Multi-Task Learning with Deep Neural Networks: A Survey

Michael Crawshaw
Department of Computer Science
George Mason University
mcrawsha@gmu.edu

SURVEY PAPER PRESENTATION

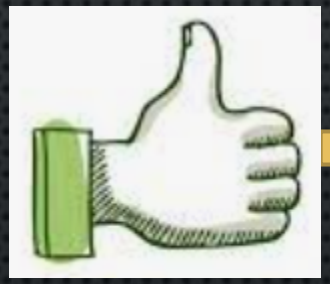
ABHISHEK BAIS
GRADUATE STUDENT, SOFTWARE ENGINEERING, SJSU

When a baby learns to walk, it acquires and assimilates general motor skills that it augments, and uses later in life to perform more complex tasks such as playing soccer

MULTI-TASK LEARNING REFLECTS HUMAN LEARNING PROCESS MORE CLOSELY THAN THE SINGLE-TASK LEARNING PROCESS

MORE FORMALLY

MULTI-TASK LEARNING IS A SUBFIELD OF MACHINE LEARNING IN WHICH MULTIPLE TASKS ARE SIMULTANOUSLY LEARNED



Advantages

- IMPROVED DATA EFFICIENCY
- REDUCED OVERFITTING THROUGH SHARED REPRESENTATIONS
- FASTER LEARNING BY LEVERAGING AUXILLARY INFORMATION

Challenges

- CHOSING TASKS TO BE LEARNT TOGETHER IS NON-TRIVIAL

WHAT DOES THE PAPER FOCUS ON?

Multi-Task Learning Architectures

Multi-Task Learning Optimization Methods

Multi-Task Relationship Learning

TASK DOMAIN	Focus on domain specific tasks Example: Computer Vision, NLP
MULTI MODAL	Handle tasks with input in more than one mode Example: Visual Question Answering
LEARNED	Learn architecture or weights of shared model Example Branched sharing, Modular sharing
CONDITIONAL	Selects parts of NN dynamically based on inputs to network Example: Neural Module Networks

MULTI-TASK LEARNING ARCHITECTURES

Hard Parameter (module weight) sharing architectures

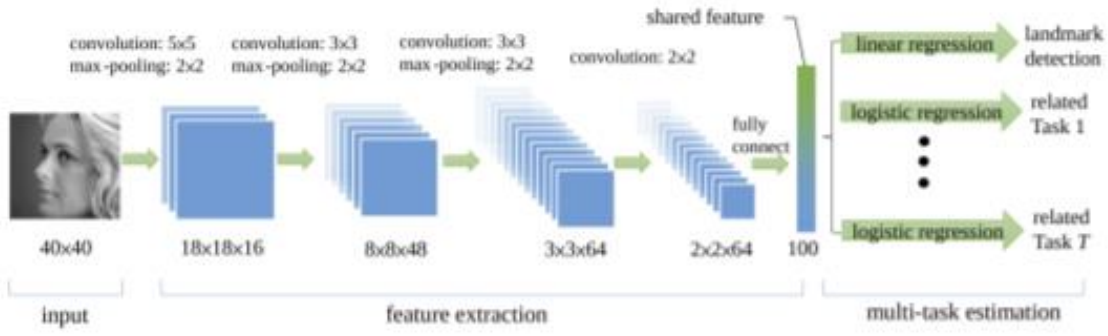


Figure 1: Architecture for TCDCN (Zhang et al., 2014). The base feature extractor is made of a series of convolutional layers which are shared between all tasks, and the extracted features are used as input to task-specific output heads.

TASK DOMAIN – COMPUTER VISION

Focus on partitioning network into task specific shared components in a way that allows generalization via shared information flow between tasks

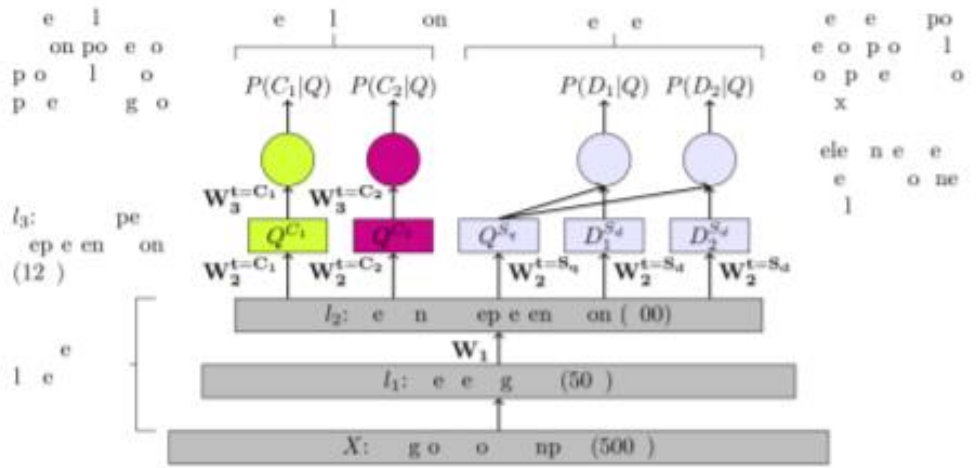


Figure 6: Network architecture of (Liu et al., 2015a). The input is converted to a bag-of-words representation and hashed into letter 3-grams, followed by a shared linear transformation and nonlinear activation function. This shared representation is passed to task-specific outhead heads to compute final outputs for each task.

TASK DOMAIN – NLP

Well suited for Multi-Task Learning due to abundance of related questions one can ask about a given piece of text, and task agnostic representations used in NLP

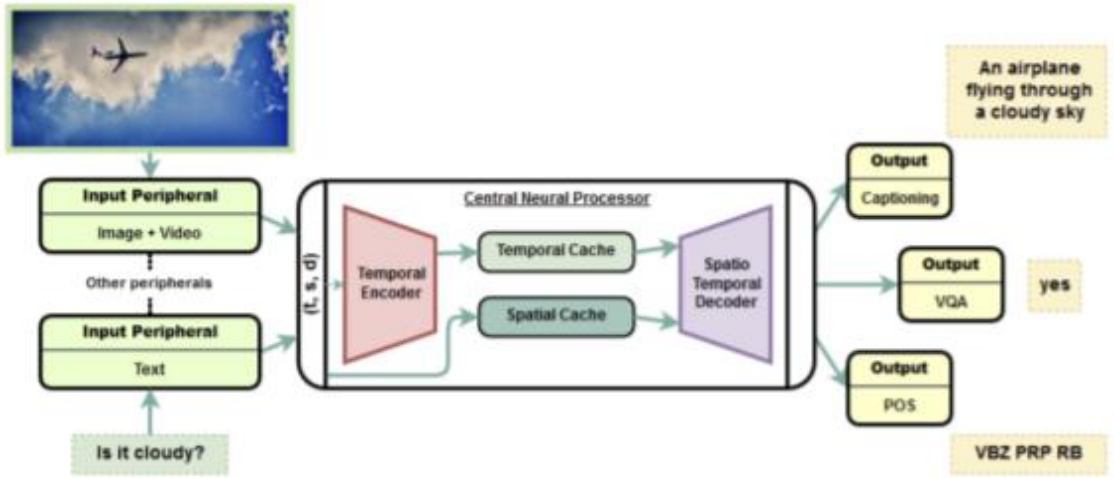


Figure 12: OmniNet architecture proposed in (Pramanik et al., 2019). Each modality has a separate network to handle inputs, and the aggregated outputs are processed by an encoder-decoder called the Central Neural Processor. The output of the CNP is then passed to several task-specific output heads.

MULTI MODAL

Handle tasks using data from multiple domains, such as vision and linguistic data for visual QA

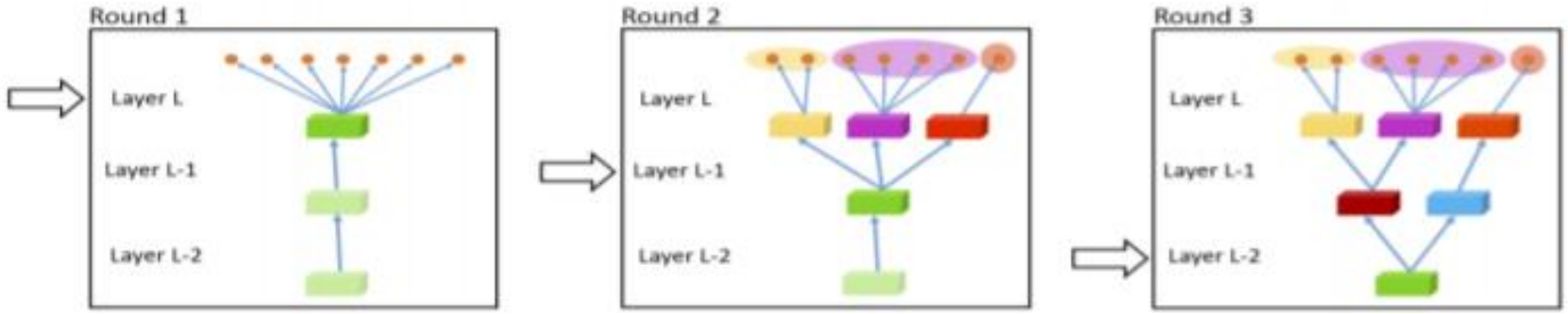


Figure 13: Learned branching architecture proposed in (Lu et al., 2017). At the beginning of training, each task shares all layers of the network. As training goes on, less related tasks branch into clusters, so that only highly related tasks share as many parameters.

LEARNED – BRANCHED SHARING

Coarse grain way to share parameters between tasks.
Once the computation graph for 2 tasks differ, they never rejoin

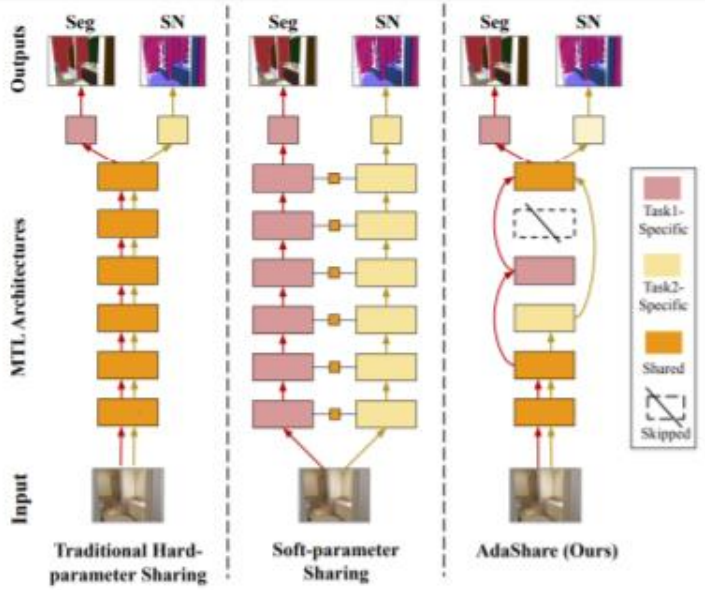


Figure 16: A learned parameter sharing scheme with AdaShare (Sun et al., 2019b). Each layer in the network is either included or ignored by each task, so that each task uses a subnetwork which is (likely) overlapping with other tasks.

LEARNED – MODULAR SHARING

Fine-grained approach where set of NN modules shared between tasks

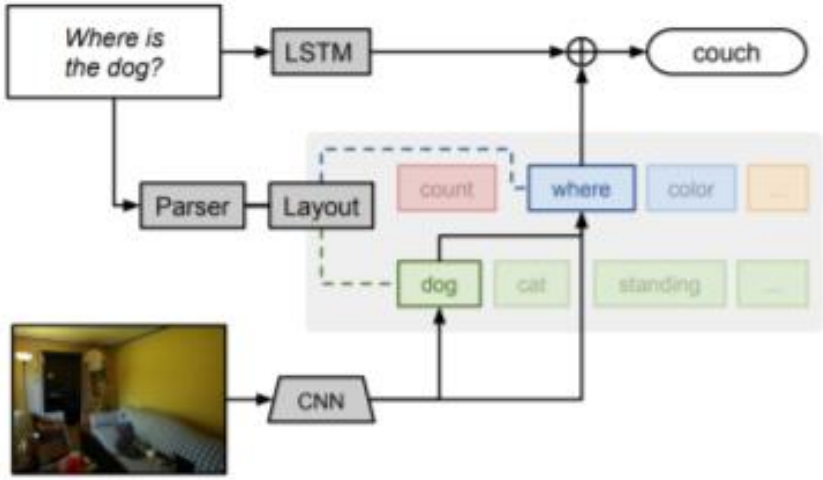


Figure 18: Example Neural Module Network execution (Andreas et al., 2016). The semantic structure of a given question is used to dynamically instantiate a network made of modules that correspond to the elements of the question.

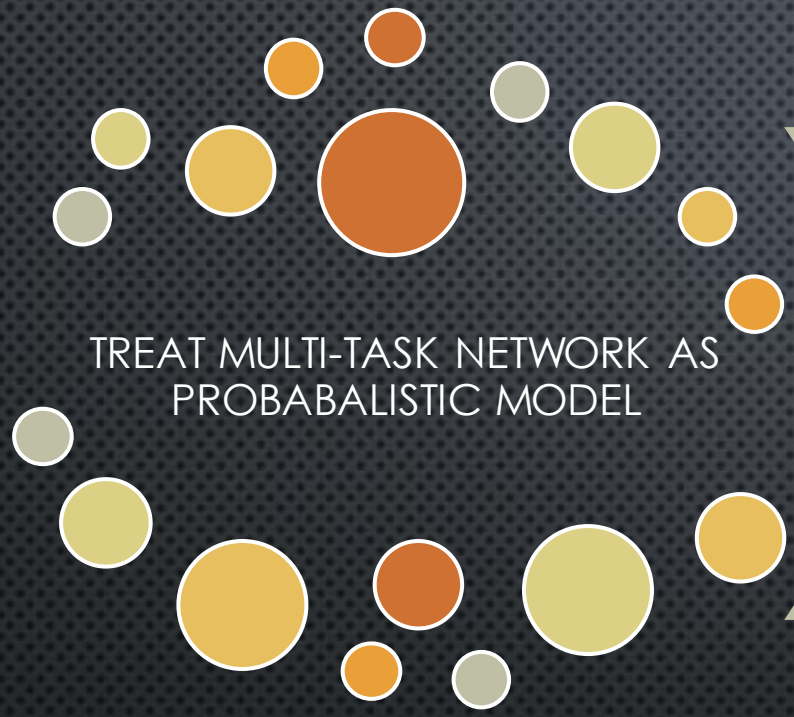
CONDITIONAL – NEURAL MODULE NETWORKS

Dynamic Architectures that select parts of NN for execution depending on input to the network

LOSS WEIGHTING	Aggregate loss functions from individual tasks into a single multi-task weighted loss function E.g., Kendall et.al., 2017
TASK SCHEDULING	Choose which task(s) to train at each training step E.g., Sharma et. al, 2017
GRADIENT MODULATION	Alleviate negative transfer i.e., conflicting task gradients E.g., GREAT by Sinha et. al., 2018
KNOWLEDGE DISTILLIATION	Distills knowledge from multiple single-task teachers to single multi-task student E.g., Distral framework by The et. al, 2017

MULTI-TASK OPTIMIZATION METHODS

Soft Parameter Sharing architectures that regularize model parameters by penalizing them based on distance from other parameters



DERIVE WEIGHTED MULTI-TASK LOSS FUNCTION BY MAXIMIZING LIKELIHOOD OF GROUND TRUTHS

Kendal et. al., 2017

LOSS WEIGHTING – BY UNCERTAINTY

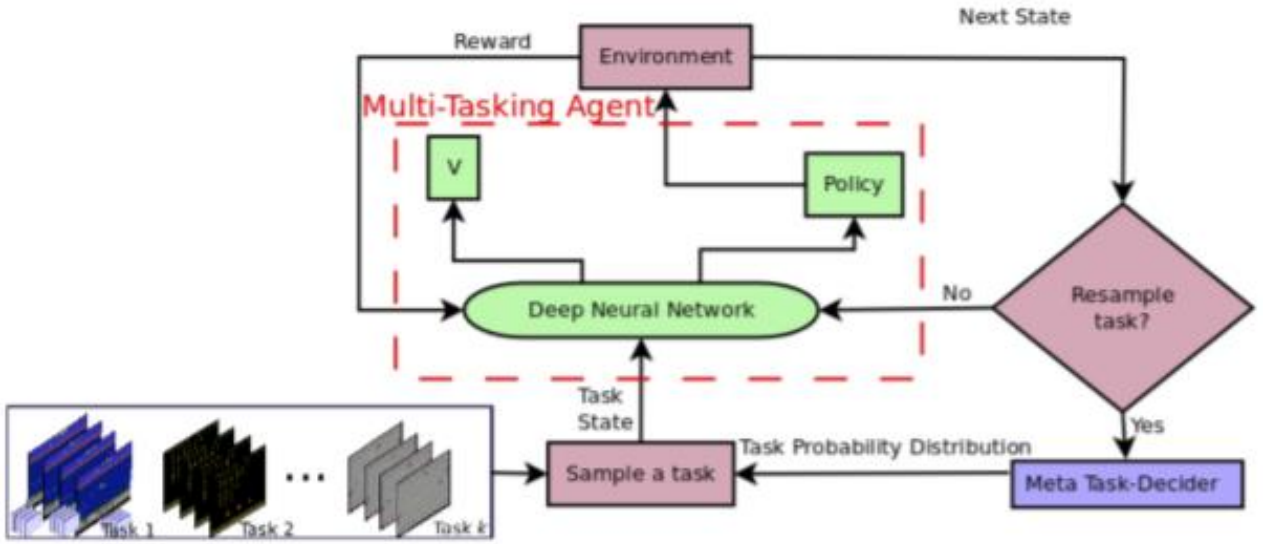


Figure 20: Task scheduling visualization from (Sharma et al., 2017). A meta task-decider is trained to sample tasks with a training signal that encourages tasks with worse relative performance to be chosen more frequently.

Sharma et. al., 2017

TASK SCHEDULING

Is a process of choosing which task to train at each training step. Common approaches are train all tasks or choose a random sample of tasks to train

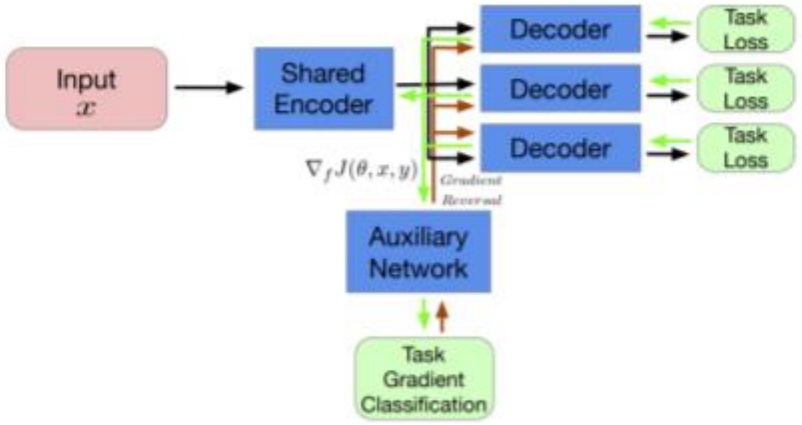


Figure 21: Multi-task GREAT model (Sinha et al., 2018). An auxiliary network takes a gradient vector for a single task's loss and tries to classify which task the gradient vector came from. The network gradients are then modulated to minimize the performance of the auxiliary network, to enforce the condition that gradients from different task functions have statistically indistinguishable distributions.

GREAT by Sinha et. al., 2018

GRADIENT MODULATION

Helps reduce "negative transfer". This happens when 2 tasks have gradients in opposite direction so following the gradient of 1 task, leads to decreased performance of the other task. GREAT ensures gradients of related tasks point in the same direction

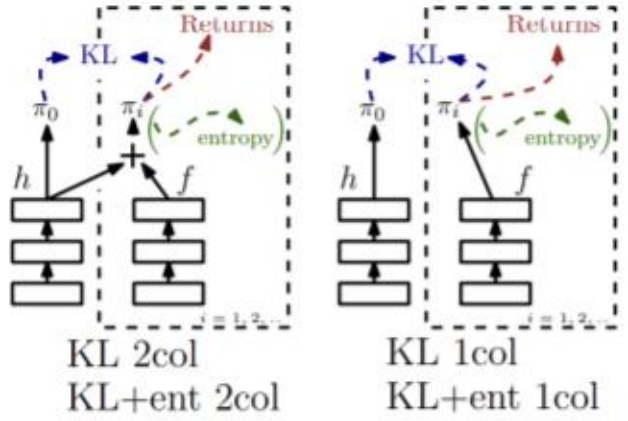


Figure 22: Two architectures from the Distral framework for RL (Teh et al., 2017). On the left is an architecture which employs both of the main ideas behind Distral: KL-regularization of single-task policies with the multi-task policy and a two-column policy for each task, where one column is shared between all tasks. On the right is an architecture which only employs KL-regularization of the single-task policies.

Distral by Teh et. al., 2017

KNOWLEDGE DISTILLATION

Used to instill a single multi-task student network with knowledge of many individual single-task teacher networks. Students are seen to outperform Teachers. DISTRAL allows for symmetric information flow from Student to Teacher

GROUPING TASKS

Partitions group of tasks into clusters so they can be trained collectively

Example: Standley et.al., 2019

TRANSFER RELATIONSHIPS

Learns transfer relationship between tasks

Example: Taskonomy by Zamir et. al., 2018

MULTI-TASK RELATIONSHIP LEARNING

Focusses on learning explicit representations of tasks or relationships between tasks

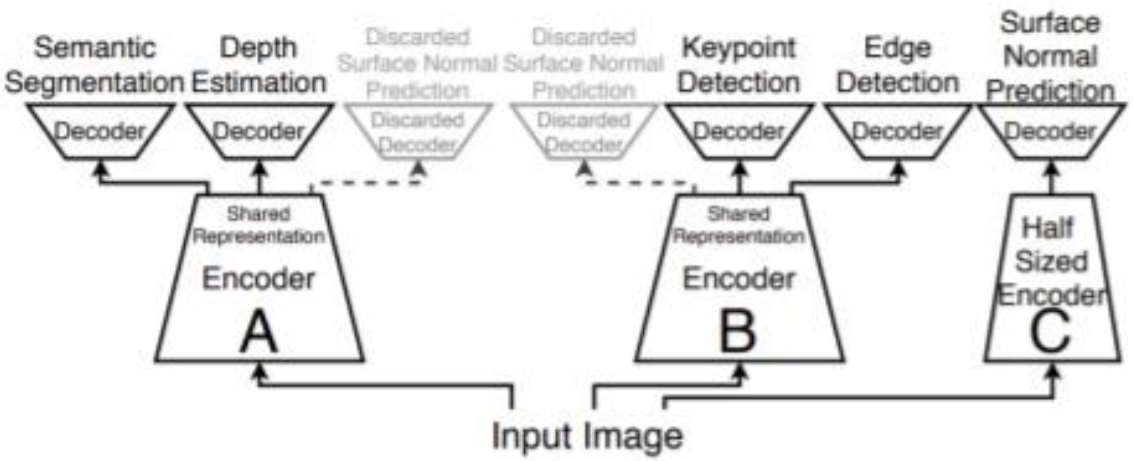


Figure 24: An example partitioning of a group of tasks into clusters with positive transfer (Standley et al., 2019).

Standley et. al., 2019

GROUPING TASKS

Provides an alternate way to avoid "negative transfer" by separating learning of 2 tasks that exhibit it from the start

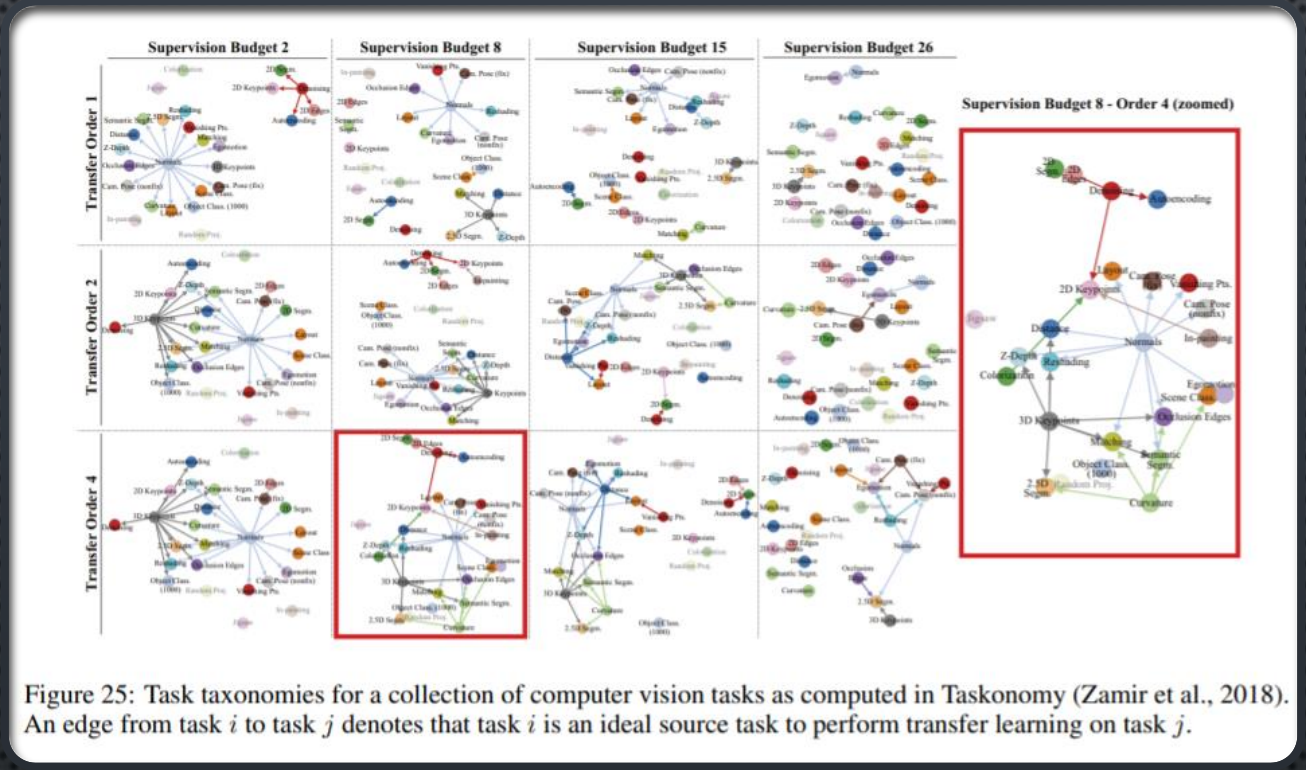


Figure 25: Task taxonomies for a collection of computer vision tasks as computed in Taskonomy (Zamir et al., 2018). An edge from task i to task j denotes that task i is an ideal source task to perform transfer learning on task j.

Zamir et. al., 2018 Taskonomy dataset with 4 million images labeled for 26 tasks

TRANSFER RELATIONSHIPS

Few (source tasks) learn from the full dataset; other tasks learn by transferring learning from the source tasks. E.g., Zamir proposed a Single-Task network to train each task, transfer relationships to other tasks

5.1 Computer Vision Benchmarks

- **NYU-v2** (Silberman et al., 2012) is a dataset of RGB-depth images from 464 indoor scenes with 1449 densely labeled images and over 400,000 unlabeled images. The labeled images are labeled for instance segmentation, semantic segmentation, and scene classification, and all images contain depth values for each pixel. All images are frames extracted from video sequences.

5.2 Natural Language Processing Benchmarks

Unless otherwise specified, it can be assumed that the text within a corpus is English.

- **Penn Treebank** (Marcus et al., 1993) is a corpus of text consisting of 4.5 million words. The text is aggregated from multiple sources including scientific abstracts, news stories, book chapters, computer manuals, and more, and contains Part-of-Speech tags and syntactical structure annotations.

5.3 Reinforcement Learning Benchmarks

- **Arcade Learning Environment** (Bellemare et al., 2013) (or ALE) is a diverse collection of hundreds of Atari 2600 games, where observations are given to the agent as raw pixels. These games were originally designed to be a challenge for the human video game player, so they present a challenge for modern RL agents in aspects such as exploration and learning with sparse rewards.

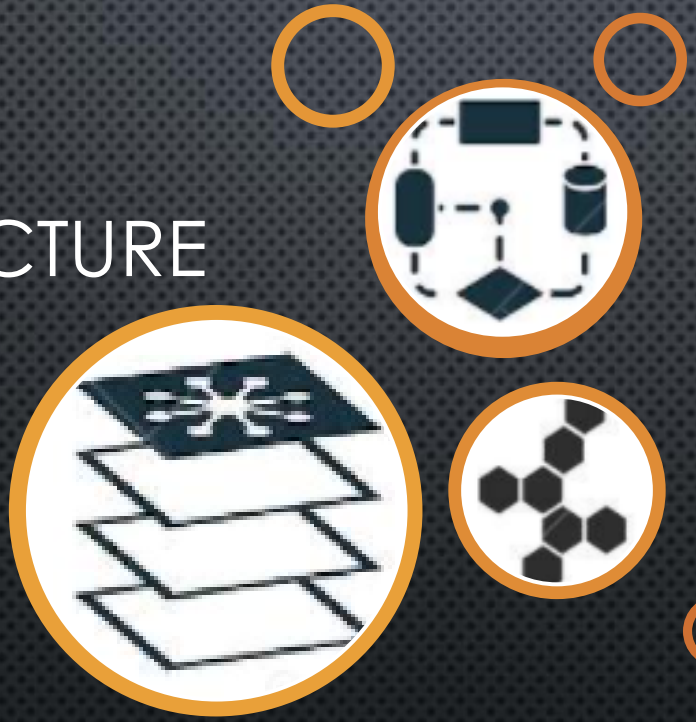
5.4 Multi-Modal Benchmarks

- **Flickr30K Captions** (Young et al., 2014) is a collection of 30,000 photographs obtained from the image hosting website Flickr, with over 150,000 corresponding captions.

COMMONLY USED MULTI-TASK BENCHMARKS

CONCLUSION

ARCHITECTURE



OPTIMIZATION METHODS

RELATIONSHIP LEARNING

DEVELOPMENTS IN MULI-TASK LEARNING IMPORTANT STEP TOWARDS DEVELOPING AI WITH MORE HUMAN LIKE QUALITIES