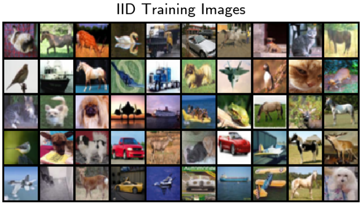
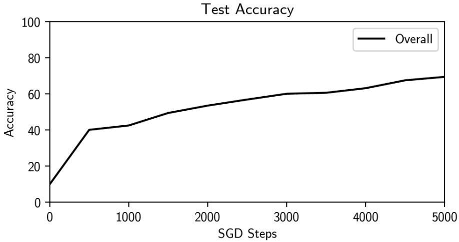
Graph-Based Continual Learning

Binh Tang & David S. Matteson (Cornell University)

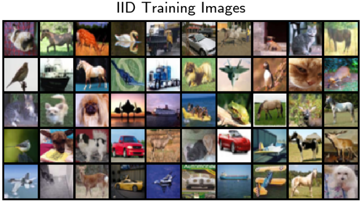
ICLR | May 2021

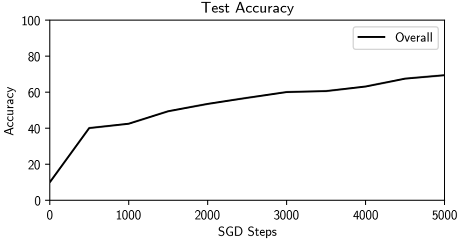
Catastrophic Forgetting

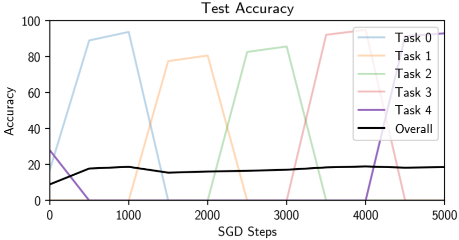




Catastrophic Forgetting





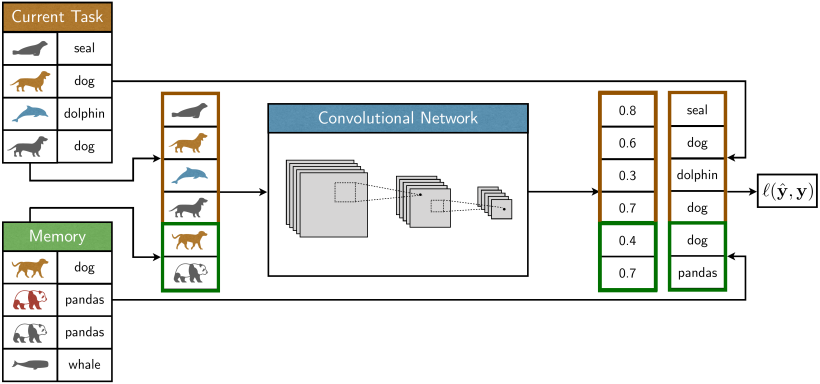


Continual Learning

- Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i, t_i)\}_{i=1}^{N}$ where the task IDs $t_i \in \mathbb{N}$ are not i.i.d., but the input-output pairs $(\mathbf{x}_i, \mathbf{y}_i)$ are conditionally i.i.d.
- We consider continuous, **online streams** of tasks in which samples from different tasks arrive at different times.
- Our goal is to learn supervised classification models that are less prone to catastrophic forgetting, performing well with or without task IDs while requiring a small memory footprint.

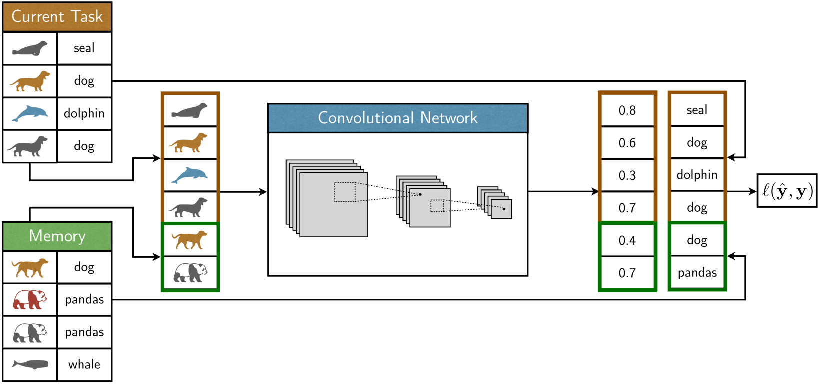
Motivations

- Rehearsal approaches store and replay samples in an episodic memory.



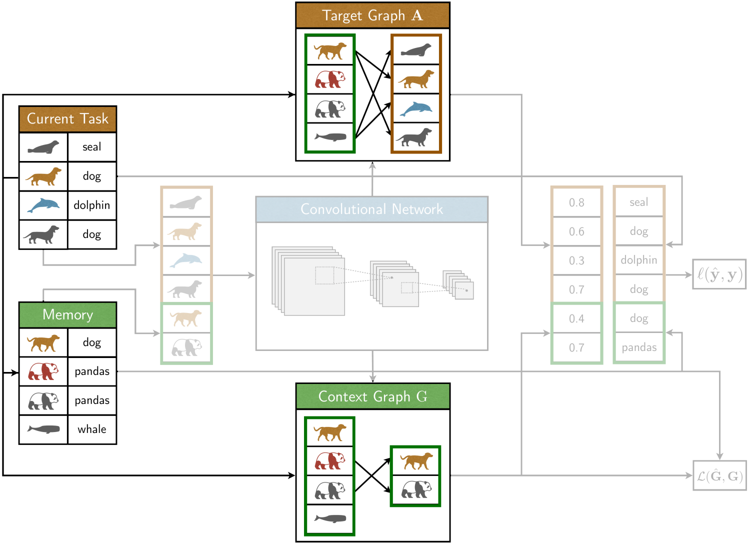
Motivations

- Rehearsal approaches store and replay samples in an episodic memory.
- Existing methods fail to utilize relational structures between samples, while relational memory is a prominent feature of biological systems.



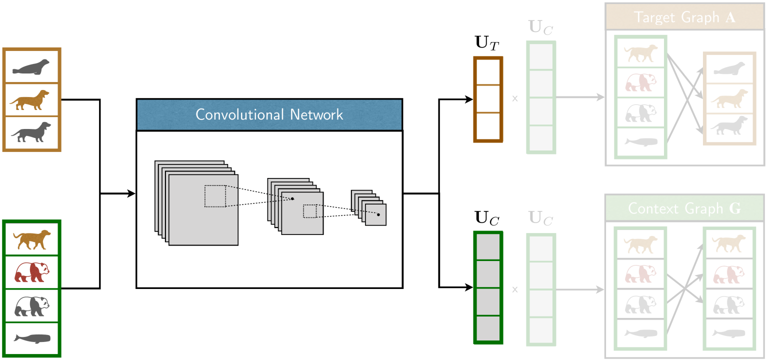
Main Ideas

- Our model encodes sample similarities via edges in random graphs.



Graph Construction

1. We use a CNN to embed context and target images into $\mathbf{U}_{\mathcal{C}} = \{\mathbf{u}_i\}_{i \in \mathcal{C}}$ and $\mathbf{U}_{\mathcal{T}} = \{\mathbf{u}_j\}_{j \in \mathcal{T}}$, respectively.



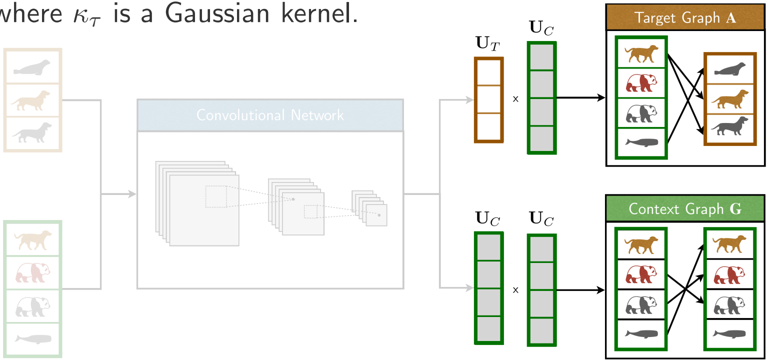
Graph Construction

2. We build a context graph $\mathbf{G}$ and a context-target graph $\mathbf{A}$. The edges are represented by independent Bernoulli random variables:

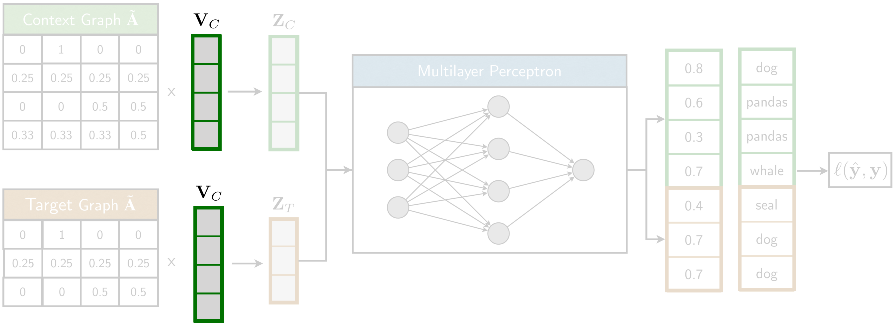
$$p(\mathbf{G} \mid \mathbf{U}_{\mathcal{C}}) = \prod_{i \in \mathcal{C}} \prod_{k \in \mathcal{C}} \mathrm{Ber}(\mathbf{G}_{ik} \mid \kappa_{\tau}(\mathbf{u}_i, \mathbf{u}_k)), \tag{1}$$

$$p(\mathbf{A} \mid \mathbf{U}_{\mathcal{T}}, \mathbf{U}_{\mathcal{C}}) = \prod_{j \in \mathcal{T}} \prod_{k \in \mathcal{C}} \mathrm{Ber}(\mathbf{A}_{jk} \mid \kappa_{\tau}(\mathbf{u}_j, \mathbf{u}_k)), \tag{2}$$

where κ_{τ} is a Gaussian kernel.

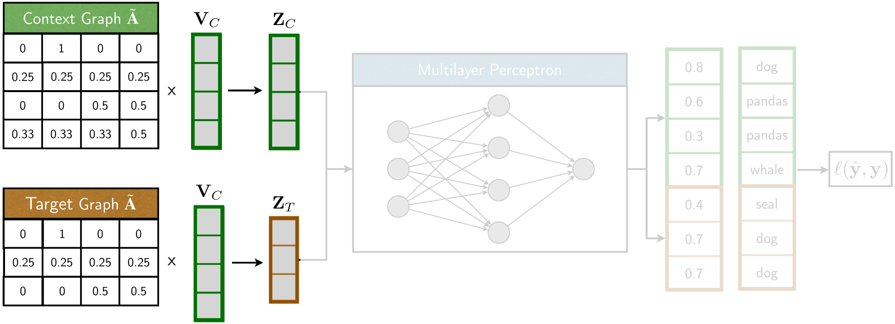


1. We use another CNN with tied weights to encode context images and context labels together into $\mathbf{V}_{\mathcal{C}} = \{\mathbf{v}_i\}_{i \in \mathcal{C}}$.



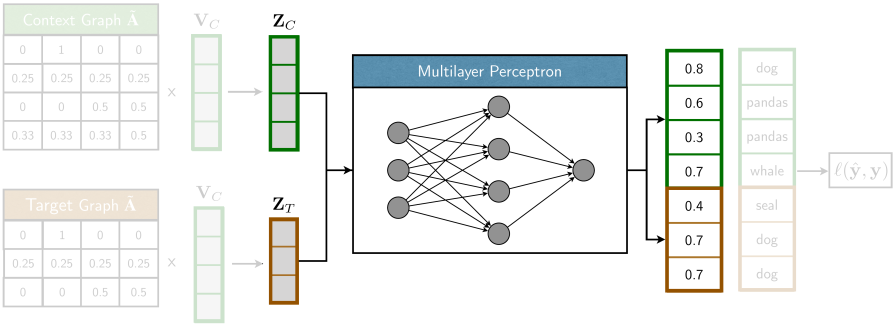
Predictive Distribution

2. Given normalized graphs $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{A}}$, we compute a context-aware representations by aggregating information from similar images.



Predictive Distribution

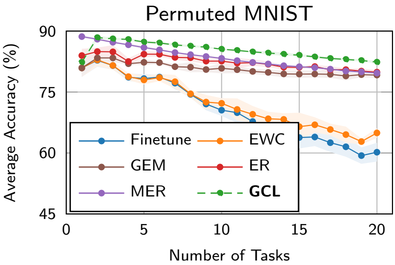
3. An MLP makes probabilistic predictions for context and target images.

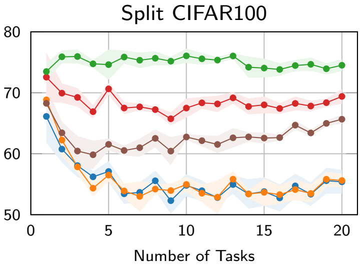


Graph Regularization

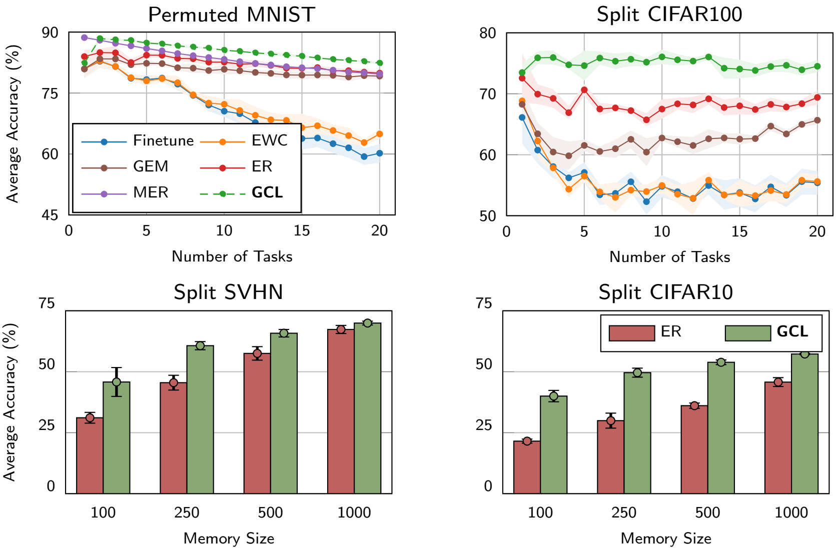
- We gradually grow the graphs $\mathbf{G}$ and $\mathbf{A}$ as new tasks arrive and save the distribution of $\mathbf{G}$ and context images to the episodic memory.
- The graph $\mathbf{G}$ potentially captures a meaningful relational structure, but replaying the context images alone ignores $\mathbf{G}$'s learned edges.
- We add a regularization term to penalize deviations from learned edges in the context graph $\mathbf{G}$.

Classification Results

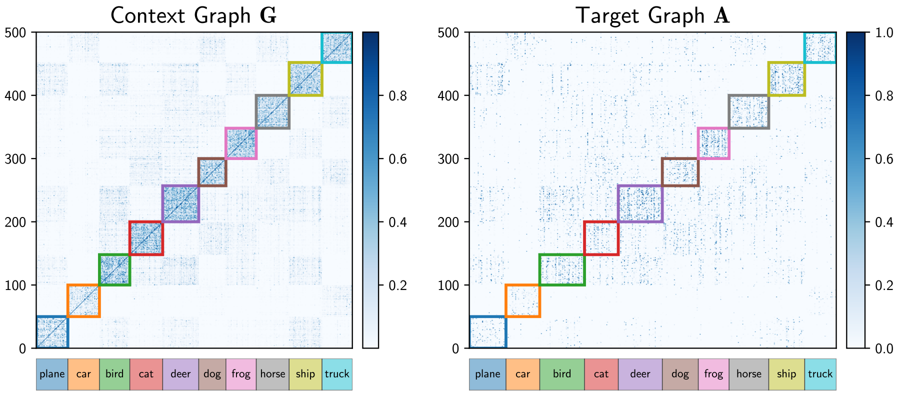




Classification Results



Learned Graphs



Conclusion & Future Work

- We introduce a graph-based approach to continual learning that exploits pairwise similarities between samples.
- Our model demonstrates an efficient use of the episodic memory and performs competitively under various settings.
- Future work include extensions to other domains (e.g image generation) and related problems (e.g. meta learning).