VariGrow: Variational Architecture Growing for Task-Agnostic Continual Learning based on Bayesian Novelty

Randy Ardywibowo[‡], Zepeng Huo[‡],
Zhangyang Wang[◇], [‡]Bobak Mortazavi, [§]Shuai Huang, [‡]Xiaoning Qian

[‡]Texas A&M University, [◇]University of Texas at Austin,
[§]University of Washington

DARPA

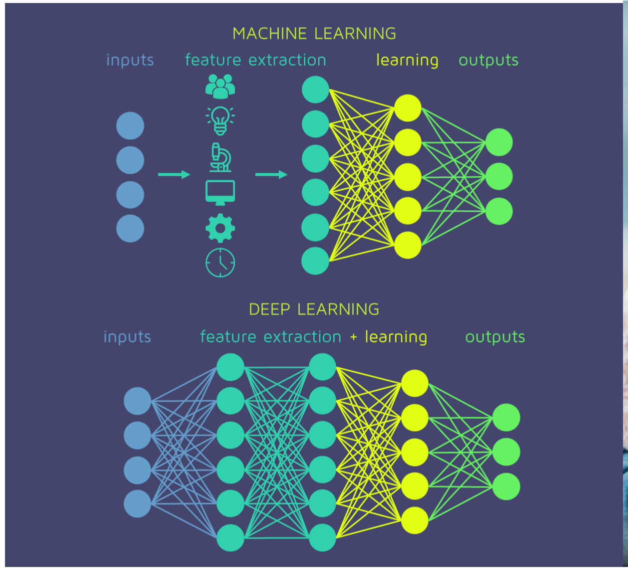
UNIVERSITY of WASHINGTON

TEXAS The University of Texas at Austin

TEXAS A&M UNIVERSITY Engineering

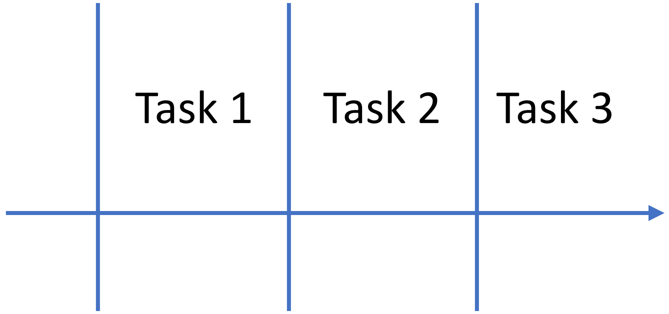
Background

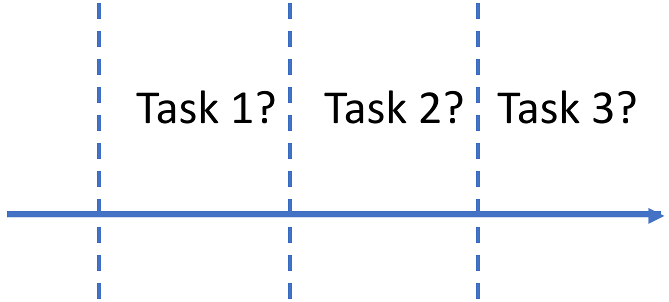
- Machine Learning (ML) often assume data to be identically and independently distributed (*iid*).
- ML agents may encounter new contexts throughout its use.
- Tend to catastrophically forget previously learned knowledge.
- Humans learn from non-*iid* data streams in widely varying contexts.



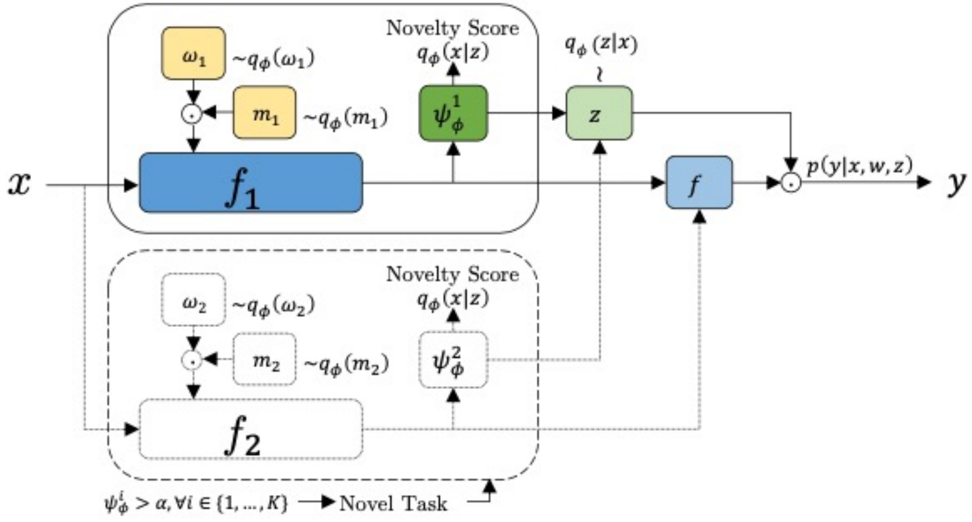
Background

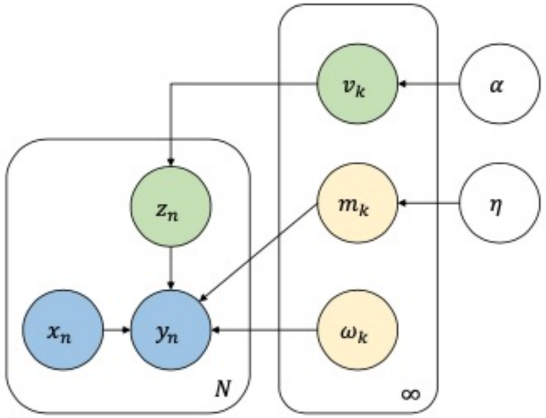
- Continual Learning aims to solve this.
- Many methods assume that the data is explicitly divided into *tasks* known both during training and testing.
- There are often no clear transition boundaries between different contexts.





VariGrow: Variational Architecture Growing





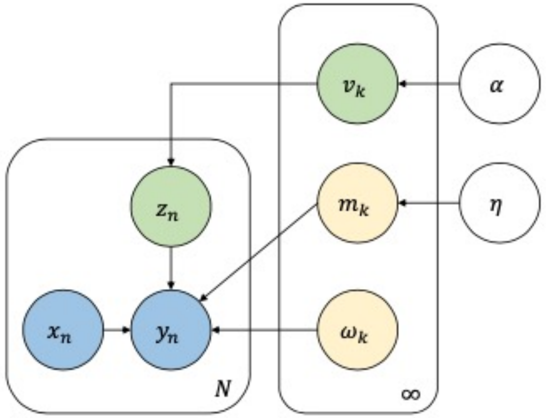
- Formulate a Bayesian nonparametric formulation to growing neural networks.
- As new contexts are detected, new experts are created, alleviating the catastrophic forgetting problem.

Methodology: Variational Inference

- The posterior and variational distribution is decomposed into different clusters representing different contexts.

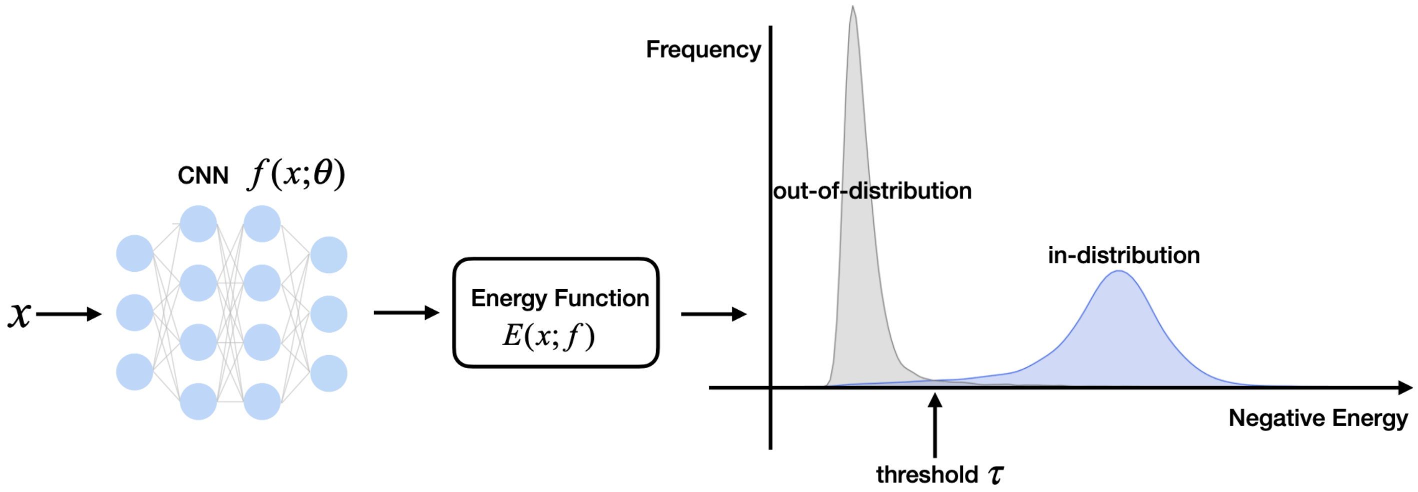
$$p(\boldsymbol{w}, \boldsymbol{z}|\mathcal{D}_{1:t}) = \prod_{i=1}^{t} p(\boldsymbol{w}_{\boldsymbol{z}_i}|\mathcal{D}_i) p(\boldsymbol{z}|\mathcal{D}_{1:t}).$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{t} q_{\boldsymbol{\phi}}(\boldsymbol{w}_{\boldsymbol{z}_i}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$$



- $q_\phi(z|x)$ and $q_\phi(w_z)$ need to be expressive and flexible.
- We define this nonparametrically, by maintaining a growing mixture $q_\phi(z|x)$ of experts $q_\phi(w_z)$.

The Mixing Distribution $q_\phi(z|x)$



$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \frac{q_\phi(\boldsymbol{x}|\boldsymbol{z})q_\phi(\boldsymbol{z})}{\sum_{i=1}^{\infty} q_\phi(\boldsymbol{x}|\boldsymbol{z}=i)q_\phi(\boldsymbol{z}=i)}.$$

- Instead of using generative models to estimate $q_\phi(x|z)$,
- We define the mixing distribution using an energy-based method.
- This allows us to detect novel contexts without using task labels.

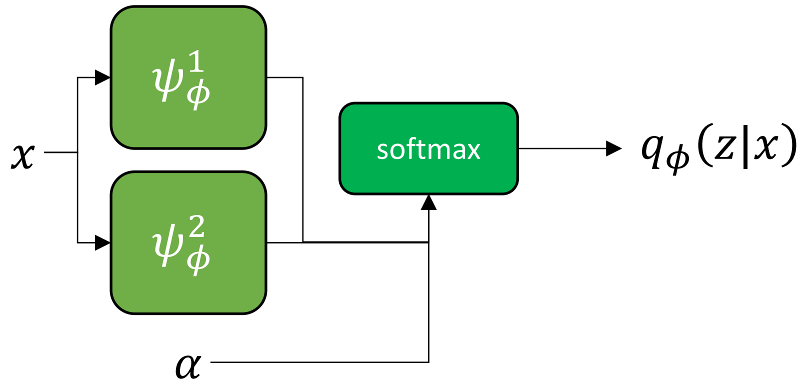
The Mixing Distribution $q_\phi(z|x)$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \frac{q_\phi(\boldsymbol{x}|\boldsymbol{z})q_\phi(\boldsymbol{z})}{\sum_{i=1}^{\infty} q_\phi(\boldsymbol{x}|\boldsymbol{z}=i)q_\phi(\boldsymbol{z}=i)}.$$

- The mixing distribution $q_\phi(z|x)$ is defined as

$$q_\phi(z=k|x) = \mathrm{Softmax}\left(\psi_\phi^k; \psi_\phi^{-k}, \alpha\right) \qquad k \le K$$

$$q_\phi(z=k|x) = \frac{1}{2^{k-K}}\mathrm{Softmax}\left(\alpha; \psi_\phi\right) \qquad k > K$$



- With the *Helmholtz free energy* ψ_ϕ^k defined by the log-posterior predictive distribution:

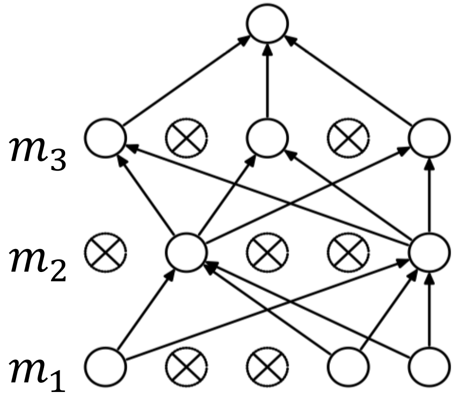
$$\psi_\phi^k(\boldsymbol{x}) = -T\log\left(\sum_{c=1}^{C}\exp(\ell_\phi^k(\boldsymbol{x},c)/T)\right) \qquad \ell_\phi^k(\boldsymbol{x},c) = \mathbb{E}_{q_\phi(\boldsymbol{w}|\boldsymbol{z}=k)}[\log p(y=c|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z}=k)]$$

The Expert Distribution $q_\phi(w|z)$

- Define a sparsifying prior for the weights of our experts to keep reasonable memory usage:

$$p(m) = \texttt{Bern}(e^{-\eta})$$

$$\mathrm{KL}(q_{\boldsymbol{\phi}}(m_k)||p(m_k)) \approx \lambda q_{\boldsymbol{\phi}}(m_k = 1) = \lambda\sigma(\phi_k)$$

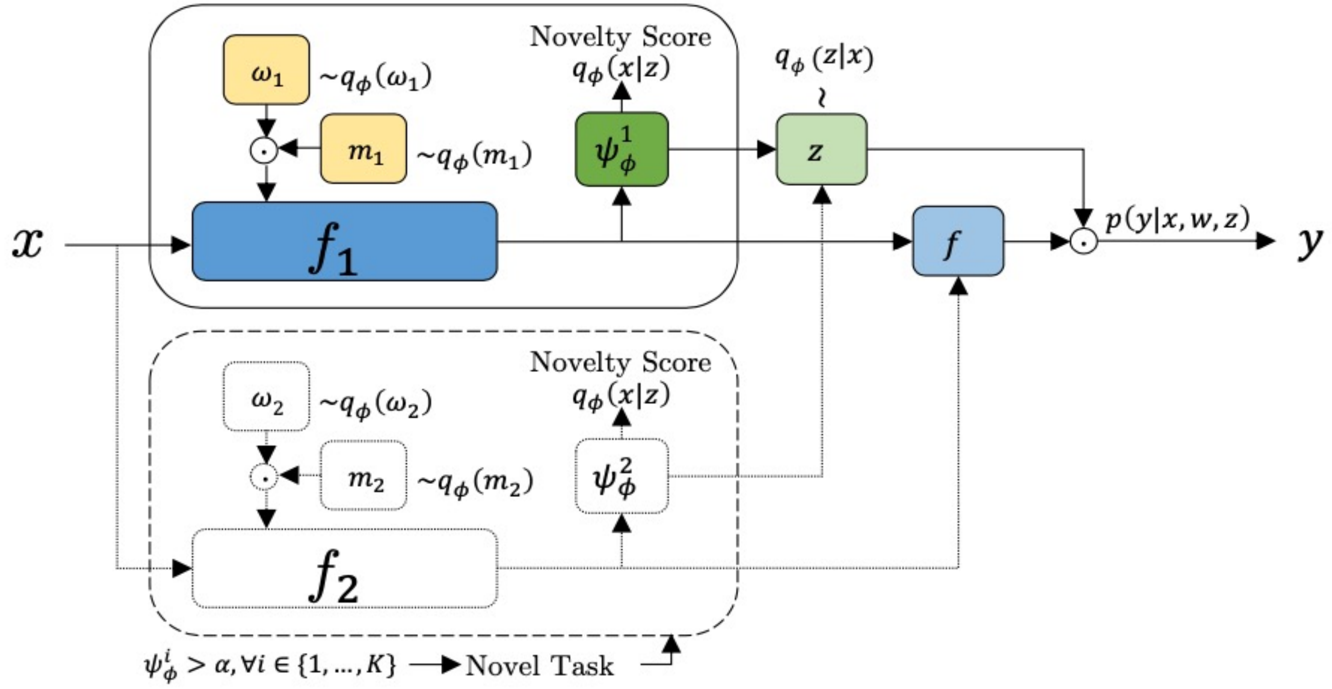


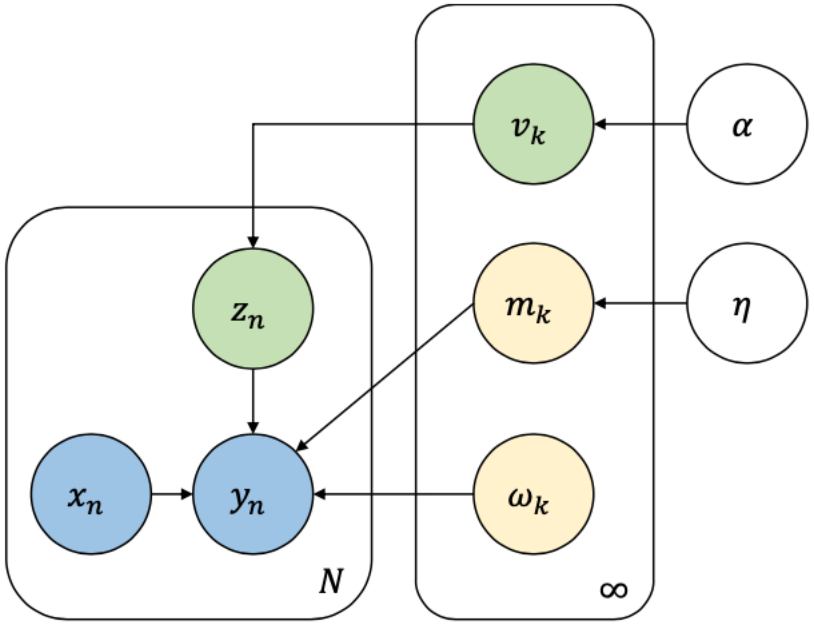
- Define the variational distribution implicitly:

$$\epsilon \sim p(\epsilon), \quad m = \zeta(\phi, \epsilon) \quad \Rightarrow \quad m \sim q_{\boldsymbol{\phi}}(m)$$

$$\epsilon \sim Logistic(0{,}1) \qquad \zeta(\phi, \epsilon) \;=\; \mathbb{1}\left[\log\left(\frac{\sigma(\phi)}{1-\sigma(\phi)}\right) + \epsilon > 0\right]$$

Methodology: Overview





- The result is an architecture that grows when novel contexts arrive and is sparse according to a prior distribution.

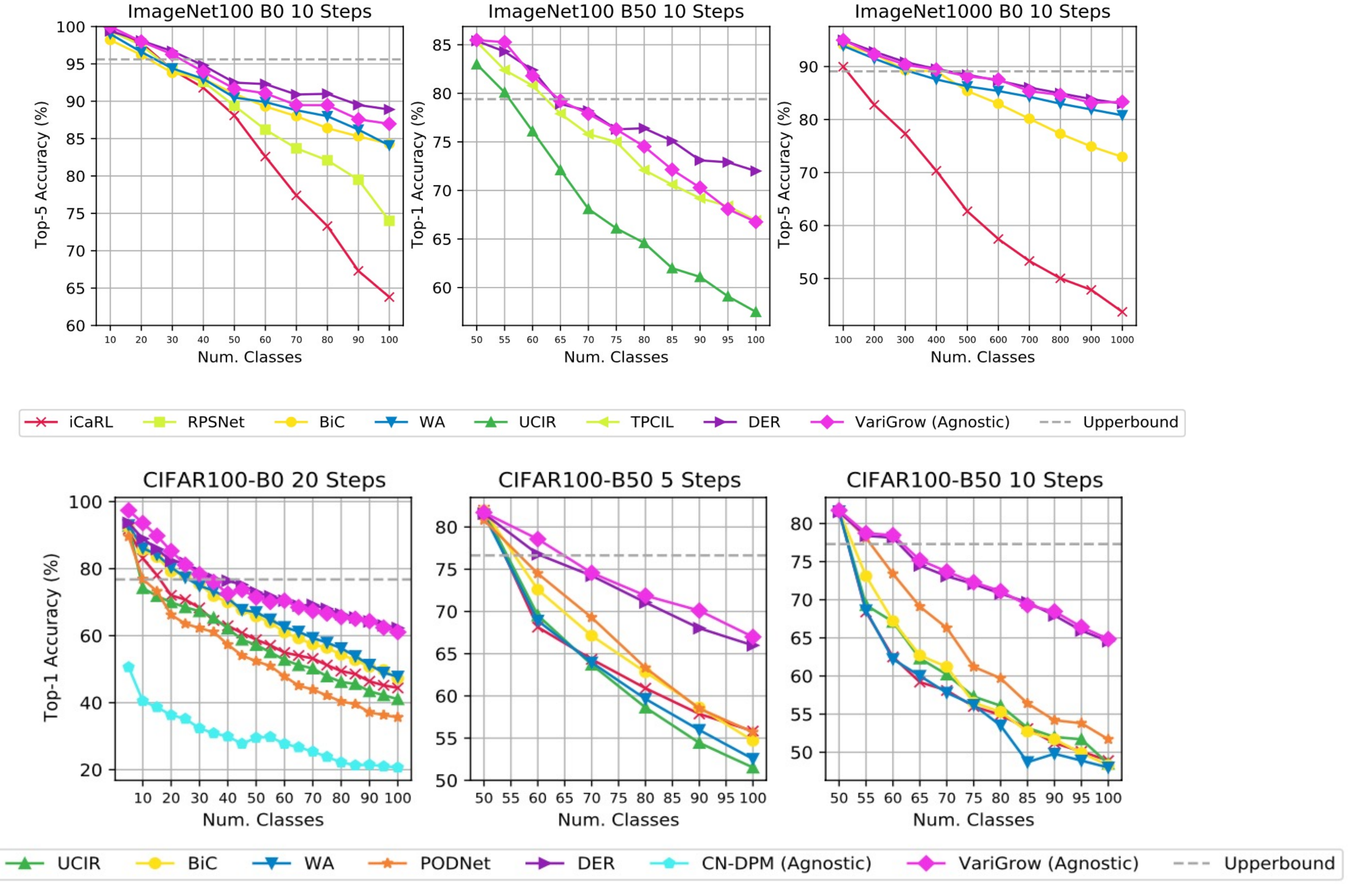
Results

- We perform various experiments on continual learning on CIFAR-100 and ImageNet.
- Our method achieves competitive accuracy even against methods that are explicitly given the task labels.

Method	5 Steps		10 Steps		20 Steps		50 Steps	
	Params.	Acc. (%)	Params.	Acc. (%)	Params.	Acc. (%)	Params.	Acc. (%)
Bound	11.2	80.40	11.2	80.41	11.2	81.49	11.2	81.74
iCaRL (Rebuffi et al., 2017)	11.2	71.14	11.2	65.27	11.2	61.20	11.2	56.08
UCIR (Hou et al., 2019)	11.2	62.77	11.2	58.66	11.2	58.17	11.2	56.86
BiC (Hou et al., 2019)	11.2	73.10	11.2	68.80	11.2	66.48	11.2	62.09
WA (Zhao et al., 2020)	11.2	72.81	11.2	69.46	11.2	67.33	11.2	64.32
PODNet (Douillard et al., 2020)	11.2	66.70	11.2	58.03	11.2	53.97	11.2	51.19
AANets (Liu et al., 2021)	11.2	67.59	11.2	65.66	-	-	-	-
RPSNet (Rajasegaran et al., 2019)	60.6	70.50	56.5	68.60	-	-	-	-
DER (Yan et al., 2021a)	2.89	75.55	4.96	74.64	7.21	73.98	10.15	72.05
CN-DPM (Lee et al., 2020) (Agnostic)	19.2	20.34	19.2	17.60	19.2	18.79	19.2	19.70
VariGrow (Agnostic)	2.97	75.50	4.88	75.04	7.30	74.03	10.25	72.21

Results

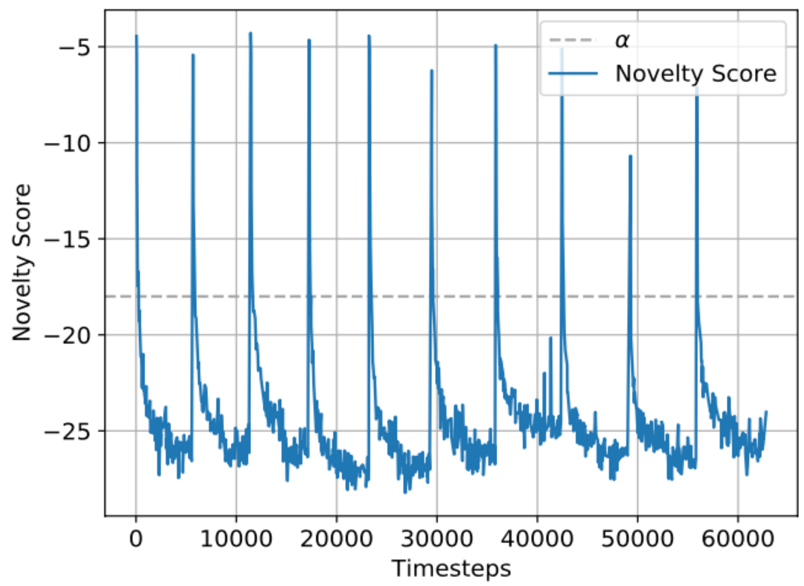
- Our method can retain knowledge of previous tasks as shown by the accuracy it retains.



Results

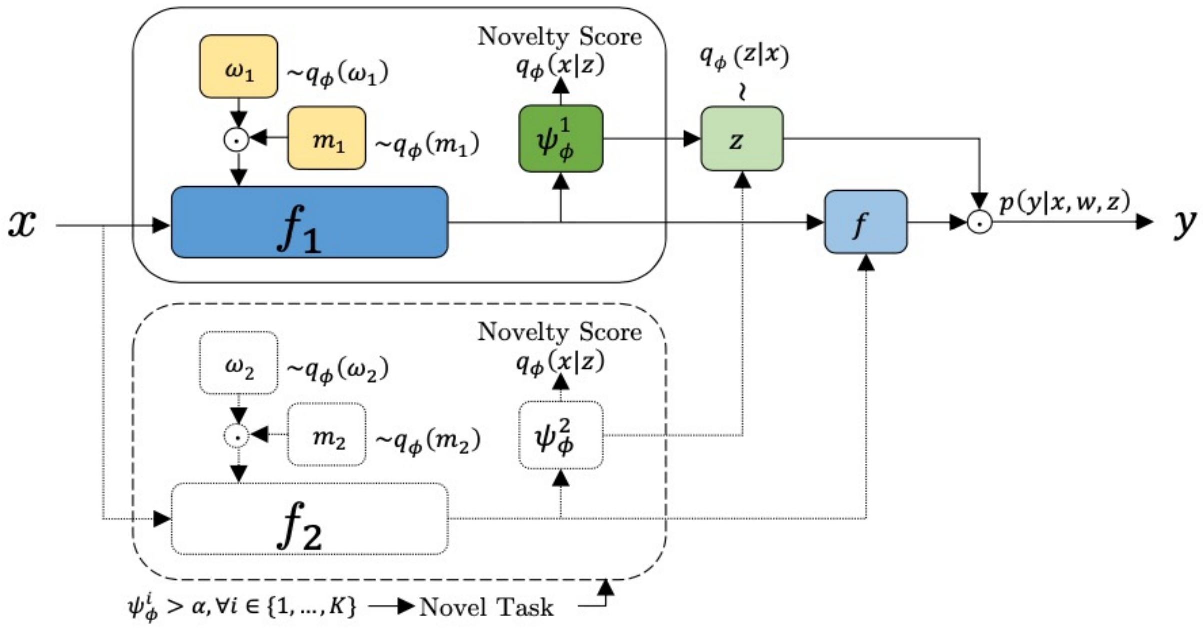
- Our model can perform well in scenarios where the contexts switching is not given
- Observing the novelty scores of our model, we can see that it can indeed detect the context switching.

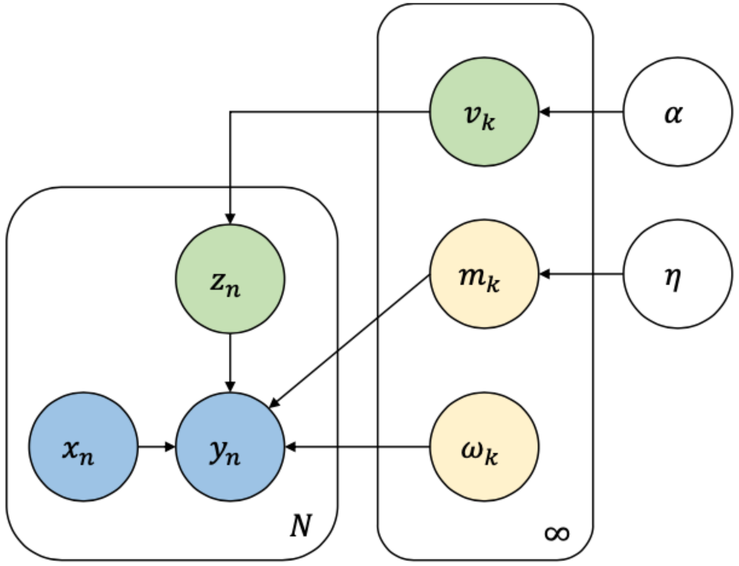
Setting	Accuracy (%)	
	5 Steps	10 Steps
Baseline	73.97	72.45
Lookback Old Tasks	71.21	70.98
Fuzzy Boundaries	70.03	69.19



Conclusion

- We developed VariGrow, a variational continual learning method that:
- Automatically detect novel contexts.
- Learn novel concepts while retaining previously knowledge.





Thank You