général, seul un petit ensemble de données est disponible pour la Phase-0. Par conséquent, l'ensemble de données est partitionné dans les proportions de 40% pour la Phase-0 et 60% pour la Phase-1, tout en conservant la distribution des classes. Nous sélectionnons 5% de vidéos non-risque pour équilibrer l'ensemble de données pour l'entraînement à la Phase-0. Les deux ensembles de données sont ensuite divisés en 80-20% pour l'apprentissage et la validation.

4.2 Résultats de l'apprentissage en continu

Les méthodes MTD sont génériques et applicables à toute sorte de réseaux de neurones artificiels ayant une seule couche de classification. Nous avons donc d'abord entraîné le "Pooling Transformer" introduit dans [7] pendant 50 époques en utilisant l'optimiseur de descente de gradient stochastique (SGD) avec un taux d'apprentissage initial de 0.001 et une décroissance exponentielle de 0.1. La meilleure précision de validation de la Phase-0 était de 76.80% et nous l'utilisons comme modèle pré-entraîné pour les expériences de la Phase-1. Pour le choix de la valeur de ϵ voir Eq. 1), nous avons essayé différentes valeurs dans l'ensemble $\{0.05, 0.02, 0.01, 0.002, 0.001\}$ pour finalement trouver la meilleur valeur : $\epsilon = 0.002$.

La figure 1 illustre la précision pour l'apprentissage en continu avec 500 échantillons. Le réseau a une précision initiale de 86.47% sur l'ensemble de validation de la Phase-1. Après 500 mises à jour, la précision du réseau diminue de 5.7% pour MTD. Le réseau conserve une meilleure performance de validation avec EMTD et la précision ne décroît que de 2.1% .

L'EMTD avec reciblage permet de compenser la dérive du modèle tout en étant plus rapide [3] que les méthodes de l'état de l'art comme ExStream [5]. Nous supposons que l'EMTD avec reciblage donnera de meilleures précisions que

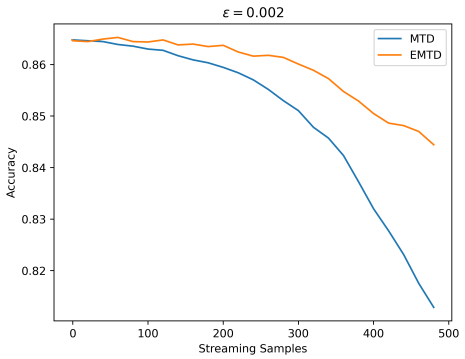

FIGURE 1 : Représentation graphique de la précision pour 500 échantillons de flux pour MTD et EMTD

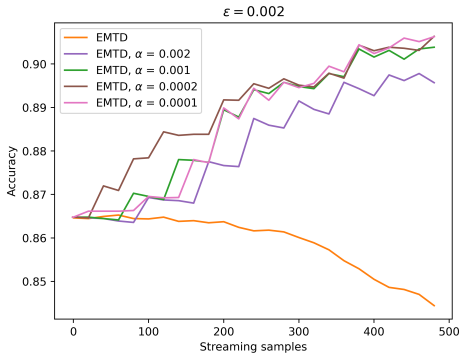

FIGURE 2 : Graphique de précision de l'EMTD pour 500 échantillons avec différents paramètres de reciblage α

l'EMTD sans reciblage. Le choix d'une "bonne" valeur α pour le reciblage EMTD (Eq. 2) peut être formulé comme un problème d'optimisation avec la contrainte que le temps d'exécution global reste inférieur par rapport aux méthodes basées sur la descente de gradient pure (ExStream) [5]. Nous le résolvons par une recherche sur grille éparse (*GridSearch*) comme nous l'avons fait pour ϵ.

La dérive du modèle est testée selon l'éq. 2 tous les 20 échantillons. Un tampon de reciblage de taille 10 est progressivement rempli et mis-à-jour (cf. section 3.2). Nous avons effectué des essais pour un ensemble des valeurs : $\alpha = \{0.001, 0.002, 0.0001, 0.0002\}$. Les résultats sont illustrés sur la figure 2 Plus la valeur α est faible, plus le reciblage est fréquent, par exemple pour $\alpha = 0.0001$, le reciblage est presque systématique tous les 20 échantillons. Concernant la précision, lorsque le reciblage est effectué avec $\alpha = 0.001$, elle a tendance à augmenter. La dérive du modèle est donc bien compensée alors que le reciblage n'est pas si fréquent.

Enfin, nous comparons les cinq méthodes : i) MTD, ii) EMTD, iii) MTD + reciblage (MTD + r), iv) EMTD + reciblage (EMTD + r) et v) ExStream sur un plus grand nombre de 2500 échantillons en continu. Nous avons utilisé un tampon de caractéristiques de taille 20 pour les tampons EMTD et ExStream. La couche du classifieur a été initialisée de manière aléatoire pour ExStream, donnant les meilleurs scores selon [3]. Pour MTD et EMTD, le classifieur a été pré-entraîné sur les données de la Phase-0. Dans [3], nous avons montré que MTD avec reciblage était très proche d'ExStream. La figure 3 illustre le fait que l'EMTD avec le meilleur paramètre de reciblage est encore plus proche : il ne diffère que de 0.35%

par rapport à ExStream, contre 3.88% pour MTD + r.

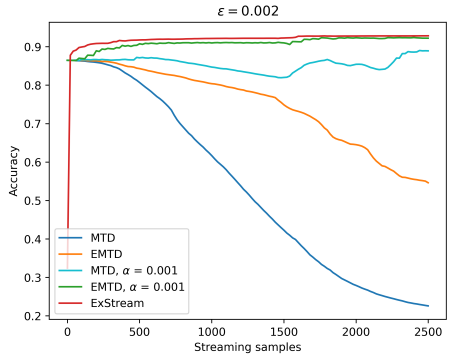

FIGURE 3 : Graphique de précision pour 2500 échantillons de flux ExStream, MTD, MTD + r, EMTD et EMTD + r

4.3 Complexité temporelle

La figure 4 montre le temps de calcul par les méthodes pour 100 échantillons et pour 2500 échantillons. Pour 2500 échantillons, ExStream s'exécute en 209.81 secondes, MTD en 182.87s, EMTD en 186.67s et EMTD avec reciblage en 187.84s. Toutes les expériences ont été réalisées sur des GPU NVIDIA A40 avec 46 Go de mémoire. La figure 4 montre que les deux tampons courts de l'EMTD n'ajoutent pas un surcoût important par rapport à la méthode MTD de base. Les méthodes basées sur la MTD sont plus rapides qu'ExStream car elles n'ont pas de calcul de gradient à chaque étape. Le reciblage nécessite une étape systématique de descente de gradient, mais elle est conditionnelle, effectuée sur un petit nombre d'échantillons du tampon de reciblage (10 dans notre cas), et peut ne pas être nécessaire pour certaines étapes.

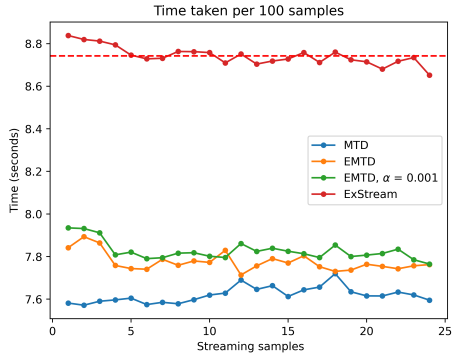

FIGURE 4 : Temps d'apprentissage pour MTD, EMTD, EMTD + reciblage et ExStream pour 2500 échantillons

5 Conclusion et perspectives

Dans ce travail, nous avons amélioré la méthode MTD en ajoutant un critère d'incertitude pour la sélection des échantillons dans le cadre de l'apprentissage continu. Nous avons appelé cette méthode EMTD et l'avons utilisée pour les Transformers de vision (ViT). L'idée centrale de l'EMTD est de choisir des échantillons qui sont plus *uncertains* pour le réseau et d'écarter simultanément les échantillons contenant moins d'informations au cours de l'entraînement. Nous avons utilisé des courts mémoires-tampons pour mettre en œuvre cette stratégie d'échantillonnage et une mémoire de faible taille d'échantillons sélectionnés par l'EMTD pour effectuer le reciblage. Nous avons comparé EMTD, MTD et ExStream et montré que EMTD avec reciblage est capable d'améliorer MTD avec reciblage. Il obtient des résultats très proches de ceux d'ExStream tout en étant plus rapide à entraîner. La perspective future de notre travail sera de se concentrer sur l'utilisation de pseudo-étiquettes au lieu des étiquettes de vérité terrain pour l'apprentissage continu.

Références

[1] Umang AGGARWAL : *Active and Incremental Deep learning with class imbalanced data*. Thèse de doctorat, University of Paris-Saclay, France, 2022.

[2] Iván González-Díaz ET.AL. : Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for alzheimer disease research. *In Proceedings of the International workshop on Multimedia indexing and information retrieval for healthcare, MIIRH@ACM, Spain, October 22*, pages 11–14, 2013.

[3] Abel Kahsay GEBRESLASSIE, Jenny BENOIS-PINEAU et Akka ZEMMARI : Streaming learning with move-to-data approach for image classification. *In International Conference on Content-based Multimedia Indexing Austria, September 14 - 16*, pages 167–173, 2022.

[4] Paul GUÉLORGET, Bruno GRILHÈRES et Titus B. ZAHARIA : Deep active learning with simulated rationales for text classification. *In ICPRAI*, volume 12068 de *Lecture Notes in Computer Science*, pages 363–379, 2020.

[5] Tyler L. HAYES, Nathan D. CAHILL et Christopher KANAN : Memory efficient experience replay for streaming learning. *In International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24*, pages 9769–9776, 2019.

[6] Tyler L HAYES, Kushal KAFLE et ET.AL. : REMIND your neural network to prevent catastrophic forgetting. *In Computer Vision–ECCV 2020, Glasgow, UK, August 23–28*, pages 466–483. Springer, 2020.

[7] Rupayan MALLICK, Jenny BENOIS-PINEAU et Akka Zemmari ET.AL. : Pooling transformer for detection of risk events in in-the-wild video ego data. *In 26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25*, pages 2778–2784, 2022.

[8] Miltiadis POURSANIDIS, Jenny BENOIS-PINEAU et Akka Zemmari ET.AL. : Move-to-data : A new continual learning approach with deep cnns, application for image-class recognition. *CoRR*, abs/2006.07152, 2020.

[9] Dequan WANG et Evan Shelhamer ET.AL. : Tent : Fully test-time adaptation by entropy minimization. *In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

[10] Felix WIEWEL et Bin YANG : Entropy-based sample selection for online continual learning. *In 28th European Signal Processing Conference, EUSIPCO 2020, Netherlands, January 18-21*, pages 1477–1481, 2020.