1 Implementation details

All models were trained using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0005 and with a batch size of 64. We stopped the training of the models whenever their accuracy on the validation set did not increase for 3 times in a row. Word embeddings had a size of 300. RNNs had two hidden layers and LSTM cells had a size of 1024. MLPs had one hidden layer of size 1024. We used the implementation released by (Johnson et al., 2017) for the LSTM+CNN+SA architecture.

2 Hyperparameter search

For EWC, we searched for the best $\lambda$ value among 100, 1000, 10000. For Rehearsal, we considered sampling size values of 100, 1000, 10000 training examples from Task A. We reported results for the models having the highest CL score computed according to the validation sets of both tasks. For EWC, the best model had $\lambda = 100$; for Rehearsal, the best model used 10000 training examples from Task A in both orders, WH→Y/N and Y/N→WH.

3 Continual Learning Evaluation Measures

Besides standard Accuracy (Acc), we consider metrics that have been introduced specifically to evaluate continual learning. In general, there is not much agreement among authors about the best metrics to evaluate continual learning models. Thus, Díaz-Rodríguez et al. (2018) propose a set of comprehensive metrics which allow to evaluate different factors of continual learning models, such as accuracy, forgetting, backward/forward knowledge transfer, memory overhead, and computational efficiency. In this paper, we focus on evaluating accuracy and forgetting across tasks. First, the authors define a measure describing the overall behavior of continual learning models. In particular, for each measure $i$ describing a particular aspect of a model, let $c_i$ (where $c_i \in [0, 1]$) be its average value and $s_i$ (where $s_i \in [0, 1]$) be its standard deviation across $r$ runs. Let $w_i \in [0, 1]$ (where $\sum_i w_i = 1$) be the weight given to measure $i$. Then, the CL score, which measures the overall score of the model across tasks, is defined. Higher values are better and the measure lies in the range $[0, 1]$. Formally, it is computed as follows:

$$CL \text{ score} = \sum_{i=1}^{[C]} w_i c_i$$

Let $R \in \mathbb{R}^{N \times N}$ be the train-test accuracy matrix, whose element $R_{i,j}$ is equal to the test accuracy on task $j$ after having trained the model up to task $i$, where $N$ is the number of tasks. In the evaluation of the CL score, we take the following measures into account:

- **Mean accuracy (Mean acc)** (Díaz-Rodríguez et al., 2018), which measures the overall accuracy of the model on the learned tasks. Higher values are better and the measure lies in the range $[0, 1]$. Formally, it is defined as:

$$\text{Mean acc} = \frac{\sum_{i \geq j} R_{i,j}}{N(N+1)/2}$$

- **Remembering (Rem)** (Díaz-Rodríguez et al., 2018), which measures how much the model remembers how to perform previously learned tasks. Higher values are better and the measure lies in the range $[0, 1]$. Formally, it is defined as:

$$\text{Rem} = 1 - \left| \min(BWT, 0) \right|,$$

where **Backward transfer (BWT)** allows to measure the influence that learning a task has...
on the performance of the previously learned tasks and it is formally defined as:

$$BWT = \frac{N^2}{N(N-1)} \sum_{i=2}^{N} (R_{i,j} - R_{j,i})$$

- **Intransigence (Int)** (Chaudhry et al., 2018), which captures how much a model is regularized towards preserving past knowledge and as a consequence less capable of learning new tasks. Lower values are better and the measure lies in the range $[-1, 1]$. Formally, intransigence on the $k$-th task is defined as:

$$I_k = a_{k,k}^* - a_{k,k},$$

where $a_{k,k}$ denotes the accuracy on task $k$ of the model trained sequentially up to task $k$ and $a_{k,k}^*$ denotes the accuracy on task $k$ of the Cumulative model trained on tasks $1, \ldots, k$. In this paper, we only measure intransigence for the second task, because we take only two tasks into account and it does not make sense to compute intransigence for the first task. Hence, $Int$ denotes $I_2$.

**4 Elastic Weight Consolidation**

Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) is a regularization approach which introduces plasticity in artificial neural networks by slowing down learning in weights which are important to solve previously learned tasks. The method takes inspiration from the human brain, in which the plasticity of synapses which are important to solve previously learned tasks is reduced. EWC adds a regularization term to the loss function allowing the model to converge to parameters where it has a low error for both tasks. In particular, if Task A and Task B have to be learned sequentially EWC, after having learned Task A, computes the Fisher Information Matrix, whose $i$-th diagonal element assesses how important parameter $i$ of the model is to solve Task A. Then, the model is trained on Task B starting from the parameters previously learned to solve Task A by minimizing the following loss function:

$$L = L_B(\theta) + \frac{\lambda}{2} \sum_i F_{i,i}(\theta_i - \theta_i^A)^2,$$

where $L_B$ is the loss function of Task B, $F_{i,i}$ is the $i$-th diagonal element of the Fisher Information Matrix, $\theta_i$ is the $i$-th parameter, $\theta_i^A$ is the optimal $i$-th parameter for Task A, and $\lambda$ controls the regularization strength, i.e. the higher it is, the more it is important to remember Task A.

**5 Confusion matrices**

Tables 1 to 5 show the confusion matrices of the Wh, Naive, Cumulative, Rehearsal, and EWC models, respectively, on the Wh $\rightarrow$ Y/N setup. Tables 1 and 7 to 10, instead, show the confusion matrices of the Y/N, Naive, Cumulative, Rehearsal, and EWC models, respectively, on the Y/N $\rightarrow$ Wh setup. In particular, predictions on these confusion matrices are grouped according to their category, so that rows represent the question type each question belongs to, columns represent the category each answer belongs to, and cells show the number of predictions the model obtains for a particular question type and answer category.

**6 Neuron activations**

Figures 1 and 2 show the neuron activations on the penultimate hidden layer of Naive model for the I) Wh $\rightarrow$ Y/N setup and the model trained independently on Y/N-q, respectively. All the visualizations of neuron activations reported in the paper are obtained by computing the vectors containing the neuron activations of the penultimate hidden layer of the model during forward propagation and by plotting the resulting vectors transformed into two dimensions through t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008).
<table>
<thead>
<tr>
<th>Wh</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6752</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6702</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6666</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6653</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>1204</td>
<td>14</td>
<td>2088</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>equal_shape</td>
<td>26</td>
<td>1150</td>
<td>2232</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>3430</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>equal_material</td>
<td>21</td>
<td>34</td>
<td>1440</td>
<td>2037</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix of the model trained independently on Wh-q.

<table>
<thead>
<tr>
<th>Naive</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6738</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>81</td>
<td>0</td>
<td>0</td>
<td>6621</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6666</td>
</tr>
<tr>
<td>query_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>148</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3410</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3430</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix of the Naive model on the WH → Y/N setup.

<table>
<thead>
<tr>
<th>Cumulative</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6752</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6702</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6665</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>query_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6653</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3410</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3430</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix of the Cumulative model on the WH → Y/N setup.

<table>
<thead>
<tr>
<th>Rehearsal</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6743</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6702</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6664</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>query_material</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>6651</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3410</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3430</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 4: Confusion matrix of the best Rehearsal model on the WH → Y/N setup.
Table 5: Confusion matrix of the best EWC model on the W→Y/N setup.

<table>
<thead>
<tr>
<th></th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6715</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>5479</td>
<td>0</td>
<td>0</td>
<td>1223</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6657</td>
</tr>
<tr>
<td>query_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1337</td>
<td>5316</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3408</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3429</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 6: Confusion matrix of the model trained independently on Y/N-q.

<table>
<thead>
<tr>
<th></th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6753</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6753</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6666</td>
</tr>
<tr>
<td>query_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6653</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3410</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3430</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 7: Confusion matrix of the Naive model on the Y/N → W→ setup.

<table>
<thead>
<tr>
<th></th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6753</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6701</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6666</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_material</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>6651</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>2732</td>
<td>38</td>
<td>229</td>
<td>310</td>
<td>0</td>
</tr>
<tr>
<td>equal_shape</td>
<td>1317</td>
<td>1144</td>
<td>346</td>
<td>603</td>
<td>0</td>
</tr>
<tr>
<td>equal_size</td>
<td>1330</td>
<td>16</td>
<td>1559</td>
<td>525</td>
<td>0</td>
</tr>
<tr>
<td>equal_material</td>
<td>1297</td>
<td>0</td>
<td>30</td>
<td>2205</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 8: Confusion matrix of the Cumulative model on the Y/N → W→ setup.
<table>
<thead>
<tr>
<th>Rehearsal</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6752</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6702</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6666</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_material</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>6651</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3309</td>
</tr>
<tr>
<td>equal_shape</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3409</td>
</tr>
<tr>
<td>equal_size</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3429</td>
</tr>
<tr>
<td>equal_material</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3532</td>
</tr>
</tbody>
</table>

Table 9: Confusion matrix of the best Rehearsal model on the Y/N → WH setup.

<table>
<thead>
<tr>
<th>EWC</th>
<th>query_color</th>
<th>query_shape</th>
<th>query_size</th>
<th>query_material</th>
<th>Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>query_color</td>
<td>6748</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_shape</td>
<td>0</td>
<td>6701</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_size</td>
<td>0</td>
<td>0</td>
<td>6666</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>query_material</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6652</td>
<td>0</td>
</tr>
<tr>
<td>equal_color</td>
<td>3110</td>
<td>9</td>
<td>17</td>
<td>173</td>
<td>0</td>
</tr>
<tr>
<td>equal_shape</td>
<td>801</td>
<td>1214</td>
<td>69</td>
<td>1326</td>
<td>0</td>
</tr>
<tr>
<td>equal_size</td>
<td>542</td>
<td>35</td>
<td>35</td>
<td>1674</td>
<td>2</td>
</tr>
<tr>
<td>equal_material</td>
<td>464</td>
<td>2</td>
<td>1</td>
<td>3065</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 10: Confusion matrix of the best EWC model on the Y/N → WH setup.
Figure 1: Analysis of the neuron activations on the penultimate hidden layer of the Naive model for the $\text{WH} \rightarrow \text{Y/N}$ setup.

Figure 2: Analysis of the neuron activations on the penultimate hidden layer of the model trained independently on $\text{Y/N-q}$. 
References


Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni. 2018. Don’t forget, there is more than forgetting: new metrics for continual learning. In Workshop on Continual Learning, NeurIPS.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Inferring and executing programs for visual reasoning. In ICCV.

