Enhancing Continual Learning in Diabetic Retinopathy: Multimodal Zero-shot Clustering and Strategic Experience Replay

Gusseppe Bravo-Rocca
Barcelona Supercomputing Center
Barcelona, Spain
gusseppe.bravo@bsc.es

Peini Liu
Barcelona Supercomputing Center
Barcelona, Spain
peini.liu@bsc.es

Jordi Guitart
Barcelona Supercomputing Center
Universitat Politècnica de Catalunya
Barcelona, Spain
jordi.guitart@bsc.es

Ajay Dholakia
Lenovo Infrastructure Solutions Group
Morrisville, NC, USA
adholakia@lenovo.com

David Ellison
Lenovo Infrastructure Solutions Group
Morrisville, NC, USA
dellison@lenovo.com

Rodrigo M Carrillo-Larco
Emory University
Atlanta, GA, USA
rmcarri@emory.edu

Abstract

Traditional deep neural networks often suffer from catastrophic forgetting when faced with domain incremental learning. To address this, we have developed an approach that generates descriptions from multimodal inputs using a Large Language Model, learning a zero-shot clustering that is used to cluster subsequent tasks with no supervision. We advance the concept of Experience Replay by integrating a strategic sampling methodology derived from these clusters, which refreshes the neural network's memory, thereby curtailing the degradation of knowledge retention across successive tasks. This sampling is integral to our Experience Replay strategy, which updates the multi-head classifier incrementally, using only a fraction of the points. When evaluated on a challenging Diabetes Retinopathy dataset, our approach not only mitigates forgetting but also complements and enhances existing continual learning strategies such as Elastic Weight Consolidation (EWC), Gradient Episodic Memory (GEM), and Learning Without Forgetting (LwF).

1. Introduction

Incremental learning emulates the adaptive learning capabilities of living beings, overcoming the limitations of traditional machine learning by enabling continuous knowledge acquisition from new data without forgetting previously learned information. This approach addresses challenges such as catastrophic forgetting and the need for models to adapt to new tasks and data distributions in real-time [4, 7, 11]. Our paper introduces an unsupervised learning framework, leveraging a Large Language Model (LLM) for interpreting multimodal data, particularly in medical imaging for diabetic retinopathy detection from fundus images (see Figure 1). By generating textual descriptions from images and metadata, our framework facilitates zero-shot clustering and Experience Replay (ER), enhancing model adaptability and efficiency on CPU infrastructure without compromising data privacy [8, 12]. The model architecture incorporates a neural network expecting embeddings as input and dynamically adds classifiers for new tasks. This design ensures scalability and flexibility, crucial for addressing domain shifts in medical imaging without retraining from scratch, thus preserving privacy and reducing performance degradation [15, 18]. Moreover, our approach augments existing continual learning strategies (EWC [7], GEM [10] and LwF [9]) by integrating CLIP models for robust image embedding, demonstrating significant performance improvements on challenging datasets like the Diabetes Retinopathy dataset [6, 7, 9, 10].

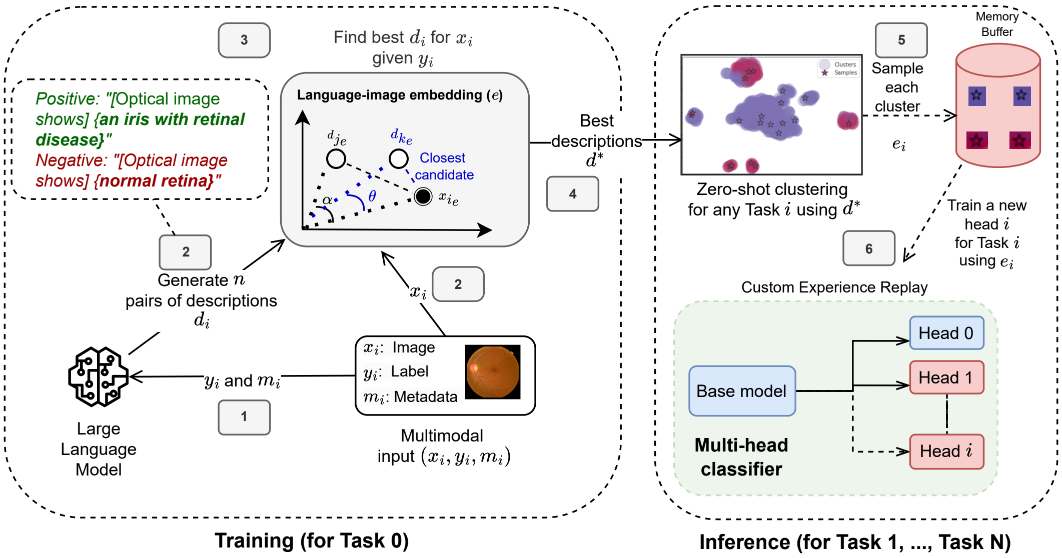
3
Find best d_i for x_i given y_i
Positive: "[Optical image shows] {an iris with retinal disease}"
Negative: "[Optical image shows] {normal retina}"
Language-image embedding (e)
d_{je}
d_{ke}
Closest candidate
α
θ
x_{ie}
Best descriptions d^*
4
2
Generate n pairs of descriptions d_i
x_i
2
x_i: Image
y_i: Label
m_i: Metadata
y_i and m_i
1
Large Language Model
Multimodal input (x_i, y_i, m_i)
Training (for Task 0)
5
Sample each cluster
Memory Buffer
Clusters
Samples
e_i
Zero-shot clustering for any Task i using d^*
Train a new head i for Task i using e_i
6
Custom Experience Replay
Base model
Head 0
Head 1
Head i
Multi-head classifier
Inference (for Task 1, ..., Task N)

Figure 1. Our method employs a LLM to generate descriptions d_i for each image x_i, using metadata m_i and y_i for initial domain learning in Task 0 (supervised phase). These descriptions underpin unsupervised zero-shot clustering, forming $|\{y_i\}|$ clusters. Key points from these clusters are buffered for replay (e_i). A multi-head classifier leverages this buffer in an Experience Replay strategy, learning the pertinent head i for predictions y, thus preserving knowledge across successive tasks (unsupervised phase).

2. Related work

This work intersects continual learning with the use of LLMs like GPT-4 and foundation models such as CLIP to enhance zero-shot learning and ensure privacy-preserving incremental learning. Here, we contextualize our contributions within the existing body of research, underscoring how our methodology diverges from and builds upon prior work.

2.1. LLMs and Zero-shot Learning

LLMs have significantly advanced zero-shot learning capabilities, allowing for nuanced understanding and generation of human-like text. Utilizing GPT-4 for generating text descriptions and CLIP for visual representations enables our approach to zero-shot clustering, a departure from the direct application of LLMs in task execution [2, 12].

2.2. Experience Replay (ER)

Rooted in simulating human memory, ER techniques, including recent innovations like Dark Experience Replay [3], aim to balance new learning with the preservation of past knowledge. Our work extends this concept by incorporating dual-memory structures to mitigate catastrophic forgetting, enhancing the efficiency of continual learning without a clear consensus on the optimal memory architecture [1, 14].

2.3. Privacy-preserving Techniques

In response to the privacy challenges in storing raw data from previous tasks, our methodology focuses on embedding exemplars, aligning with the need for privacy in fields like medical image analysis. This strategy not only complies with privacy regulations but also maintains the utility of experience replay for effective learning [13, 16, 17].

2.4. CLIP Embeddings for Continual Learning

We adapt CLIP for use as an embedding tool in our ongoing continual learning framework, avoiding the constraints of directly refining it for sequential tasks. This approach preserves zero-shot learning capabilities and addresses catastrophic forgetting, leveraging LLM-generated descriptions for optimal exemplar selection [5].

3. Approach

Our approach employs GPT-4 and CLIP for zero-shot clustering and experience replay in continual learning, enabling exemplar identification without storing raw images, enhancing privacy and efficiency.

3.1. Zero-shot Clustering with LLM and CLIP

We apply GPT-4 to generate textual descriptions and use CLIP for both visual and textual embeddings, facilitating zero-shot clustering, as outlined in Algorithms 1 and 2. This setup enables classifying images into predefined categories without explicit prior training, leveraging the semantic depth of generated descriptions.

Algorithm 1: Embedding Generation for Zero-shot Clustering

```
Data: Set of images $\{I_1, I_2, ..., I_n\}$, Set of textual
      descriptions $\{D_1, D_2, ..., D_m\}$ generated by
      GPT-4
Result: Normalized embeddings $\mathbf{X}_i$ for images and
        $\mathbf{T}_j$ for text
for each image $I_i$ and description $D_j$ do
    $\mathbf{X}_i \leftarrow \frac{\text{CLIP}_{\text{image}}(I_i)}{\|\text{CLIP}_{\text{image}}(I_i)\|}$
    $\mathbf{T}_j \leftarrow \frac{\text{CLIP}_{\text{text}}(D_j)}{\|\text{CLIP}_{\text{text}}(D_j)\|}$
```

Algorithm 2: Zero-shot Clustering via Cosine Similarity

```
Data: Normalized embeddings $\mathbf{X}_i$ for images, $\mathbf{T}_j$
      for text
Result: Label assignments $L_i$ for each image based
        on highest cosine similarity
for each image embedding $\mathbf{X}_i$ do
    $S_{ij} \leftarrow \cos(\mathbf{X}_i, \mathbf{T}_j) = \frac{\mathbf{X}_i \cdot \mathbf{T}_j}{\|\mathbf{X}_i\|\|\mathbf{T}_j\|}$
    $L_i \leftarrow \text{argmax}_j\, S_{ij}$
```

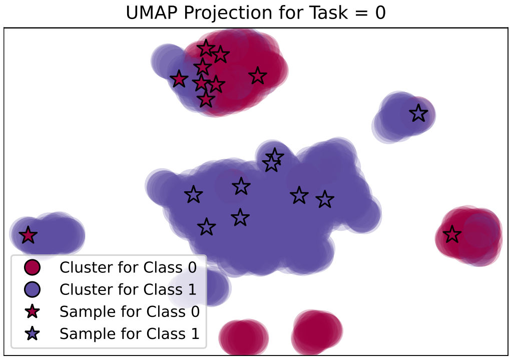

Figure 2. Two clusters (one for each class) with ten samples from the memory buffer (zeroshot exemplars) for embeddings belonging to Task 0 (fundus images with uniform quality).

3.2. Stratified Sampling for Experience Replay

To ensure a balanced and representative experience replay buffer post-clustering, we use stratified sampling (see Algorithm 3). This method guarantees a diverse set of experiences is retained for learning, crucial for maintaining performance across tasks. In Figure 2, we see the chosen samples for each cluster.

3.3. Experience Replay Algorithm

Our ER algorithm is crucial for mitigating catastrophic forgetting. It refreshes the model's knowledge by revisiting a subset of previously encountered data, now selected through zero-shot clustering and stratified sampling. This method is shown in Algorithm 4.

Algorithm 3: Stratified Sampling for Experience Replay

```
Data: Multimodal documents $\mathcal{M}$, Zero-shot labels
      $\mathcal{Z}$
Result: Sampled subset $\mathcal{S}$ ensuring class balance
for each label $l \in \mathcal{Z}$ do
    $\mathcal{M}_l \leftarrow \{m \in \mathcal{M} | \text{label}(m) = l\}$
    if $|\mathcal{M}_l| > n_{neighbors}$ then
        $\mathcal{S}_l \leftarrow \text{sample}(\mathcal{M}_l, n_{\text{neighbors}})$
    else
        $\mathcal{S}_l \leftarrow \mathcal{M}_l$
$\mathcal{S} \leftarrow \bigcup_{l \in \mathcal{Z}} \mathcal{S}_l$
```

Algorithm 4: Experience Replay Strategy

```
Data: Updated replay buffer $\mathcal{B}$, new data points $\mathcal{D}$
Result: Refreshed model knowledge via selective
        replay
for each new data point $d \in \mathcal{D}$ do
    $\mathcal{M}_d \leftarrow \text{convert\_to\_multimodal\_document}(d)$
    $\mathcal{B} \leftarrow \text{update\_buffer}(\mathcal{B}, \mathcal{M}_d, \text{strategy})$
for each training epoch do
    $\mathcal{S}_{\text{replay}} \leftarrow \text{sample\_from\_buffer}(\mathcal{B})$
    Train model on $\mathcal{S}_{\text{replay}}$ combined with current
      task data
```

As outlined in Algorithms 1, 2, 3, and 4, our approach integrates embedding generation, zero-shot clustering, and stratified sampling to dynamically update the experience replay buffer, enabling efficient and privacy-aware continual learning.

4. Experiments

We evaluated our approach through experiments focused on domain incremental learning for diabetic retinopathy detection. Our experiments were designed to mimic real-world scenarios, testing the model's robustness under various conditions.

Testbed: The experimental setup included Ubuntu 22.04 LTS on hardware equipped with dual Intel® Xeon® Platinum 8360Y CPUs at 2.40GHz, with 256 GB RAM. Software tools comprised Docker image intel/oneapi-aikit for the Intel® AI Analytics Toolkit, avalanche-lib for continual learning, PyTorch and torchvision for deep learning, and intel-extension-for-pytorch and scikit-learn-intelex for optimized computations.

Dataset: We employed a modified version (meant for incremental learning and binary classification) of the APTOS 2019 Blindness Detection dataset [6], featuring 3,662 retina images for detecting diabetic retinopathy, developed by Aravind Eye Hospital, India.

Methodology: Our methodology involved preparing three tasks to simulate different imaging conditions (as described in Figure 3), using a neural network model with a multi-head classifier (meant for each task). Model training encompassed continual learning strategies like Naive (fine-tuning), EWC, LwF, and GEM, with evaluation based on Average Mean Class Accuracy (AMCA), defined as:

$$AMCA = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{C}\sum_{c=1}^{C} a_{c,t}\right) \qquad (1)$$

where T represents the number of test intervals (three in our dataset), and C is the count of classes (two in our dataset).

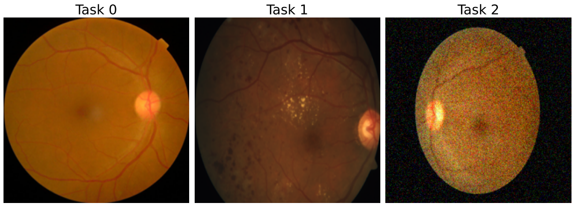

Figure 3. Fundus images representing different tasks with varying image quality and conditions. From left to right: Task 0 shows a fundus photograph with uniform image quality; Task 1 is an image with some variation in lighting; Task 2 displays an image with added Gaussian noise to simulate a challenging imaging condition.

Results: The experiments were conducted over five runs with different random seeds to make them statistically significant. Tables 1 and 2 demonstrate that our approach, especially when using our experience replay strategy with the zero-shot exemplars buffer, consistently improve traditional strategies in terms of AMCA across various sample sizes. The following explains a bit more these numbers:

Table 1. AMCA values for Naive and GEM strategies across various sample sizes per class

Samples	**Naive**		**GEM**	
	Ours	**Original**	**Ours**	**Original**
15	**0.924**	0.904	**0.924**	0.919
20	**0.924**	0.909	**0.924**	0.919
25	**0.925**	0.913	**0.924**	0.918
30	**0.925**	0.916	**0.923**	0.919
50	**0.926**	0.922	**0.924**	0.921

Naive + Our Approach: The slight AMCA increase with Naive strategy implies basic incremental learning benefits from zero-shot capabilities. Our method likely enhanced class representation through stratified sampling.

Table 2. AMCA values for LwF and EWC strategies across various sample sizes per class

Samples	**LwF**		**EWC**	
	Ours	**Original**	**Ours**	**Original**
15	**0.923**	0.917	**0.917**	0.899
20	**0.922**	0.917	**0.919**	0.902
25	**0.922**	0.917	**0.919**	0.906
30	**0.924**	0.917	**0.919**	0.908
50	**0.924**	0.918	**0.918**	0.911

EWC + Our Approach: EWC mitigates knowledge loss by penalizing significant weight modifications. Improved AMCA suggests our approach's embeddings enhanced crucial weight identification, aiding knowledge preservation and new information integration.

LwF + Our Approach: LwF employs knowledge distillation to retain learned information. Our approach possibly provided better embeddings, enabling effective performance maintenance on old tasks while adopting new ones, as indicated by higher AMCA.

GEM + Our Approach: GEM aims for gradient alignment across tasks to prevent forgetting. Our approach seemingly improved task retention through enriched embeddings in the replay buffer, enhancing gradient alignment and AMCA.

5. Conclusion

Our experiments validate the benefit of integrating zero-shot learning with experience replay in continual learning for medical imaging. We demonstrated that LLMs can enhance domain incremental learning, particularly through zero-shot clustering and description-based experience replay, ensuring high model performance across diverse scenarios. We contributed a framework that applies these concepts to diagnose diabetic retinopathy, proving its effectiveness in practical settings and potential for broad applicability. Future work should explore the extension of our approach to more complex data, refine the clustering algorithm for larger datasets, and combine our method with other learning strategies to further boost performance and efficiency. Scalability and ethical considerations around privacy and bias also warrant further investigation. Besides, it needs further evaluations before it can be adopted in clinical practice.

References

[1] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system, 2022. arXiv preprint arXiv:2201.12604. 2

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., 2020. 2

[3] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., 2020. 2

[4] Matthias Delange, German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. A survey on continual learning for deep neural networks. In *arXiv preprint arXiv:2007.00487*, 2021. 1

[5] Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models, 2023. arXiv preprint arXiv:2310.16226. 2

[6] Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection, 2019. 1, 4

[7] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1

[8] Pratyush Kumar and Muktabh Mayank Srivastava. Example mining for incremental learning in medical imaging, 2018. 1

[9] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 1

[10] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6470–6479. Curran Associates Inc., 2017. 1

[11] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 1

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. arXiv preprint arXiv:2103.00020. 1, 2

[13] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017. 2

[14] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference, 2019. arXiv preprint arXiv:1810.11910. 2

[15] Joan Serra, Dídac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 1

[16] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2994–3003. Curran Associates Inc., 2017. 2

[17] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–910, 2015. 2

[18] Jianshu Zhang, Yankai Fu, Ziheng Peng, Dongyu Yao, and Kun He. Core: Mitigating catastrophic forgetting in continual learning through cognitive replay, 2024. arXiv preprint arXiv:2402.01348. 1