

Continual Learning

Catastrophic Forgetting

Saliou BARRY, Zhile ZHANG, Carine MOUBARAK
Sorbonne University

November 28, 2024

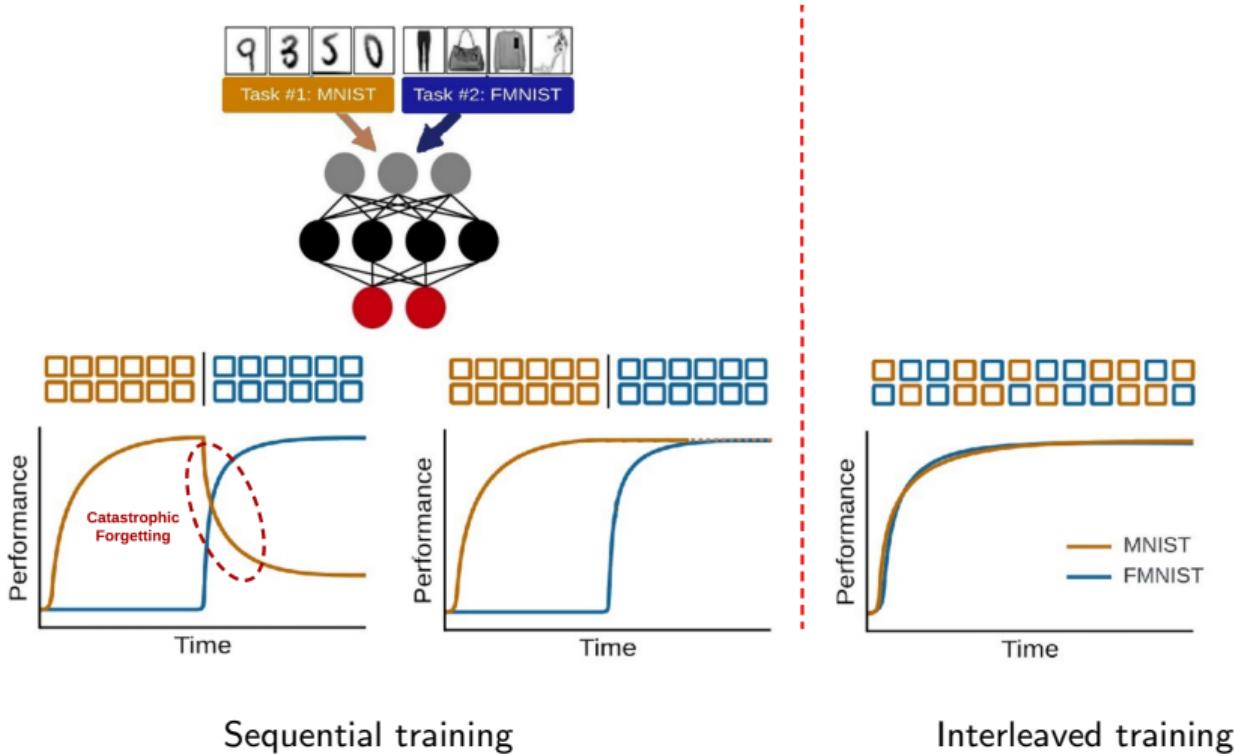
Overview

1. Méthodes de Référence

2. Notre Contribution

3. Expérimentations

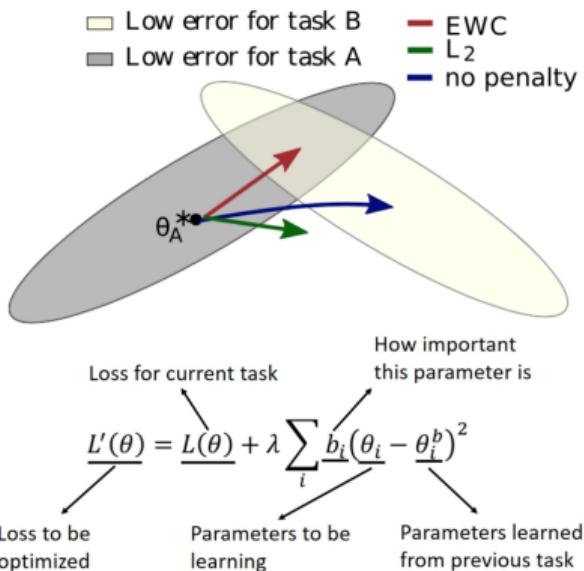
Catastrophic forgetting



Approches

- Méthodes de Régularisation
 - Elastic weight consolidation (EWC),
 - Synaptic Intelligence (SI),
 - Learning without forgetting(LwF)
- Méthodes de Rejeu (ou Replay)
 - Replay Memory
 - Generative Replay,
 - Experience Replay
- Méthodes de Masquage et d'Isolation de Paramètres
 - Supermasks,
 - Hard Attention to Task (HAT),
 - Progressive Neural Networks
- Méthodes Basées sur les Distillations de Connaissances
 - Knowledge Distillation (KD),
 - Dark Experience Replay (DER++)
- Méthodes Basées sur le Pruning et les Adapters :
 - Adapter Modules,
 - Pruning Sélectif

Elastic Weight Consolidation (EWC)



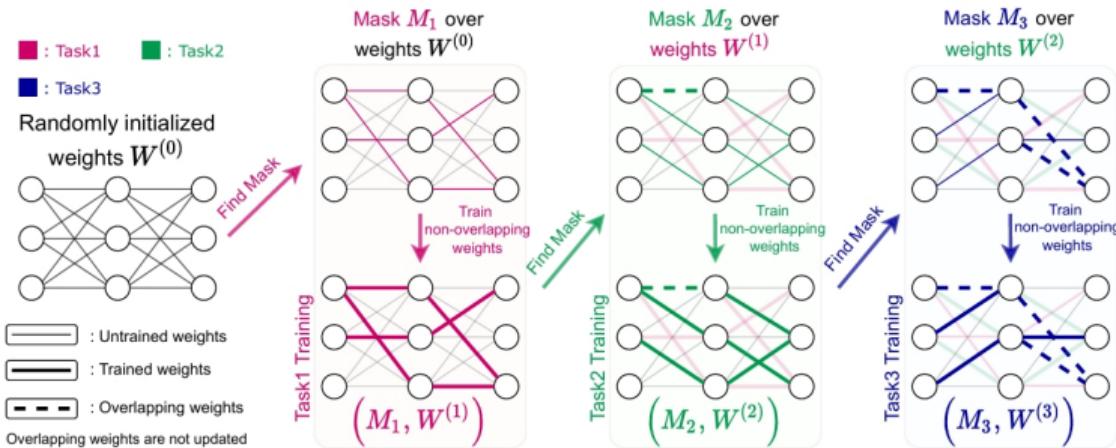
Avantages

- Compromis entre plasticité et stabilité
- Pas de Stockage d'anciennes données

limites

- Moins efficace lorsque les nouvelles tâches sont très différentes des précédentes
- Hypothèse de Gaussianité
- Complexité de mise à l'échelle
- Difficile d'estimer précisément l'importance des paramètres

Supermasks



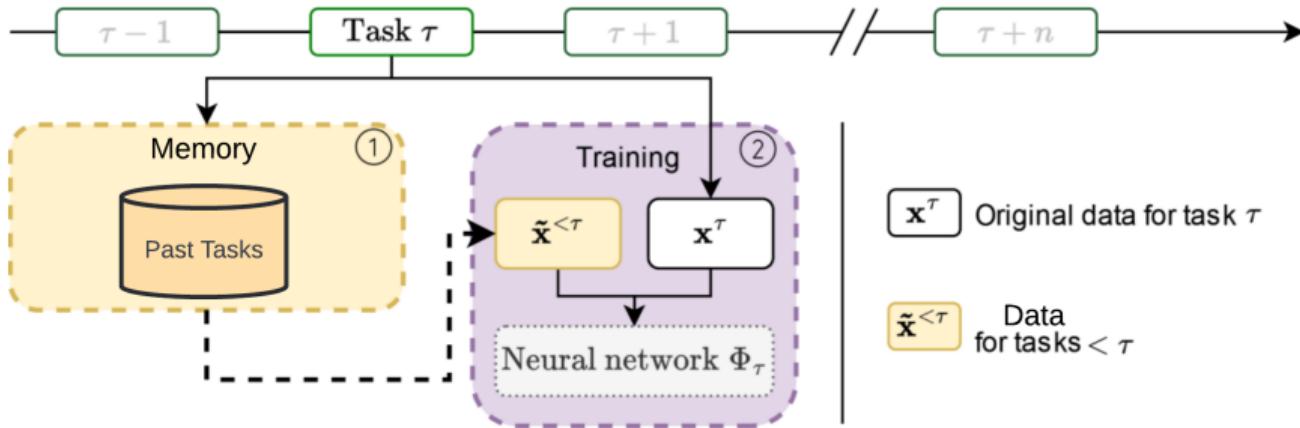
Avantages

- Utilisation efficace des sous-réseaux
- Simplicité

Limites

- Hypothèse sur l'identité des tâches
- Limitation en nombre de tâches.
- Pas de transfert de connaissances

Replay Memory



Avantages

- Préservation des connaissances passées.
- Flexible

Limites

- Augmentation du temps d'entraînement
- Sélection des échantillons
- Dépendance à la taille de la mémoire

Neuro-Layered Adaptive Memory (NLAM)

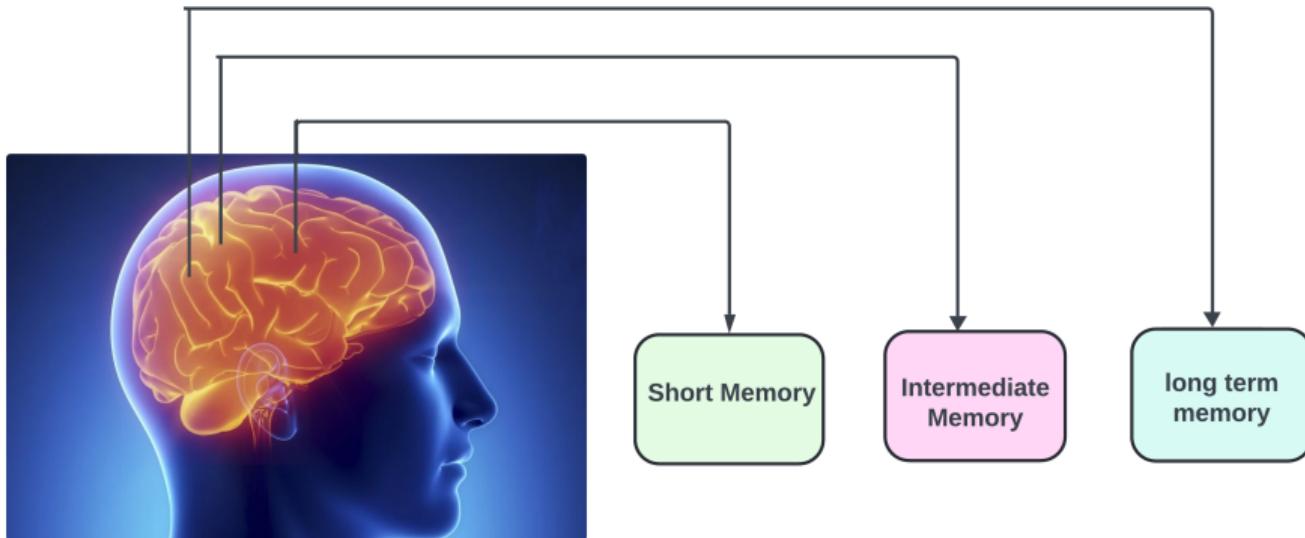
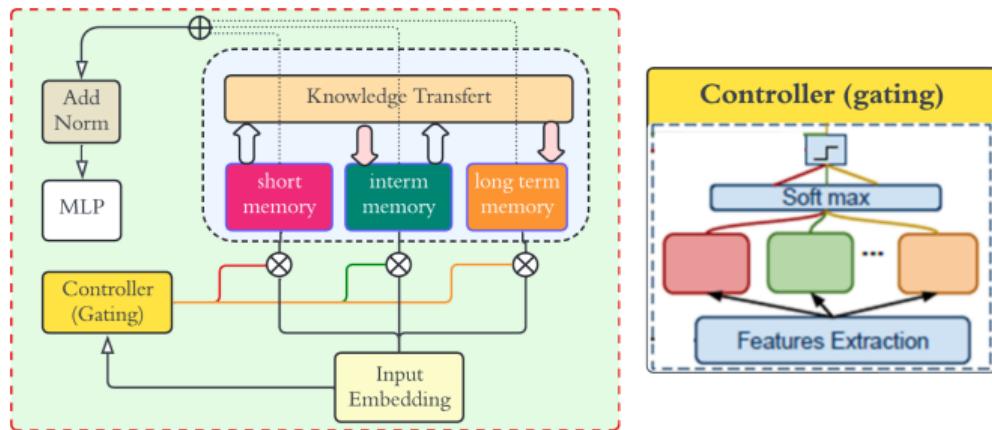


Schéma global de l'architecture NLAM.



$$sortie_i = w_i \cdot \text{Bloc}_i$$

Où

- w_i : Pondération du bloc.
- Bloc_i : Représentation du bloc i .
- $sortie_i$: Sortie du bloc i ;

- Le **Controller** pondère les blocs en fonction de leur importance relative pour la tâche.
- Les sorties des blocs sont **combinées** et passées dans un dense layer.

Knowledge Transfert

Permet de transférer les connaissances apprises d'un bloc à l'autre après convergence , permettant une consolidation des connaissances.

La convergence est atteinte lorsque la perte $\Delta Loss < \epsilon$ (*où ϵ est un seuil*), indiquant que l'apprentissage pour une tâche est stabilisé.

Mise à jour des Blocs :

$$B_{t+1} = \alpha \cdot B_t + (1 - \alpha) \cdot V$$

Où:

- B_t : Bloc à l'instant t , contenant les connaissances actuelles.
- V : Représentation des connaissances transférées.
- α : Contrôle l'importance de la mémoire précédente par rapport au nouvel ajout.
- B_{t+1} : Bloc mis à jour après l'intégration de V .

Short Memory

Les vecteurs générés par les LSTM représentent des dépendances temporelles à court terme.

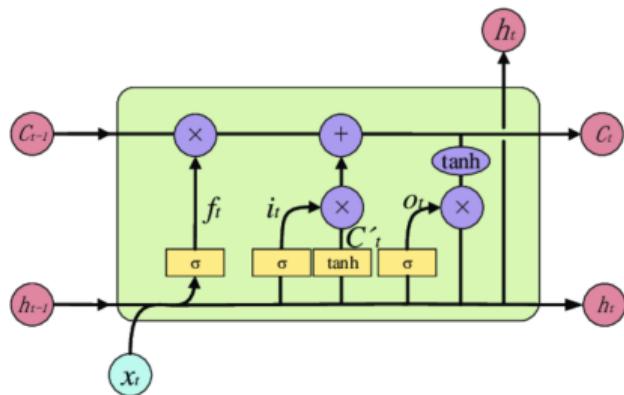


Figure: LSTM

- Entrée : l'input embedding avec la pondération du contrôleur
- Sortie : Vecteurs de représentations temporelles .
→ But principal :
- Modéliser des dépendances séquentielles à court term.

Intermediate Memory

Les Transformers permettent d'extraire des représentations globales et contextuelles.

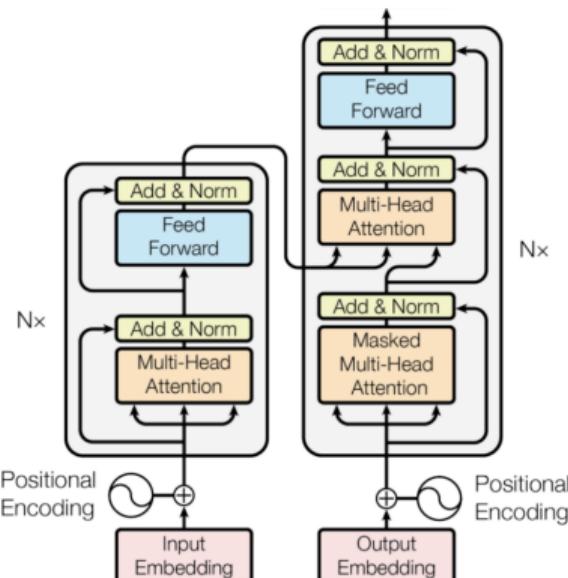
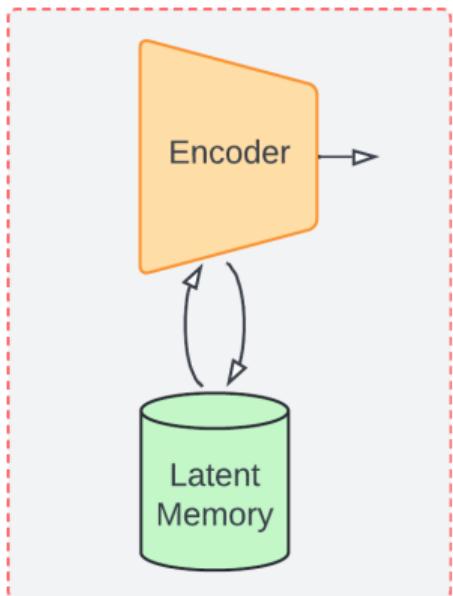


Figure: Transformers

- Entrée : l'input embedding avec la pondération du contrôleur
- Sortie : Représentations globales et contextuelles.
→ But principal :
- Extraire des dépendances à moyen terme.

Long term Memory



- Entrée : l'input embedding avec la pondération du contrôleur
- Sortie : Représentation abstraite consolidée.
→ But principal :
Stocker des caractéristiques abstraites des tâches.

Figure: MemoryEncoder

→ Avantages

- Fournit une représentation hiérarchique multi-niveaux.
- Évite le stockage des données brutes
- Permet le transfert de connaissances entre niveaux.
- Pas d'hypothèse sur la nature des tâches

→ Limites

- Coût computationnel élevé

→ Amélioration

- Rajout d'un mécanisme d'oublié

Dataset

- **ASC (Phone, Camera, Restaurant)** : Classer des avis selon leur sentiment (*positif, négatif, neutre* pour *Restaurant*, uniquement *positif* et *négatif* pour *Phone* et *Camera*).
- **ACL** : Classifier des phrases contenant une citation en fonction de leur intention (*background, motivation, etc.*) (**5 classes**).
- **AI** : Identifier la relation exprimée dans une phrase entre deux entités (*used for, part of, etc.*) (**7 classes**).
- **PubMed** : Classifier des interactions entre une molécule chimique et une protéine. (**13 classes**).

Choix des métriques

- **Acc (Précision)** : Proportion des prédictions correctes.
- **MF1 (Macro-F1)** : La moyenne des F1 pour les classes d'une tâche :

$$\text{MF1}_{\text{tâche}} = \frac{1}{N} \sum_{i=1}^N F1_i;$$

- **Forward transfer** : Mesure comment les tâches précédentes aident les nouvelles.

$$\text{FT}_{T_i} = \text{Acc}_{T_i}^{\text{aprèsCL}} - \text{Acc}_{T_i}^{\text{avantCL}}$$

- **Backward transfer** : Mesure l'impact d'une nouvelle tâche sur l'accuracy d'une tâche précédente.

$$\text{BT}_{T_j} = \text{Acc}_{T_j}^{\text{aprèsCL}} - \text{Acc}_{T_j}^{\text{avantCL}}$$

Résultats expérimentaux pour le modèle NLAM

Tâches Bloc	Restaurant	Camera	Restaurant → Camera (CL) rest_CL	Camera (CL) cam_CL	Backward	Forward
Bloc 1	82.31	84.67	30.45	83.72	-51.86	-0.95
Bloc 2	85.36	86.12	40.78	85.88	-44.58	-0.24
Bloc 3	86.10	87.11	39.68	86.77	-46.42	-0.34
Bloc 1 & 2	87.56	88.20	56.33	88.16	-31.23	-0.04
Bloc 1 & 3	87.80	88.98	61.2	88.54	-26.60	-0.44
Bloc 2 & 3	89.56	90.24	79.89	90.02	-9.67	-0.22
Bloc 1 & 2 & 3	92.00	90.50	92.33	92.49	+0.33	+1.99

Table: Performances en précision du modèle NLAM avec diverses configurations de mémoire sur deux catégories (Restaurant et Camera)

Comparaison des modèles

Tâches	Rest.		ACL		AI		Phone		PubMed		Cam.	
Model	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc	MF1	Acc
NLAM Non-CL	85.01	92.03	71.02	76.02	73.00	75.29	82.51	89.05	76.09	-	85.5	92.5
NLAM-CL	86.5	93.0	72.01	75.03	73.5	77.50	86.20	90.5	78.0	-	89.5	94.0
NCL	79.52	86.54	68.39	72.87	67.94	75.71	84.10	86.33	72.49	-	85.71	90.70
KD	78.05	85.59	69.17	73.73	67.49	75.09	82.12	84.99	72.28	-	81.91	88.69
EWC	80.98	87.64	65.94	71.17	65.04	73.58	82.32	85.13	71.43	-	85.35	89.14
DER++	79.00	86.46	67.20	72.16	63.96	73.54	83.22	85.61	72.58	-	87.10	91.47
HAT	76.42	85.16	60.70	68.79	47.37	65.69	72.33	79.13	69.97	-	74.04	85.14
HAT Adapter	79.29	86.70	68.25	72.87	64.84	73.67	81.44	84.56	71.61	-	82.37	89.27
DAS	80.34	87.16	69.36	74.01	70.93	77.46	85.99	87.70	72.80	-	88.16	92.30

Table: Comparaison des modèles à travers les tâches (MF1 et Précision)

Conclusion

Notre modèle : Une solution à l'oubli catastrophique !

- **Objectif atteint** : La perte d'informations entre les tâches est éliminée.
- **Limitation** : Les calculs sont exigeants pour un grand nombre de tâches.
- **Perspectives** : Explorer des optimisations pour améliorer l'efficacité.

Merci pour votre attention !

Références

-  Zixuan Ke and Yijia Shao and Haowei Lin and Tatsuya Konishi and Gyuhak Kim and Bing Liu (2023)
Continual Pre-training of Language Models,
-  Prateek Yadav and Mohit Bansal (2023)
Exclusive Supermask Subnetwork Training for Continual Learning,
-  Gido M. van de Ven and Nicholas Soures and Dhireesha Kudithipudi (2024)
Continual Learning and Catastrophic Forgetting,
-  James Seale Smith and Lazar Valkov and Shaunak Halbe and Vyshnavi Gutta and Rogerio Feris and Zsolt Kira and Leonid Karlinsky (2024)
Adaptive Memory Replay for Continual Learning,