Catastrophic Forgetting in Deep Learning

Julio Hurtado
CPD7000

Table of Contents

What is Machine Learning?

Artificial Intelligence

Learn from data

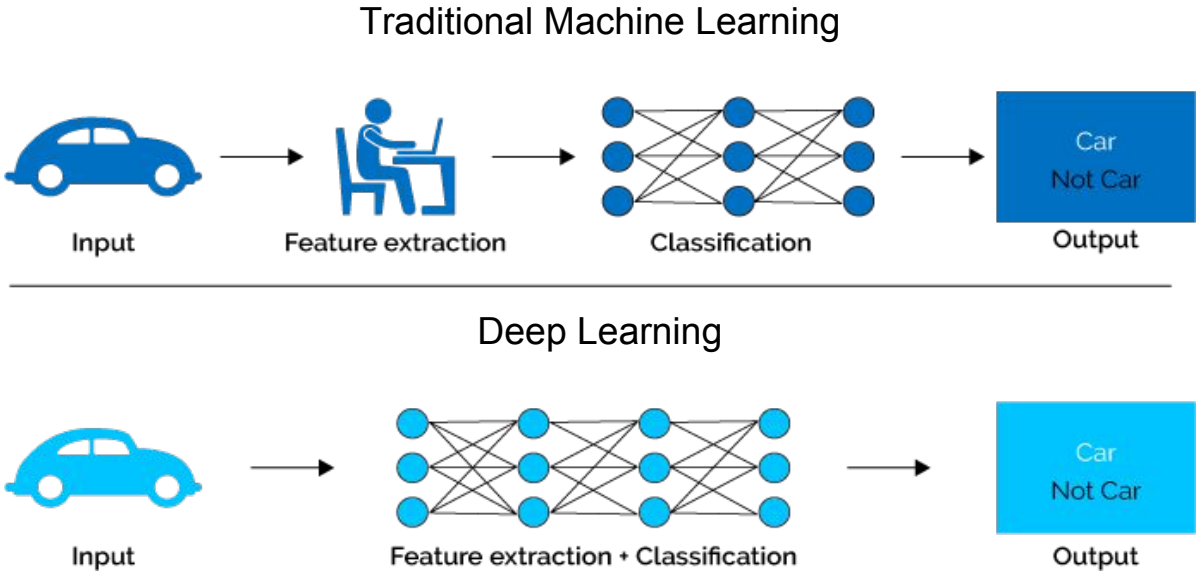
Big Data

Find Patterns in data

Unstructured data

Supervised Learning

Deep Learning



Deep Learning

Advantages

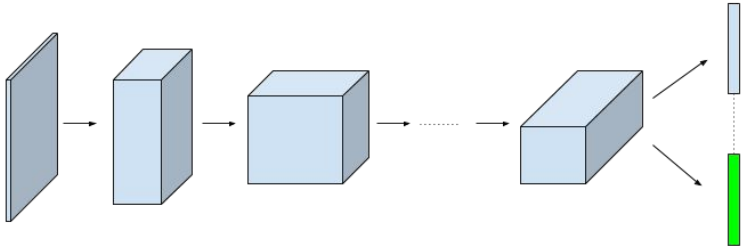
- Significantly outperforms other solutions in multiple domains
- Reduces the need for feature engineering
- Is an architecture that can be adapted to new problems relatively easily

Disadvantages

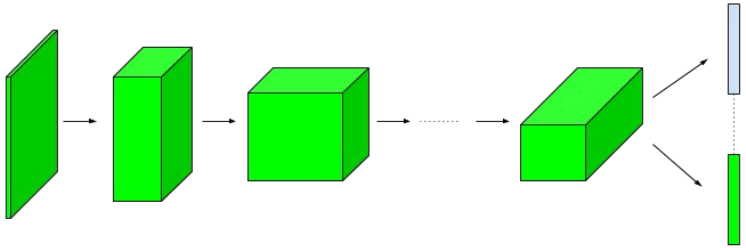
- Is extremely computationally expensive to train
- Required a large amount of data
- Determining the topology/flavor/training method/hyperparameters for deep learning is a black art with no theory to guide you

Transfer Learning

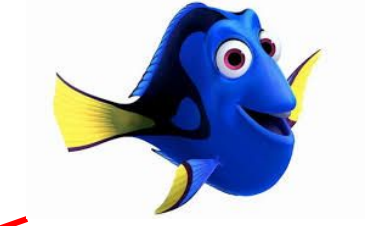
Feature extraction



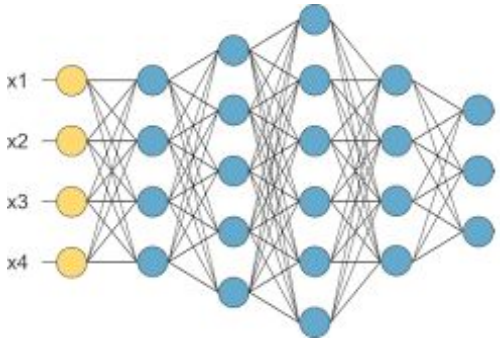
Fine-tuning

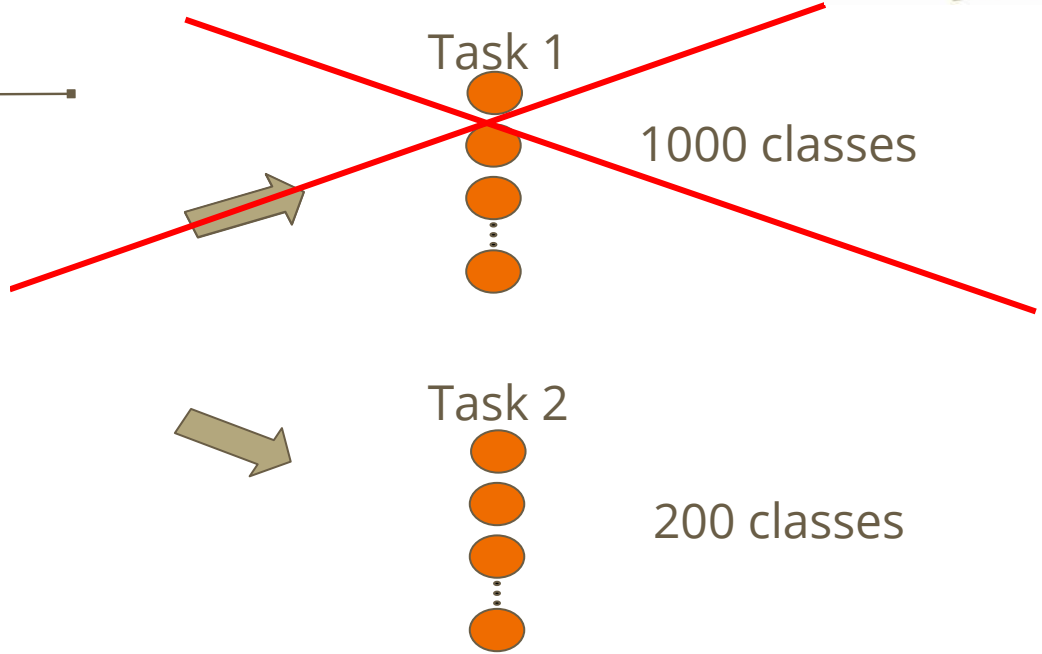


Catastrophic Forgetting



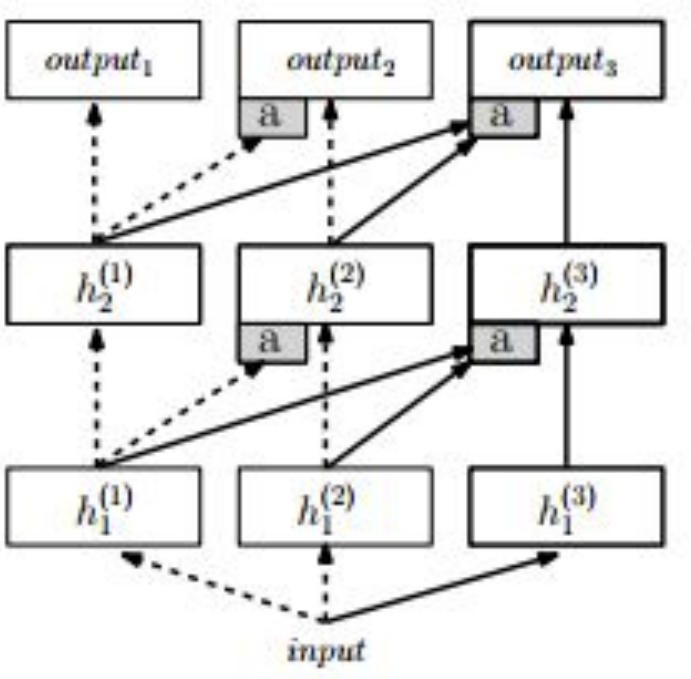
Feature Extraction





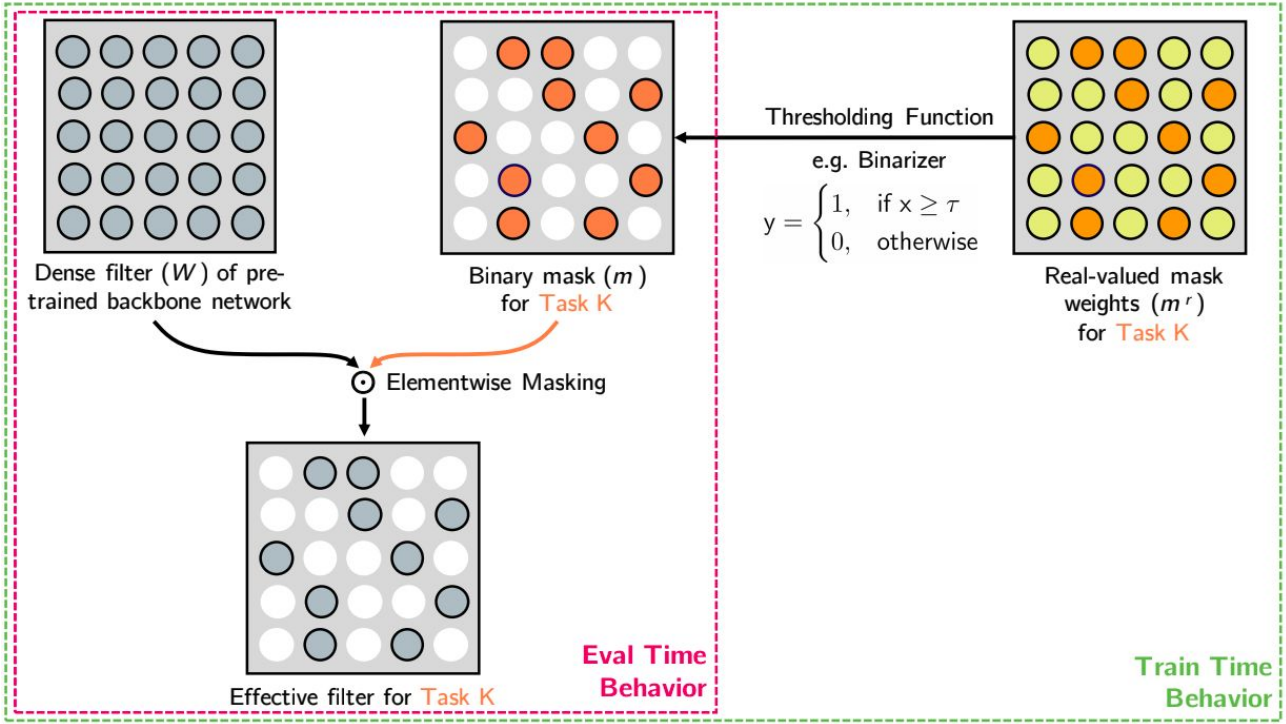
State of the Art

Progressive Neural Networks

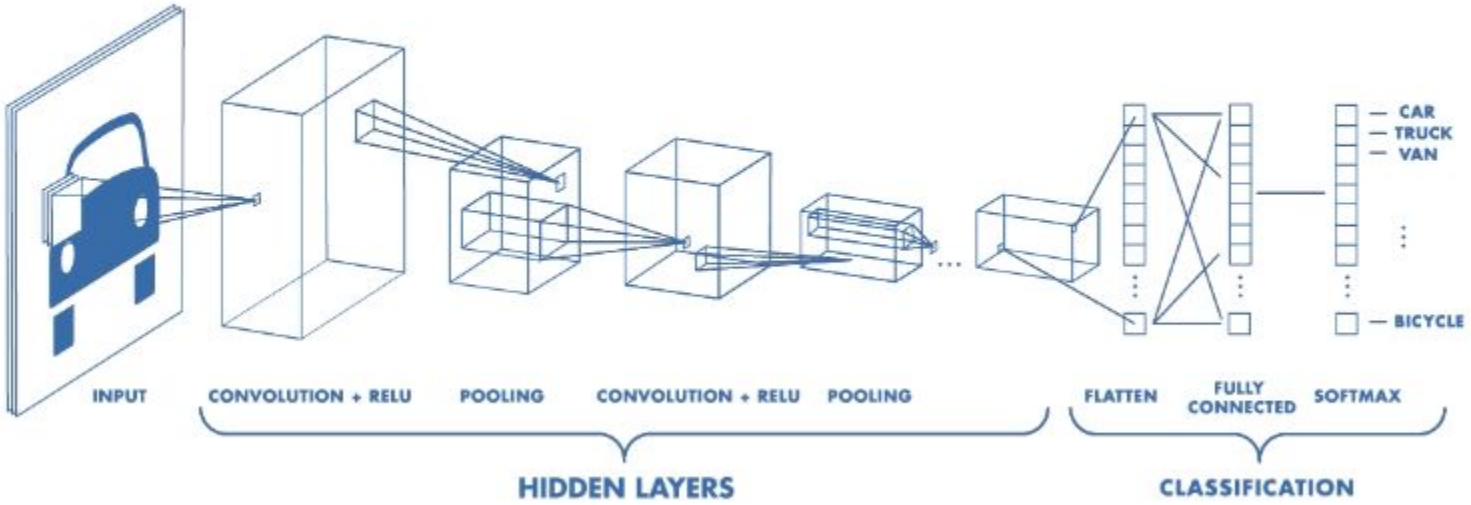


State of the Art

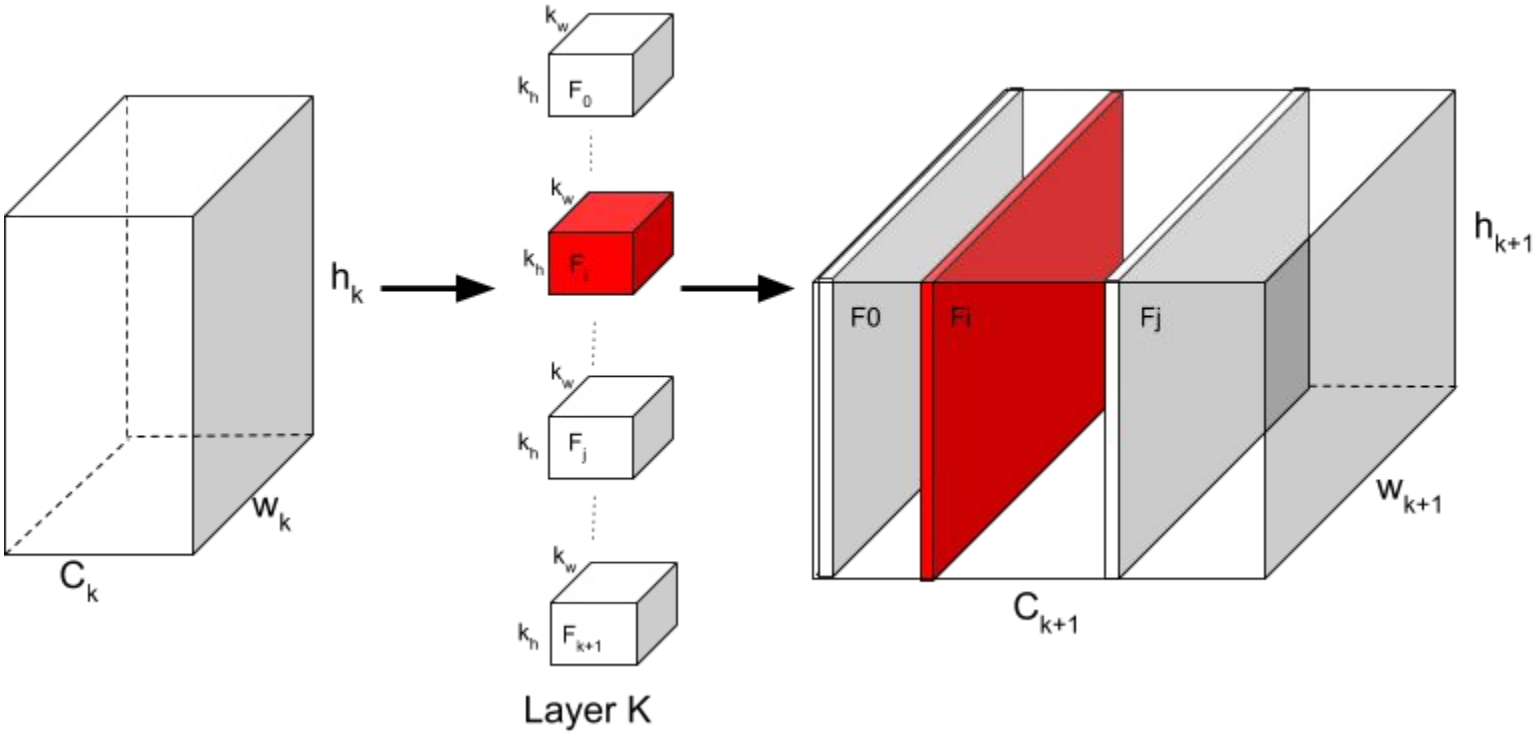
PiggyBack



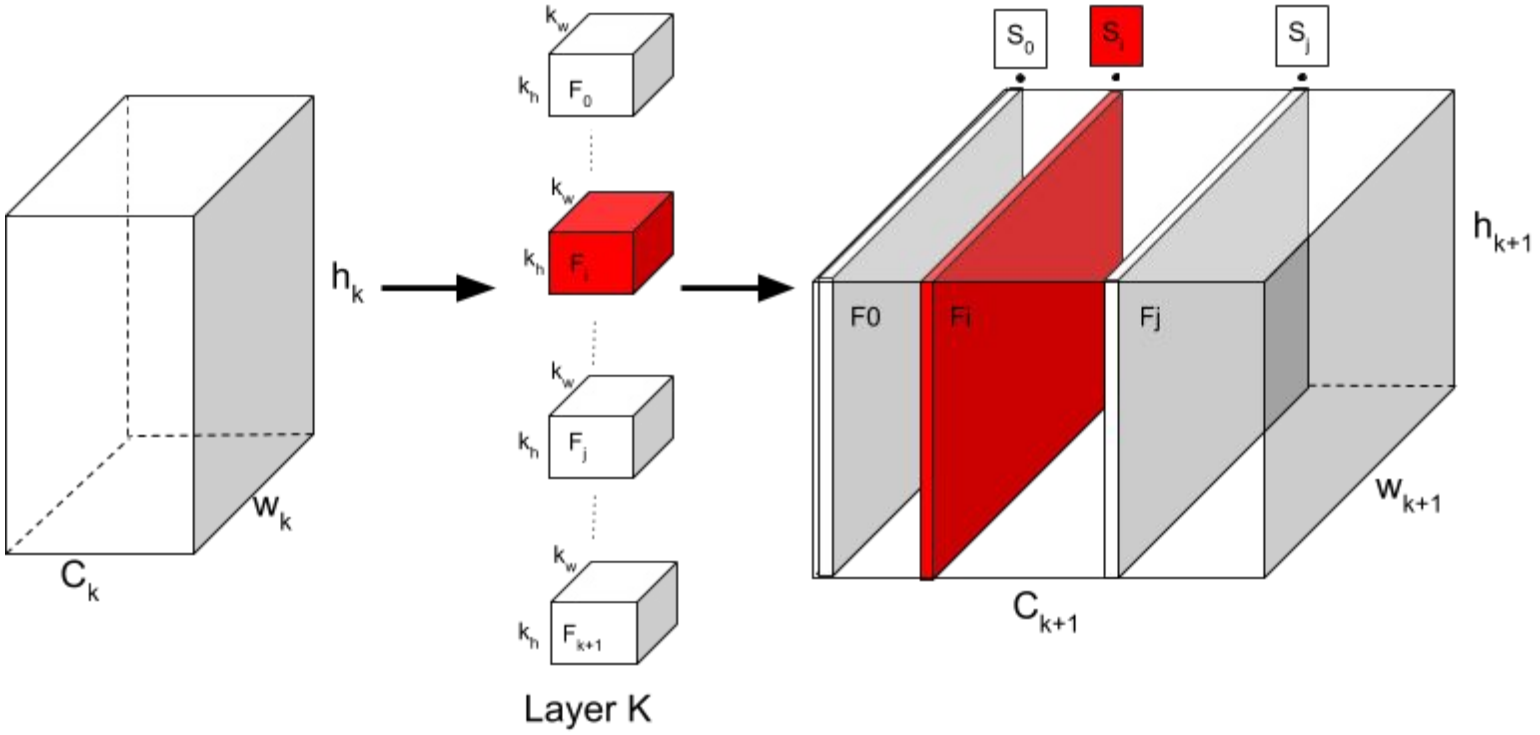
How does a CNN works



Our Solution



Our Solution

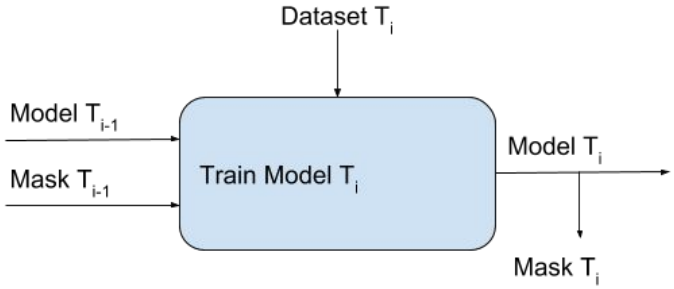


Our Solution

Mask T_{-1} initialize con 1s

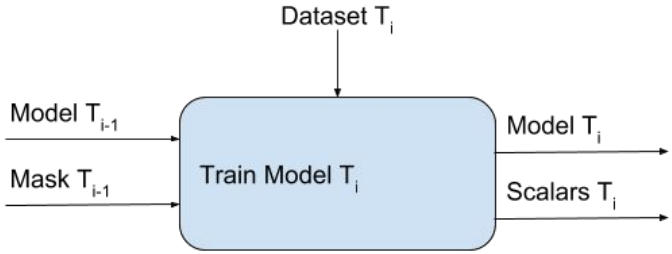
Group-sparse regularization

$$\Gamma_l(\Theta^l) = (1-\beta^l)\frac{1}{2}\left\|\hat{\Theta}^l\right\|_F^2 + \frac{\beta^l}{K^l}\sum_{k=1}^{K^l}\delta\left(\sum_{t=1}^{K^{l-1}}\left\|\hat{\Theta}^l_{k,t,\cdot}\right\|_\cdot > 0\right)$$



Mask Learning

$$\Gamma_l(\Theta^l, S^l) = (1-\beta^l)\frac{1}{2}\left\|\hat{\Theta}^l\right\|_F^2 + \beta^l\sum_{k=1}^{K^l}|S^l_k|$$



Mask T_i =((Scalars T_0 + … + Scalars T_i) > τ)

Meta Learning

- Are the filters learned in previous task useful for new tasks?
  - In a similar or different context
- Learning to learn
  - To learn more generic filters
  - Without affecting the accuracy of the model
- Are filters learned in new tasks useful for previous tasks?
  - Assuming that we have the data from the previous tasks

Future Works

- Continue doing experiment
  - How many filter use each task?
  - How many task we can train with the same model?
- Change the way we find the none binary Mask
- Find the way to add a meta learning model to our model
  - How do we know if the model is generalizing well

References

- Yosinski, Jason, et al. "How transferable are features in deep neural networks?." Advances in neural information processing systems. 2014.
- Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
- Rusu, Andrei A., et al. "Progressive neural networks." arXiv preprint arXiv:1606.04671 (2016).
- Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." Proceedings of the National Academy of Sciences 114.13 (2017): 3521-3526.
- Wang, Yu-Xiong, Deva Ramanan, and Martial Hebert. "Growing a brain: Fine-tuning by increasing model capacity." IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
- Mallya, Arun, Dillon Davis, and Svetlana Lazebnik. "Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Información de Contacto

- Email: jahurtado [at] uc.cl
- Twitter: @JuliousHurtado

Catastrophic Forgetting in Deep Learning

Julio Hurtado
CPD7000