Advanced Machine Learning Hw 3

Graph Metanetworks and Unlearning

Alessio Palma and Leonardo Plini
Sapienza University of Rome
palma@di.uniroma1.it leonardo.plini@uniroma1.it

Who we are?

PhD student in Data Science at Sapienza University of Rome

PhD student in Artificial Intelligence at Sapienza University of Rome and National Institute of Nuclear Physics

Alessio Palma and Leonardo Plini
Sapienza University of Rome
alessio.palma@uniroma1.it leonardo.plini@uniroma1.it

PINlab
PERCEPTION AND INTELLIGENCE

SAPIENZA
UNIVERSITÀ DI ROMA

PinLab

Faculty

Fabio Galasso

Head of the lab & Associate Professor

Indro Spinelli

Assistant Professor

Senior Collaborators

Guido Maria D'Amely di Melendugno, Ph.D.

Post-Doc Research Associate

Alessandro Flaborea, Ph.D.

Start-up'er at ItalAI

Luca Franco, Ph.D.

Manager at ItalAI

PINlab PERCEPTION AND INTELLIGENCE

SAPIENZA UNIVERSITÀ DI ROMA

Students

Giovanni Ficarra
Ph.D. Student

Laura Laurenti
Ph.D. Student

Paolo Mandica
Ph.D. Student

Alessio Sampieri
Ph.D. Student

Luca Scofano
Ph.D. Student

Luca Collorone
Ph.D. Student

Stefano D'Arrigo
Ph.D. Student

Edoardo De Matteis
Ph.D. Student

Massimiliano Pappa
Ph.D. Student

Daniele Trappolini
Ph.D. Student

Matteo Gioia
Ph.D. Student

Leonardo Plini
Ph.D. Student

Valentino Sacco
Ph.D. Student

Aurora Bassani
Ph.D. Student

Simone Facchiano
Ph.D. Student

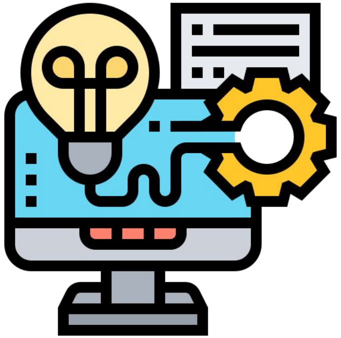
Matteo Migliarini
Ph.D. Student

Alessio Palma
Ph.D. Student

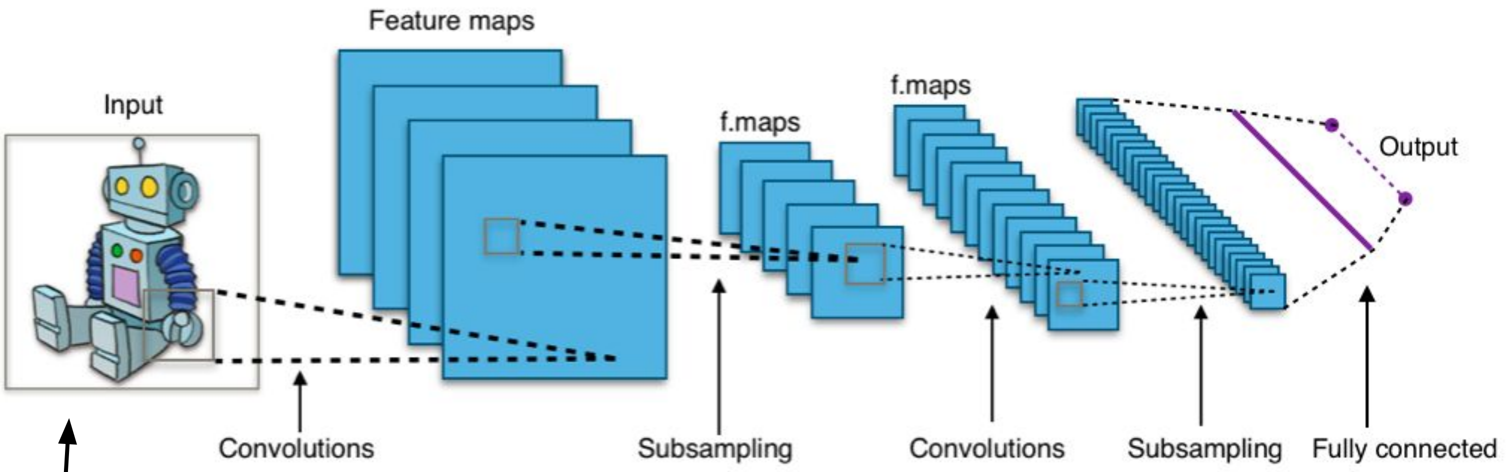
Luca Romani
Ph.D. Student

PINlab
PERCEPTION AND INTELLIGENCE

Graph Metanetworks

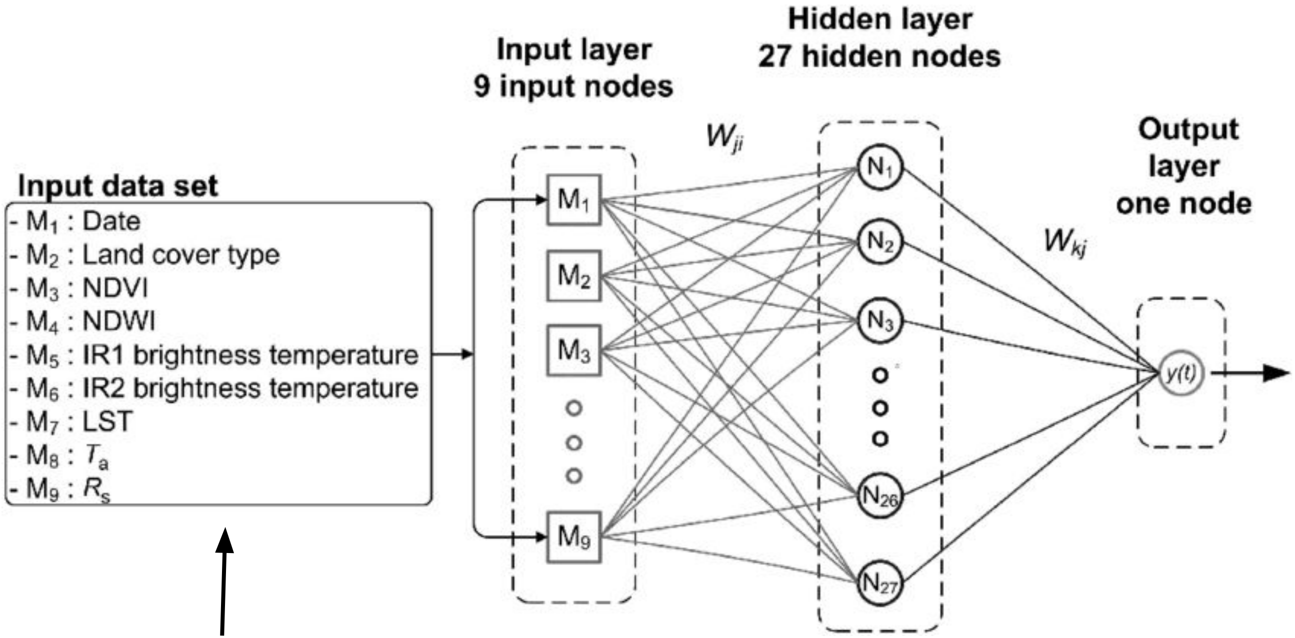


Usual deep learning data



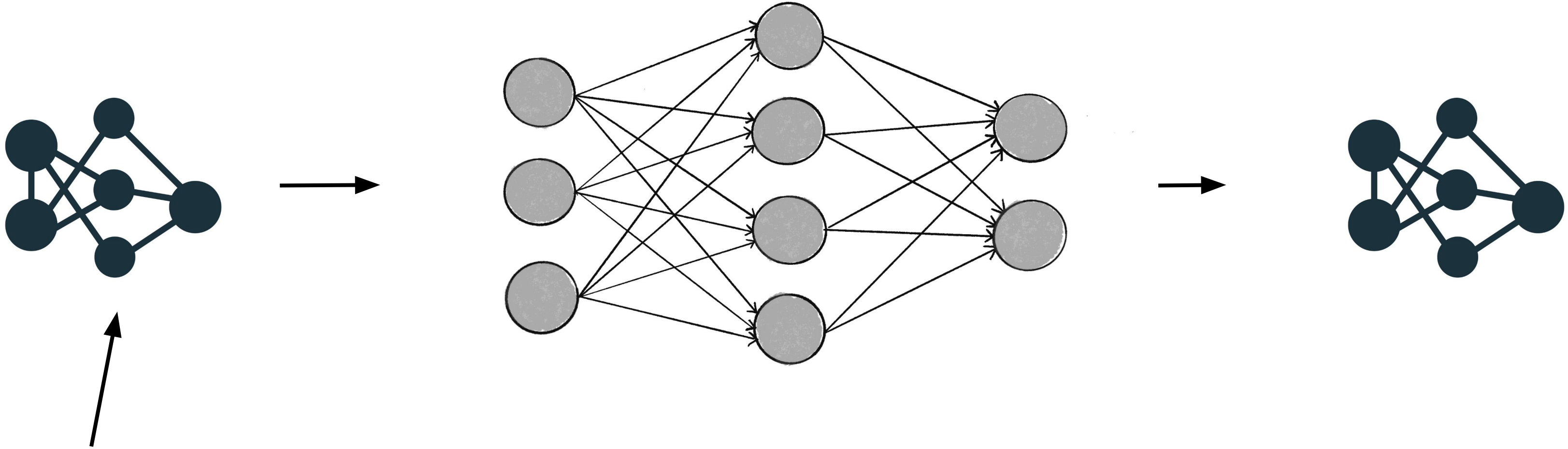
Images

Usual deep learning data



Feature vector about real world data

Metanetworks



Neural network

Metanetworks

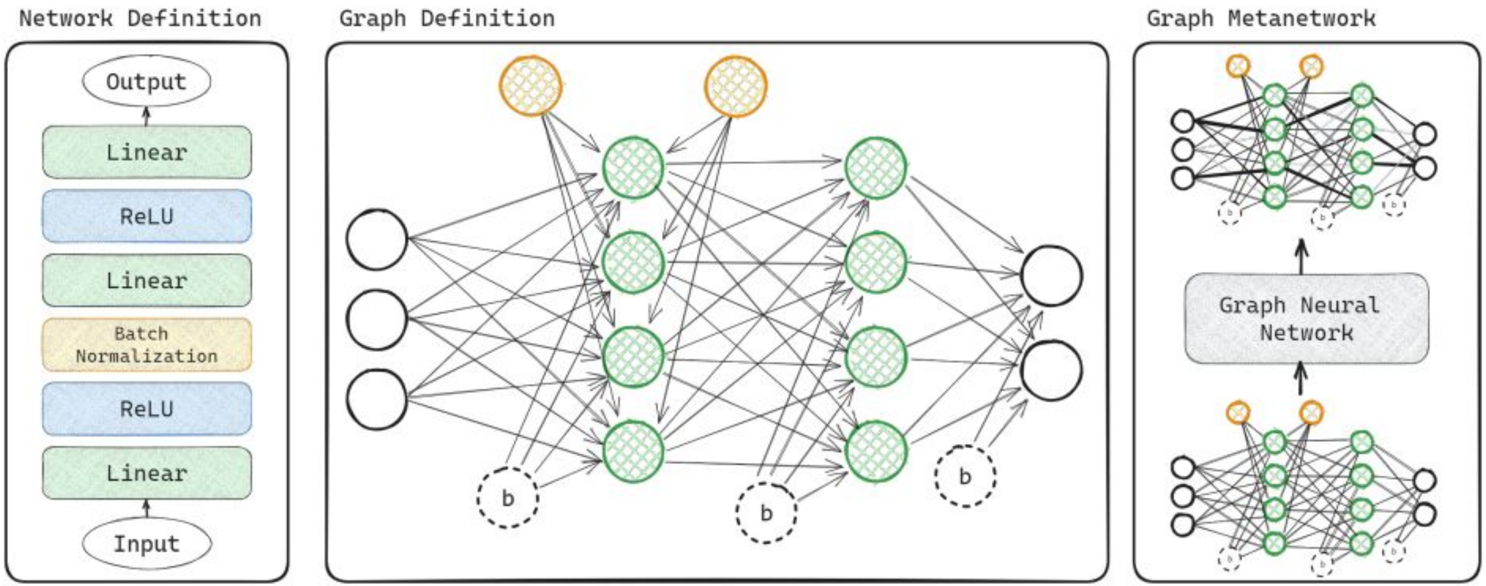
Neural network

Metanetworks

Common limitations:

- they are often able to process only specific architectures, such as MLPs or CNNs, hence struggling with generalization;
- they consider parameters of neural networks as flattened 1D tensors, hence losing all the structural information of the original network.

Graph Metanetworks for Processing Diverse Neural Architectures



Graph Metanetworks for Processing Diverse Neural Architectures

- They encode any neural network as a **parameter graph**:
  - Nodes: represent neurons or neuron groups;
  - Edges: represent parameters (weights) between neurons. Each parameter is associated with a single edge.

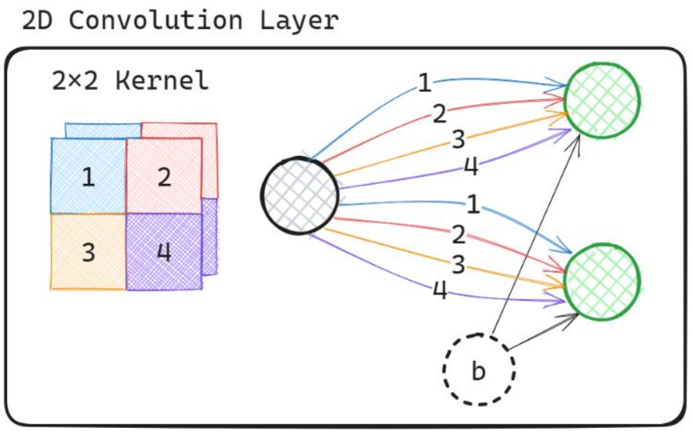
- Parameter graphs are then fed into standard Graph Neural Networks (MPNNs) for processing.

- The output embeddings produced by the MPNN can then be used for various downstream tasks.

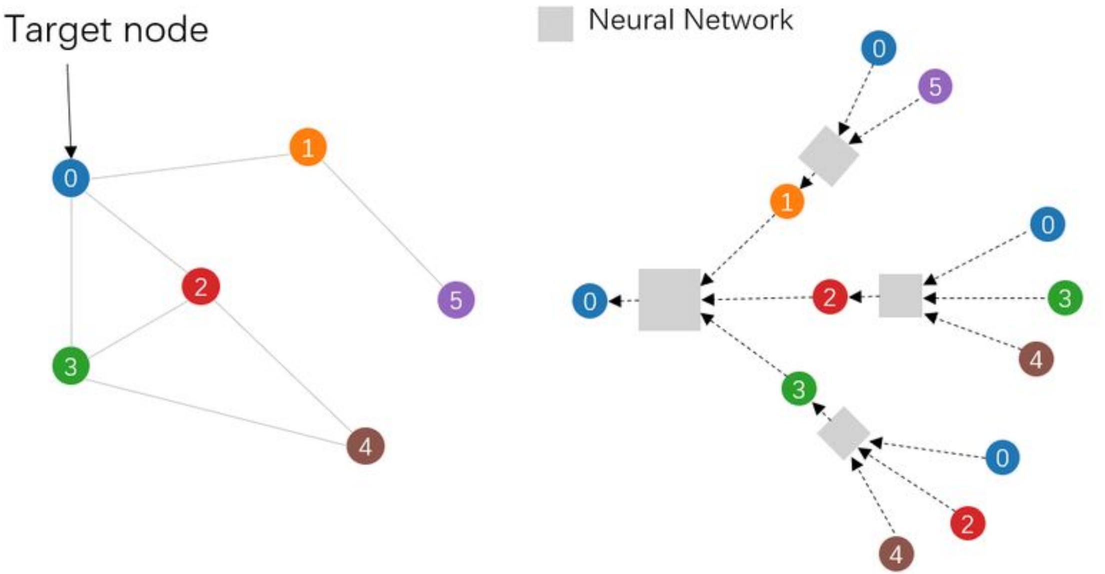
- Examples:
  - A graph-level task can be predicting the network accuracy on a dataset;
  - An edge-level task can be modifying parameters to modify the network functionality.

Example of a parameter graph

- For the convolutional layer case, the parameter graph construction allocates **one node for each input and output channel**. We then have parallel edges between each input and output node **for each spatial location in the filter kernel**, making this a multigraph. One bias node is added, and the bias parameters are encoded as edges from that bias node to each output channel of the layer.



Digression: GNNs…



(a) Input graph (b) Neighborhood aggregation

…aka Message Passing Neural Networks (MPNNs)

- A MPNN is a generalization of a GNN that **updates node, edge, and global features all together**. For a graph, let $\mathbf{v}_i$ be the feature vector of node i, $\mathbf{e}_{(i,j)}$ a feature vector of the directed edge (i,j), **u** the global feature vector associated to the entire graph, and let E be the set of edges in the graph. The directed edge (i,j) represents an edge starting from j and ending at i. Since we allow multigraphs, where there can be several edges (and hence several edge features) between a pair of nodes (i,j), we let $E_{(i,j)}$ denote the set of edge features associated with (i,j).

$$v_i \leftarrow \mathrm{MLP}_2^v\left(v_i, \sum_{j:(i,j)\in E_{(i,j)}} \mathrm{MLP}_1^v(v_i, v_j, e_{(i,j)}, u), u\right)$$

$$e_{(i,j)} \leftarrow \mathrm{MLP}^e(v_i, v_j, e_{(i,j)}, u)$$

$$u \leftarrow \mathrm{MLP}^u\left(\sum_i v_i, \sum_{e\in E} e, u\right)$$

Unlearning

Unlearning

Some definitions:

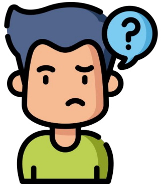
- **Learning**: to acquire the knowledge or skills through study or experience
- **Unlearning**: to lose or discard knowledge that is false or outdated
- **Relearning**: to learn again. This is when diversity breeds innovation, possibility and opportunity

Unlearning

?

What do we mean by Unlearning in machine learning and computer vision?

Unlearning



As a human, you can forget but unlearning is impossible

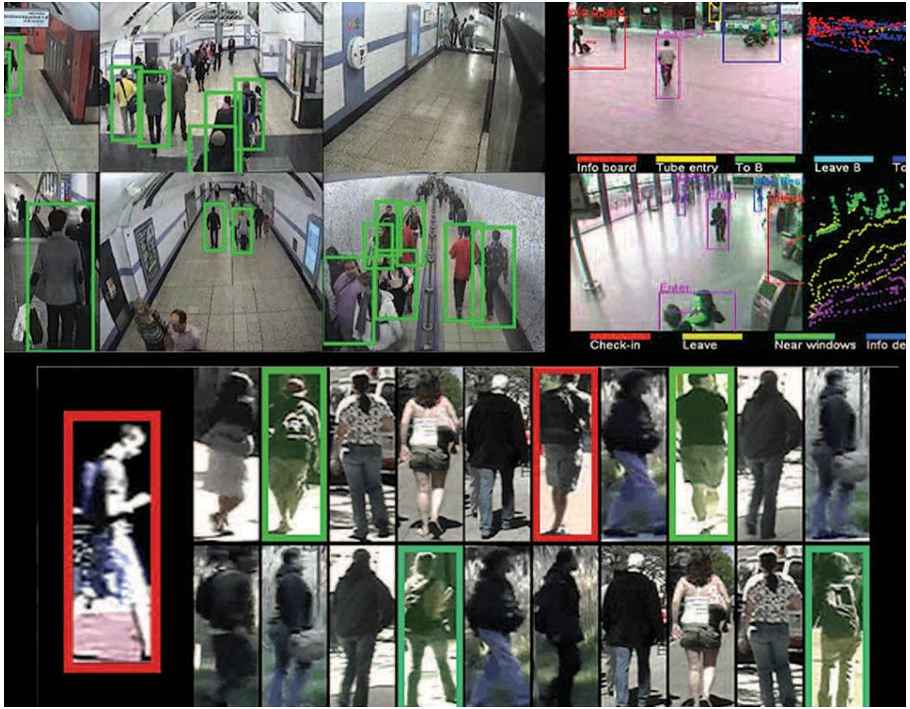
Neuroscientists say that “ Unlearning is simply impossible. You can’t really remove something from your mind unleash there is some sort of brain damage or extreme forms of mind control”

Unlearning

But machines can be forced to forget

In principle, we could ask a recognition system to forget a certain person. But is this filtering or unlearning?

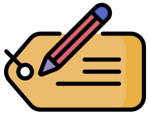
Unlearning is not filtering out
Unlearning is deleting some knowledge



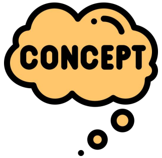
Unlearning

Machine Unlearning

- Capability to complementary remove/forget some data and the related knowledge without changing the performance on the rest



- Forgetting some labels of the dataset



- Removing unwanted concepts in the knowledge representation

Unlearning

Many reason to forget labels

1. **LEGAL**: data are affected by privacy issues or copyrights constraint

2. **ETHICAL**: data can be biased and create ethical unbalance

3. **EPISTEMOLIGIC**: data are useless, obsolete or unwanted for the model

Unlearning

There are cases in which AI systems produce toxic or fake contents like nudity or brutal images of war.

We want to:

1. identify fake/ toxic content
2. avoid the generation of toxic content
3. unlearn the knowledge of toxic content making its generation impossible

Unlearning

We have many types of unlearning

- **Data points**: Removing certain data points from the training set, such as mislabeled data
- **Features**: Deletion of a subset of misleading features, such as gender or race
- **Classes of Data**: Erasure of entire classes, such as user removal
- **Concepts**: Removing the knowledge of emerging concepts or undefined classes
- **Tasks**: Removal of a specific task, such as asking a robot to forget an assistance behavior after the recovery of a patience, for privacy purposes

Unlearning

Unlearning has been proposed initially for legal/privacy reasons.
Now it is studied for understanding the limits of pretrained models.

Unlearning has a double goal:

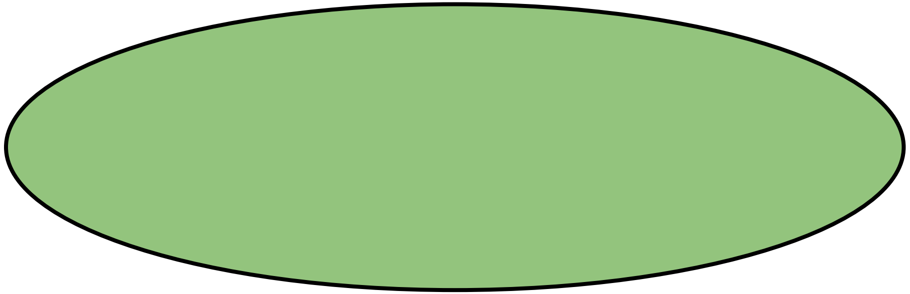
- The goal is to ”untrain” the model, for eliminating the impact of unwanted datapoints
- Reaching weights similar to those of models trained without such data.

When the model re-trained without the unwanted data, it is called **exact unlearning** **or perfect unlearning**

Unlearning

Dataset

$\mathbf{D} = \{x_i\}^{N}_{i=1}$

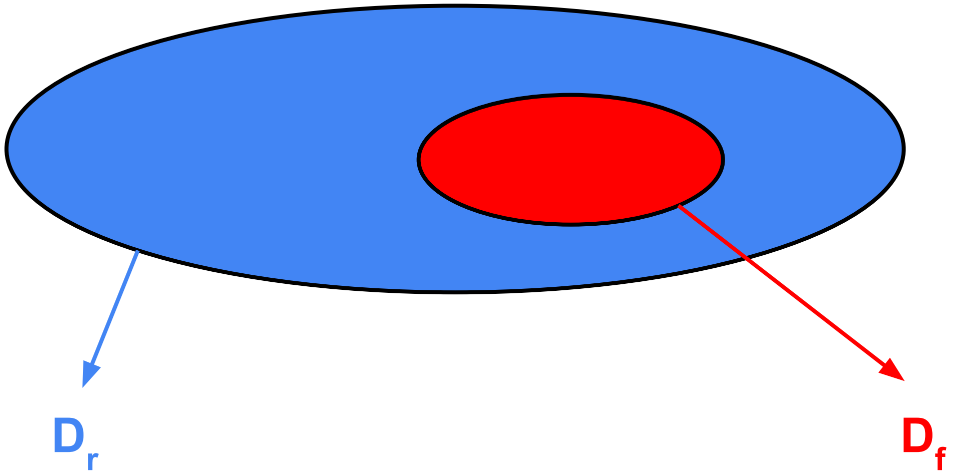


Unlearning

Dataset

$$\mathbf{D} = \{x_i\}^N_{i=1}$$

$$\mathbf{D_r} = \mathbf{D} \,/\, \mathbf{D_f}$$



Unlearning

We can define 3 models:

- $\mathbf{F_w}$ **(x)** the model trained on D
- $\mathbf{F_{w'}}$ **(x)** the unlearned model
- $\mathbf{F_{w^*}}$ **(x)** the perfect unlearned model by retraining with Retain Dataset

Unlearning

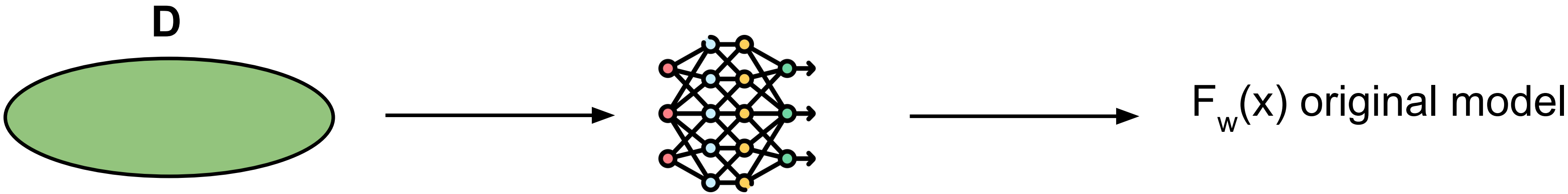
Forget Property: delete knowledge associated with the data to be unlearned

$F_{w'}(x) = F_{w^*}(x)$ for any x in D_f

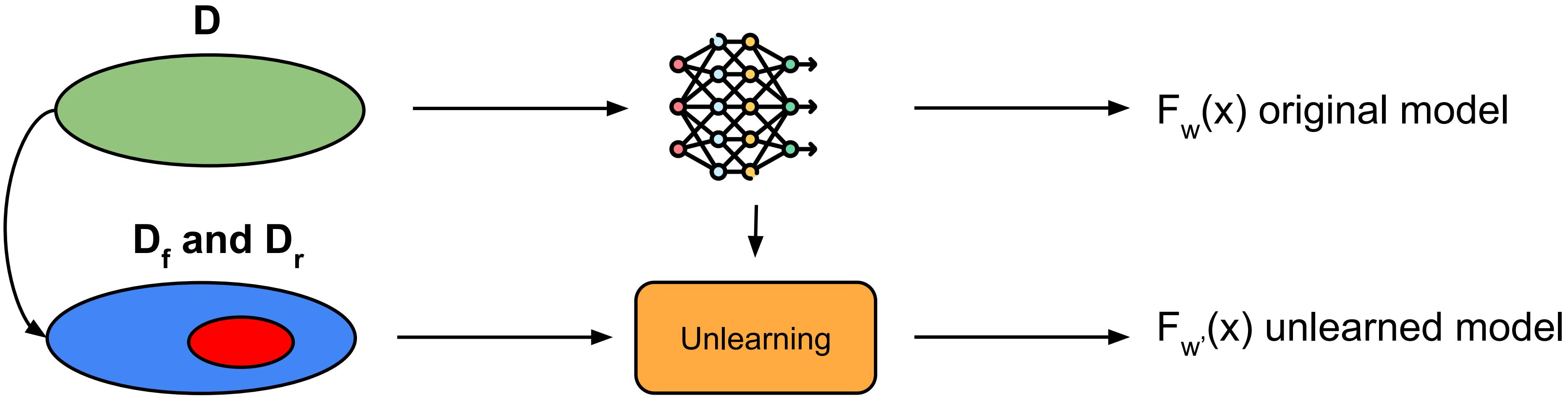
Retain Property: mantain knowledge associated with the data to retain

$F_{w'}(x) = F_w(x)$ for any x in D_r

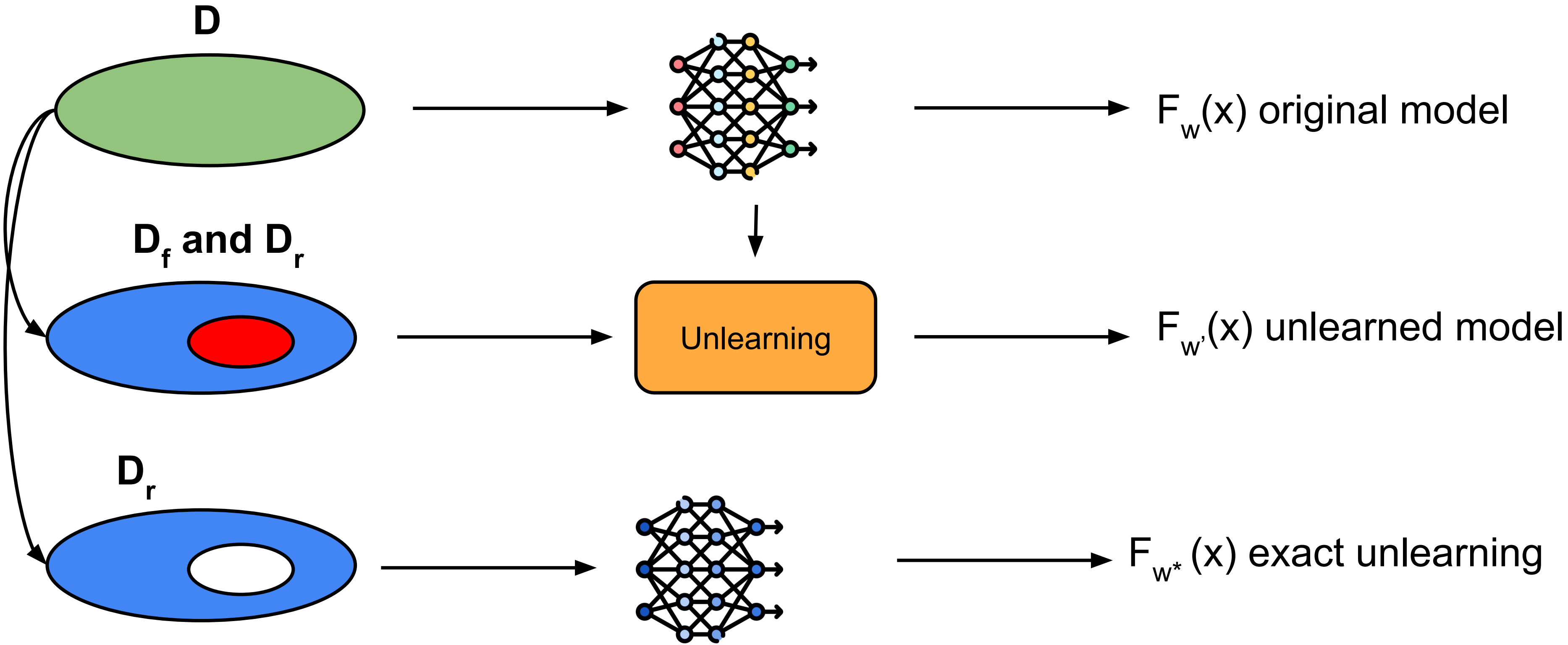
Unlearning



Unlearning


D
D_f and D_r
Unlearning
$F_w(x)$ original model
$F_{w'}(x)$ unlearned model

Unlearning



D
D_f and D_r
D_r
Unlearning
$F_w(x)$ original model
$F_{w'}(x)$ unlearned model
$F_{w^*}(x)$ exact unlearning

References

- Derek Lim, et al. (2024). “Graph Metanetworks for Processing Diverse Neural Architectures”. In The 12th International Conference on Learning Representations;
- Peter W. Battaglia, et al. (2018). “Relational inductive biases, deep learning, and graph networks”. arXiv arXiv:2209.02299;
- “Learning, Unlearning and Relearning”, R. Cucchiara, ELLIS doctoral Symposium 2024;
- T. T. Nguyen, et al. “A survey of machine unlearning”. arXiv arXiv:2209.02299 (2022);
- H. Xu et al “Machine Unlearning: A Survey”. ACM Survey 2023;
- “Google’s AI ‘Reimagine’ tool helped us add wrecks, disasters, and corpses to our photos” blogpost by Allison Johnson.