The 9th Advanced Course on Data Science & Machine Learning

- June 8-12, 2026 -

Lecture:

Multi-Task Deep Learning

Prof. Giuseppe Di Fatta
Faculty of Engineering
Free University of Bozen-Bolzano, Italy
Giuseppe.DiFatta@unibz.it

June 12, 2026

Overview

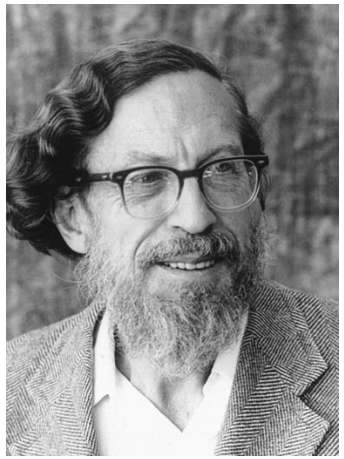
- First intuition: the Stein's Paradox in Statistics
- Multi-Task Learning (MTL) and its Applications
- Multi-Task Deep Learning (MTDL or simply MTL)
  - Internal mechanisms, approaches and challenges
- MTL as Multi-Objective Optimisation
  - Gradient Surgery
- Some final theoretical insights
- Conclusions

Stein's Paradox in Statistics

unibz

- The best estimation of unobserved quantities is their observed averages.
  - The mean minimises the squared error over the samples.
  - The sample mean may differ from the population mean, but cannot do any better from a sample set of a single variable.

- There is a better estimator than sample average when estimating **multiple independent** Gaussian random variables:
  - Arguably the first form of multi-task learning



Charles Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution", TR, Stanford University, 1956

Stein's Paradox: Example

- Baseball players' performance
  - **Hit**: the batter strikes the ball and safely reaches or passes the first base

$$\text{Batting average} = \frac{\text{hits}}{\text{bat times}}$$

  - unobserved quantity: true batting performance in a season
  - observed quantity: batting average in the first N bat times

Willie Mays' 3000th hit, July 18, 1970

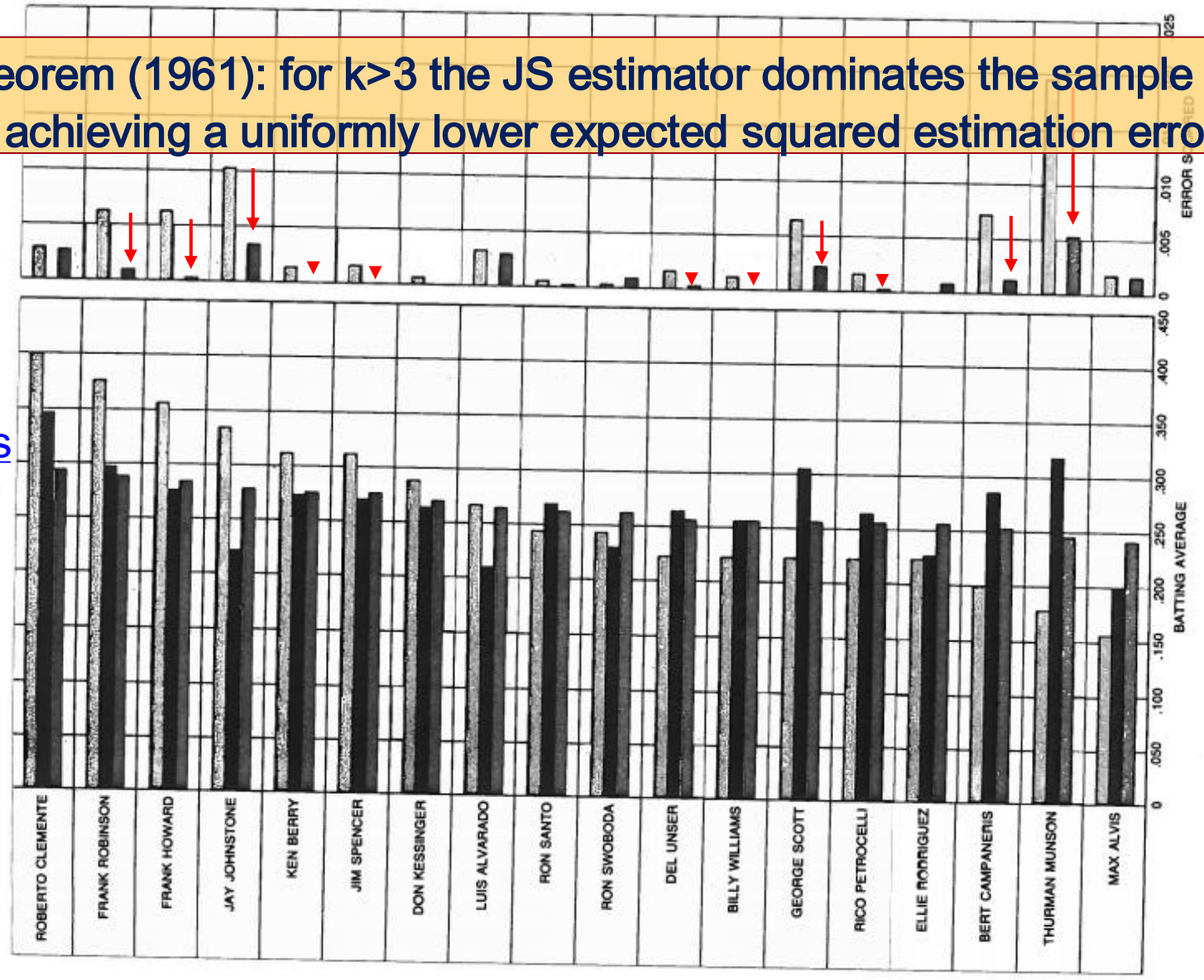
- **"Paradox"**: the best estimator for each player's expected performance is **NOT** their individual observed batting average (the unbiased estimator).

1970 Major-League Baseball Players

James-Stein Theorem (1961): for k>3 the JS estimator dominates the sample average vector by achieving a uniformly lower expected squared estimation error.

Squared Errors

Batting Averages



Source: Bradley Efron and Carl Morris, "Stein's Paradox in Statistics", 1977

The Shrinking Factor

- The shrinking factor in James-Stein estimator is a form of **multi-task regularisation** for averages.
- For a number k (k>3) of independent variables (players):

$$\mu_i^{js} = \bar{\mu} + c(\mu_i - \bar{\mu})$$

$$c = 1 - \frac{(k-3)\sigma^2}{\sum(\mu_i - \bar{\mu})^2}$$

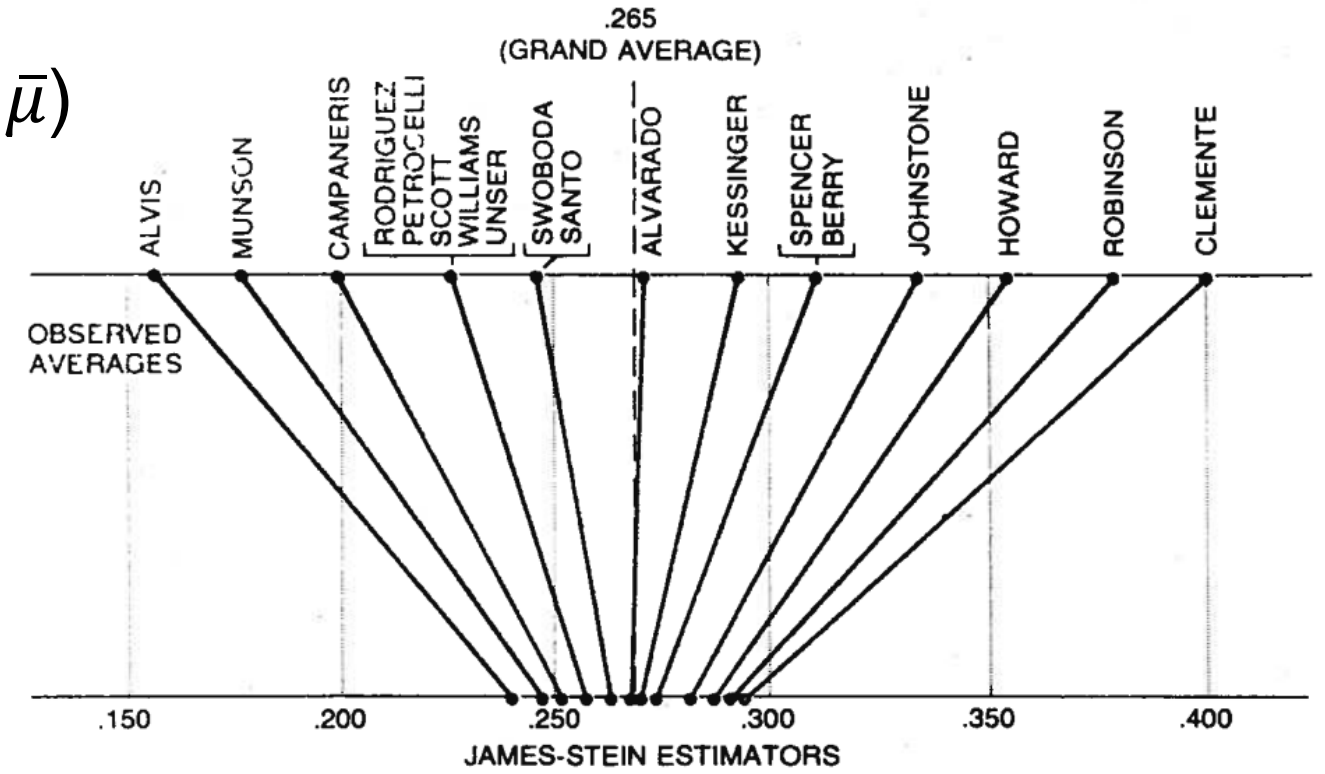
where:

μ_i^{js} *: JS estimator*

$\bar{\mu}$: grand average

μ_i: single average

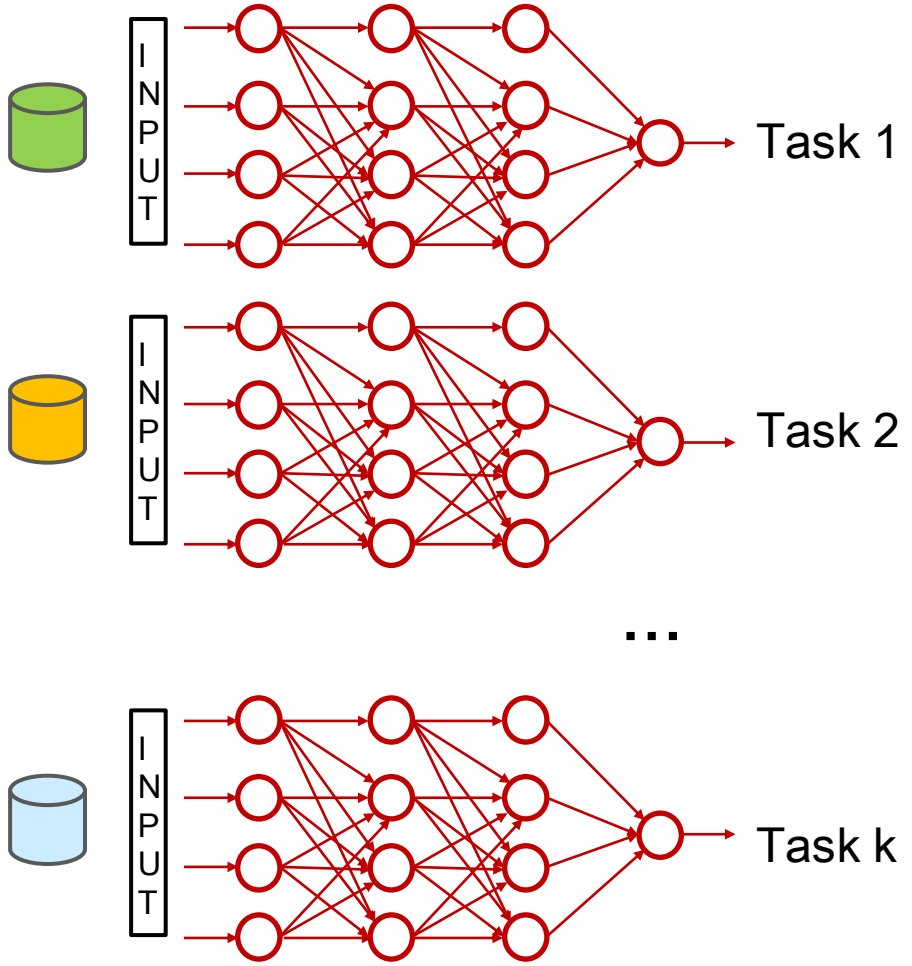
k, σ: known and fixed



Source: Bradley Efron and Carl Morris, "Stein's Paradox in Statistics", 1977

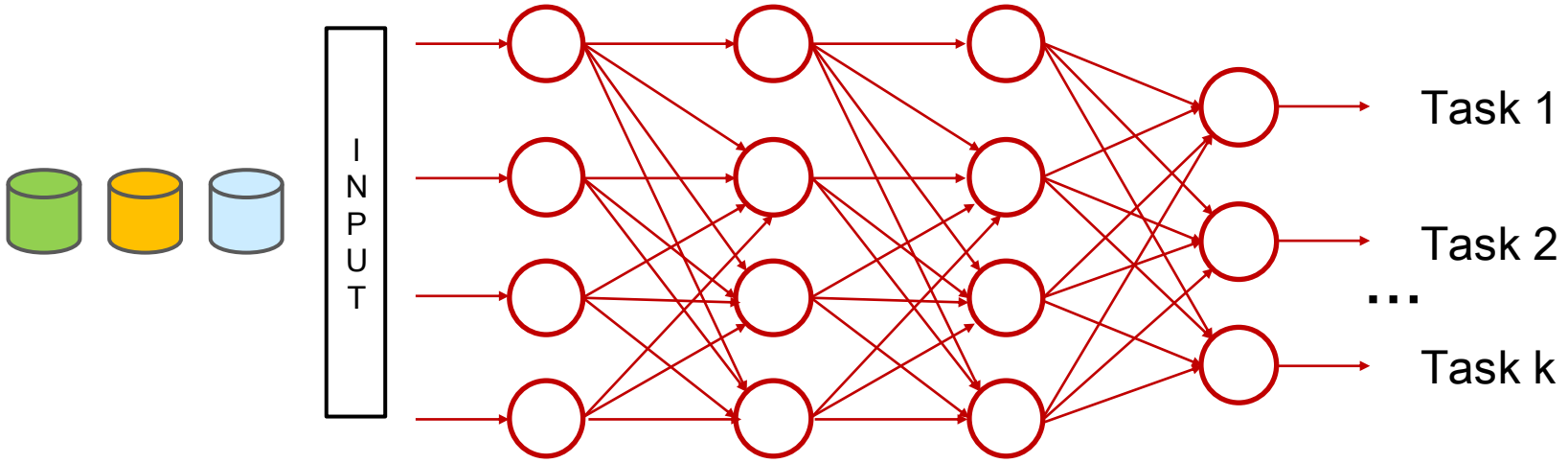
Single-Task Learning in ANN

- many binary classifiers



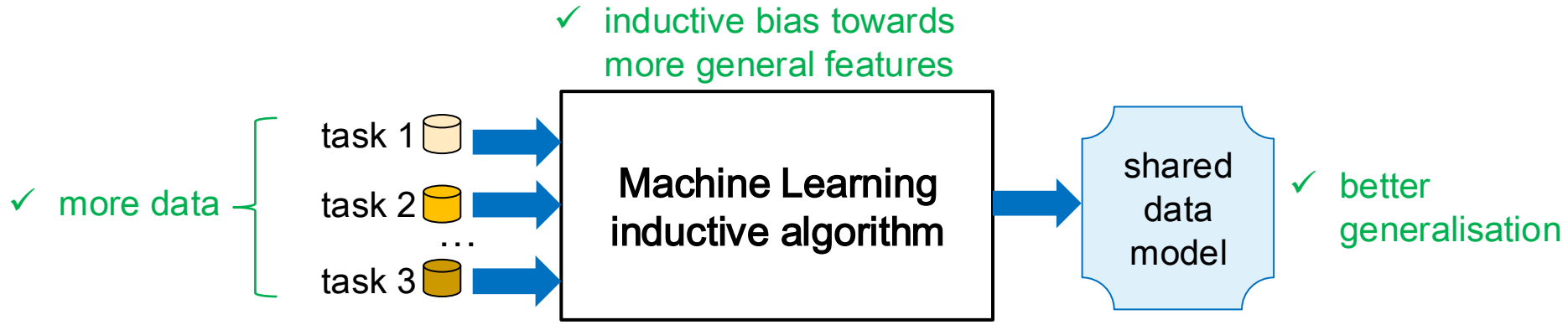
Multi-Task Learning (MTL) in ANN

- a single shared model, simple architecture



Multi-Task Learning (MTL)

- In MTL, tasks are trained **in parallel** using a **shared representation**.
- An **inductive transfer** mechanism improves **generalisation** performance by leveraging domain information from **related tasks**.
- The training data for many tasks work as an **inductive bias**: learning all tasks **concurrently** helps each task to be learned better and reduces the risk of overfitting on any of them.

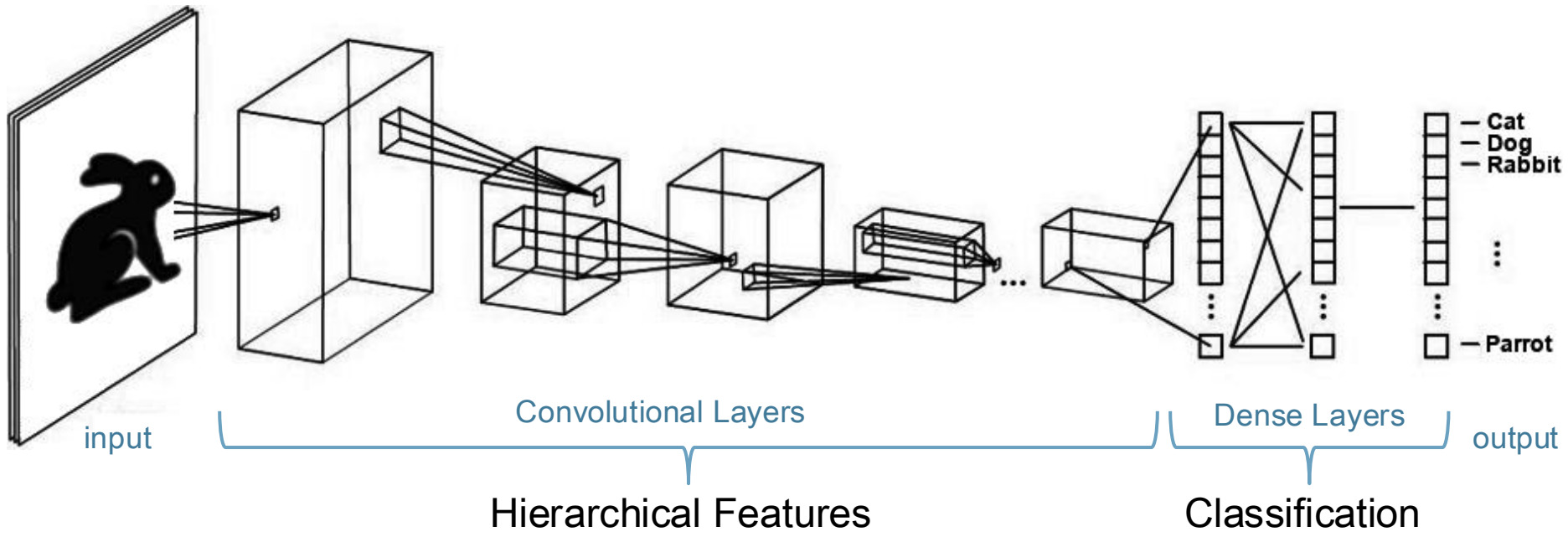


"MTL is a collection of ideas, techniques, and algorithms, not one algorithm."

Rich Caruana, "Multitask Learning", Machine Learning, 1997

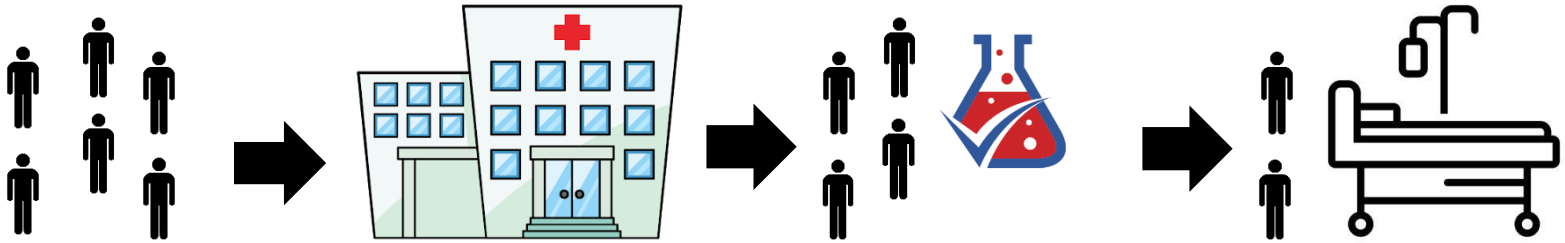
Example: Multinominal Classification

- Different class labels as multiple tasks
- Object recognition from images: cat, dog, rabbit, etc.
- CNN layers: from generic features to specific ones
- Dense layers for classification



Example: Medis Pneumonia Data

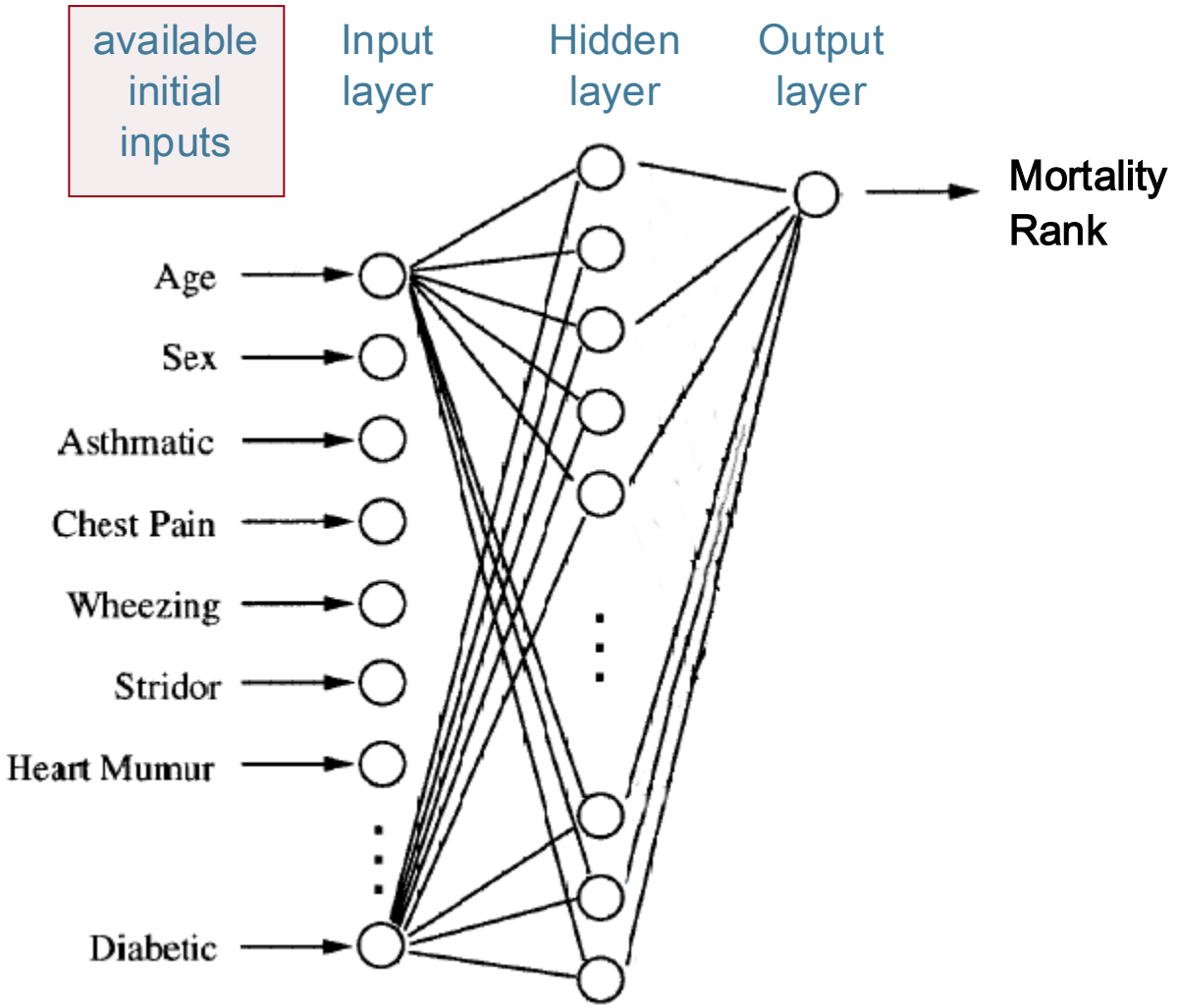
- Predict mortality risk for hospitalisation



Symptoms, risk factors	Laboratory tests	Intensive care
• Age • Asthmatic • Chest pain • Wheezing • Stridor • Hearth murmur • Diabetic • etc.	• Hematocrit • White blood cell count • Potassium • etc.	• Mortality risk

STL on Medis Pneumonia Data

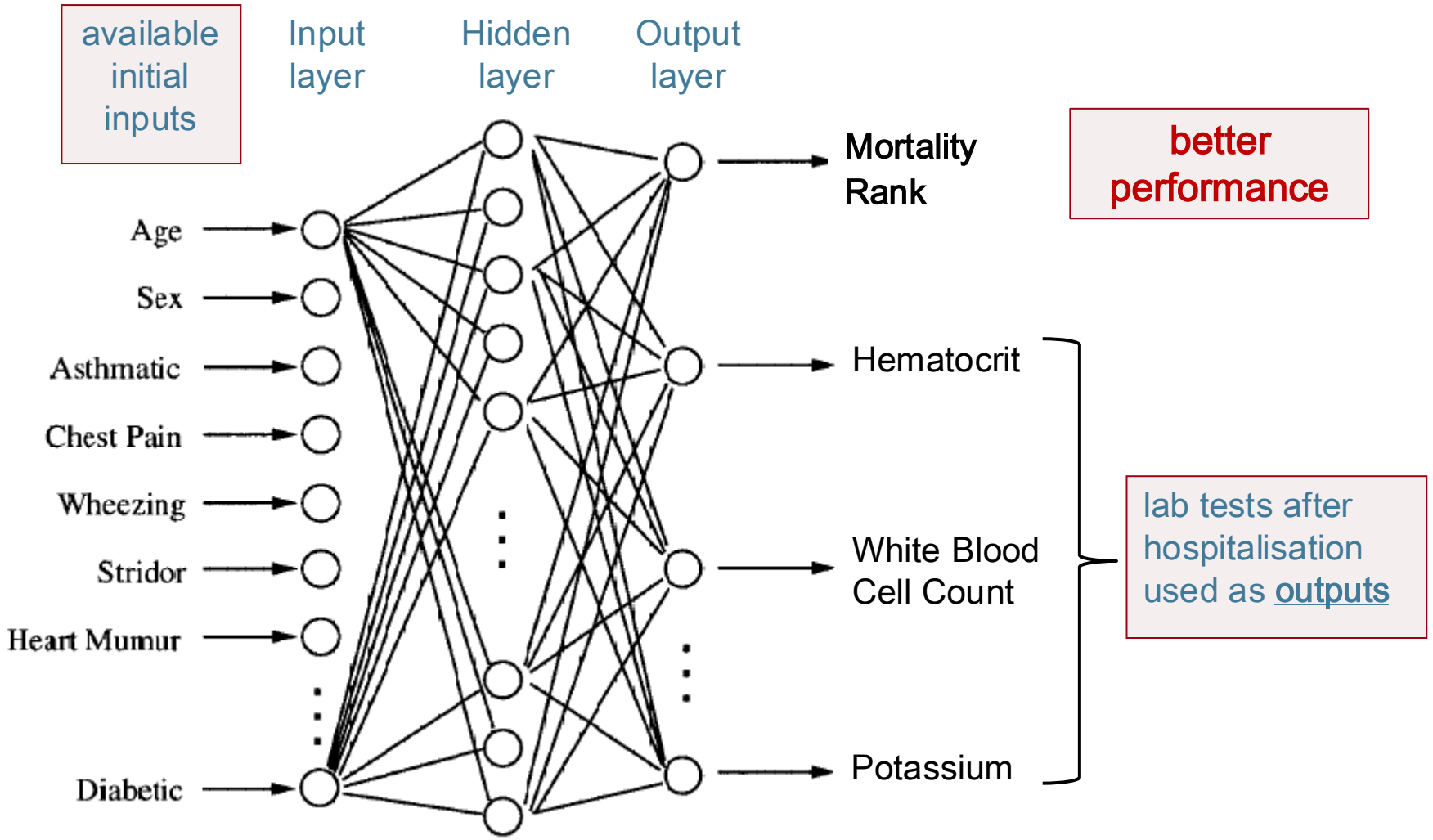
- STL: predict mortality risk (rank) for hospitalisation decision



Rich Caruana, "Multitask Learning", Machine Learning, 1997

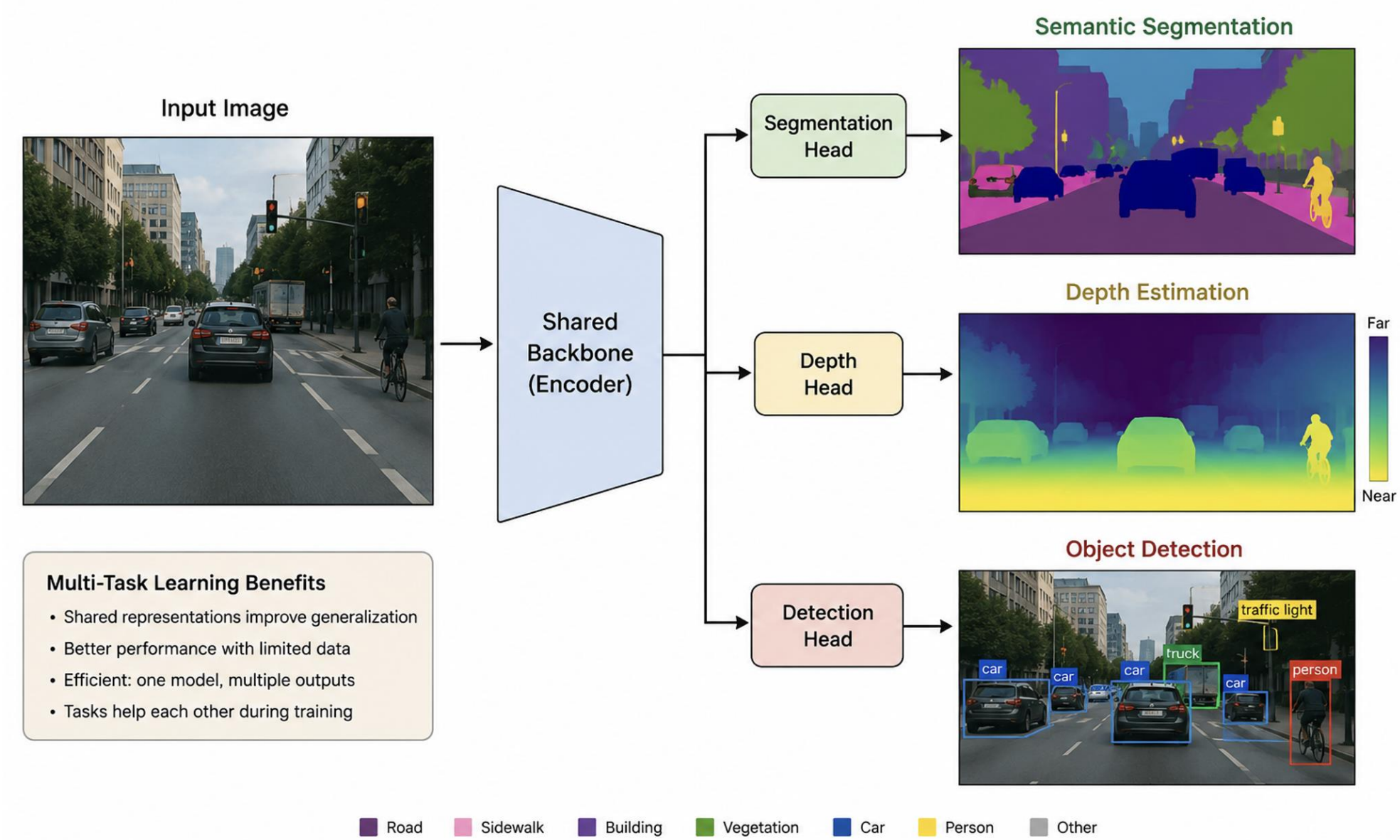
MTL on Medis Pneumonia Data

- MTL: predict mortality risk for hospitalisation decision plus auxiliary tasks



Rich Caruana, "Multitask Learning", Machine Learning, 1997

Example: Multi-Task Learning in Vision Tasks



Multi-Task Learning Benefits

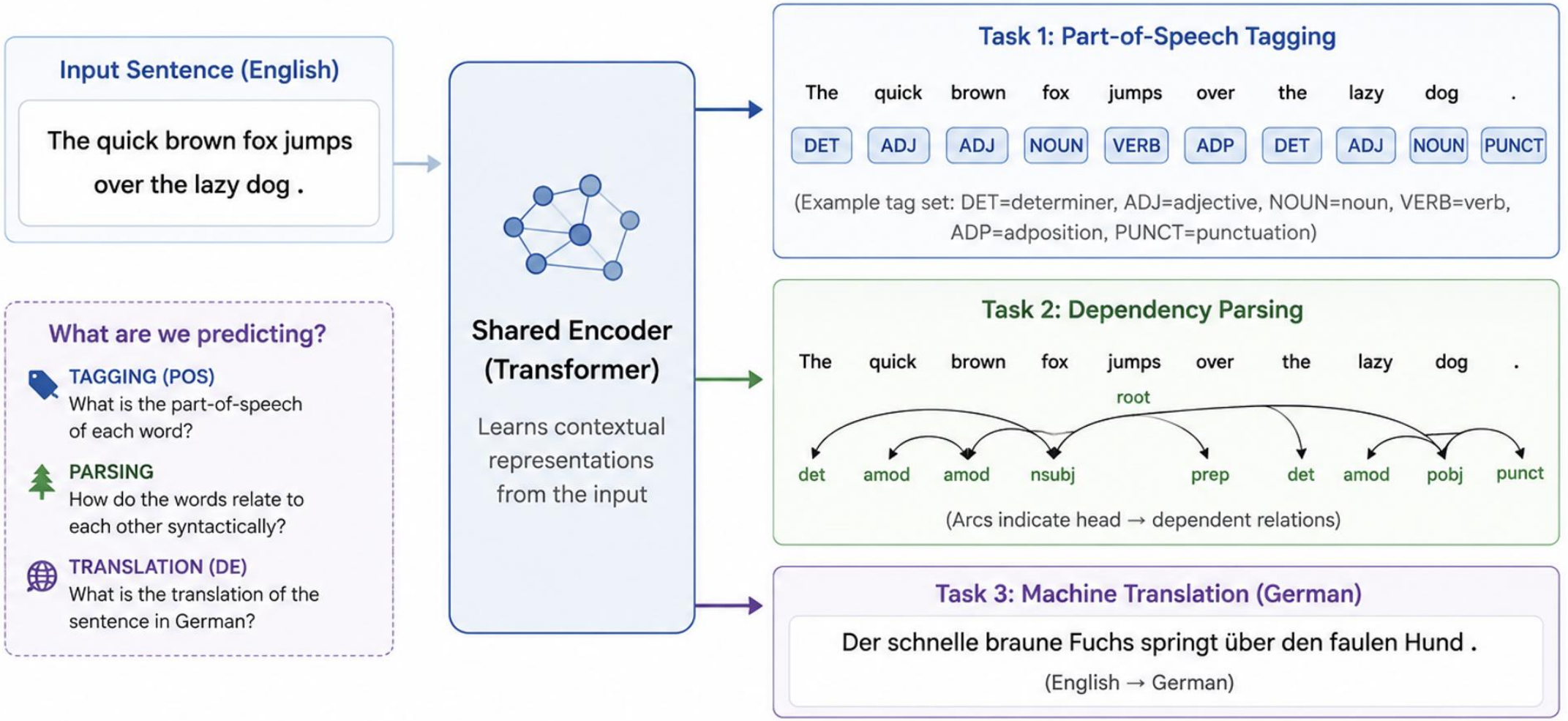
- Shared representations improve generalization
- Better performance with limited data
- Efficient: one model, multiple outputs
- Tasks help each other during training

Segmentation

Depth

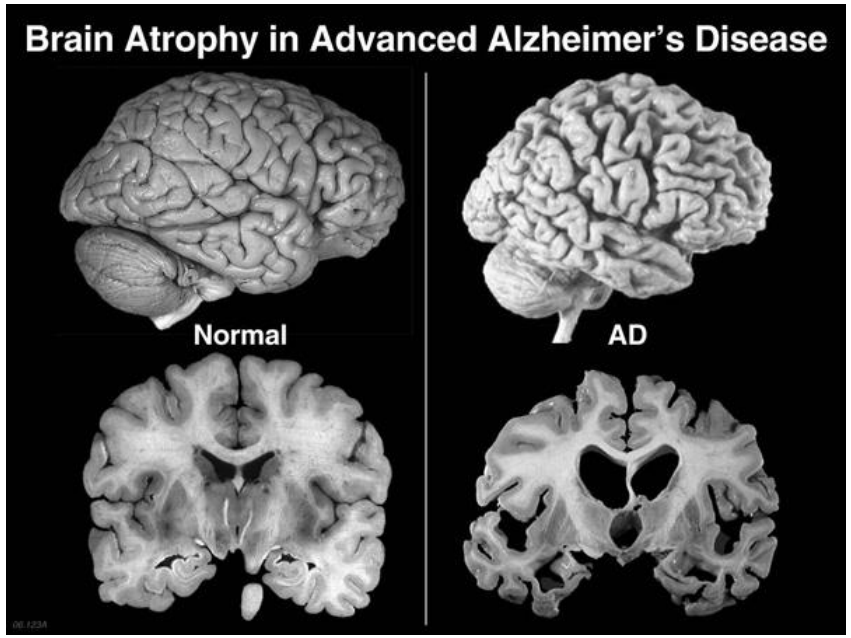
Detection

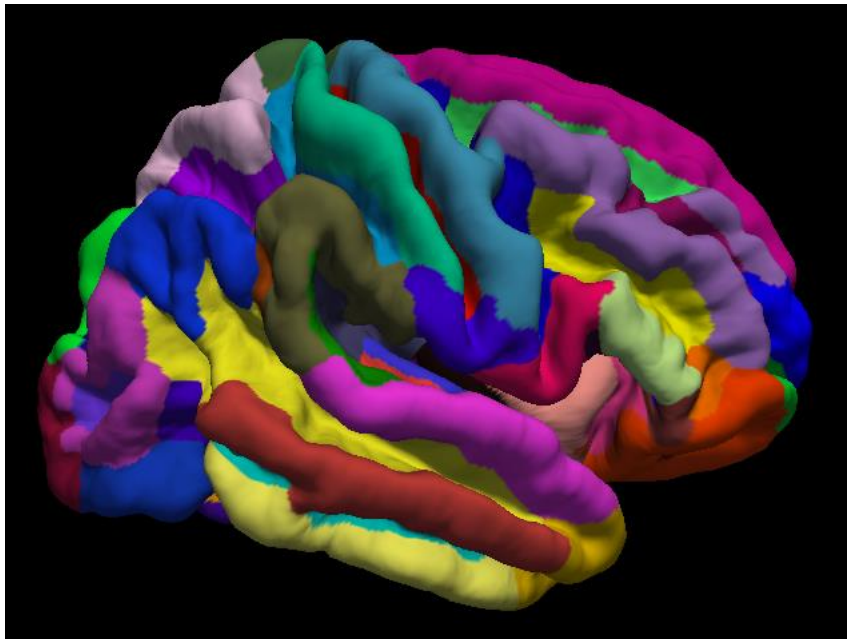
NLP MTL: Tagging + Parsing + Translation



Example: Multinomial Classification for Disease Prediction

- Neurodegenerative diseases: multi-class and multi-label classification (e.g., co-morbidity):
  - Alzheimer's disease, Fronto-temporal dementia, Parkinson's disease, etc.





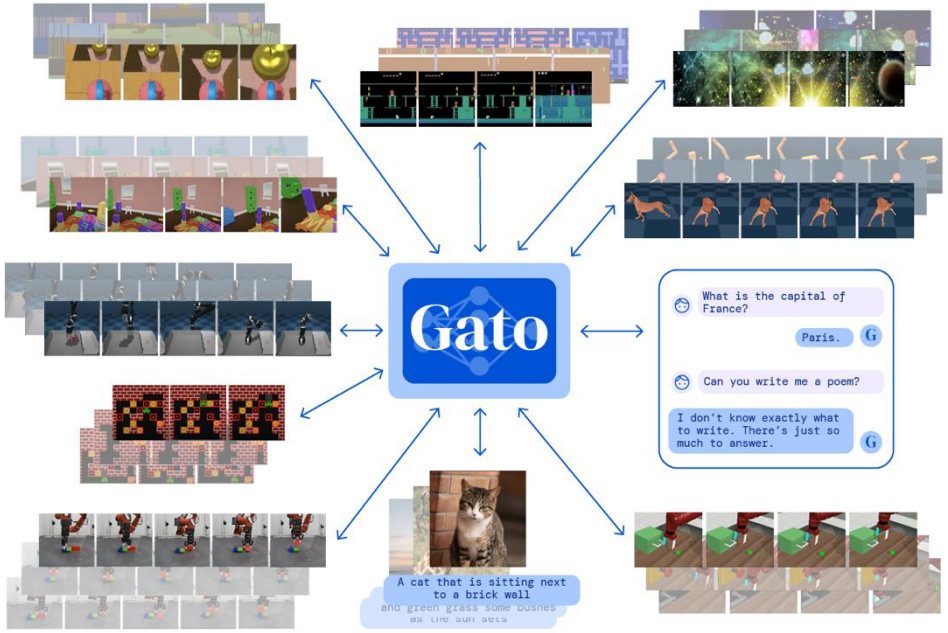
T. Borsani, G. Nicosia, G. Di Fatta, Multi-Task Deep Learning for the Multi-Label Prediction of Neurodegenerative Diseases, 2026 IEEE Conference on Artificial Intelligence (CAI), May 8-10, 2026, Granada, Spain

GATO - A Generalist Agent for 604 Tasks

unibz

May 12, 2022: Google DeepMind introduced GATO, a multi-modal, multi-task agent

Multi-embodiment generalist policy: “the same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens.”



https://www.deepmind.com/blog/a-generalist-agent

- Described as "general purpose" AI and a step toward Artificial General Intelligence

Inside MTL: Cooperation and Competition among Tasks

unibz

MTL as a paradigm shift: from learning tasks independently to learning them jointly

- Why does MTL improve generalisation?
- When does task interaction become detrimental?

Topics:

- Shared representations and positive transfer
- Sequential vs concurrent MTL
- Task competition, negative transfer, and gradient conflicts

MTL Internal Mechanisms

unibz

1. **Shared Representations and Inductive Bias**
   - Learning multiple tasks jointly acts as a regulariser, encouraging representations that capture general structure in the data and improving generalisation.
2. **Data Amplification**
   - Tasks with limited or noisy data benefit from information provided by other tasks, effectively increasing the amount of training signal.
3. **Eavesdropping**
   - Features that are difficult to learn from one task alone can be learned through related tasks and transferred via shared representations.
4. **Implicit Discovery of Task Relatedness**
   - Through backpropagation in shared layers, the model can automatically exploit relationships among tasks without requiring explicit knowledge of their relatedness.

Transfer Learning Approaches

unibz

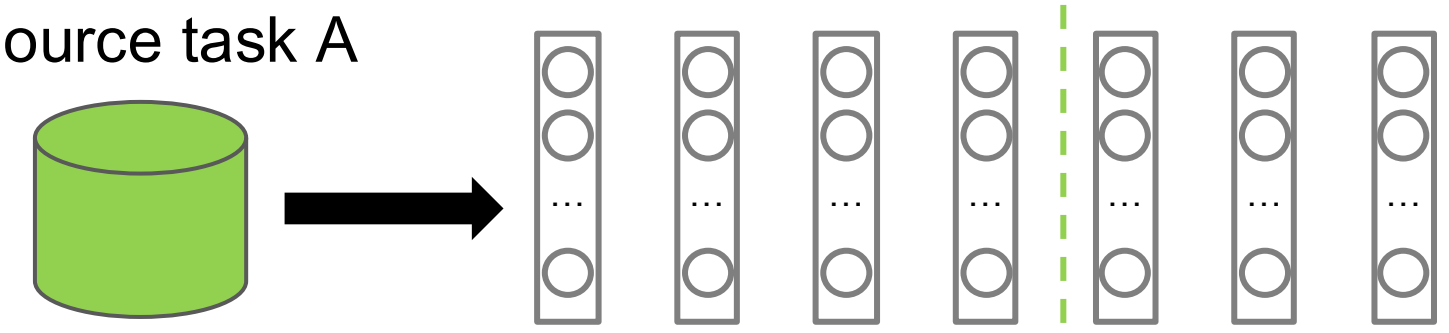
- **Sequential Transfer Learning**
  - Tasks are learned sequentially, with knowledge transferred from a source task to a target task.
  - The goal is to improve performance on the target task by leveraging representations learned from a previous task.

- **Concurrent Transfer Learning**
  - Multiple tasks are learned simultaneously using a shared model. Knowledge is transferred through the shared representation.
  - The model seeks to learn representations that are beneficial across all tasks.

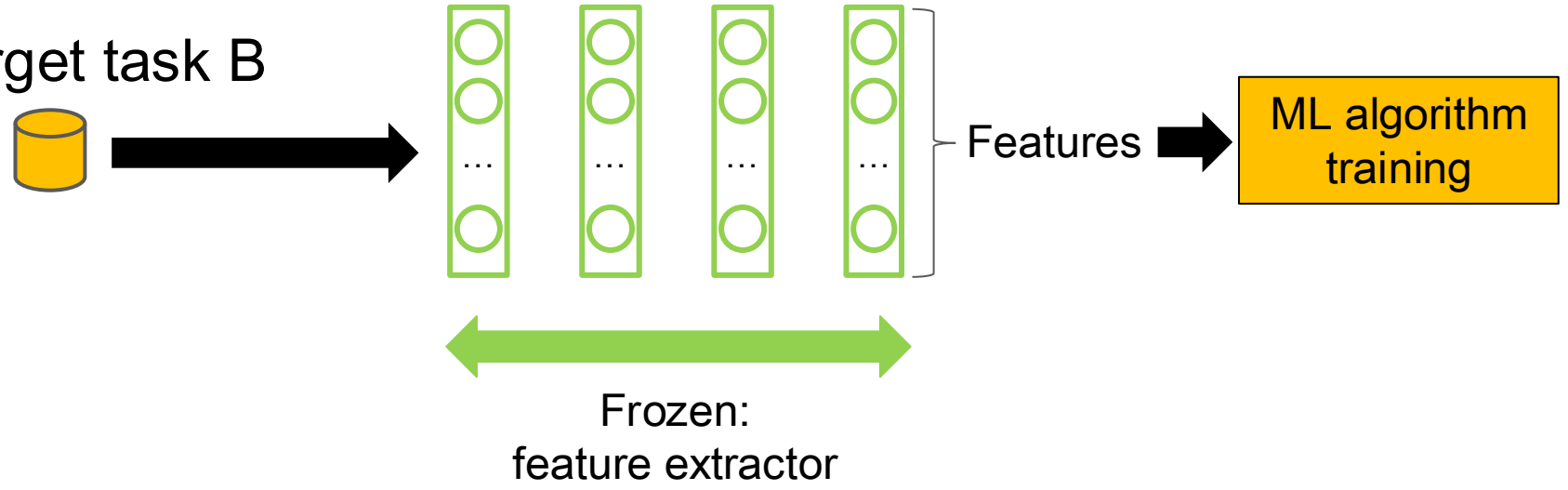
- **Multiform Transfer Learning**
  - Multiple formulations of the same problem are learned jointly (*task engineering*).
  - A single-task problem is reformulated as several related tasks, allowing the model to exploit multi-task learning mechanisms such as shared representations and inductive bias.

Sequential Transfer Learning: Feature Extraction

- Source task A

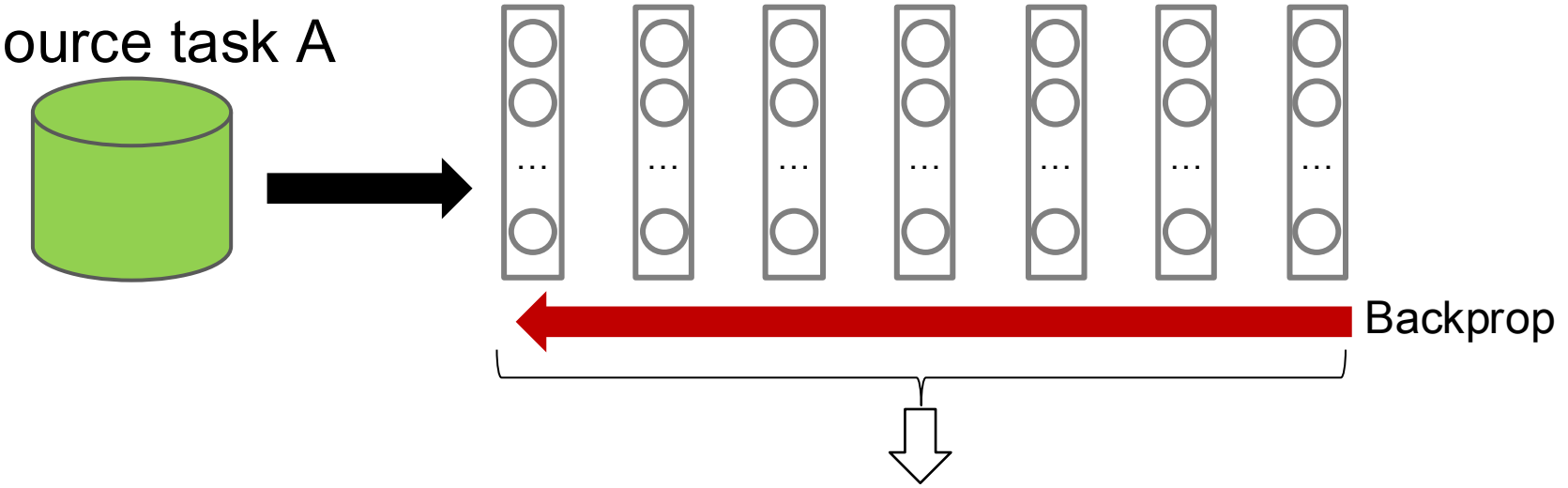


- Target task B

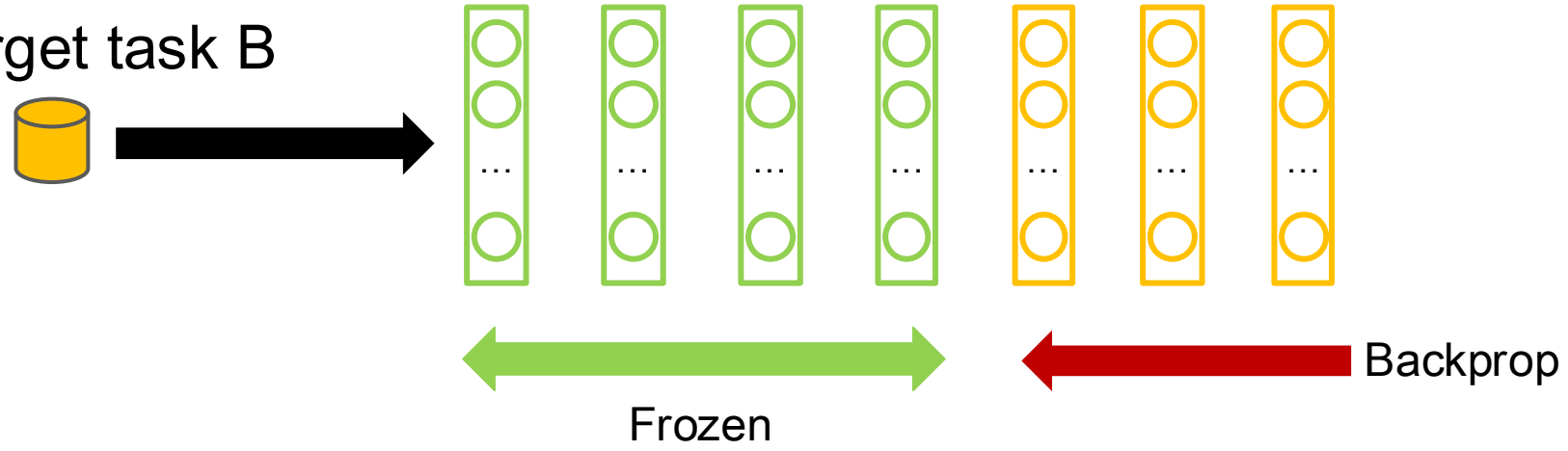


Sequential Transfer Learning: Fine Tuning

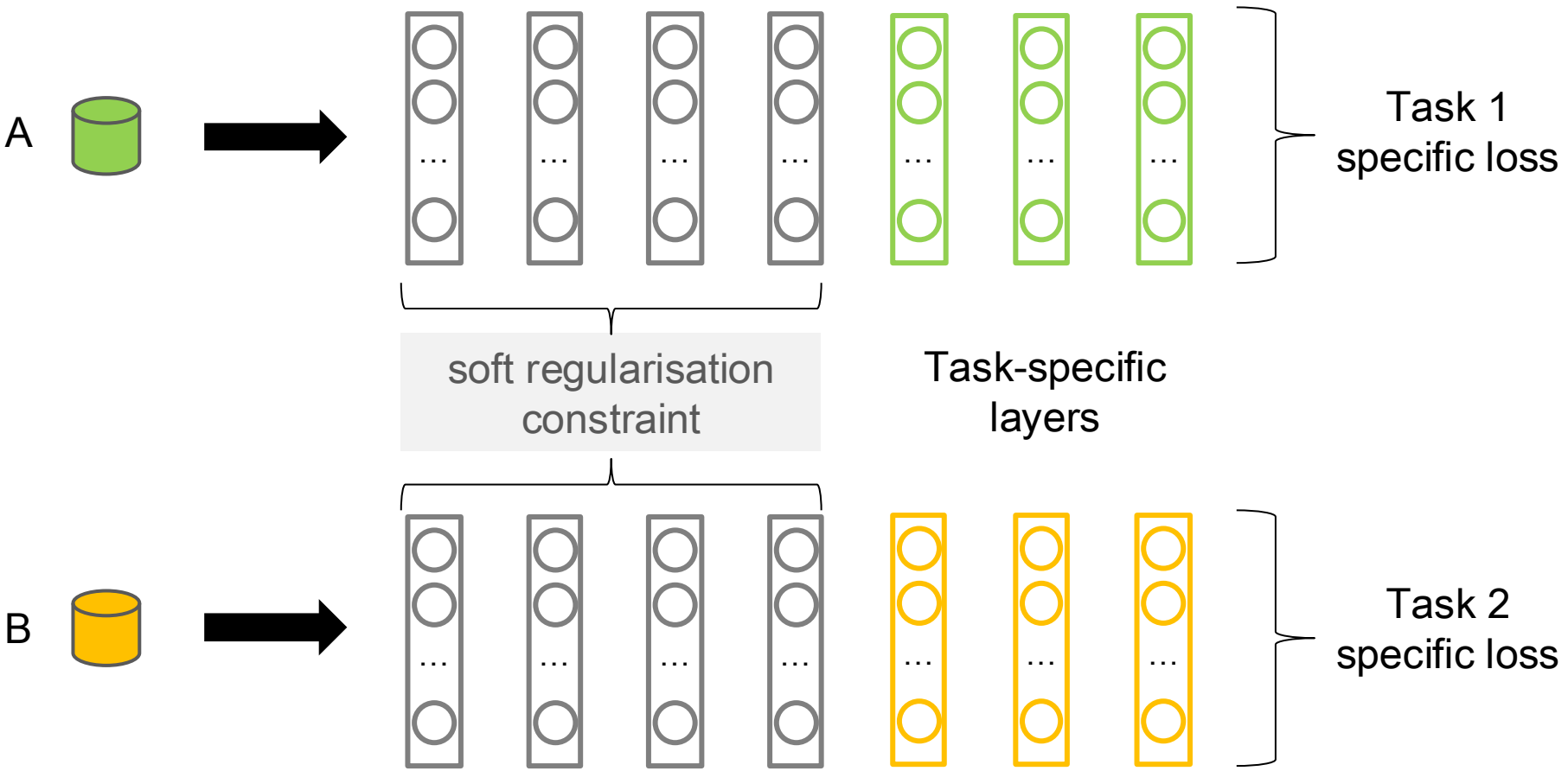
- Source task A



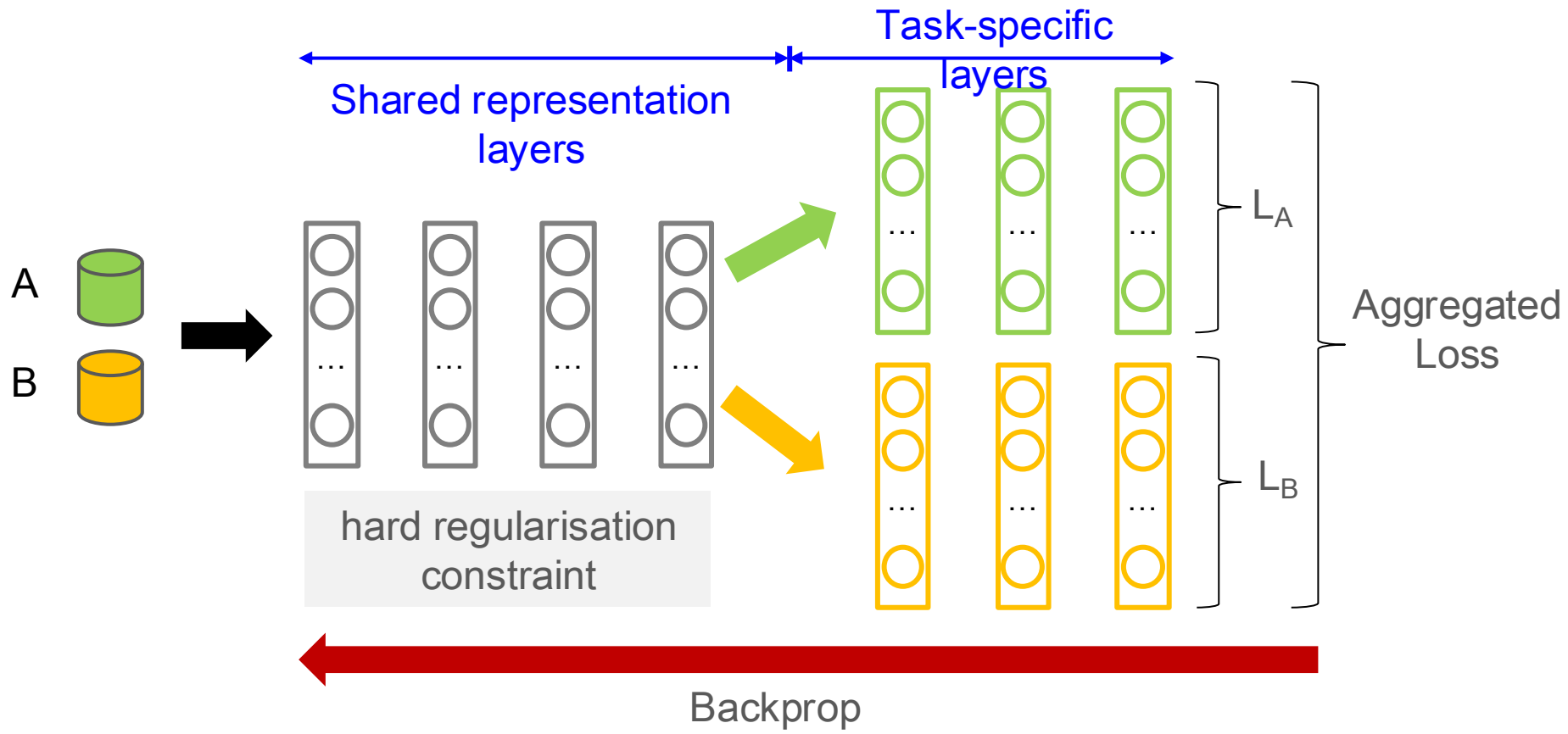
- Target task B



MTL Soft Regularisation



MTL Hard Regularisation



MTL Challenges

- DNN architectures and learning strategies
  - Architecture design, initialisation strategies, loss aggregation approaches
- Feature transferability
  - Which features should be shared?
  - Issues: negative-transfer in feature layers and under-transfer in classifier layers
- Task relatedness
  - How to select related tasks?
  - How to measure task similarity?
  - How to embed known task relations?
  - Can MTL be employed to learn unknown or unexpected task relations?

Transfer Learning in Pre-Trained Models

unibz

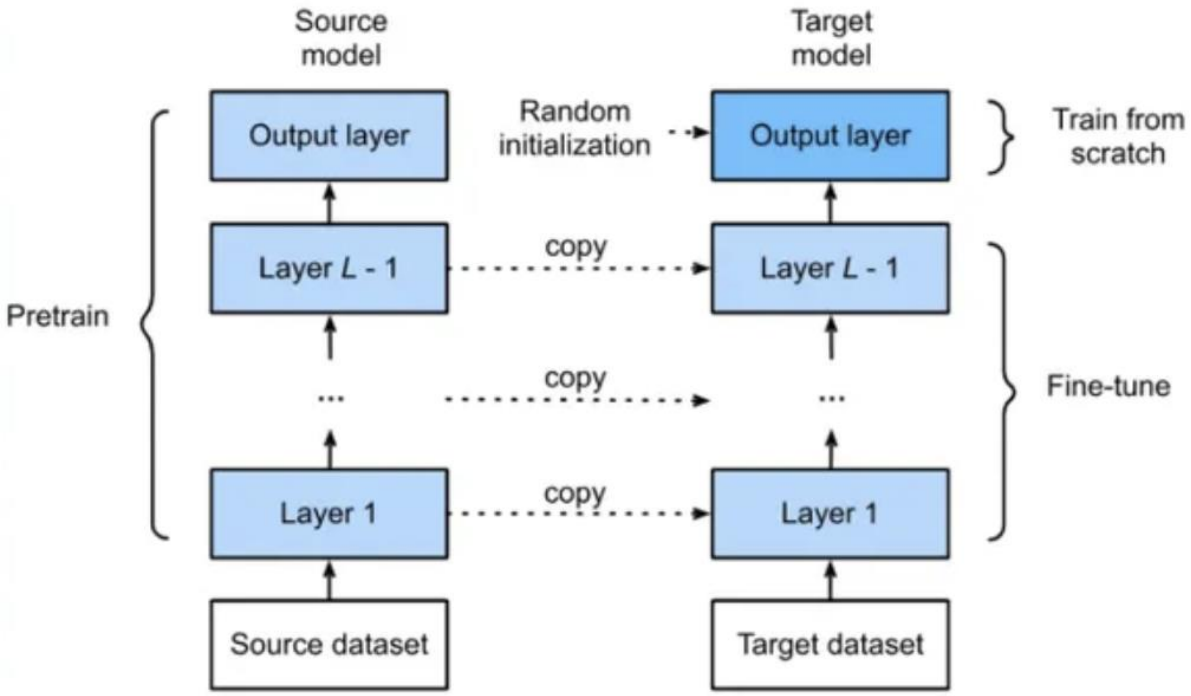
- **Pre-Trained Models (PTMs)** are models trained on large-scale data and later adapted to downstream tasks through transfer learning.
- **Foundation Models** are large PTMs trained on massive datasets, typically using self-supervised objectives, and capable of supporting many tasks.

- Model customisation (e.g., fine-tuning, LoRA, adapters) is a form of **transfer learning**.
- Many **multimodal foundation models** leverage **multi-task learning** principles by learning from multiple modalities and objectives simultaneously.

Transfer Learning and Fine Tuning

- Transfer learning refers to repurposing a model trained for one task to suit a new but related task. A pre-trained model can be used as a **feature extractor** for building a specialised model
  - For example, given a pre-trained CNN model, the lower layers are **frozen** and used to extract features on which the final layers (head) are retrained for the specific image type for a new application.

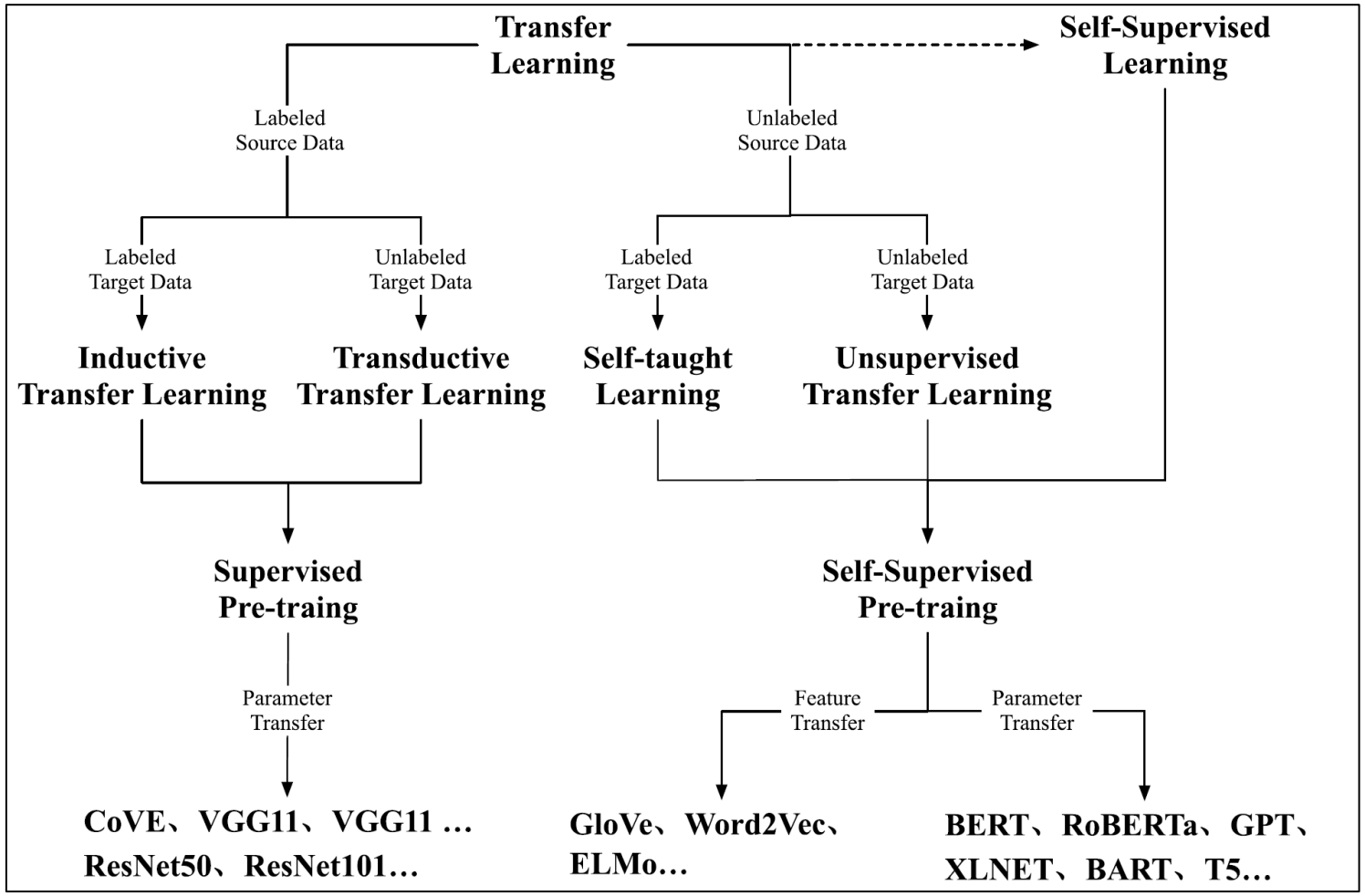
- Fine-tuning involves retraining a pre-trained model for a new task or for improving its accuracy by adjusting its weights.



Pre-Training Approaches

A wide range of pre-training and self-supervised approaches

Large-scale pre-training enables models to acquire general knowledge from massive amounts of unlabelled data and then adapt efficiently to downstream tasks.



Source: Pre-trained models: Past, present and future, Xu Han et al., AI Open, V.2, 2021

Unsupervised Multi-Task Learning in LLMs

- Traditional MTL:
  - multiple datasets
  - multiple losses
  - shared representation

LLMs show **strong transfer across tasks** and **zero-shot** performance on many **implicit tasks**:

- QA, summarization, translation, reading comprehension, commonsense reasoning, etc.

- "Language Models are **Unsupervised** Multitask Learners" (GPT-2, A. Radford et al., 2019)
  - one dataset: WebText with 8M docs and ~40GB of text
  - one objective: next-token prediction
  - one model: 1.5B parameters

Multi-task behaviour emerges from LLMs without explicit multi-task supervision.

The Emergence of In-Context Learning

- "Language Models are **Few-Shot Learners**" [Brown et al., 2020]
- GPT-3 demonstrated that sufficiently large language models could perform many tasks **through prompting alone**.

Zero-shot: instruction only

vs

One-shot: one example

vs

Few-shot: several examples

- The few-shot version often produced more accurate translations.
- This was **surprising** because the model was **not explicitly retrained for the task**.

In-Context Learning vs Fine-Tuning

unibz

- Large transformer models exhibit in-context learning.
  - ➢ Importantly, the model's weights do not change. The learning occurs entirely within the context window.

- Few-shot prompting often provided substantial improvements over zero-shot prompting. **So, why not fine-tuning the model with instruction-task patterns?**

→ **In instruction-tuned models zero-shot performance improved dramatically.**

E.g.,

- InstructGPT
- ChatGPT
- Claude
- Gemini
- Llama-Instruct

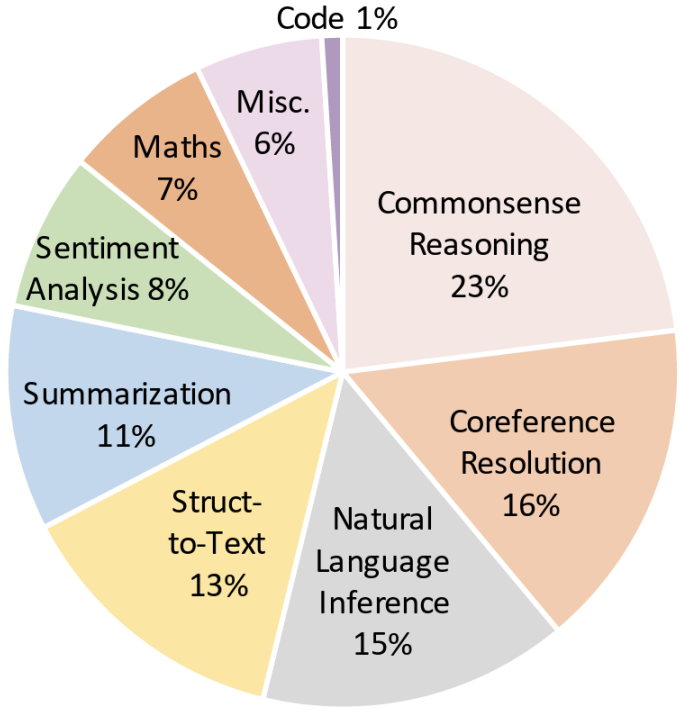
Note: in Retrieval-Augmented Generation (RAG) the model receives relevant documents at inference time. This complements zero-shot or few-shot prompting.

Supervised Multi-Task Learning in LLMs

unibz

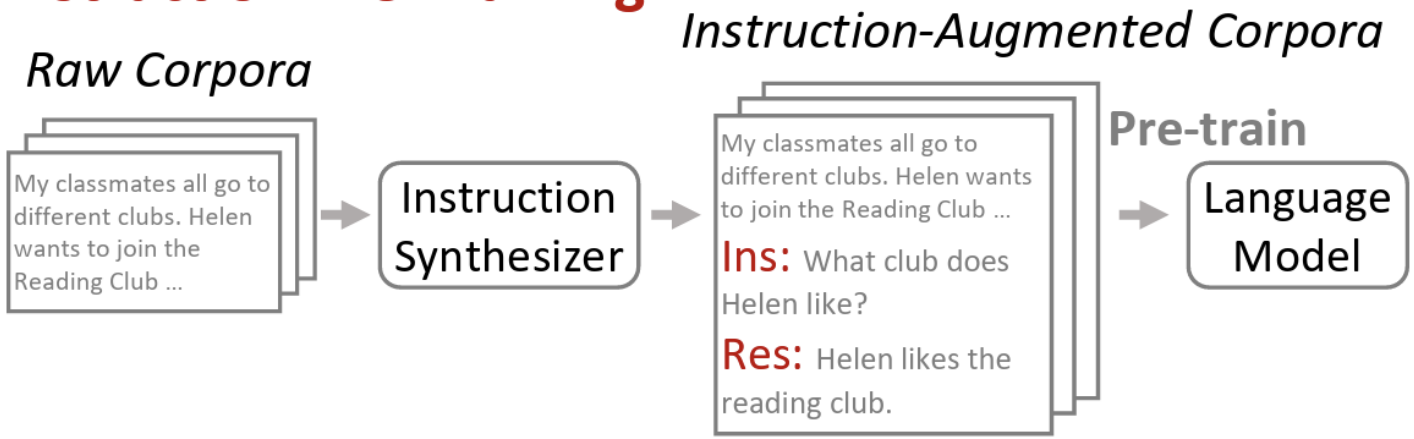
- "Instruction Pre-Training: Language Models are **Supervised** Multitask Learners", [D. Cheng et al., EMNLP 2024]

task = instruction-response pair



Vanilla Pre-Training:

Instruction Pre-Training:



Instruction Pre-Training data augmentation via an instruction synthesizer

LLM Emerging Capabilities

From Scaling to Understanding

Foundation models acquire broad latent abilities from large-scale pre-training.

- These abilities are learnt from signals spanning many implicit and explicit tasks.
- Adaptation specialises and activates capabilities for downstream tasks.

How do some capabilities emerge only after fine-tuning or instruction tuning?

- Current practice remains largely empirical and recipe-driven.

- Multi-task learning can provide a principled framework to study:
  - ability learning and capability emergence,
  - knowledge transfer and task interference,
  - specialisation and generalisation.

- Future progress may require integrating MTL with insights from cognitive development, pedagogy, and learning sciences (e.g., curriculum learning).

Multi-Task/Multi-Objective Optimisation

unibz

- MTL is intrinsically a multi-objective problem
  - different task objectives may be consistent or may conflict with each other.
    1. minimisation of a single objective function as linear combination
    2. trade-off of many objective functions

- Optimisation approaches
  - Using prior knowledge, then **Bayesian optimisation** methods can be applied:
    - knowledge transfer/sharing for a faster automatic hyperparameter optimization in ML
    - can improve the efficiency of training and the generalization capability of models
  - No prior knowledge: **Evolutionary approaches** for a unified search space for all tasks.
    - modes of knowledge transfer include shared genetic makeup, direct genetic crossover, other methods w/o direct solution crossover.
  - Find a trade-off using **Pareto Multi-Objective Optimisation**
    - Pareto optimal solutions (Pareto front)

G. Di Fatta, G. Nicosia, V. Ojha, P. Pardalos, Multi-Task Deep Learning as Multi-Objective Optimization, Encyclopedia of Optimization, 1-10, Springer, 2023.

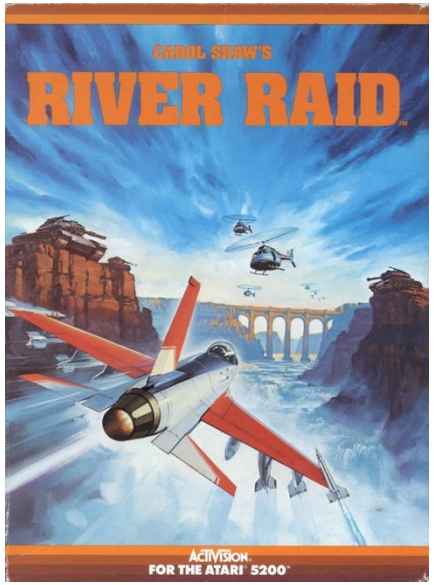
Pareto Multi-Task Deep Learning

unibz

- Population-based algorithms, such as Neuroevolution, can be naturally extended to multi-task learning.
  - Use ES to concurrently optimise many tasks and many objectives with a single DNN model.
- **Multi-Task Multi-Objective Deep Neuroevolution with a Pareto optimisation approach**
  - Tasks selected from related Atari 2600 games
  - Prior knowledge used to define multiple utility functions
  - Analysis of the underlying training dynamics with standard techniques and with the Hypervolume indicator and the Kullback-Leibler divergence
- **Results: a single model trained on multiple games outperforms models trained on individual games.**

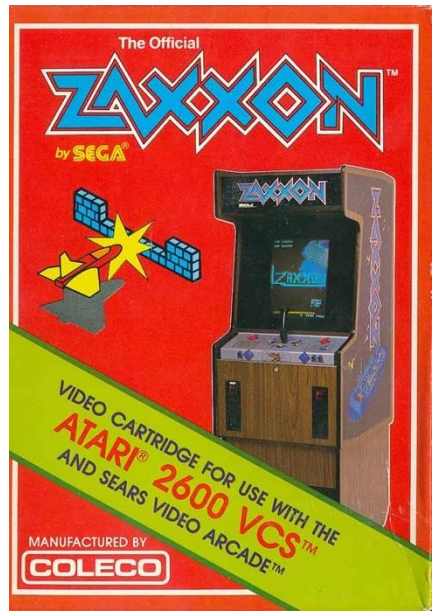
D. Dyankov, S. Riccio, G. Di Fatta, G. Nicosia, Multi-task learning by pareto optimality, LOD 2019

Experimental Analysis on Two Atari Games



- Network architecture:
  - input layer, three convolutional layers, a fully connected layer, a fully connected output layer (18 actions)

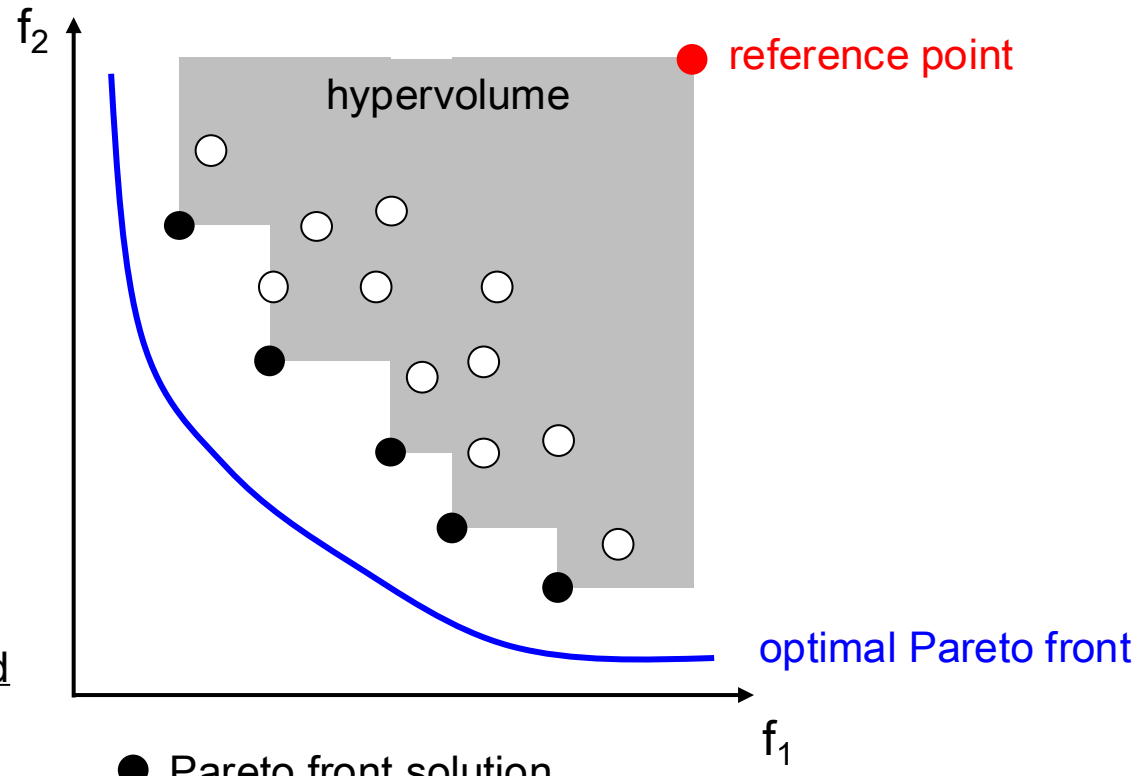
- Two games with similar dynamics:
  - River Raid and Zaxxon



Pareto Front and Hypervolume

- The **Pareto front** is the set of solutions {θ'} not dominated by any other solution:

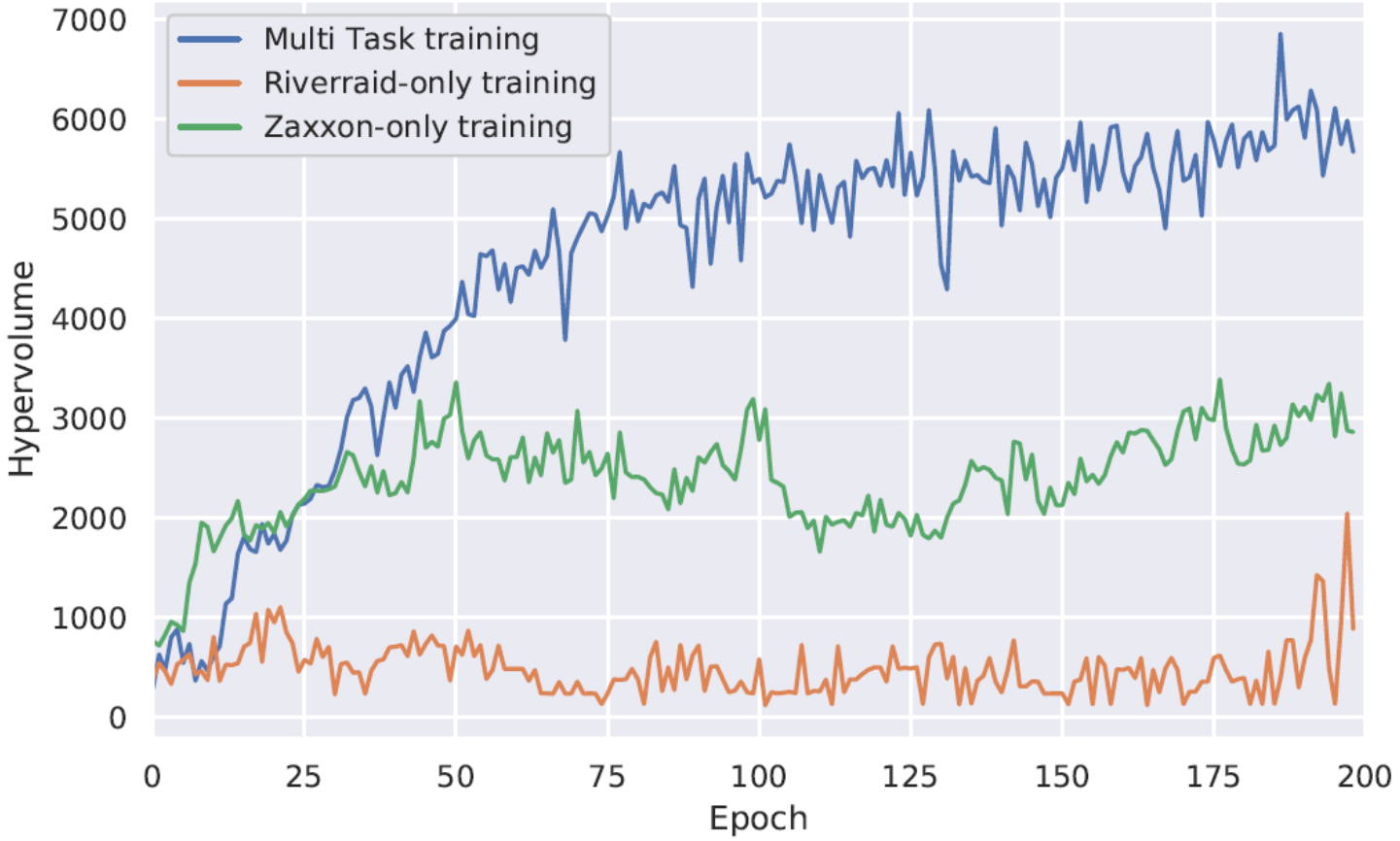
$$P(\Theta) = \{\theta' \in \Theta \mid \nexists \theta \in \Theta : \theta \preceq \theta'\}$$



f_1 and f_2 are objective functions to be minimised

Pareto Front Hypervolume

- The Pareto front obtained with multi-task ES covers a larger area than the two single-task single-objective networks.
  - The new algorithm is finding strategies able to master both tasks at the same time.

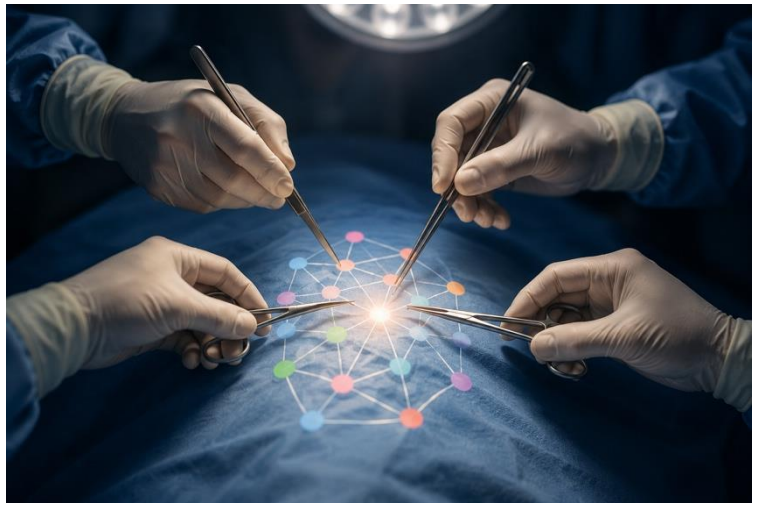


S. D. Riccio, D. Dyankov, G. Jansen, G. Di Fatta, and G. Nicosia, Pareto Multi-Task Deep Learning, ICANN 2020.

Gradient Conflicts and Negative Transfer

Different tasks may generate conflicting gradients during training:

- one task may improve while another degrades,
- optimisation becomes unstable,
- shared representations can suffer from negative transfer.



➢ **How can multiple objectives be concurrently optimised without destructive interference?**

Gradient Surgery Methods	Main Idea
Uncertainty Weighting (Kendall et al., 2018)	Adapt task weights from predictive uncertainty
GradNorm (Chen et al., 2018)	Balance gradient magnitudes dynamically
PCGrad (Yu et al., 2020)	Project conflicting gradients to reduce interference
CAGrad (Liu et al., 2021)	Conflict-averse multi-objective optimisation
ConFIG (Saha et al., 2025)	Compute conflict-free updates aligned with all objectives
SAM-GS (Borsani et al., 2025)	Similarity-aware gradient surgery with momentum modulation

Angle-based Gradient Conflicts in PCGrad

In angle-based conflicts a simple sum of task-specific gradients leads to suboptimal training.

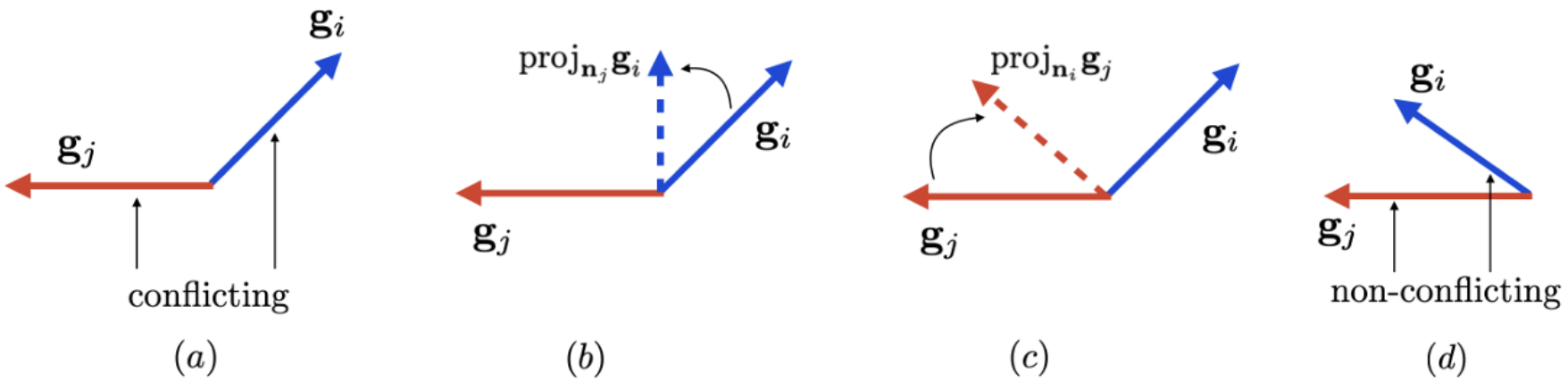
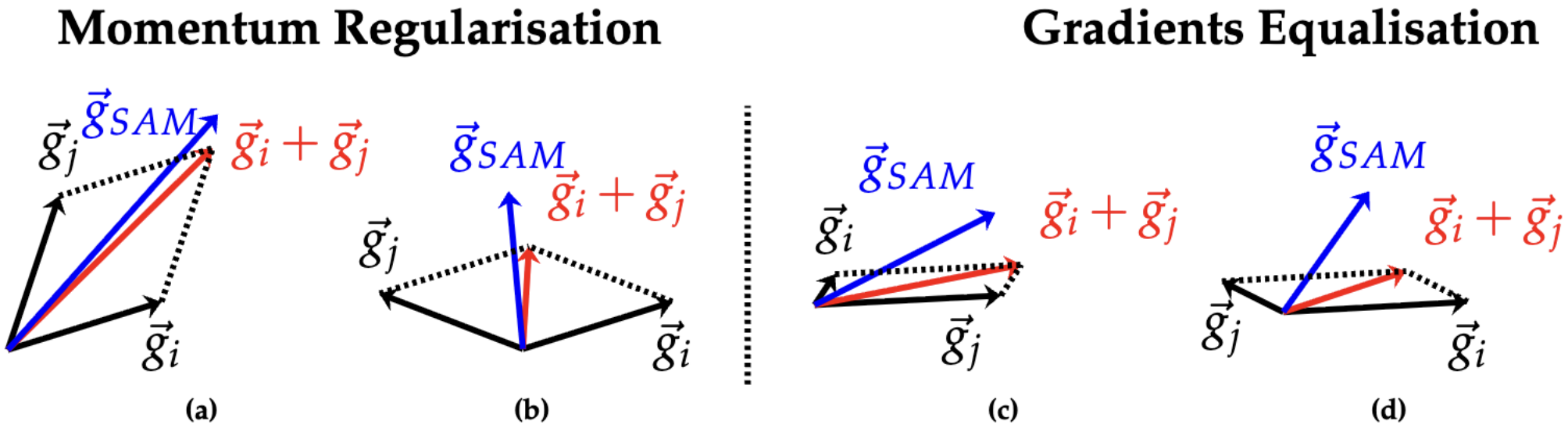


Image source: Yu et al. (2020) PCGrad

Similarity-Aware Momentum Gradient Surgery (SAM-GS)

- In MTL task gradients may have conflicting directions and/or different magnitudes, causing task dominance and interference, thus degrading the training process.

➢ SAM-GS solution: momentum regularisation and gradient equalisation



T. Borsani, A. Rosani, G. Nicosia, G. Di Fatta, **Gradient Similarity Surgery in Multi-Task Deep Learning**, ECML-PKDD 2025.

2D Toy Problem: 2 local and 1 global minima

- The tragic triad: Conflicting Gradients, Dominating Gradients, High Curvature

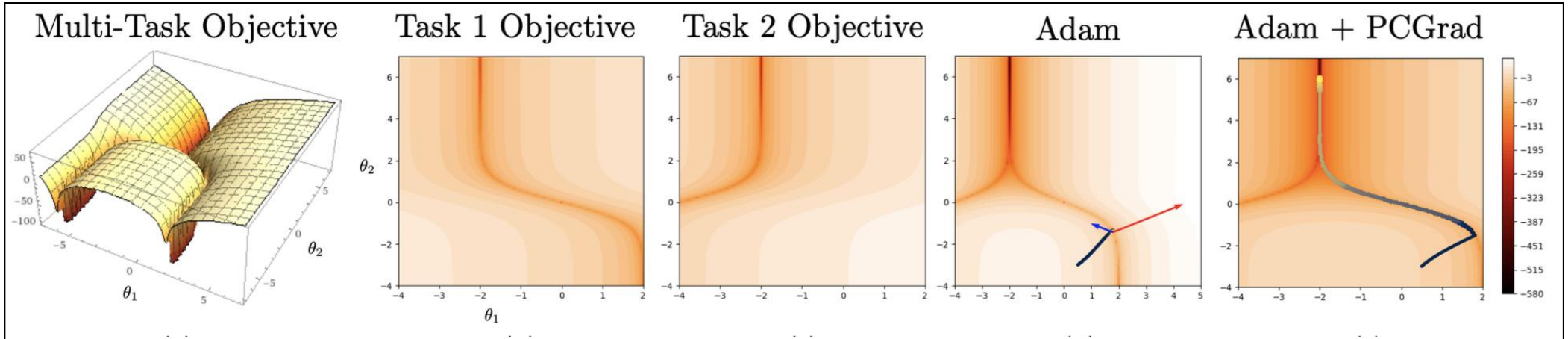
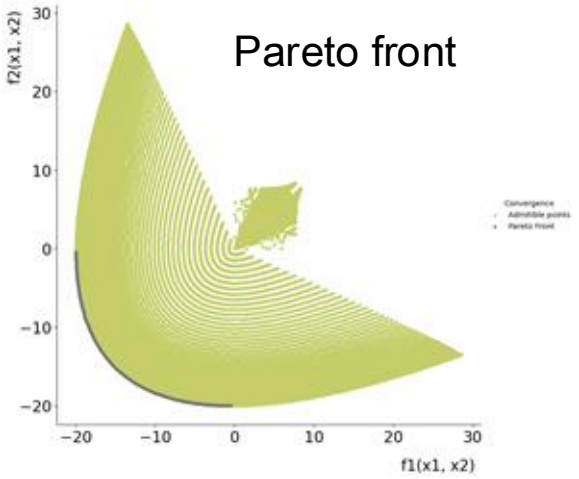
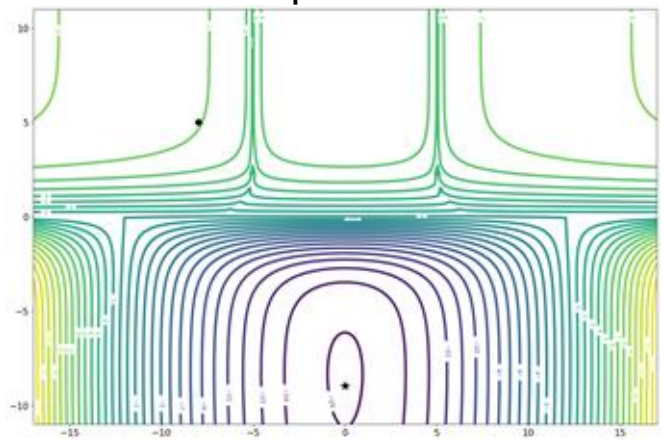


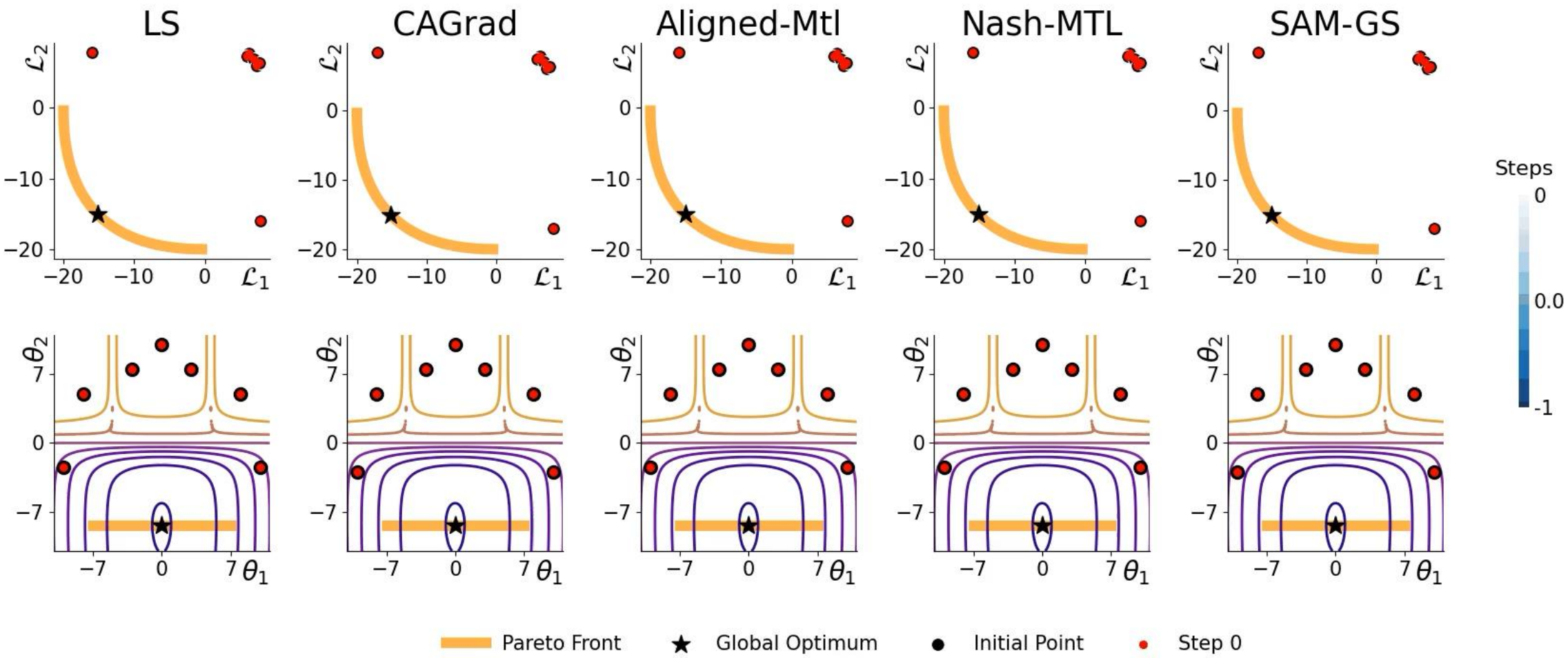
Image from PCGrad - T. Yu et al., - Gradient Surgery for Multi-Task Learning, NIPS 2020





T. Borsani, A. Rosani, G. Nicosia, G. Di Fatta, **Gradient Similarity Surgery in Multi-Task Deep Learning**, ECML 2025

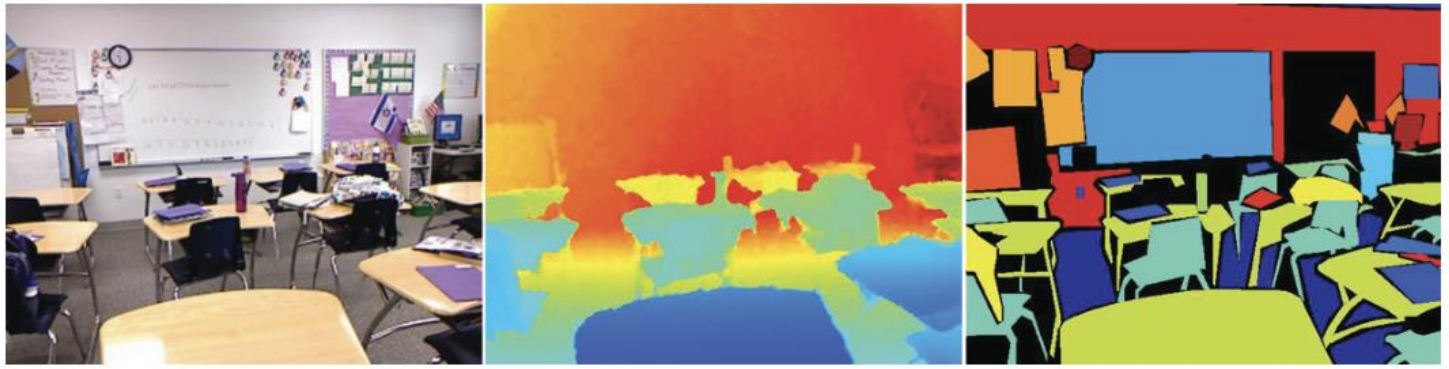
SAM-GS (2025)



T. Borsani, A. Rosani, G. Nicosia, G. Di Fatta, **Gradient Similarity Surgery in Multi-Task Deep Learning**, ECML 2025

NYU-V2 Dataset: Three Tasks

Environment	Indoor scenes
Sensor	Microsoft Kinect (RGB + Depth)
Labeled RGB-D images	1,449
Indoor scenes	464
Unlabeled frames	407,024
Annotations	Semantic labels, instance labels, depth maps
Resolution	640 × 480 pixels



Three common tasks on the NYU-Depth V2 (NYU-V2) dataset:

- **Semantic Segmentation**: assign a semantic class label to every pixel in an image (e.g., wall, floor, bed, chair, table, sofa, window, etc.) → segmentation map
- **Depth Estimation**: predict the distance from the camera to every pixel in the image → depth map
- **Surface Normal Computation**: estimate the orientation of surfaces at every pixel, 3D unit vectors → normal map

Together, these three tasks provide a comprehensive understanding of indoor scenes by combining semantic information (segmentation) and geometric information (depth and surface normals).

NYU-V2 Dataset Results

unibz

Table 1: NYU-V2 results

	Segmentation		Depth		Surface Normal						
					Angle Dist ↓		Within t° ↑				
	mIoU↑	Pix Acc ↑	Abs Err ↓	Rel Err ↓	Mean	Median	11.25	22.5	30	**MR** ↓	$\Delta m\%$ ↓
Single-Task Learning (STL) STL	38.3	63.76	0.6754	0.278	25.01	19.21	30.14	57.2	69.15		
LS	39.29	65.33	0.5493	0.2263	28.15	23.96	22.09	47.5	61.08	11.4	5.59
SI	38.45	64.27	0.5354	0.2201	27.6	23.37	22.53	48.57	62.32	10.3	4.39
RLW	37.17	63.77	0.5759	0.241	28.27	24.18	22.26	47.05	60.62	13.8	7.78
DWA	39.11	65.31	0.551	0.2285	27.61	23.18	24.17	50.18	62.39	10.2	3.57
UW	36.87	63.17	0.5446	0.226	27.04	22.61	23.54	49.05	63.65	10.0	4.05
MGDA	30.47	59.9	0.607	0.2555	24.88	19.45	29.18	56.88	69.36	7.4	1.38
PCGRAD	38.06	64.64	0.555	0.2325	27.41	22.8	23.86	49.83	63.14	10.6	3.97
GradNorm	20.09	64.64	0.7200	0.2800	**24.83**	**18.86**	**30.81**	**57.94**	**69.73**	7.2	7.22
GradDrop	39.39	65.12	0.5455	0.2279	27.48	22.96	23.38	49.44	62.87	9.6	3.58
CAGrad	39.79	65.49	0.5486	0.225	26.31	21.58	25.61	52.36	65.58	7.1	0.2
IMTL-G	39.35	65.6	0.5426	0.2256	26.02	21.19	26.2	53.13	66.24	6.3	-0.76
Nash-MTL	40.13	65.93	0.5261	0.2171	25.26	20.08	28.4	55.47	68.15	4.2	-4.04
FAMO	38.88	64.9	0.5474	0.2194	25.06	19.57	29.21	56.61	68.98	4.8	-4.1
Aligned-MTL	**40.82**	66.33	0.5300	0.2200	25.19	19.71	28.88	56.23	68.54	3.6	-4.93
Similarity-Aware Momentum Gradient Surgery (SAM-GS) SAM-GS	40.79	**66.46**	**0.5251**	**0.2169**	25.03	19.65	29.26	56.35	68.78	**2.4**	**-5.3**

Theoretical Bound of the Generalisation Error

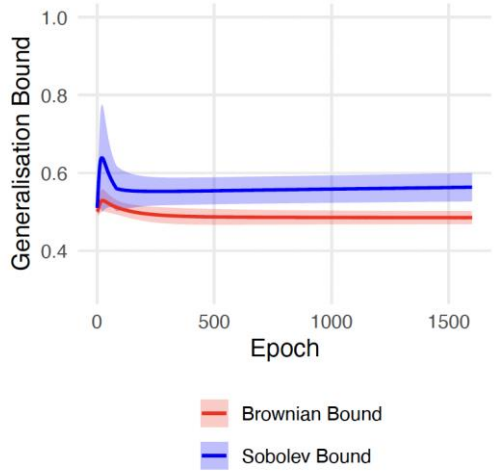
- MTL shared representations often improve accuracy and data efficiency.
  - Existing bounds of the generalisation error typically depend on network size and depth.
  - These bounds are often too loose to explain the strong performance observed in practice.

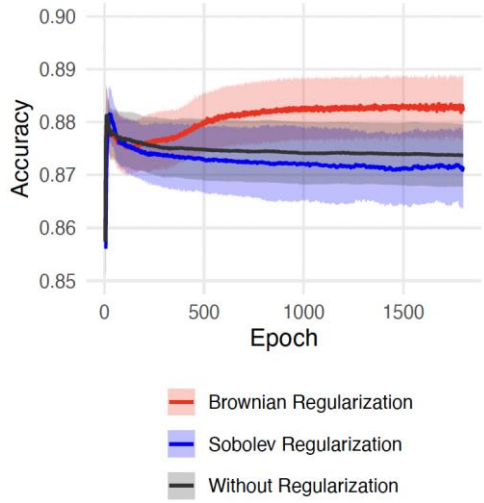
- **Key Idea**
  - Represent neural network layers as mathematical operators.
  - Analyse information propagation through the network rather than only weight magnitudes.
  - Use Koopman-operator-based analysis to obtain tighter generalisation guarantees.
- **Main Outcomes**
  - First Koopman-based generalisation framework for multi-task deep neural networks.
  - Derived tighter bounds than existing approaches by exploiting task relationships and network structure. It has advantages such as scalability and reduced dependence on network width.
  - Improved understanding of why multi-task networks generalise well.

M. Mohammadigohari, G. Di Fatta, G. Nicosia, P.M. Pardalos,
"On the Koopman-Based Generalization Bounds for Multi-Task Deep Learning", LOD 2025.

Bounds of the Generalisation Error - Comparison

Authors	Rate
Neyshabur et al., 2015	$\frac{2^L \prod_{j=1}^{L} \|W_j\|_F}{\sqrt{n}}$
Golowich, Rakhlin, and Shamir, 2018	$\left(\prod_{j=1}^{L} \|W_j\|_F\right) \min\left\{n^{-1/4}, \sqrt{\frac{L}{n}}\right\}$
Ju, Li, and Zhang, 2022	$\frac{\sum_{j=1}^{L} \theta_j \|W_j\|_F}{\sqrt{n}}$
Single-task Koopman bound Hashimoto et al., 2024	$\frac{\|g\|_{H_L}}{\sqrt{n}} \prod_{j=1}^{L} \frac{G_j E_j \|W_j\|^{s_j - 1}}{\det\left(W_j^\top W_j\right)^{1/4}}$
Multi-task Koopman bound	$\sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n}} \|g\|_{H^{sL}(\mathbb{R}^d, \mathbb{R}^m)} \prod_{j=1}^{L} \frac{G'_j E'_j \|W_j\|^{s_j - 1}}{\det\left(W_j^\top W_j\right)^{1/4}}$
improved bounds	
Improved theoretical bound	$\sqrt{\frac{\kappa \operatorname{Tr}(\mathbf{M})}{n}} \|g\|_{H^{(\mathrm{B})}(\mathcal{X}, \mathbb{R}^m)} \prod_{j=1}^{L} \frac{G'_j E'_j \|\mathbf{W}_j\|}{\det\left(W_j^\top W_j\right)^{1/4}}$





M. Mohammadigohari, G. Di Fatta, G. Nicosia, P.M. Pardalos,
"On the Koopman-Based Generalization Bounds for Multi-Task Deep Learning", LOD 2025.

Conclusions

unibz

- Machine Learning and weak AI have achieved remarkable success on isolated tasks; MTL pushes toward **more general learning architectures**.
- Transfer learning has become a cornerstone of modern AI, enabling knowledge acquired from large-scale datasets for many downstream tasks.
  - The widespread adoption of **pre-trained models** has made **transfer learning** a key enabler of industrial-scale deployment in vision, speech, and language.

- Multi-Task Learning goes beyond transferring features: it seeks to exploit the structure and relationships among tasks during learning.
  - Single-task learning discovers and exploits structure in the data.
  - Multi-task learning discovers and exploits structure in the tasks.

References

- Caruana (1997) — Multitask Learning
- Ruder (2017) — An Overview of Multi-Task Learning in Deep Neural Networks
- Misra et al. (2016) — Cross-Stitch Networks for Multi-task Learning
- Ruder et al. (2017) — Sluice Networks
- Ma et al. (2018) — Multi-gate Mixture-of-Experts for Multi-task Learning
- Kendall et al. (2018) — Multi-Task Learning Using Uncertainty to Weigh Losses
- Chen et al. (2018) — GradNorm
- Sener & Koltun (2018) — Multi-Task Learning as Multi-Objective Optimization
- Zamir et al. (2018) — Taskonomy
- A. Radford et al., 2019 — Language Models are Unsupervised Multitask Learners
- Dyankov et al. (2019) — Multi-task learning by pareto optimality
- Brown et al. (2020) — Language Models are Few-Shot Learners
- Riccio et al. (2020) — Pareto multi-task deep learning
- Standley et al. (2020) — Which Tasks Should Be Learned Together?
- Vandenhende et al. (2020) — Multi-Task Learning with Deep Neural Networks: A Survey
- Yu et al. (2020) — Gradient Surgery for Multi-Task Learning (PCGrad)
- Liu et al. (2021) — Conflict-Averse Gradient Descent (CAGrad)
- TaskPrompter (2023) — Prompt-based Multi-Task Dense Prediction
- Yu et al. (2024) — Unleashing the Power of Multi-Task Learning
- MTLoRA (2024) — Parameter-Efficient Multi-Task Learning with LoRA
- Cheng et al., EMNLP 2024, Instruction Pre-Training: Language Models are Supervised Multitask Learners
- Liu et al. (2025) — ConFIG: Towards Conflict-free Training of Physics Informed Neural Networks
- Borsani et al. (2025) — SAM-GS: Similarity-Aware Momentum Gradient Surgery for Multi-Task Deep Learning