Meta Learning

MIT
Iddo Drori, Fall 2020

Multi-Task Learning (MTL) Progress and Motivation

Learning 57 Atari Games

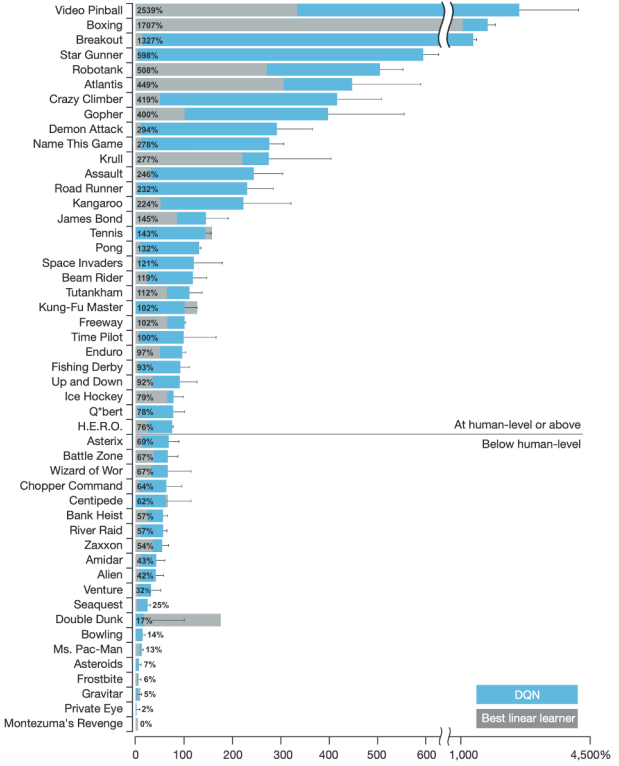
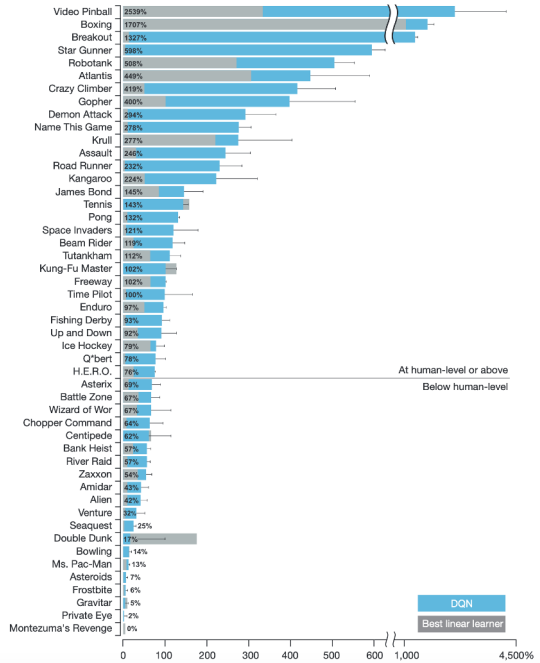


Figure source: Human-level control through deep reinforcement learning, Mnih et al, Nature 2015

Progress in Atari Games

2015

2018



Montezuma's revenge and pitfall were at random performance in 2015 and super human in 2018, all 57 games are at super-human performance in 2020

Figure source: Human-level control through deep reinforcement learning, Mnih et al, Nature 2015

Learning 57 Fields

Task	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM
Astronomy	Solar system, galaxies, asteroids, ...	STEM
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other
College Biology	Cellular structure, molecular biology, ecology, ...	STEM
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM
High School European History	Renaissance, reformation, industrialization, ...	Humanities
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences
High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
High School US History	Civil War, the Great Depression, The Great Society, ...	Humanities
High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
Management	Organizing, communication, organizational structure, ...	Other
Marketing	Segmentation, pricing, market research, ...	Other
Medical Genetics	Genes and cancer, common chromosome disorders, ...	Other
Miscellaneous	Agriculture, Fermi estimation, pop culture, ...	Other
Moral Disputes	Freedom of speech, addiction, the death penalty, ...	Humanities
Moral Scenarios	Detecting physical violence, stealing, externalities, ...	Humanities
Nutrition	Metabolism, water-soluble vitamins, diabetes, ...	Other
Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...	Humanities
Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...	Humanities
Professional Accounting	Auditing, reporting, regulation, valuation, ...	Other
Professional Law	Torts, criminal law, contracts, property, evidence, ...	Humanities
Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...	Other
Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...	Social Sciences
Public Relations	Media theory, crisis management, intelligence gathering, ...	Social Sciences
Security Studies	Environmental security, terrorism, weapons of mass destruction, ...	Social Sciences
Sociology	Socialization, cities and community, inequality and wealth, ...	Social Sciences
US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...	Social Sciences
Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...	Other
World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...	Humanities

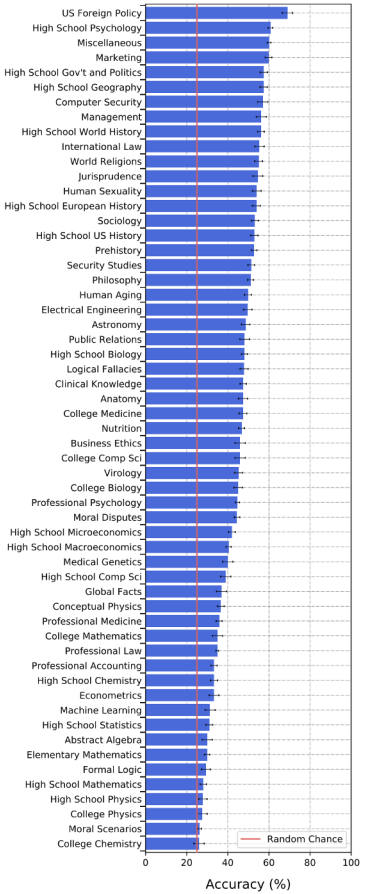
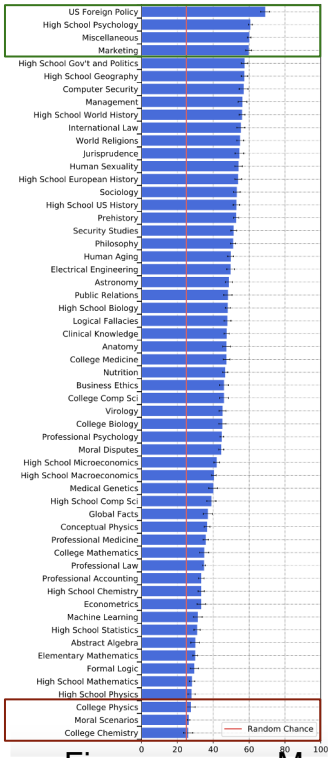


Figure source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 2020

Expected Progress in Learning 57 Fields

2020 → 2023



2020: Learning US Foreign policy performance is at 70%. College Chemistry and Physics are the hardest being slightly above random performance using GPT-3. Learning machine learning has slightly better performance.

Expected progress:
College Chemistry and Physics will be superhuman in 2023. All fields will be super-human in 2025.

Learning to learn courses is already happening.

Figure source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 2020

Abstract Algebra

Find all c in Z_3 such that Z_3[x]/(x^2 + c) is a field.,0,1,2,3,B

"Statement 1 | If aH is an element of a factor group, then |aH| divides |a|. Statement 2 | If H and K are subgroups of G then HK is a subgroup of G.","True, True","False, False","True, False","False, True",B

Statement 1 | Every element of a group generates a cyclic subgroup of the group. Statement 2 | The symmetric group S_10 has 10 elements.,"True, True","False, False","True, False","False, True",C

Statement 1| Every function from a finite set onto itself must be one to one. Statement 2 | Every subgroup of an abelian group is abelian.,"True, True","False, False","True, False","False, True",A

Find the characteristic of the ring 2Z.,0,3,12,30,A

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

Anatomy

What is the embryological origin of the hyoid bone?,The first pharyngeal arch,The first and second pharyngeal arches,The second pharyngeal arch,The second and third pharyngeal arches,D

Which of these branches of the trigeminal nerve contain somatic motor processes?,The supraorbital nerve,The infraorbital nerve,The mental nerve,None of the above,D

The pleura,have no sensory innervation.,are separated by a 2 mm space.,extend into the neck.,are composed of respiratory epithelium.,C

In Angle's Class II Div 2 occlusion there is,excess overbite of the upper lateral incisors.,negative overjet of the upper central incisors.,excess overjet of the upper lateral incisors.,excess overjet of the upper central incisors.,C

Which of the following is the body cavity that contains the pituitary gland?,Abdominal,Cranial,Pleural,Spinal,B

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

Astronomy

You are pushing a truck along a road. Would it be easier to accelerate this truck on Mars? Why? (Assume there is no friction),It would be harder since the truck is heavier on Mars.,It would be easier since the truck is lighter on Mars.,It would be harder since the truck is lighter on Mars.,It would be the same no matter where you are.,D

Where do most short-period comets come from and how do we know?,The Kuiper belt; short period comets tend to be in the plane of the solar system just like the Kuiper belt.,The Kuiper belt; short period comets tend to come from random directions indicating a spherical distribution of comets called the Kuiper belt.,The asteroid belt; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the asteroid belt.,The Oort cloud; short period comets tend to be in the plane of the solar system just like the Oort cloud.,A

Say the pupil of your eye has a diameter of 5 mm and you have a telescope with an aperture of 50 cm. How much more light can the telescope gather than your eye?,10000 times more,100 times more,1000 times more,10 times more,A

Why isn't there a planet where the asteroid belt is located?,A planet once formed here but it was broken apart by a catastrophic collision.,There was not enough material in this part of the solar nebula to form a planet.,There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.,Resonance with Jupiter prevented material from collecting together to form a planet.,D

Why is Mars red?,"Because the surface is covered with heavily oxidized (""rusted"") minerals.",Because the atmosphere scatters more light at bluer wavelengths transmitting mostly red light.,Because Mars is covered with ancient lava flows which are red in color.,Because flowing water on Mars's surface altered the surface minerals several billion years ago.,A

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

Massachusetts Institute of Technology

Business Ethics

"Beyond the business case for engaging in CSR there are a number of moral arguments relating to: negative _____, the _____that corporations possess and the _____ of business and society.","Externalities, Power, Independence","Publicity, Insubstantial resources, Mutual dependence","Publicity, Power, Independence","Externalities, Power, Mutual dependence",D

"_____ is the direct attempt to formally or informally manage ethical issues or problems, through specific policies, practices and programmes.",Corporate social responsibility,Business ethics management,Sustainability,Environmental management,B

"To ensure the independence of the non-executive board members, they are a number of steps which can be taken, which include non-executives being drawn from _____ the company, being appointed for a _____ time period as well as being appointed _____.","Outside, Limited, Independently","Inside, Limited, Intermittently","Outside, Unlimited, Intermittently","Inside, Unlimited, Independently",A

"Three contrasting tactics that CSO's can engage in to meet their aims are _____ which typically involves research and communication, _____, which may involve physically attacking a company's operations or _____, often involving some form of _____.","Non-violent direct action, Violent direct action, Indirect action, Boycott","Indirect action, Instrumental action, Non-violent direct action, Information campaign","Indirect action, Violent direct action, Non-violent direct-action Boycott","Non-violent direct action, Instrumental action, Indirect action, Information campaign",C

"In contrast to _____, _____ aim to reward favourable behaviour by companies. The success of such campaigns have been heightened through the use of _____, which allow campaigns to facilitate the company in achieving _____ .","Buycotts, Boycotts, Blockchain technology, Charitable donations","Buycotts, Boycotts, Digital technology, Increased Sales","Boycotts, Buyalls, Blockchain technology, Charitable donations","Boycotts, Buycotts, Digital technology, Increased Sales",D

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

Clinical Knowledge

The energy for all forms of muscle contraction is provided by:,ATP.,ADP.,phosphocreatine.,oxidative phosphorylation.,A

What is the difference between a male and a female catheter?,Male and female catheters are different colours.,Male catheters are longer than female catheters.,Male catheters are bigger than female catheters.,Female catheters are longer than male catheters.,B

In the assessment of the hand function which of the following is true?,Abduction of the thumb is supplied by spinal root T2,Opposition of the thumb by opponens policis is supplied by spinal root T1,Finger adduction is supplied by the median nerve,Finger abduction is mediated by the palmar interossei,B

How many attempts should you make to cannulate a patient before passing the job on to a senior colleague?,4,3,2,1,C

Glycolysis is the name given to the pathway involving the conversion of:,glycogen to glucose-1-phosphate.,glycogen or glucose to fructose.,glycogen or glucose to pyruvate or lactate.,glycogen or glucose to pyruvate or acetyl CoA.,C

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College Biology

Which of the following represents an accurate statement concerning arthropods?,They possess an exoskeleton composed primarily of peptidoglycan.,They possess an open circulatory system with a dorsal heart.,They are members of a biologically unsuccessful phylum incapable of exploiting diverse habitats and nutrition sources.,"They lack paired, jointed appendages.",B

"In a given population, 1 out of every 400 people has a cancer caused by a completely recessive allele, b. Assuming the population is in Hardy-Weinberg equilibrium, which of the following is the expected proportion of individuals who carry the b allele but are not expected to develop the cancer?",1/400,19/400,20/400,38/400,D

"The presence of homologous structures in two different organisms, such as the humerus in the front limb of a human and a bird, indicates that",the human and bird are polyphyletic species,a human's and bird's evolution is convergent,the human and bird belong to a clade,the human and bird developed by analogy,C

"According to the pressure-flow model of movement of phloem contents, photosynthate movement from source to sink is driven by",an ATP-dependent pressure-flow pump,a water-pressure potential gradient,transpiration,apoplastic diffusion,B

Which of the following contain DNA sequences required for the segregation of chromosomes in mitosis and meiosis?,Telomeres,Centromeres,Nucleosomes,Spliceosomes,B

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College Chemistry

Which of the following statements about the lanthanide elements is NOT true?,The most common oxidation state for the lanthanide elements is +3.,Lanthanide complexes often have high coordination numbers (> 6).,All of the lanthanide elements react with aqueous acid to liberate hydrogen.,The atomic radii of the lanthanide elements increase across the period from La to Lu.,D

A 0.217 g sample of HgO (molar mass = 217 g) reacts with excess iodide ions according to the reaction shown above. Titration of the resulting solution requires how many mL of 0.10 M HCl to reach equivalence point?,1.0 mL,10 mL,20 mL,50 mL,C

"Predict the number of lines in the EPR spectrum of a solution of 13C-labelled methyl radical (13CH3•), assuming the lines do not overlap.",4,3,6,24,A

"3 Cl−(aq) + 4 CrO_4^2−(aq) + 23 H+(aq) → 3 HClO2(aq) + 4 Cr3+(aq) + 10 H2O(l). In the reaction shown above, Cl−(aq) behaves as",an acid,a base,a catalyst,a reducing agent,D

"Which of the following lists the hydrides of group-14 elements in order of thermal stability, from lowest to highest?",PbH4 < SnH4 < GeH4 < SiH4 < CH4,PbH4 < SnH4 < CH4 < GeH4 < SiH4,CH4 < SiH4 < GeH4 < SnH4 < PbH4,CH4 < PbH4 < GeH4 < SnH4 < SiH4,A

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College CS

Which of the following regular expressions is equivalent to (describes the same set of strings as) (a* + b)*(c + d)?,a*(c + d)+ b(c + d),a*(c + d)* + b(c + d)*,a*(c + d)+ b*(c + d),(a + b)*c +(a + b)*d,D

"A certain pipelined RISC machine has 8 general-purpose registers R0, R1, . . . , R7 and supports the following operations.

ADD Rs1, Rs2, Rd Add Rs1 to Rs2 and put the sum in Rd

MUL Rs1, Rs2, Rd Multiply Rs1 by Rs2 and put the product in Rd

An operation normally takes one cycle; however, an operation takes two cycles if it produces a result required by the immediately following operation in an operation sequence. Consider the expression AB + ABC + BC, where variables A, B, C are located in registers R0, R1, R2. If the contents of these three registers must not be modified, what is the minimum number of clock cycles required for an operation sequence that computes the value of AB + ABC + BC?",5,6,7,8,B

"The Singleton design pattern is used to guarantee that only a single instance of a class may be instantiated. Which of the following is (are) true of this design pattern?

I. The Singleton class has a static factory method to provide its instance.

II. The Singleton class can be a subclass of another class.

III. The Singleton class has a private constructor.",I only,II only,III only,"I, II, and III",D

"A compiler generates code for the following assignment statement.

G := (A + B) * C - (D + E) * F

The target machine has a single accumulator and a single-address instruction set consisting of instructions load, store, add, subtract, and multiply. For the arithmetic operations, the left operand is taken from the accumulator and the result appears in the accumulator. The smallest possible number of instructions in the resulting code is",5,6,7,9,D

"Consider a computer design in which multiple processors, each with a private cache memory, share global memory using a single bus. This bus is the critical system resource. Each processor can execute one instruction every 500 nanoseconds as long as memory references are satisfied by its local cache. When a cache miss occurs, the processor is delayed for an additional 2,000 nanoseconds. During half of this additional delay, the bus is dedicated to serving the cache miss. During the other half, the processor cannot continue, but the bus is free to service requests from other processors. On average, each instruction requires 2 memory references. On average, cache misses occur on 1 percent of references. What proportion of the capacity of the bus would a single processor consume, ignoring delays due to competition from other processors?",1/50,1/27,1/25,2/27,B

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College Math

"Let V be the set of all real polynomials p(x). Let transformations T, S be defined on V by T:p(x) -> xp(x) and S:p(x) -> p'(x) = d/dx p(x), and interpret (ST)(p(x)) as S(T(p(x))). Which of the following is true?",ST = 0,ST = T,ST = TS,ST - TS is the identity map of V onto itself.,D

"A tank initially contains a salt solution of 3 grams of salt dissolved in 100 liters of water. A salt solution containing 0.02 grams of salt per liter of water is sprayed into the tank at a rate of 4 liters per minute. The sprayed solution is continually mixed with the salt solution in the tank, and the mixture flows out of the tank at a rate of 4 liters per minute. If the mixing is instantaneous, how many grams of salt are in the tank after 100 minutes have elapsed?",2,2 - e^-2,2 + e^-2,2 + e^-4,D

"Let A be a real 2x2 matrix. Which of the following statements must be true?

I. All of the entries of A^2 are nonnegative.

II. The determinant of A^2 is nonnegative.

III. If A has two distinct eigenvalues, then A^2 has two distinct eigenvalues.",I only,II only,III only,II and III only,B

"Suppose that f(1 + x) = f(x) for all real x. If f is a polynomial and f(5) = 11, then f(15/2)",-11,0,11,33/2,C

"Let A be the set of all ordered pairs of integers (m, n) such that 7m + 12n = 22. What is the greatest negative number in the set B = {m + n : (m, n) \in A}?",-5,-4,-3,-2,B

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College Medicine

Glucose is transported into the muscle cell:,via protein transporters called GLUT4.,only in the presence of insulin.,via hexokinase.,via monocarbylic acid transporters.,A

Which of the following is not a true statement?,Muscle glycogen is broken down enzymatically to glucose-1-phosphate,Elite endurance runners have a high proportion of Type I fibres in their leg muscles,Liver glycogen is important in the maintenance of the blood glucose concentration,Insulin promotes glucose uptake by all tissues in the body,D

"In a genetic test of a newborn, a rare genetic disorder is found that has X-linked recessive transmission. Which of the following statements is likely true regarding the pedigree of this disorder?",All descendants on the maternal side will have the disorder.,Females will be approximately twice as affected as males in this family.,All daughters of an affected male will be affected.,There will be equal distribution of males and females affected.,C

"A high school science teacher fills a 1 liter bottle with pure nitrogen and seals the lid. The pressure is 1.70 atm, and the room temperature is 25°C. Which two variables will both increase the pressure of the system, if all other variables are held constant?","Increasing temperature, increasing moles of gas","Increasing temperature, increasing volume","Decreasing volume, decreasing temperature","Decreasing moles of gas, increasing volume",A

An expected side effect of creatine supplementation is:,muscle weakness.,gain in body mass.,muscle cramps.,loss of electrolytes.,B

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

College Physics

A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is,4,5,6,20,A

For which of the following thermodynamic processes is the increase in the internal energy of an ideal gas equal to the heat added to the gas?,Constant temperature,Constant volume,Constant pressure,Adiabatic,B

"One end of a Nichrome wire of length 2L and cross-sectional area A is attached to an end of another Nichrome wire of length L and cross- sectional area 2A. If the free end of the longer wire is at an electric potential of 8.0 volts, and the free end of the shorter wire is at an electric potential of 1.0 volt, the potential at the junction of the two wires is most nearly equal to",2.4 V,3.3 V,4.5 V,5.7 V,A

A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is,4,5,6,20,A

"The muon decays with a characteristic lifetime of about 10^-6 second into an electron, a muon neutrino, and an electron antineutrino. The muon is forbidden from decaying into an electron and just a single neutrino by the law of conservation of",charge,mass,energy and momentum,lepton number,D

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

Cyber

SHA-1 has a message digest of,160 bits,512 bits,628 bits,820 bits,A

"_____ can modify data on your system – so that your system doesn't run correctly or you can no longer access specific data, or it may even ask for ransom in order to give your access.",IM – Trojans,Backdoor Trojans,Trojan-Downloader,Ransom Trojan,D

What is ethical hacking?,""""Hacking"" ethics so they justify unintended selfish behavior","Hacking systems (e.g., during penetration testing) to expose vulnerabilities so they can be fixed, rather than exploited",Hacking into systems run by those whose ethics you disagree with,"A slang term for rapid software development, e.g., as part of hackathons",B

Exploitation of the Heartbleed bug permits,overwriting cryptographic keys in memory,a kind of code injection,a read outside bounds of a buffer,a format string attack,C

The _____ is anything which your search engine cannot search.,Haunted web,World Wide Web,Surface web,Deep Web,D

Source: Measuring Massive Multitask Language Understanding, Hendrycks et al, 9.7.2020

EE

"In an SR latch built from NOR gates, which condition is not allowed","S=0, R=0","S=0, R=1","S=1, R=0","S=1, R=1",D

"In a 2 pole lap winding dc machine , the resistance of one conductor is 2Ω and total number of conductors is 100. Find the total resistance",200Ω,100Ω,50Ω,10Ω,C

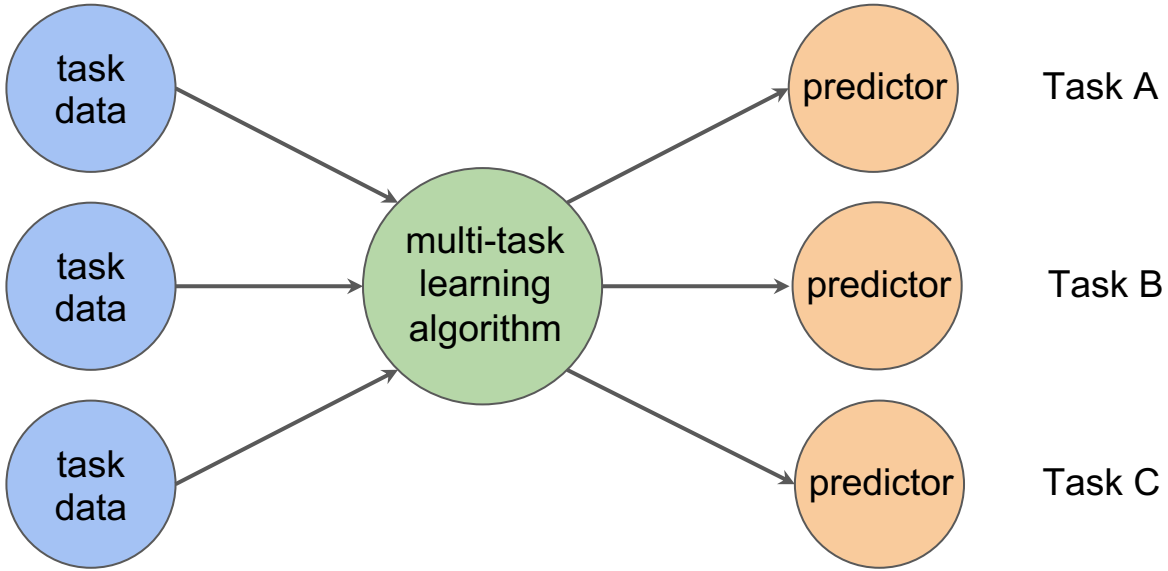
"The coil of a moving coil meter has 100 turns, is 40 mm long and 30 mm wide. The control torque is 240*10-6 N-m on full scale. If magnetic flux density is 1Wb/m2 range of meter is",1 mA.,2 mA.,3 mA.,4 mA.,B

"Two long parallel conductors carry 100 A. If the conductors are separated by 20 mm, the force per meter of length of each conductor will be",100 N.,0.1 N.,1 N.,0.01 N.,B

A point pole has a strength of 4π * 10^-4 weber. The force in newtons on a point pole of 4π * 1.5 * 10^-4 weber placed at a distance of 10 cm from it will be,15 N.,20 N.,7.5 N.,3.75 N.,A

Multi-Task Learning (MTL)

Multi-Task Learning



Multi-Task Learning: Self Driving Cars

- Multiple tasks: detect cars, pedestrians, signs, lights, curbs, lanes, cross walks, etc.
  - Tasks (100)
  - sub-tasks

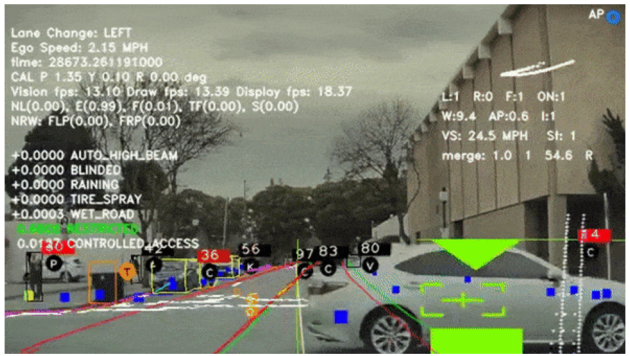
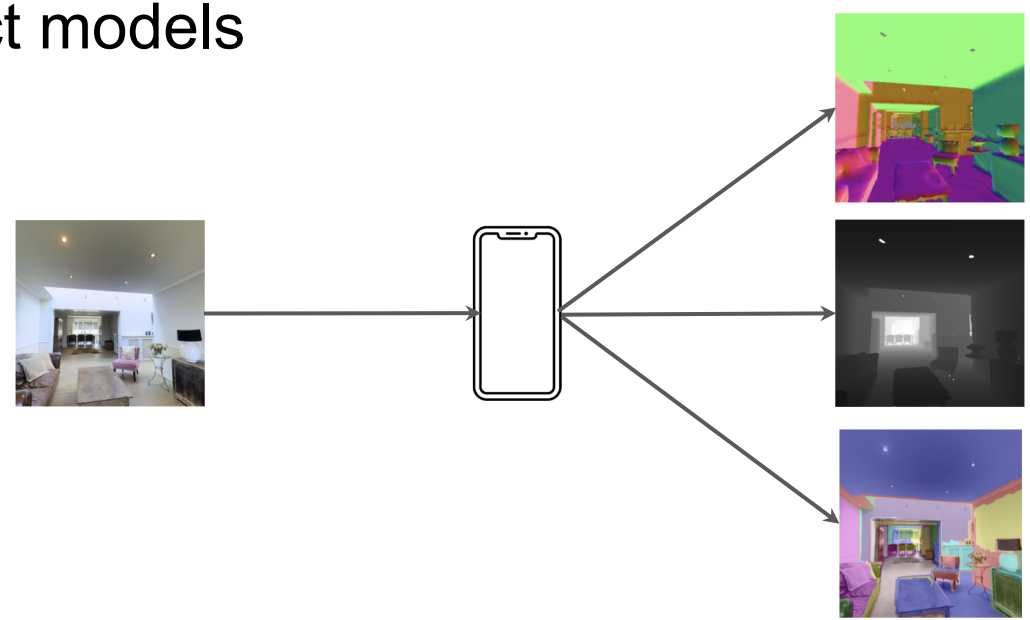


Figure Source: Tesla AutoPilot

Multi-Task Learning: Edge Devices

- High performance, prediction accuracy
- Efficient computation, training and inference time
- Compact models



Multi-Task Learning (MTL) Architectures

- One neural network for learning multiple tasks: all-in-one
- Separate networks for each task: individual prediction
- Hybrid approach

- Combinatorial optimization problem:
  - Bipartite matching of tasks to networks

Multi-Task Learning (MTL) Questions

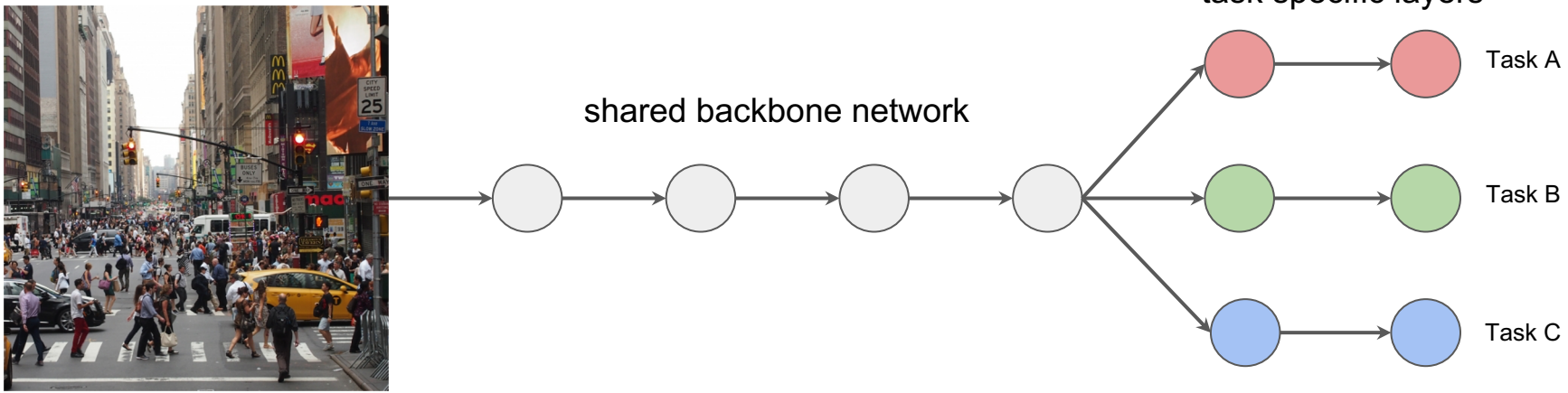
- How tasks influence one another? Does each task help the other tasks? or is there negative transfer?
- How to share weights between different tasks?
- How does network size influence MTL?
- How does dataset size and distribution of number of samples per task influence MTL?
- Are the tasks similar? Heterogeneous?

Multi-Task Learning

- Multiple heterogeneous tasks: different importance, difficulty, number of samples, noise level

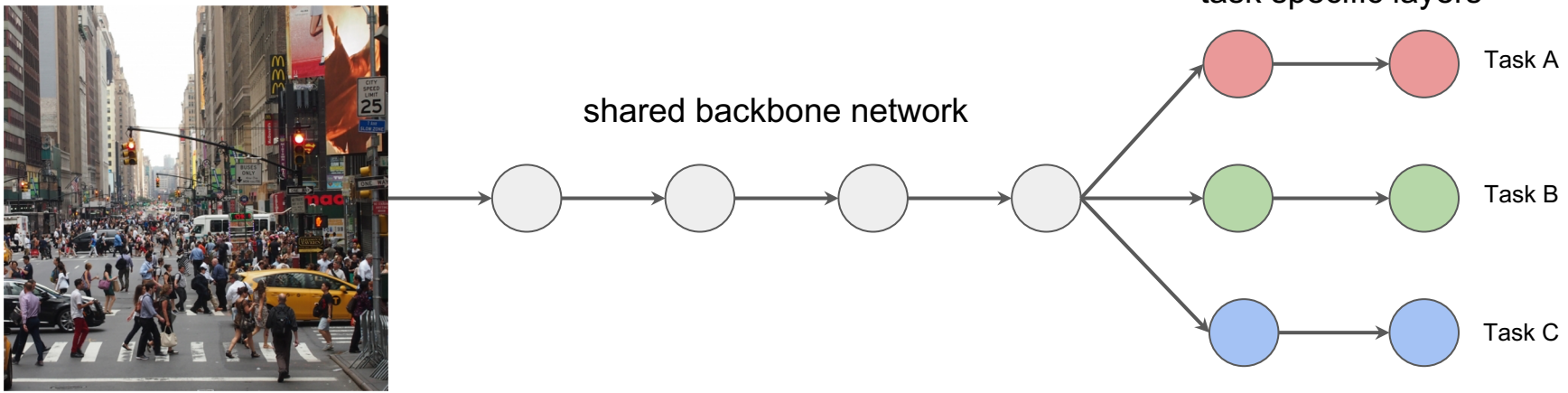
Shared backbone for multiple tasks with multiple heads

- Input example: $\boldsymbol{x}$
- Tasks: $\boldsymbol{t}$ = 1..T
- Output of task $\boldsymbol{t}$: $\boldsymbol{y_t}$
- Number of data points N
- Dataset of iid data points $\{\boldsymbol{x}^i, \boldsymbol{y}^i_1,\ldots,\boldsymbol{y}^i_T\}$ for i = 1...N



Shared backbone for multiple tasks with multiple heads

- Shared backbone network $\boldsymbol{f}$
- Shared backbone parameters $\boldsymbol{\theta}_s$
- Task-specific decoder network $\boldsymbol{g_t}$ with task-specific parameters $\boldsymbol{\theta_t}$
- Task-specific loss: $\mathcal{L}_t(\boldsymbol{\theta}) := \mathcal{L}_t(\boldsymbol{\theta}_s, \boldsymbol{\theta_t}) := 1/N \sum_i \mathcal{L}_t(\boldsymbol{g_t}(\boldsymbol{f}(\boldsymbol{x}^i; \boldsymbol{\theta}_s); \boldsymbol{\theta_t}), \boldsymbol{y}^i_t)$
- Linear scalarization total multi-task loss: $\mathcal{L}(\boldsymbol{\theta}) = \sum_t \alpha_t \mathcal{L}_t(\boldsymbol{\theta})$



Linear Scalarization for MTL

- Total multi-task loss

$$\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) = \sum_t \alpha_t \mathcal{L}_t(\boldsymbol{\theta})$$

- Advantages?
- Disadvantages?

Linear Scalarization

- Total multi-task loss is linear weighted combination

$$\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) = \sum_t \alpha_t \mathcal{L}t(\boldsymbol{\theta})$$

- Advantages? simple
- Disadvantages?
  - Selecting weights? develop loss weighting strategies?
  - Performance dependent on weights
  - Only handles convex part of Pareto front

Linear Scalarization

- Total multi-task loss is a linear weighted combination

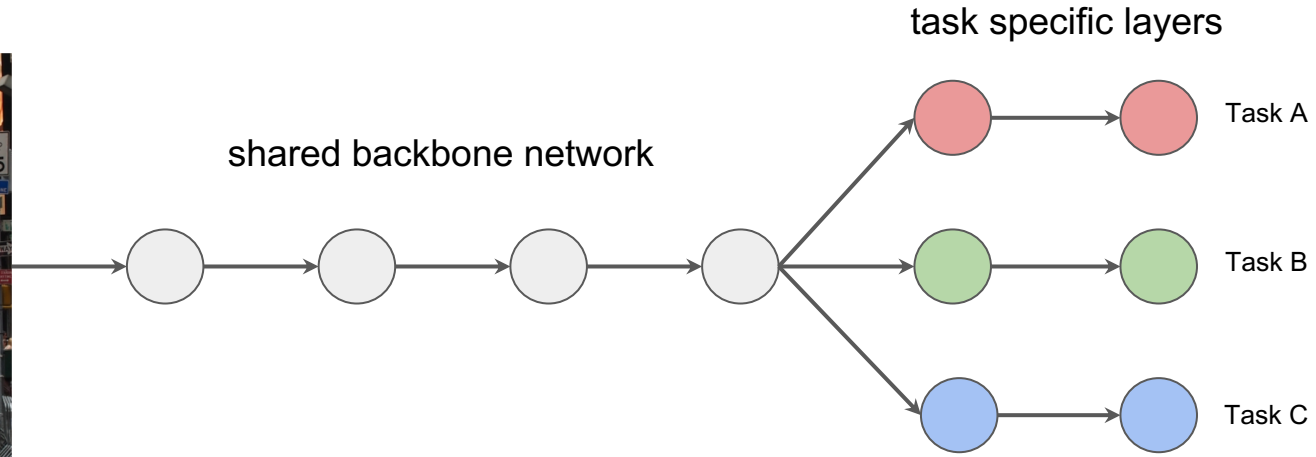
$$\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) = \sum_t \alpha_t \mathcal{L}t(\boldsymbol{\theta})$$

- Justification for linear scalarization: solutions may not be comparable. For example solution $\boldsymbol{\theta}$ may be better for task t_1 whereas solution $\boldsymbol{\theta}$‘ is better for task t_2 :

  two solutions $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$‘ s.t. $\mathcal{L}t_1(\boldsymbol{\theta}s,\boldsymbol{\theta}t_1) < \mathcal{L}t_1(\boldsymbol{\theta}’s,\boldsymbol{\theta}’t_1)$ and $\mathcal{L}t_2(\boldsymbol{\theta}s,\boldsymbol{\theta}t_2) > \mathcal{L}t_2(\boldsymbol{\theta}’s,\boldsymbol{\theta}’t_2)$ for tasks t_1 and t_2

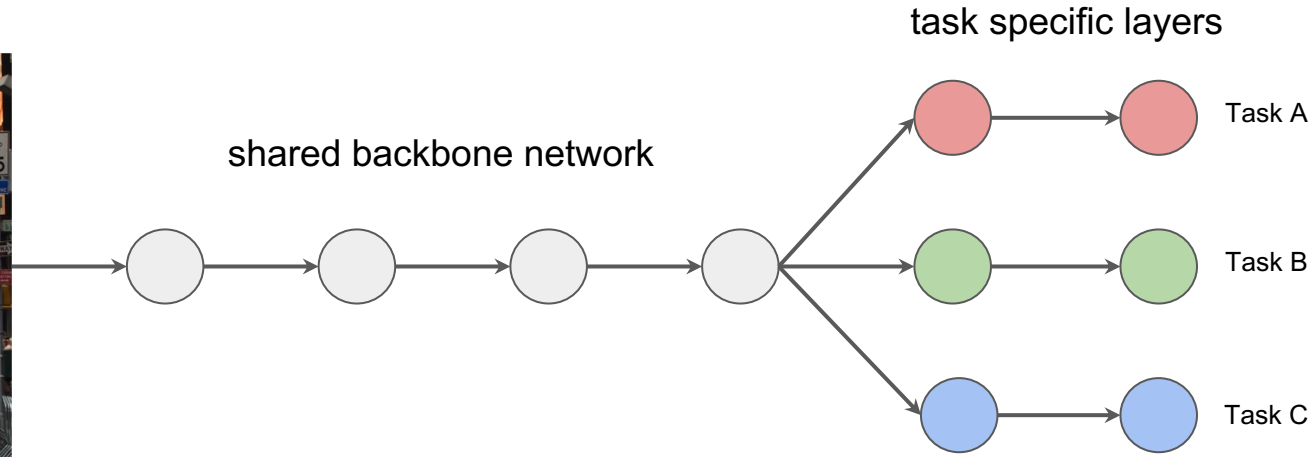
Shared backbone for multiple tasks with multiple heads

- Sharing weights in early layers, coupled
- Split network into backbone and task-specific layers
- Advantages?
- Disadvantages?



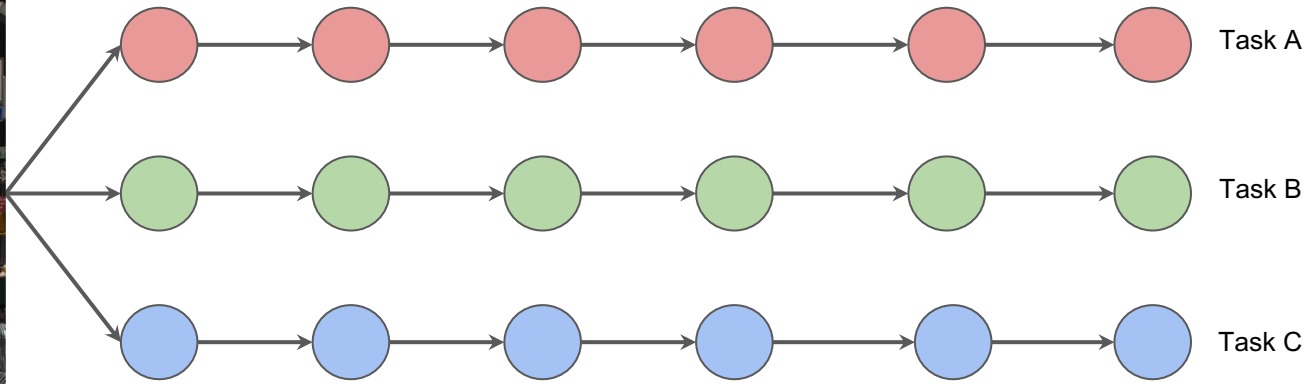
Shared backbone for multiple tasks with multiple heads

- Sharing weights in early layers, coupled
- Split network into backbone and task-specific layers, where to split?
- Advantages? efficient runtime
- Disadvantages? over-sharing, negative transfer



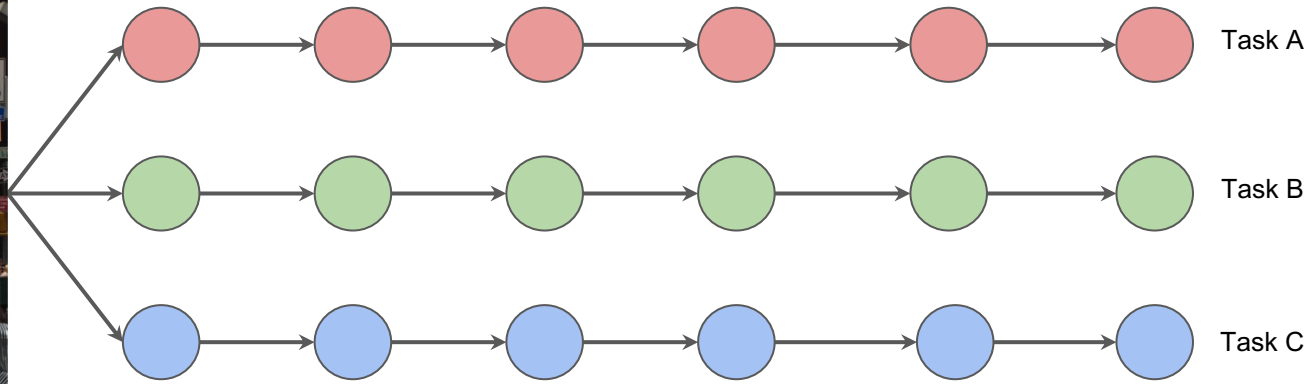
Individual network for each task

- No sharing weights
- Decoupled functionality
- Advantages?
- Disadvantages?



Individual network for each task

- No sharing weights
- Decoupled functionality
- Advantages? no negative transfer
- Disadvantages? inefficient runtime, does not scale well with number of tasks



Negative Transfer

- Why does training individual networks often work better than a shared network?

Negative Transfer

- Relationships between tasks determines if a shared architecture works
- One task may dominate training
- Tasks may learn at different rates
- Gradients may conflict

Multi-Task Learning and Adversarial Attacks

- Models trained on multiple tasks at once are more robust to adversarial attacks on individual tasks

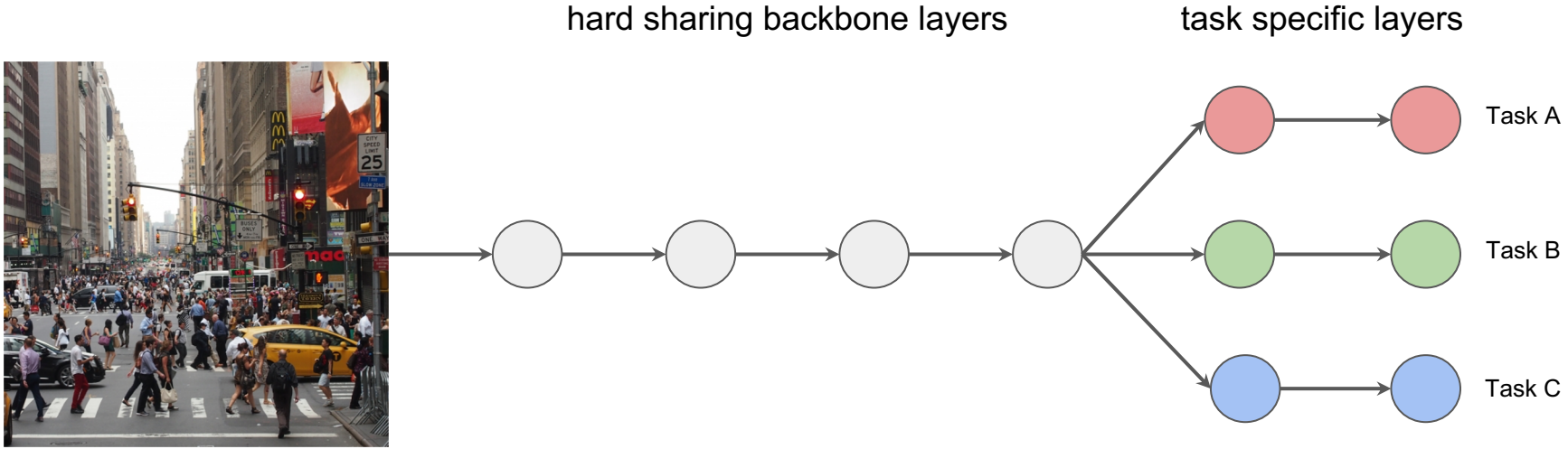
MTL Architectures

Architectures

- Hard parameter sharing
- Soft parameter sharing
- Ad-hoc sharing
- Learning to route, branch

Hard Parameter Sharing

- Sharing information in early layers, over-sharing.
- Split network into task-specific layers, where to split?
- Define loss function



Multi-Objective Optimization

- Optimize collection of possibly conflicting objectives:

$$\min_{\boldsymbol{\theta}_s,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T} \mathcal{L}(\boldsymbol{\theta}_s,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T) = \min_{\boldsymbol{\theta}_s,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T} (\mathcal{L}_1(\boldsymbol{\theta}_s,\boldsymbol{\theta}_1),\ldots,\mathcal{L}_T(\boldsymbol{\theta}_s,\boldsymbol{\theta}_T))$$

Multi-Objective Optimization

- Tasks t = 1..T
- Neural network parameters x

- Multi-objective function: $f(x)$: $\mathbb{R}^n$ -> $\mathbb{R}^T$

$$f(x) = (f_1(x),\ldots,f_T(x))$$

- Objective function of task t is task-specific loss:

$$f_t(x): \mathbb{R}^n \rightarrow \mathbb{R}$$

Pareto Optimal

- For any $\boldsymbol{x}$,$\boldsymbol{y}$ in $\mathbb{R}^n$ $\boldsymbol{x}$ dominates $\boldsymbol{y}$ iff $\boldsymbol{f}(\boldsymbol{x})$ <- $\boldsymbol{f}(\boldsymbol{y})$
- A point $\boldsymbol{x}$ is Pareto optimal if it is not dominated by any other point
- A point $\boldsymbol{x}$ is locally Pareto optimal if it is not dominated by any point in a neighborhood of $\boldsymbol{x}$

Pareto Frontier

- Point C is not on the Pareto frontier because it is dominated by points A and B
- Points A and B are not dominated by any other point, and are therefore on the Pareto frontier.

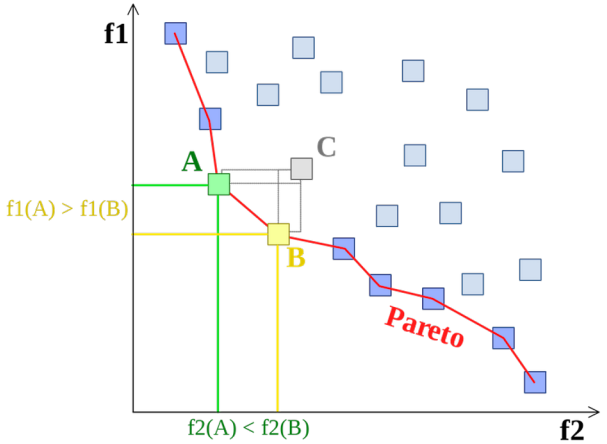


Figure source: Wikipedia

Pareto Stationary

- If each $f_t(x)$ is continuously differentiable a point x is Pareto stationary if there exists α in $\mathbb{R}^T$ such that

$$\alpha_t >= 0, \sum_t \alpha_t = 1 \text{ and } \sum_t \alpha_t \nabla f_t(x) = 0$$

- All Pareto optimal points are Pareto stationary.

MTL Algorithm

- Gradient descent on task-specific parameters
- Solving optimization problem:

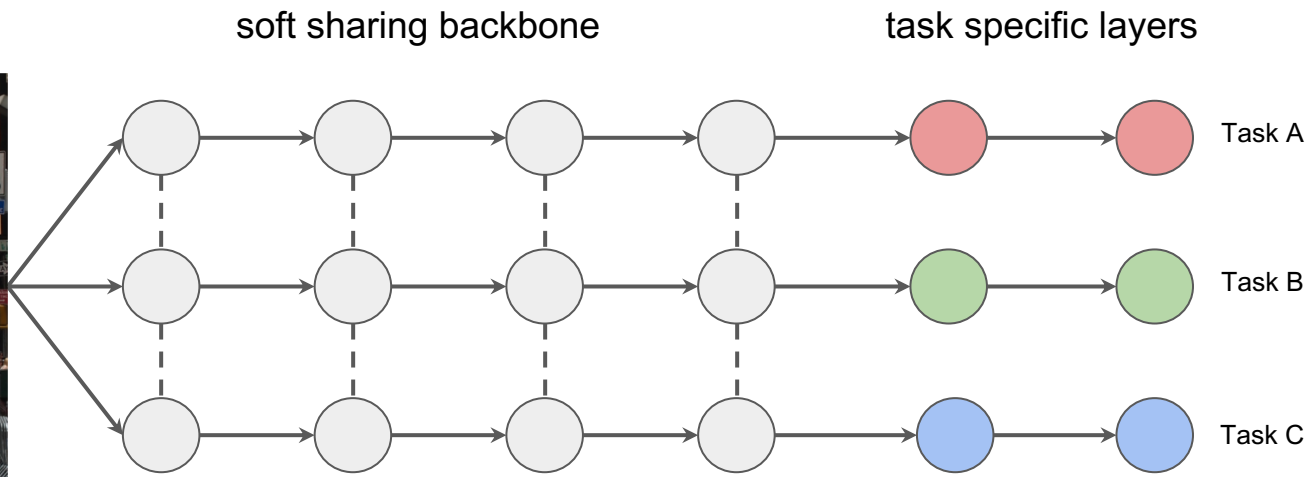
$$\min_{\alpha_1,\ldots,\alpha_T} \{ \| \sum_t \alpha_t \nabla f_t(x) \| \mid \sum_t \alpha_t = 1, \alpha_t >= 0 \text{ for all } t \}$$

$$\min_{\alpha_1,\ldots,\alpha_T} \{ \| \sum_t \alpha_t \nabla_{\theta_s} \mathcal{L}_t(\theta_s, \theta_t) \| \mid \sum_t \alpha_t = 1, \alpha_t >= 0 \text{ for all } t \}$$

Source: Multi-task learning as multiobjective optimization, Sener and Koltun, 2018

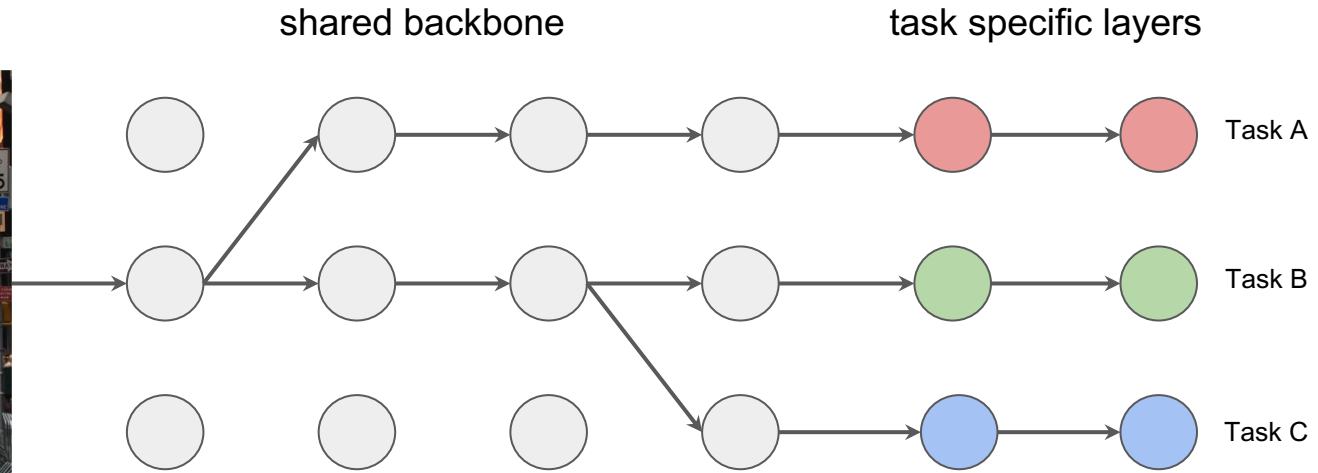
Soft Parameter Sharing

- Sharing information in early layers
- Does not scale well with number of tasks



Ad-hoc Sharing

- Compute task relatedness
- Iteratively group network
- Better performance than soft or hard sharing



Learn Shared Architecture

- Directed acyclic graph
- Nodes represent computational operations
- Edges represent data flows
- Differential branching operations

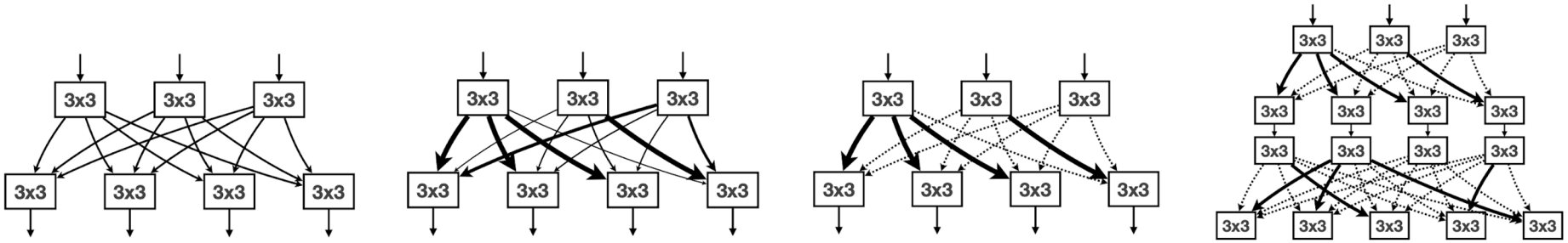


Figure source: Learning to branch for multi-task learning, Guo et al, 2020

Layer Routing

- Learn separate execution paths for different tasks

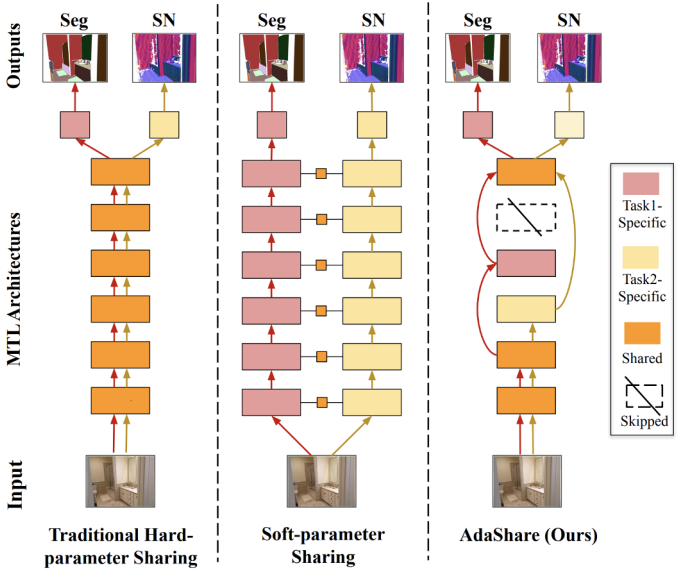


Figure source: AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning, Sun et al, 2019

Taskonomy Dataset

- 4.5 million indoor scenes from 600 buildings
- 26 diverse tasks, every image is labeled for all tasks

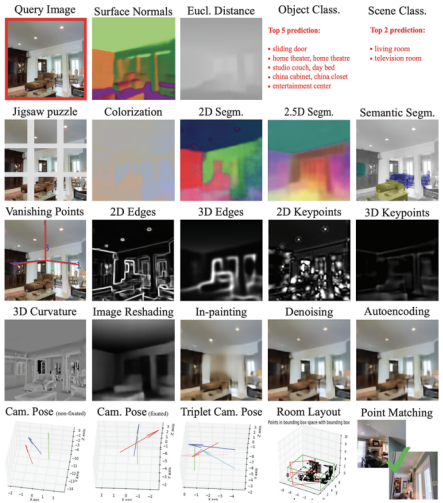
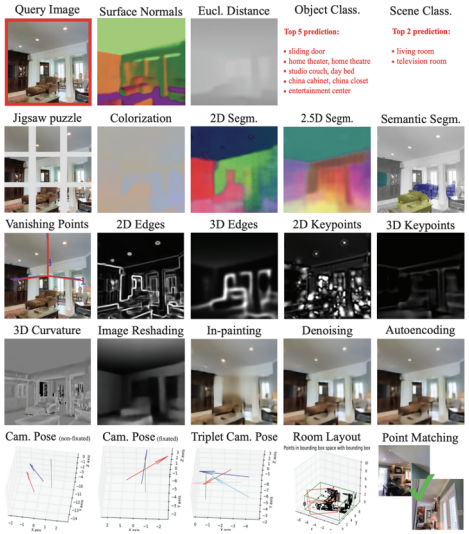


Figure source: Taskonomy: Disentangling Task Transfer Learning, Zamir et al, 2018

Transfer Relationships between tasks

- Taskonomy



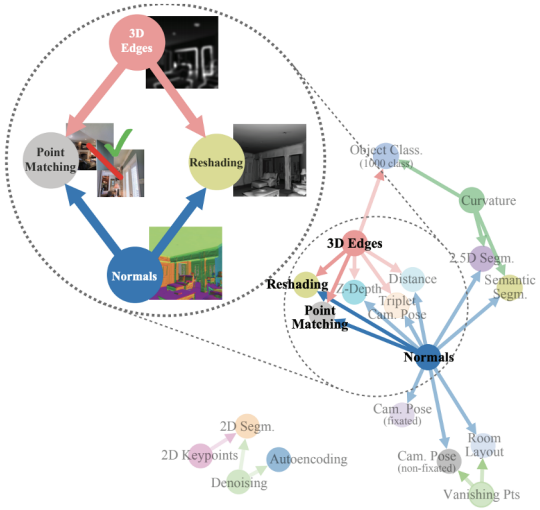


Figure source: Taskonomy: Disentangling Task Transfer Learning, Zamir et al, 2018

Multi-Task Learning

- Which tasks should and should not be learned together in one network when employing multi-task learning?

		Relative Performance On					
		SemSeg	Depth	Normals	Keypoints	Edges	Average
Trained With	SemSeg	–	-5.41%	-11.29%	-4.32%	-34.64%	-13.92%
	Depth	4.17%	–	-3.55%	3.49%	3.76%	1.97%
	Normals	8.50%	2.48%	–	1.37%	12.33%	6.17%
	Keypoints	4.82%	1.38%	-0.02%	–	-5.26%	0.23%
	Edges	3.07%	-0.92%	-4.42%	1.37%	–	-0.23%
	Average	5.14%		-4.82%		-5.95%	-1.15%

Figure source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

Multi-Task Learning

- Transfer relationships may not predict multi-task relationships

	Depth	Normals	Keypoints	Edges
SemSeg	1.740%	1.828%	0.723%	0.700%
Depth		1.915%	0.406%	0.468%
Normals			0.089%	0.118%
Keypoints				0.232%

transfer learning affinities

	Depth	Normals	Keypoints	Edges
SemSeg	-0.62%	-1.39%	0.25%	-15.78%
Depth		-0.54%	2.43%	1.42%
Normals			0.67%	3.95%
Keypoints				-1.95%

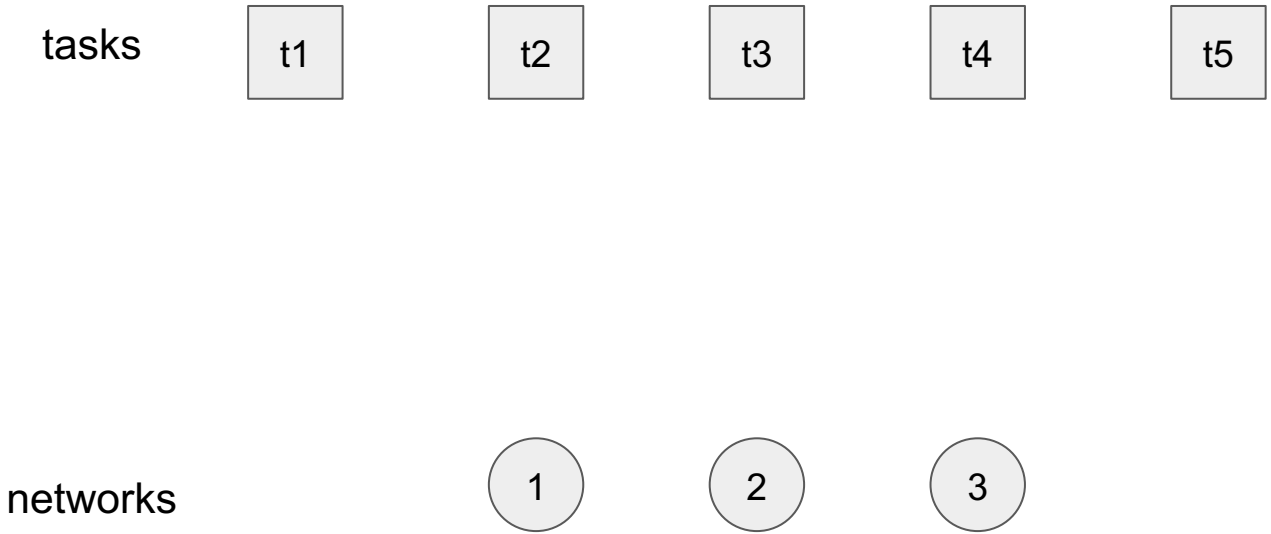
MTL learning affinities

Figure source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

MTL: Combinatorial Optimization Problem

MTL: Combinatorial Optimization Problem

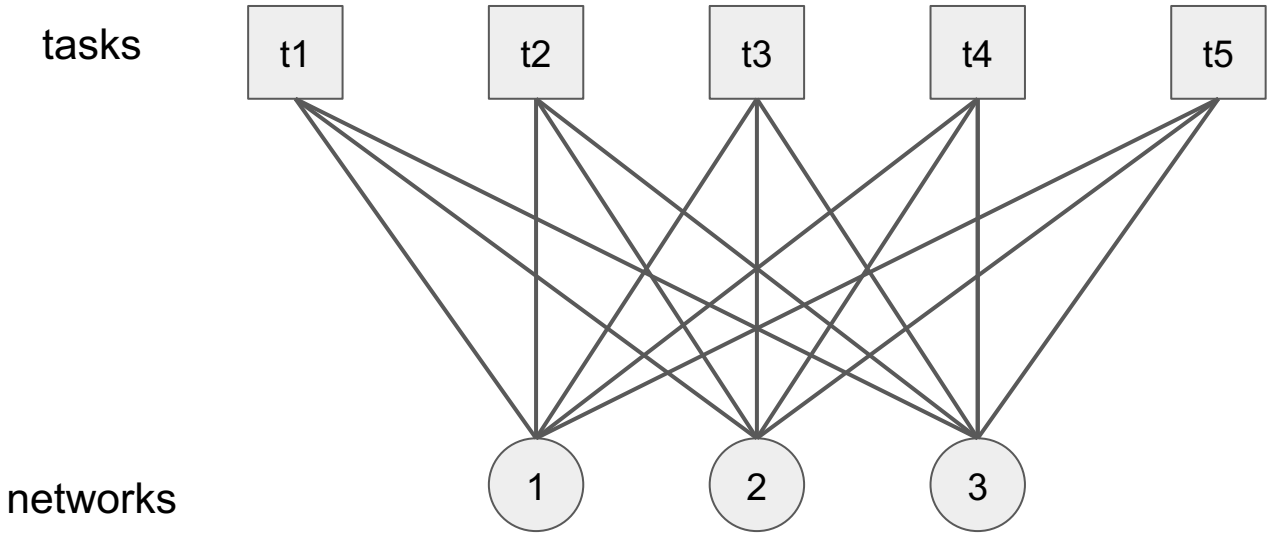
- Bipartite matching of tasks to networks given budget



Source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

MTL: Combinatorial Optimization Problem

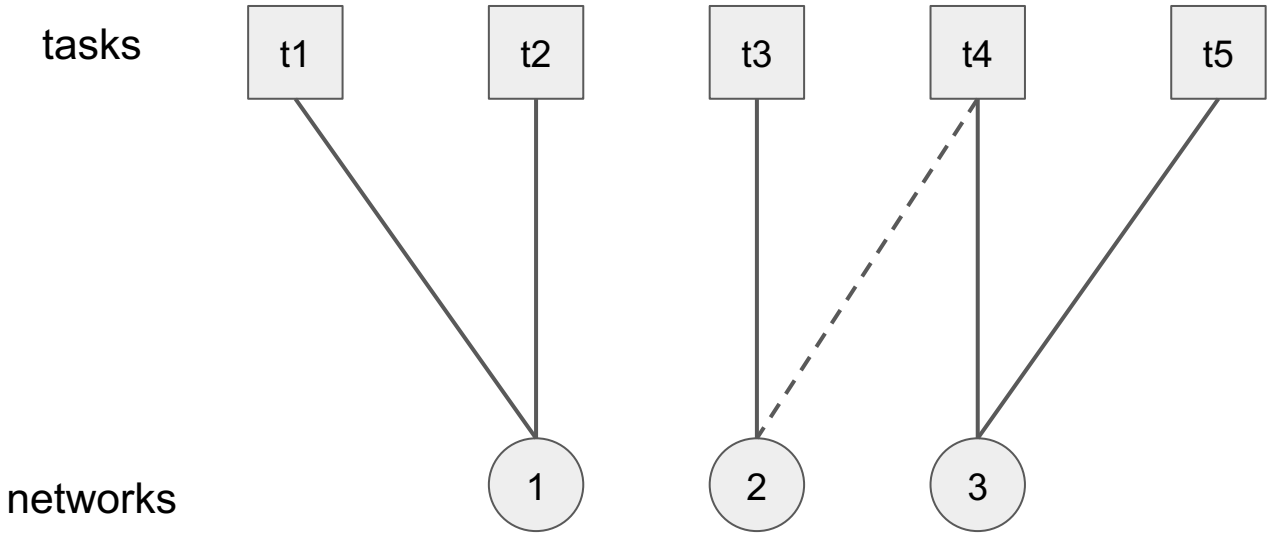
- NP-hard



Source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

MTL: Combinatorial Optimization Problem

- Approximate solution



Source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

MTL: Combinatorial Optimization Problem

- Tasks $T = \{t_1,\ldots,t_k\}$
- Inference budget b, total time to complete all tasks
- Neural network n, with inference cost time c_n
- Loss for each task $L(n, t_i)$

  infinity if network does not solve task
- Solution S is a set of networks that together solve all tasks

Source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

MTL: Combinatorial Optimization Problem

- Computation cost of solution: $cost(S) = \sum_{(n \text{ in } S)} c_n$
- Loss of solution on task is lowest loss on task among S

$$L(S, t_i) = \min_{(n \text{ in } S)} L(n, t_i)$$

- Overall performance of solution

$$L(S) = \sum_{(t_i \text{ in } T)} L(S, t_i)$$

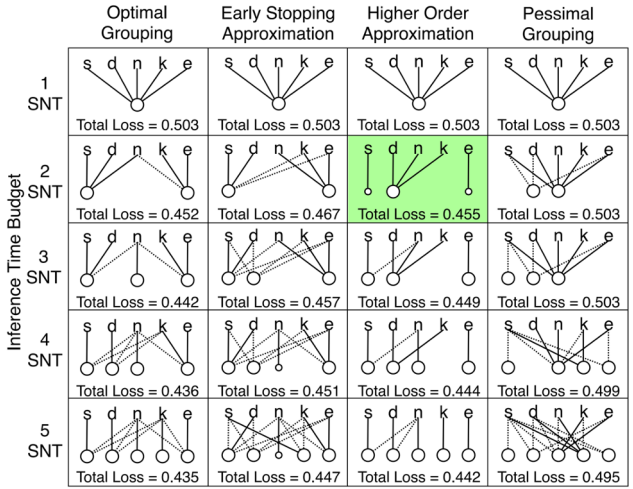
- Find solution with lowest overall loss and cost within budget

$$S^* = \text{argmin}_{S:cost(S)<b} L(S)$$

Source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

Multi-Task Learning

- Combinatorial optimization problem:
  - Bipartite matching of tasks to networks given budget
  - NP-hard problem
  - Approximation



Algorithm 1 Get Best Networks

Input: C_r, a running set of candidate networks, each with an associated cost $c \in \mathbb{R}$ and a performance score for each task the network solves. Initially, $C_r = C_0$
Input: $S_r \subseteq C_0$, a running solution, initially $\emptyset$
Input: $b_r \in \mathbb{R}$, the remaining time budget, initially b

function GETBESTNETWORKS(C_r, S_r, b_r)
  $C_r \leftarrow$ FILTER(C_r, S_r, b_r)
  $C_r \leftarrow$ SORT(C_r) ▷ Most promising networks first
  $Best \leftarrow S_r$
  for $n \in C_r$ **do**
    $C_r \leftarrow C_r \setminus n$
    $S_i \leftarrow S_r \cup \{n\}$
    $b_i \leftarrow b_r - c_n$
    $Child \leftarrow$ GETBESTNETWORKS(C_r, S_i, b_i)
    $Best \leftarrow$ BETTER($Best, Child$)
  return $Best$

function FILTER(C_r, S_r, b_r)
  Remove networks from C_r with $c_n > b_r$.
  Remove networks from C_r that cannot improve S_r's performance on any task.
  return C_r

Figure source: Which Tasks Should Be Learned Together in Multi-task Learning? Standley et al, 2020

Meta Learning

MIT
Iddo Drori, Fall 2020