Hildesheim University and Fraunhofer IAIS

Thesis Proposal

How does Continual Learning affect FLAVA

MSc. in Data Analytics

Jatin Karthik Raghu (313301)

University Supervisor
Dr. Ujjwal, Prof. Dr. Niels Landwehr

Fraunhofer Supervisor
Dr. Benny Stein

1 Introduction

The amount of data produced today, by both humans and machines, considerably exceeds the capacity of humans to understand and make complex decisions on that data. This has given rise to the new age AI models which essentially deals with large amounts of data and learn inherently from this unlabeled data to further improve the results set by previous baseline models. Models trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning are called *foundation models*. As defined in [2]; foundation models are a new paradigm for building AI in which one model is trained on a huge amount of data and adapted to many applications.

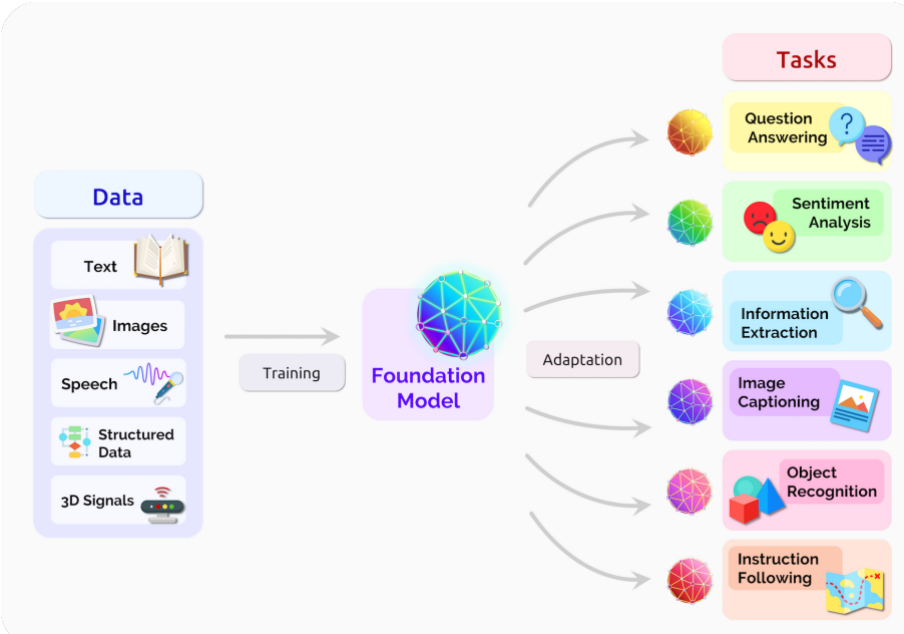

Figure 1: Basic Layout of Foundational model (On page 6 figure 2 of [2])

In traditional machine and deep learning paradigms generally distinguish the processes of knowledge training and knowledge inference, where the model is required to complete its training on a pre-prepared dataset within a limited time which can then be used for inference. As compared to traditional behavior of data in ML and DL training and inference; in the real-world data is dynamic and ever evolving. Therefore, the idea of *Continual Learning (CL)* emerged, which focuses on retaining previously learned knowledge while learning new tasks or information over time. Continual learning addresses two main problems; firstly, learning from data with multiple categories in sequential order can easily lead to the problem of catastrophic forgetting (forgetting previously acquired knowledge) [14]. Secondly, when learning from new data of the same category in sequential order, this can also lead to the

problem of concept drift [8]. [1]

2 Motivation

From the Introduction section, we can clearly understand that CL is quintessential after model creation for effective deployment in the real-world. Therefore, with the new age of Vision Language Model (VLM) and other LLM models being used more frequently; there arises a need to check if these large models are either prone to drift and can further inculcate new tasks with ease. Large foundational models have become prevalent to everyday use with the advent of GPT-3 (commonly used as chatgpt); thus, re-training such huge models as data evolves is very impractical and inefficient. This is where the main motivation behind applying CL to foundation models arise.

Furthermore, most of these foundation models have been trained on data which is not publicly available like CLIP[15](Contrastive Language-Image Pre-Training), GPT (Generative Pre-Trained model) etc. and the source code for such models haven't been made public too. Hence, we shall focus on a specific foundation model-FLAVA[2][16] and mainly research upon how FLAVA responds to new tasks via zero shot evaluation (no fine tuning for the new task). As to why exactly only the FLAVA VLM model is chosen; because the training time and number of parameters for FLAVA is much less than other VLM models like CLIP and FLIP (please refer to fig. 2). Secondly, the model is available publicly and no such CL based experimentation has been reported on FLAVA to the best of my knowledge. Therefore, exploring how FLAVA would react to new data with time with and without CL methods applied on it is the main motivation for my thesis.

3 Related Work

3.1 Vision Language Models

In recent years, pre-trained models have progressed at a dizzying speed in tandem with the evolution of transformers. Due to the significance of single-modal language/vision pre-training, some pioneering works [5, 10, 12] have recently attempted to explore the joint representation of language and vision by pre-training large-scale models on vision and language modalities, which are referred to as Vision-and-Language Models (VLMs). Models in the VLM space can be broadly divided into two categories: (i) dual encoders where the image and text are encoded separately followed by a shallow interaction layer for downstream tasks [9]; and (ii)

[1] "Task" in this document refers to new data or information unless otherwise mentioned

[2] FLAVA is a sub-project of Torchmultimodal repository licensed by BSD. The source code is available at: `https://github.com/facebookresearch/multimodal/tree/main/examples/flava`

Foundational Models	(Data size on image classification task)	No. of Parameters (all wrt image classification task)	Training Time	Inference Time
CLIP	400M	~ 3.6B	32 epochs on 400M data on ImageNet took 5600 GPU days	
Visual GPT-3		175B+	More than 7500 GPU days	
FLORENCE	900M	893M	10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU	
FLIP	340 M	2B - 12.8B	Similar setup as CLIP takes 2000 GPU days	
FLAVA	70M	240M	Less than CLIP and FLIP but exact time not mentioned	

Figure 2: Comparison of FLAVA against other VLM models

fusion encoder(s) with self-attention spanning across the modalities [4]. In totality considering both these categories, there are three primary components in VLMs, namely vision encoder (VE), text encoder (TE), and modality fusion module (MF). VE and TE are pre-trained with images and texts, respectively. MF is pre-trained using image-text pairs and amalgamates the output embeddings of VE and TE. On the other hand transformer-based VLMs like SimVLM [19], ALIGN [9], and CLIP [15] have also shown impressive results and are being researched deeply too.

3.1.1 FLAVA VLM

FLAVA [16], a foundational language and vision alignment model, learns strong representations through joint pre-training on both unimodal and multimodal data while encompassing crossmodal alignment objectives and multimodal fusion objectives. FLAVA adopts the ViT (Vision Transformer) architecture for the visual encoder, text encoder, as well as multimodal encoder. Unlike the other VLMs, FLAVA is pre-trained on Public Multimodal Datasets, which consist of 70M image-text pairs.

Giving a small brief into the details of FLAVA architecture:

- Image Encoder: Vision Transformer (ViT-B/16) is used which takes input images splits it into patches and outputs a list of hidden image state vectors with the classification token.

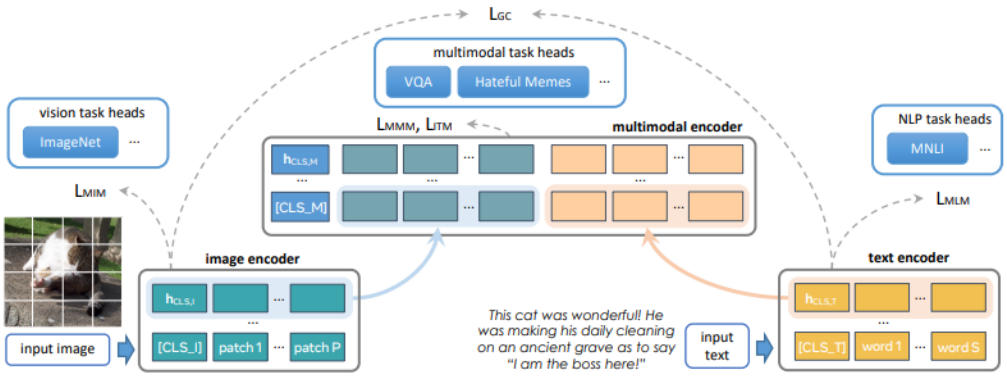

Figure 3: Overview of FLAVA Architecture

- Text Encoder: For the sake of similarity, ViT-B/16 transformer is used here too which takes piece of text as input, tokenizes it and returns a list of hidden state vectors with classification token.
- Multimodal Encoder: Two learned linear projections over each hidden state vector are concatenated into a single list, which allows cross-attention between the projected unimodal image and text representations and fusing the two modalities.

3.2 Continual Learning

A large portion of CL literature focuses on incremental learning, which can be further divided into three different scenarios – task, domain and class incremental learning [18]. These scenarios are referred to as CL scenarios and every CL method mentioned further details upon all of these variations of scenarios based on data and task category requirement. The main goal of continual learning research is to mitigate the problem of *catastrophic forgetting (CF)*, i.e., the tendency of neural networks to forget existing knowledge when learning new tasks with novel input patterns. Continual learning has been investigated in many areas, such as classification, reinforcement learning, and online learning. Main solutions can be grouped into three categories: replay-based; architecture-based and regularization-based methods. Memory replay methods use exemplars that are either stored in the memory, or synthesized using generative techniques. Hence a dedicated memory buffer is required in this category [1]. Architecture-driven CL methods either dynamically expand a network, or divide into sub-networks to cater for the new tasks [20]. Finally, regularization methods avoid catastrophic forgetting by limiting the plasticity of the model parameters that are important for previous tasks [11].

While all the aforementioned approaches show promising results in different CL scenarios, we specifically explore regularization and memory replay-based methods,

given their popularity in recent literature. Thus, we study the behaviors of pre-trained models on these methods.

3.3 Continual Learning in Foundation model

There have been some recent developments in this particular domain. Replay-based methods have been extensively studied over four different foundation models in [21]. The authors of [21] have also set up the evaluation metrics to better compare the replay-based CL methods on several foundation models. Furthermore, the CLIP foundation model has been delved into deeply and various new CL based methods on CLIP have been explored. Most of them are bound by the memory buffers because they create new image and text encoders for every new task [7]. Research has been done on other pre-trained models like [3, 13]. Considering all the literature mentioned, there has not been enough research done on CL on FLAVA specifically. Therefore, the main focus of this work would be to investigate pre-trained model of FLAVA on different CL methods specifically replay and regularization-based methods due to their extensive usage.

4 Problem Formulation

4.1 General Continual Learning Formulation

Consider a sequence of tasks $D = \{D_1, D_2...., D_T\}$ where t^{th} task $D_t = \{x_i^t, y_i^t\}_{i=1}^{N_t}$ contains tuples of input samples $x_i^t \in X$ and its corresponding label $y_i^t \in Y$. The goal is to optimize the model $f_\theta : X \rightarrow Y$ parameterized by θ, such that it predicts the label $y = f_\theta(x) \in Y$ given an unseen test sample x from an arbitrary task. During task t, the data from the previous distribution D_{t-1} is not available or restricted.

Here, assume X(t) and Y(t) are the input and output data distributions. Based on the input P(X(t)) and output P(Y(t)) distributions of task t, with $P(X(t)) \neq P(X(t+1))$, continual learning can be classified into four popular settings with slightly different assumptions i.e. task-incremental, class-incremental, domain-incremental and task-free CL. The common task- , class-, domain-incremental settings assumes task data D_t arrives in a sequence such that $t = \{1, 2...., T\}$. At task t, the class-incremental setting defines output space for all observed class labels $Y(t) \subset Y(t+1)$ with $P(Y(t)) \neq P(Y(t+1))$. Different from class-incremental setting, task incremental settings defines $Y(t) \neq Y(t+1)$ with $P(Y(t)) \neq P(Y(t+1))$ requires tasks label t to indicate isolated output heads Y(t) . Different from task- and class-incremental settings where each task has different classes, domain incremental setting is defined as $Y(t) = Y(t+1)$ with $P(Y(t)) = P(Y(t+1))$ and $P(X(t)) \neq P(X(t+1))$ such that it contains a set of images drawn from a different domain, but has the same set of classes for every task. The more challenging setting is task-free (or task-agnostic) setting, where the task data D_t changes smoothly and the task identity t is un-

known [6, 17]. It is important to note that during task t, the data from the previous distribution D_{t-1} is not available or restricted.

4.2 Continual Learning with FLAVA

FLAVA model involves an image encoder to extract unimodal image representations, a text encoder to obtain unimodal text representations, and a multimodal encoder to fuse and align the image and text representations for multimodal reasoning. It uses a ViT-B/16 transformer for both image encoder and text encoder (Please refer to [16] and fig. 3 for more details on the architecture).

The FLAVA model is trained with a global contrastive loss which promotes the similarities between image and text embeddings belonging to the same image-caption pair, so that both get aligned in the joint feature space. Let us denote the FLAVA model as F = E (visual), E(text), where E(visual) and E(text) are image and text encoders respectively. If we consider image classification as the task category, i.e. y_i and we have prompt template p, such as "a photo of a category" to form a category-specific text input $\{p; y_i\}$, so we can pass these $\{p; y_i\}$ pairs through text and image encoders respectively to get their embeddings and then we can get a similarity score to compute how close the prompt is for the respective task image.

We can further extend this for image-pair retrieval as in huge pre-trained models like FLAVA image-text pairing can be categorized as a classification task where each text is classified to an image and even a small change in hyperparameter tuning would result in catastrophic forgetting. Therefore, the main goal is to perform zero-shot evaluation in such a way that the pre-trained capabilities can be preserved via CL methods.

4.3 Dataset Foundation

Considering that FLAVA has been pre-trained on many open source datasets, it is better to limit the experimentation to smaller datasets to look into the actual performance analysis of the proposed change rather than implement a fully functional new Foundational model altogether. Therefore, the main datasets at focus would be the commonly used CL datasets which are splitMNIST, permMNIST, CIFAR10 and CIFAR100.

4.4 Evaluation Metrics

1. UT-Accuracy: Updated model via a CL method's accuracy.
2. ZS-Accuracy: Zero shot test accuracy of the model before update.
3. BwT (Backward Transfer): is the average drop of accuracy on previous sessions

after fine-tuning with the current task.

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} S_{T,i} - S_{i,i}$$

Here $S_{T,i}$ is the accuracy of current session 'S' after CL fine tuning and $S_{i,i}$ is the accuracy of task i after the i^{th} session training

5 Approach

FLAVA mainly has three components, out of which we shall focus on the image and text encoder for the Image classification task. One of the preliminary focuses would be to prune the FLAVA model and check if the pre-trained components of image, text and multimodal encoder can be used individually (Refer to [16] for more details on FLAVA architecture).

As shown in fig. 3 of FLAVA architecture , the image encoder gets an image as input and outputs the embeddings of the image and similarly the text encoder outputs the text embeddings. we would like to freeze these two encoders for future purpose. Now suppose our first Image classification task category is based on CIFAR100 dataset, then we shall run a zero-shot evaluation for Image classification using the embeddings from the frozen Image encoder. After which, we shall apply various CL methods upon this frozen image encoder like EWC (Elastic Weight Consolidation), ER (Experience Replay) (refer to section 6 for all the planned CL approaches to be investigated).

Therefore, analyzing FLAVA on two baseline regularization and replay-based methods (as mentioned) in section 6 would be the primary goal. Furthermore, evaluating these CL based approaches against zero-shot evaluation of FLAVA would provide us with interesting results as to how better or worse do huge pre-trained models perform with and without CL setting. Therefore, this same thing can also be further extended to the text encoder and NLP based task categories depending on the time and experiments.

6 Goal of Thesis

i) One of the first steps towards experimenting with the FLAVA model is Model Pruning. Check if it is possible to only use a component of the FLAVA model to perform only vision or NLP task categories specifically.
ii) Compare non-memory based CL methods specifically regularization methods like EWC(Elastic Weight Consolidation) and LwF(Learning without Forgetting) against memory based ones namely replay-methods like ER(Experience

Replay) and GR (Generative Replay) (on Image Classification Task category only first).
iii) Analyzing and comparing these two main categories of CL methods on FLAVA will help us decide upon whether a single frozen component like Image encoder of FLAVA is better performing with or without CL on zero shot evaluation.
iv) Identify the bottlenecks of CL in VLM's (Vision Language Models) and specify the reasons of failure or success depending on the experiments.

7 Project Plan

Time	**Writing/Research**
Jul	– Thesis topic planning and Literature review – Streamlining goals – First Idea Talk
Aug - Oct	– Initial pre-processing and start experimentation – Model pruning on FLAVA to prune the Image encoder and testing this pruned image encoder on datasets mentioned in section 4.3 – Investigating CL methods mentioned in Thesis Goals – Incorporating the Image encoder of FLAVA with these various CL methods starting with LwF
Nov - Dec	– Evaluating results and performing more experiments by hyperparameter tuning – Thesis drafting
Jan	– Thesis submission and Defence

Table 1: Planned Time Table

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[5] Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, and Shrikanth Narayanan. Vault: Augmenting the vision-and-language transformer with the propagation of deep language representations. *arXiv preprint arXiv:2208.09021*, 2022.

[6] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[7] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022.

[8] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[11] 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.

[12] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv*

preprint arXiv:1908.03557, 2019.

[13] Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019.

[14] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[17] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022.

[18] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[19] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[20] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

[21] Yao Zhang, Haokun Chen, Ahmed Frikha, Yezi Yang, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. Cl-crossvqa: A continual learning benchmark for cross-domain visual question answering. *arXiv preprint arXiv:2211.10567*, 2022.