Adaptation du problème de questions-réponses visuelles à un contexte d'apprentissage continu

Alexis Lechat, Stéphane Herbin, Frédéric Jurie

Adaptation du problème de questions-réponses visuelles à un contexte d'apprentissage continu

Alexis LECHAT[1], Stéphane HERBIN[1], Frédéric JURIE[2]

[1]ONERA - The French Aerospace Lab
6 chemin de la Vauve aux Granges, BP 80100, FR-91123 PALAISEAU cedex, France

[2]Normandie Univ, UNICAEN, ENSICAEN, CNRS UMR GREYC
6 Boulevard Maréchal Juin, CS 45053, 14050 CAEN cedex 4, France

alexis.lechat@onera.fr, stephane.herbin@onera.fr, frederic.jurie@unicaen.fr

Résumé – Le problème de questions-réponses visuelles (QRV) constitue un nouveau défi en intelligence artificielle. Il requiert du système des capacités de raisonnement alliant compréhension visuelle et linguistique. Nous proposons d'adapter le problème QRV à une configuration d'apprentissage continu, paradigme dont l'objectif est de créer des modèles capables d'apprendre et de s'adapter dans un environnement évolutif. L'étude réalisée ici s'intéresse à un protocole d'apprentissage séquentiel de tâches définies via un partitionnement sémantique des questions. Ce protocole est analysé sur un modèle QRV courant combiné à une solution incrémentale de répétition sur mémoire épisodique.

Abstract – Visual question answering (VQA) is one of the new challenge for artificial intelligence. It is a complete task since it requires reasoning skills combining both visual and linguistic knowledge. In this paper, we provide a way to train a VQA model in a continuous learning setting, a paradigm suitable for incrementing the knowledge of adaptive systems evolving in a changing environment. We define a task incremental protocol whose tasks are built on a semantic question clustering. This method is accredited on a common VQA model combined with an episodic memory for rehearsal learning.

1 Introduction

Les réseaux de neurones profonds constituent l'état de l'art de nombreux problèmes de vision par ordinateur. Les meilleures performances sont généralement obtenues en suivant un paradigme particulier : une phase d'apprentissage et une phase de validation. Le réseau résultant est ainsi spécialisé au domaine défini par les données d'apprentissage. L'une des principales faiblesses de ce schéma de conception est la difficulté à étendre les compétences initiales du réseau, par exemple par l'ajout de nouvelles classes à discriminer. L'apparition d'un phénomène couramment dénommé *oubli catastrophique* [1], i.e. le fait que de nouvelles données ou tâches puissent provoquer l'oubli de ce qui a été appris antérieurement, pose problème.

Une solution immédiate serait de ré entraîner complètement un réseau à partir d'une base intégrant les nouvelles données. Dans de nombreuses applications, cependant, le système exploitant le réseau évolue dans un environnement réel et doit absorber un flux continu. Il apparaît donc utile de disposer d'un modèle capable de s'adapter aux variations de son milieu et d'apprendre à partir de nouvelles expériences sans dégrader ses acquis. C'est l'enjeu de l'apprentissage incrémental ou apprentissage continu qui cherche à contrôler l'oubli catastrophique, mais aussi à favoriser les transferts entre nouveaux savoirs et connaissances apprises afin d'accélérer le processus d'entraînement tout en renforçant progressivement les performances globales.

Une application de questions-réponses visuelles (QRV) [2] consiste à répondre à une question en langage naturel sur le contenu d'une image. Ce problème est par nature multimodal, car il requiert du système un raisonnement sur des connaissances à la fois visuelles et linguistiques [3], et multitâches, car il implique des capacités de classification, de détection d'objet ou d'attribut, de localisation, de comptage, etc.

L'augmentation du nombre de données QRV accessibles [4] suscite l'intérêt pour un système dont la connaissance puisse s'incrémenter. En particulier, il semble naturel qu'un système QRV rencontre, au cours de son existence, de nouveaux contextes visuels, de nouveaux termes de vocabulaire ou de nouvelles tournures grammaticales.

La contribution de cet article est d'introduire un schéma d'apprentissage incrémental sur un système QRV. Nous proposons notamment de simuler l'incrémentation du nombre de tâches réalisables par le modèle par un partitionnement sémantique des questions. Nous évaluons ce protocole sur l'ensemble VQAv2 [5].

2 Travaux connexes

2.1 Gérer l'oubli catastrophique

Les approches d'apprentissage continu dans la littérature se focalisent sur la maîtrise de l'oubli catastrophique. Trois catégories de stratégies ont été proposées :

— Architecture adaptative : l'architecture même du réseau est modifiée pour traiter l'oubli. Une approche courante est d'agrandir progressivement le réseau pour augmenter sa capacité [6].

— Régularisation : la fonction de coût est contrainte pour consolider les poids liés aux connaissances apprises, par exemple la méthode de l'"Elastic Weight Consolidation" [7].
— Répétition : une mémoire épisodique sauvegarde une partie de l'information rencontrée pendant l'apprentissage. Cette information est exploitée et mise à jour périodiquement afin de renforcer les connaissances apprises lors de l'entraînement sur de nouveaux exemples (e.g. iCaRL [8]).

L'apprentissage continu est un sujet récent et les protocoles d'évaluations sont en voie d'uniformisation [9]. Les applications se limitent encore principalement à des contextes simples (classification sur un faible nombre de classes) et à des ensembles de données restreints (MNIST, CIFAR, CORe50).

2.2 Les différents types d'apprentissage continu

L'apprentissage continu repose sur la façon dont est organisé le flux entrant de données. On suppose que la base d'apprentissage s'incrémentera tout au long de la vie du modèle, mais sans a priori sur la quantité, la dimension, ou la distribution des données ou sur le nombre de classes.

Plusieurs approches sont utilisées pour simuler un flux de données. Parmi les plus communes [9, 10], on retrouve :

— incrémentation de classes : le réseau apprend successivement sur toutes les instances d'une classe avant d'avoir accès à la suivante, et ne reverra plus les anciennes.
— incrémentation de tâches : les données sont ordonnées en tâches apprises séquentiellement par le réseau.
— incrémentation de domaines : cela correspond à apprendre sur des données dans un nouveau domaine de représentation comparé à ce qui a déjà été vu (e.g. passer d'images réelles à des images artistiques...).

Dans la section suivante, nous discutons des particularités du QRV lors de la mise en place d'un protocole incrémental avant de proposer une méthode de partitionnement des données pour une approche d'incrémentation de tâches.

3 QRV incrémental

3.1 Organisation des données pour le QRV

Les données se présentent sous la forme de triplets $(\texttt{Image}, \texttt{Question}, \texttt{Réponse})$ (noté (v, q, a)). On travaille ici avec le jeu de données VQAv2 (Figure 1).

À partir de cet ensemble, comme mentionné dans la section précédente, plusieurs logiques sont possibles pour simuler un flux continu de données. En se ramenant à un problème de classification (une réponse correspond à une classe), en associant chaque question à une tâche ou en répartissant les images par domaine, il est possible d'envisager l'une des approches incrémentales classiques. Cependant, le format multimodal des données permet également de définir d'autres variantes spécifiques au QRV.

Nous proposons d'étudier dans ce travail un partitionnement *sémantique* des questions comme logique de structuration du flux de données. L'idée sous-jacente est de simuler un apprentissage incrémental de tâches complexes d'interprétation visuelle, et d'étudier la capacité des algorithmes à le réaliser.

Figure 1 – *Exemples de l'ensemble VQAv2 [5].*

3.2 Partitionnement sémantique des questions

Les types de tâches étant très dépendants de certains termes tels que les mots interrogatifs, un partitionnement sémantique regroupe naturellement les questions concernant des tâches similaires. On simule ainsi un problème d'apprentissage séquentiel de diverses tâches comme le dénombrement, la détection d'attribut ou la vérification d'existence.

Afin d'obtenir un résultat équilibré, nous proposons de réaliser le partitionnement des questions sur un critère de distance après plongement sémantique. On utilise pour cela la représentation GloVe [11] qui à chaque mot associe un vecteur de dimensions 300. Chaque question est encodée en un unique vecteur de dimension 300 en moyennant sur l'ensemble des mots qui la compose.

On considère l'union des questions d'entraînement et de validation (215 723 questions différentes). On obtient ainsi notre subdivision en tâches pour l'apprentissage ainsi que les différents sous-ensembles de validation spécifiques à chaque tâche. Dans nos expérimentations, la base VQAv2 est partitionnée en 20 tâches via un algorithme des K-moyennes utilisant la distance euclidienne.

Bien que simple, cette méthode permet d'obtenir des tâches sémantiquement homogènes. Les premiers mots de la question sont les plus déterminants lors du partitionnement. Par exemple, toutes les questions commençant par "Is there..." sont regroupée dans la même tâche, de même pour les questions débutant par "What is the..." ou "Does the [...] have...", ceci étant notamment dû à la fréquence prédominante de ces termes dans le corpus. Néanmoins, l'encodage GloVe permet aussi un regroupement lexical pertinent.

4 Expérimentations

4.1 Modèle de prédiction de réponse

Dans le cadre de notre étude, on s'intéresse principalement au comportement d'un modèle QRV dans un contexte d'apprentissage continu. On utilise pour cela un modèle générique inspiré de l' état de l'art QRV [12] auquel on peut ajouter

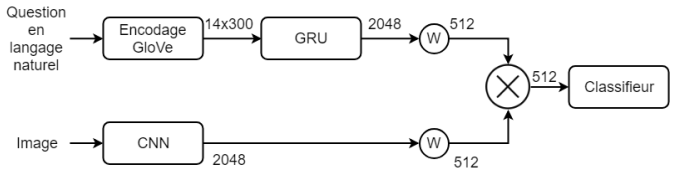

Figure 2 – *Chaîne de traitement du modèle QRV de fusion multimodal utilisant des descripteurs visuels globaux [12].*

une mémoire épisodique pour mettre en place une stratégie de répétition.

Ce modèle est constitué de deux branches, l'une encodant l'information visuelle contenue dans les images à partir de caractéristiques profondes (ResNet) apprises sur la base ImageNet, l'autre encodant la sémantique des questions à partir d'un réseau récurrent (GRU). Ces deux branches sont ensuite combinées par produit pour prédire la réponse, considérée comme la sortie d'un classifieur (Figure 2).

Ce modèle associant des encodages multimodaux et des mécanismes de fusion est typique de la plupart des approches proposées dans la littérature, dans son architecture et ses composants. Il est utilisé dans notre étude comme référence pour révéler les phénomènes caractéristiques d'un apprentissage incrémental de tâches de QRV.

Le modèle implémenté ici atteint une précision de 53% sur la validation du VQAv2 lors d'un entraînement en batch standard.

4.2 Évaluation sur le partitionnement par question du VQAv2

Le protocole suivi dans nos expérimentations est le suivant. Chaque tâche est montrée au réseau de manière séquentielle. Celui-ci réalise une unique passe d'apprentissage ("epoch") sur le sous-ensemble avant de passer à la tâche suivante. De manière similaire au protocole proposé dans [13], on étend le nombre de sorties possibles du classifieur prédisant les réponses au fur et à mesure que de nouvelles classes/réponses sont rencontrées.

Trois types d'approches sont évaluées.

Dans une première, on s'intéresse à un modèle entraîné naïvement sur les 20 tâches consécutives sans modification des paramètres d'apprentissage. Aucune solution n'est mise en place pour lutter contre l'oubli catastrophique.

Une deuxième approche exploite une mémoire épisodique de taille prédéfinie M qui accumule une partie des données observées. À chaque nouvelle tâche, le modèle apprend à la fois sur les nouveaux échantillons et sur ceux stockés en mémoire. Le nombre d'éléments conservés est équilibré entre les tâches : à la n-ième tâche, la mémoire stocke au plus $\frac{M}{n}$ triplets (v, q, a) par tâche observée. La sélection des éléments à mémoriser est aléatoire. Nous avons testé pour $M = 4000$ et $M = 40000$ ce qui correspond environ à 1% et 10% du nombre total de données d'entraînement (443 757). Cette approche simple de mémorisation montre des performances équivalentes, voire supérieures à d'autres méthodes (régularisation, intelligence synaptique) sur des problèmes d'incrémentation de classes [10].

On évalue aussi la performance obtenue dans le cas d'une répétition totale. Cette configuration correspond à une mémoire de taille infinie permettant de sauvegarder la totalité des données vues pour les repasser au réseau à chaque nouvelle tâche. Cela nous donne la borne supérieure atteignable par un apprentissage incrémental avec répétition sur une mémoire épisodique.

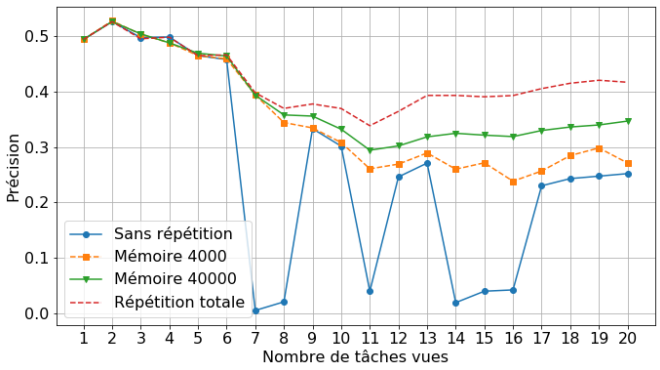

Figure 3 – *Évolution de la précision en fonction du nombre de tâches vues. Les modèles sont testés sur un ensemble de validation qui s'incrémente au fil des tâches.*

4.3 Résultats

L'entraînement des 4 modèles est optimisé avec la méthode de descente de gradient stochastique. Le taux d'apprentissage est fixé à 0.1 et reste constant. Nous n'avons pas choisi de méthode de pas adaptatif pour l'optimisation (Adam, Adamax, RMSProp, etc.) pour mieux révéler les phénomènes.

À la tâche n, on évalue le réseau sur l'ensemble de validation composé de l'union des sous-ensembles associés aux tâches 1 à n. La mesure de précision utilisée est celle définie pour le challenge VQA [2].

Les résultats sont disponibles en Figure 3. Le modèle sans répétition montre un phénomène d'oubli particulier. La performance globale du réseau se dégrade avec l'ajout de tâche, comme attendu dans ce contexte incrémental. Cependant, on constate que certaines tâches provoquent un oubli catastrophique violent. Le faible score à ces itérations prouve qu'en plus d'oublier les anciennes tâches, le réseau n'arrive pas à apprendre la nouvelle. La performance est cependant capable de remonter à la tâche suivante. Cela s'explique par la forte corrélation entre certaines tâches du problème de QRV et à notre partitionnement.

En effet, le VQAv2 est constitué à plus de 40% de questions fermées (classes "Yes" et "No"). La distribution des questions ouvertes est très hétérogène et une grande partie des réponses ont moins de dix exemples. On retrouve ce biais dans notre partitionnement avec les tâches 7, 8, 11 et 15 qui comportent principalement des questions "What..." alors que les 14 et 16 sont des tâches liées au dénombrement: "How many...". Le reste des tâches comporte une plus grande proportion de questions fermées. C'est la cause principale des oscillations de la précision. Le réseau est capable d'apprendre rapidement la classification binaire *oui/non* et cela se répercute par un gain de performance sur l'ensemble des tâches contenant des questions fermées. Cependant, une tâche composée uniquement de classes rares provoque un oubli brutal de la connaissance acquise du modèle. La permutation de l'ordre des tâches pendant

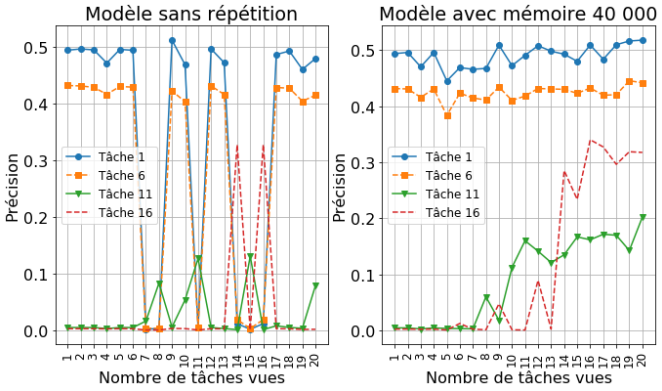

Figure 4 – *Comparaison de l'évolution de la précision pour 4 tâches différentes entre un modèle sans mémoire (à gauche) et le modèle avec une mémoire de taille 40 000 (à droite). Chaque tâche voit sa précision mesurée sur son ensemble de validation propre.*

l'apprentissage n'impacte pas le résultat observé : ce sont les mêmes tâches qui provoquent une chute des performances.

L'ajout d'une mémoire épisodique stabilise le comportement du réseau et augmente ses performances proportionnellement à sa taille.

La Figure 4 montre l'évolution de la précision du modèle sur 4 tâches tout au long de son cycle de vie. La forte corrélation inter tâche, en particulier pour les questions fermées, est ici aussi illustrée par la précision élevée du réseau sur des tâches qu'il n'a pas encore rencontrées (par exemple, la tâche 6). D'autre part, la comparaison entre le modèle sans mémoire et celui avec une mémoire de taille 40 000 met clairement en valeur l'impact de la répétition. Les tâches sur lesquelles le réseau performe correctement voient leur précision se stabiliser et le modèle réussit à s'améliorer sur les tâches les plus difficiles dont les résultats étaient négligeables dans le cas sans mémoire.

5 Conclusion

La diversité et la complexité du problème QRV motivent l'intérêt de disposer d'un modèle capable d'apprendre de façon continue. Dans cet article, nous avons proposé une démarche d'évaluation incrémentale reposant sur un partitionnement sémantique des questions. Cette méthode permet d'organiser le flux de données d'apprentissages en tâches qui sont montrées au modèle de manière séquentielle et de comparer les performances des algorithmes.

L'étude réalisée ici sur la base VQAv2 permet d'illustrer les spécificités du QRV par rapport aux problèmes de classification usuellement utilisés dans la littérature de l'apprentissage continu. Les différents niveaux de corrélation entre tâches ont un impact important sur les performances. Il peut être positif — certaines tâches profitant des apprentissages passés —, mais également fortement négatif et provoquer un oubli catastrophique global sur les tâches anciennes. Des méthodes simples exploitant une mémoire épisodique permettent de limiter ces phénomènes.

Pour la suite de nos travaux, nous nous intéressons à d'autres démarches pour le partitionnement des tâches (en prenant en compte par exemple le domaine visuel) pour analyser le comportement de l'apprentissage continu, et conduire à de nouvelles méthodes pour contrôler l'oubli catastrophique des modèles de QRV.

References

[1] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville et Y. Bengio. *An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks*. ICLR, 2014.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick et D. Parikh *VQA: Visual Question Answering*. ICCV, 2015.

[3] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman F.-F. Li, C. L. Zitnick et R. B. Girshick. *Inferring and Executing Programs for Visual Reasoning*. ICCV, 2017.

[4] D. A. Hudson et C. D. Manning. *GQA: a new dataset for compositional question answering over real-world images*. CVPR, 2019.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra et D. Parikh. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*. CVPR, 2017.

[6] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu et R. Hadsell *Progressive Neural Networks*. arXiv:1606.04671, 2016.

[7] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran et R. Hadsell *Overcoming catastrophic forgetting in neural networks*. PNAS, 2017.

[8] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, et C. H. Lampert. *iCaRL: Incremental Classifier and Representation Learning*. CVPR, 2017.

[9] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, et S. Wermter. *Continual Lifelong Learning with Neural Networks: A Review*. Neural Networks, February 2018.

[10] Y.-C. Hsu, Y.-C. Liu et Z. Kira. *Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines*. arXiv:1810.12488, 2018.

[11] J. Pennington, R. Socher et C. D. Manning. *GloVe: Global Vectors for Word Representation*. Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, 2014.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. CVPR, 2018.

[13] D. Maltoni et V. Lomonaco. *Continuous Learning in Single-Incremental-Task Scenarios*. Neural networks : the official journal of the International Neural Network Society, 2018.