Task-Agnostic Experts Composition for Continual Learning

Luigi Quarantiello[1,*], Andrea Cossu[1] and Vincenzo Lomonaco[1]

[1]Computer Science Department, University of Pisa, Largo Bruno Pontecorvo 3, 56127, Pisa - Italy

Abstract

Compositionality is one of the fundamental abilities of the human reasoning process, that allows to decompose a complex problem into simpler elements. Such property is crucial also for neural networks, especially when aiming for a more efficient and sustainable AI framework. We propose a compositional approach by ensembling *zero-shot* a set of expert models, assessing our methodology using a challenging benchmark, designed to test compositionality capabilities. We show that our *Expert Composition* method is able to achieve a much higher accuracy than baseline algorithms while requiring less computational resources, hence being more efficient.

Keywords

Compositionality, Continual Learning

1. Introduction

Current AI research has given significant attention on developing large foundation models [1]. Such architectures, typically composed by numerous non-linear layers and billions of parameters, exhibit remarkable results across diverse application domains. Nonetheless, these achievements come with a substantial computational demand, thereby impacting noticeably on the environment through increased carbon emissions [2]. For this reason, we believe that a change of direction should be promoted, by focusing on the development of more efficient and sustainable learning-based solutions.

One possible approach in this sense comes by applying the principle of compositionality. The standard definition of compositionality [3] reads as follows:

> *The meaning of a compound expression is a function of the meanings of its parts and of the way they are syntactically combined.*

In other words, it refers to the ability of recognizing the whole as the sum of its parts. This concept has been studied in the context of AI for more than 30 years [4, 5], starting by taking inspiration from biological properties of the human brain, which appears to function in a highly compositional way. In computer vision applications, in particular, it can be seen as the capacity of classifying an image by detecting the objects present in the scene.

One of the main advantages that comes from applying the compositionality property is the one of knowledge reuse. In fact, one could have a set of pretrained models, each able to recognize particular patterns or objects, and reuse them for different applications, selecting only the needed networks and composing their answers. Such composition could be *zero-shot*, *i.e.* directly applying the models without further training, or *few-shot*, allowing for a short training procedure with only a small set of samples.

Following this approach leads to a much more sustainable AI framework, since it allows to effectively exploit the knowledge acquired by the models, reusing them multiple times, while reducing the need for long and expensive training processes. Compositional models, given that they focus on small sub-components, are also *by design* invariant to larger scale changes, such as variations in the background of an image or in the relations among objects. Additionally, by simply adjusting the models composition,

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy

*Corresponding author.

luigi.quarantiello@phd.unipi.it (L. Quarantiello); andrea.cossu@unipi.it (A. Cossu); vincenzo.lomonaco@unipi.it (V. Lomonaco)

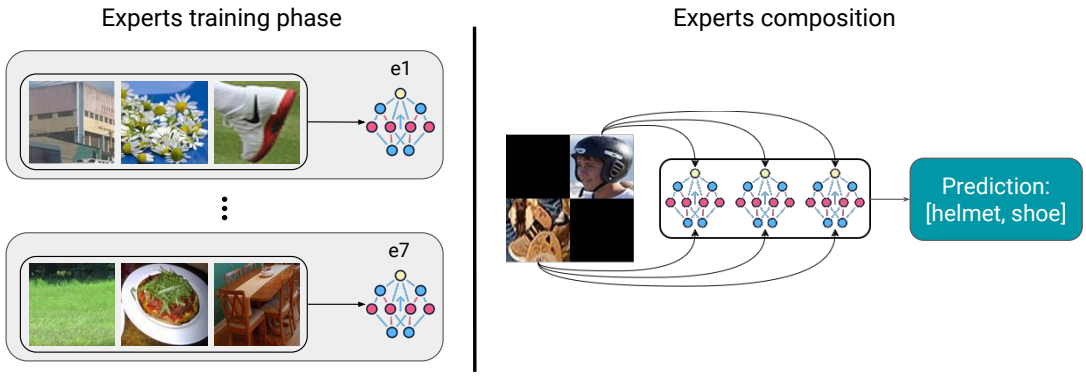

Figure 1: Left: training of the expert models, each on a different subset of cropped images from GQA. Right: the experts composition is tested on CGQA.

such architectures are able to fit seamlessly to different applications and data distributions; these makes them much more versatile and robust than their "*monolithic*" counterparts.

In this work, we mainly explore the composition of knowledge coming from different expert models. We will show that, by enforcing compositionality, we are able to largely surpass the performance of known approaches, using a compositional benchmark to asses our method.

2. Experts Composition

To show the benefits of enforcing compositionality to neural models, our proposed methodology consists mainly in two steps: first, the training of expert models, and then their application in a compositional scenario. Ideally, especially considering that the number of pretrained models available online is always increasing, it could be possible to simply download the needed networks, thus avoiding the training process at all. Unfortunately, in the case of our work, we could not find adequate *ready-to-use* models. Hence, we opted to train our own networks, nonetheless keeping them as small as possible to maintain an efficient approach.

To train the experts, we employed cropped images from the GQA [6] dataset. We selected 21 objects from the dataset, we used the information about their bounding boxes to crop the original images and we resized them to a resolution of 98x98 pixels (Figure 1, left). We assigned 3 unique classes to each model, thus resulting in 7 experts. Each model is trained to specialize on its classes, while a generic "*other*" class label is used as a container of the remainder of classes. In this way, an expert can both provide an answer when prompted with samples from its classes, but also detect when an image is out of its scope.

These trained models were then tested on the highly compositional CGQA benchmark [7], that contains samples made by a 2x2 grid, in which 2 cropped object images taken from GQA are inserted (Figure 1, right). CGQA was originally designed for Continual Learning [8], a Machine Learning paradigm in which models are trained on streams of different tasks in dynamic environments; nonetheless, the benchmark can be easily adapted for offline testing. In the paper, two main scenarios are presented, the task-incremental and class-incremental learning cases. We will compare our method against the second one, which is the most challenging setting, since no task identifier is provided and a single-head classifier must be employed.

Our approach can be formalized as follows: for each query image, we extract its composing quadrants; then, the prediction for each quadrant is computed as:

$$\forall \text{ expert}, \quad p = \arg\max \ \text{expert}(q)$$

where q is a quadrant from an image. Lastly, the entire prediction for a compositional image is given by the concatenation of the predictions on its quadrants:

$$p_{\text{comp}} = p_1 \cup p_2$$

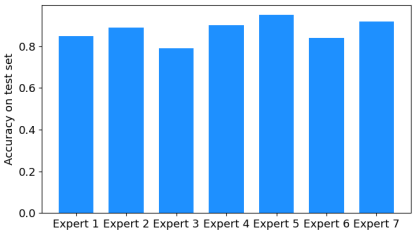

Figure 2: The accuracy of the expert models.

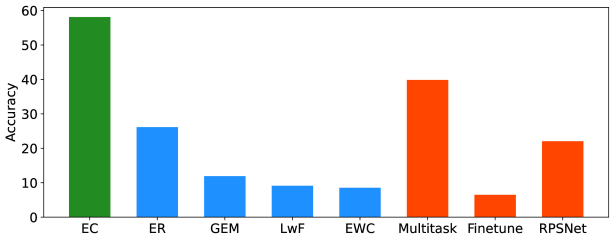

Figure 3: Accuracy of our method Experts Composition (EC) *wrt* baselines. In blue, CL methods using a pretrained backbone. In red, results taken from [7].

In this fashion, no additional training procedure is needed to recognize the objects composition, since the experts are employed *zero-shot*. This results in a substantial saving in training effort with respect to related approaches, which instead need to be fine-tuned for each experience, incurring also in loss of accuracy due to the catastrophic forgetting phenomenon.

With the same models, we also explored the use of a finetuned approach to enforce compositionality. Specifically, we used the few-shot *sys* stream provided by CGQA, defining 300 different tasks, called experiences. For each experience, we select the top-performing experts, we freeze them and we train a single-layer linear classifier on top of their concatenated features. To determine which experts to use, we perform a forward pass over all models using the training data, collecting the sum of the top logits per model. This value represents how much current data are *in-distribution* with the training data of a given model.

3. Experimental Results

3.1. Training of the Experts

Firstly, we trained a set of neural networks to define the expert models. Our main objective was to design an efficient and sustainable method; therefore, we selected the ResNet-18 architecture [9], the smallest version of the well-known, *state-of-the-art* computer vision model. As discussed in the previous section, we trained our models on images of objects from the GQA dataset.
The resulting dataset contains 21 different classes; on average, each class has about 11.000 samples. It was then divided so that each expert is trained to specialize on 3 classes, thus resulting in a set of 7 models. Following the dataset splitting of GQA, on average, each ResNet was trained on about 38.000 images, with about 2000 images as validation set, and 3800 test images.

In Figure 2, the accuracy of these models on their respective test set is plotted. The employed architectures have 18 layers, with 3x3 convolutional kernels; we used Adam as optimizer, with a learning rate of 1e-3. The resulting models have a similar behavior, with an average accuracy of 88%. Conversely, in terms of computational demand, we used a NVIDIA Tesla T4 GPU, taking about 13 minutes for each training loop. Therefore, we were able to achieve optimal performance over a large set of images, while using few resources and in a short time.

3.2. Experts Composition

Successively, we tested the composition of our pretrained experts on the CGQA benchmark. In its original paper, the authors used the dataset with Continual Learning techniques, namely ER [10], EWC [11], LwF [12], GEM [13] and RPSnet [14], as well as the *Multitask* and *Finetune* baselines.
For our experiments, we used the data from the *con* stream, which contains combinations of objects from the 21 different classes. The dataset contains 100 different combinations, *i.e.* labels, with 100 test instances for each combination.

In Figure 3, we report the results we obtained with our EC method, together with the ones from

the baseline approaches. For some of the baselines, to compare the approaches more fairly, instead of using randomly initialized networks, we pretrained a ResNet on the entire dataset extracted from GQA, designing a single expert across all 21 object classes. For the other approaches, we took the accuracy results directly from the original paper.
As it is clear from the plot, our approach abundantly surpass all the other algorithms, obtaining an accuracy over the test set of 58%. A second important observation is that, being an un-trained composition, the performance of our approach does not change over time, and it is already optimal from the first experience. Conversely, the other methods need to be trained across the experiences, being able to reach their top accuracy only at the end. For these experiments, we used the Avalanche library [15].
The best behavior is given by the ER algorithm, which achieves 26% of accuracy; in addition, on average, each of the four baselines we reproduced took 3 hours and 45 minutes to complete the training, while our experts do not require any additional fine-tuning phase. In other words, our approach turns out to be more efficient than the other ones from a computational viewpoint, but also more robust to shifts in the data distribution, which are common especially in real-world scenarios.

We also tried to compose the expert models using a custom classification head, trained using few samples for each class. We set the number of experiences to 300, as in the CGQA paper, and took the average accuracy over the test set. We experimented the composition of 3, 5 and 7 experts, reporting the obtained results in Table 1. Such method has much worse performance, especially when employing few experts. This is probably due to the fact that the training samples are too few to train an effective classifier. We plan to investigate deeper the problem, since we believe that, in some setting, zero-shot composition is not adequate and a certain degree of fine-tuning is needed to adapt to the problem.

Number of Experts	Avg. Test Set Accuracy (%)
3	18.82 ± 0.98
5	35.77 ± 1.68
7	42.82 ± 1.45

Table 1
Results from the fine-tuned expert composition.

4. Conclusion

In this work, we presented a study about enforcing compositionality in neural models. Such property plays a crucial role in the context of a sustainable AI framework, since it allows to reuse pretrained models for different applications, saving the computational resources needed for additional training procedure. We trained different expert models on a set of objects taken from the GQA dataset, and then tested their composition on CGQA, an highly compositional benchmark. We assessed our methodology against several baseline approaches, obtaining better results both in terms of performance and efficiency. Specifically, we managed to nearly double the accuracy achieved by other methods, while consuming much less computational resources and training time.

For this work, we employed a challenging benchmark for compositional approaches, but it is nonetheless a synthetic and artificial dataset. As future work, we plan to expand our method to more realistic and complex settings.
Moreover, we consider this work as an initial step, and we believe that additional endeavours can be spent to deeper explore such approaches, in different directions. As an example, it could be interesting to study how to obtain expert systems from pretrained architectures, to further enhance the efficiency of this kind of solutions. Lastly, we want to experiment more on fine-tuned compositions, to adapt the method for settings in which zero-shot composition is not enough.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI ChatGPT for grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[2] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021).

[3] M. Werning, W. Hinzen, E. Machery, The Oxford handbook of compositionality, OUP Oxford, 2012.

[4] R. Braham, J. Hamblen, The design of a neural network with a biologically motivated architecture, IEEE transactions on neural networks 1 (1990) 251–262.

[5] T. Hussain, Modularity within neural networks, Queens University Kingston, Ontario, Canada (1995).

[6] D. Hudson, C. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.

[7] W. Liao, Y. Wei, M. Jiang, Q. Zhang, H. Ishibuchi, Does continual learning meet compositionality? new benchmarks and an evaluation framework, Advances in Neural Information Processing Systems 36 (2024).

[8] G. Parisi, R. Kemker, J. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural networks 113 (2019) 54–71.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. Dokania, P. Torr, M. Ranzato, On tiny episodic memories in continual learning, arXiv preprint arXiv:1902.10486 (2019).

[11] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences 114 (2017) 3521–3526.

[12] Z. Li, D. Hoiem, Learning without forgetting, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 2935–2947.

[13] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, Advances in neural information processing systems 30 (2017).

[14] J. Rajasegaran, M. Hayat, S. Khan, F. Khan, L. Shao, Random path selection for continual learning, Advances in neural information processing systems 32 (2019).

[15] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. Hayes, M. De Lange, M. Masana, J. Pomponi, G. Van de Ven, et al., Avalanche: an end-to-end library for continual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3600–3610.