OVER-VIEW OF TRANSFER LEARNING

HAPPY BUZAABA
ABI-RIKEN AIP, University of Tsukuba

Foundations of Transfer Learning

[1] As humans, we find it easy to transfer knowledge we have learned from one domain or task to another. When we encounter a new task, we don' t have to start from scratch. Instead, we use our previous experience to learn and adapt to that new task faster and more accurately.

[2] **Transfer Learning**: refers to a situation where what has been learned from one setting (eg., distribution $P1$) is exploted in another setting (say, distribution $P2$).

Assumption: Many of the factors that explain the variations in $P1$ are relevant to the variations that need to be captured to learn $P2$.

[1]. Pan & Yang,., A Survey on Transfer Learning, IEEE 2010
[2]. Book: DEEP LEARNING Ian Godfellow, Yoshua Bengio, and Aaron Courville page 526

Fundamental Questions in Transfer Learning

1. What information is useful and transferable from source domain to the target domain?

2. What is the best way of transfering the information?

3. How to avoid transfering information that is detrimental to the desired outcome?

Note: To answer these questions, we consider similarities between the feature spaces, models & tasks of the source and target domains.

Pan & Yang: A survey on Transfer Learning. IEEE 2010

Notation in Transfer Learning

- Domain $\boldsymbol{D}$: $\boldsymbol{D} = \{\boldsymbol{\mathcal{X}}, \boldsymbol{P}(\boldsymbol{X})\}$,

  $\boldsymbol{\mathcal{X}}$ is the feature space,

  $\boldsymbol{P}(\boldsymbol{X})$ is marginal probability distribution, where $\boldsymbol{X} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \boldsymbol{\mathcal{X}}$.

If two domains are different, either $\boldsymbol{\mathcal{X}_s} \neq \boldsymbol{\mathcal{X}_T}$, or $\boldsymbol{P}(\boldsymbol{X_s}) \neq \boldsymbol{P}(\boldsymbol{X_T})$

- Task $\boldsymbol{\mathcal{T}}$: Given a specific $\boldsymbol{D}$, a task $\boldsymbol{\mathcal{T}} = \{\boldsymbol{\mathcal{Y}}, \boldsymbol{f}(\cdot)\}$

  $\mathcal{Y}$ is label space, & $\boldsymbol{f}(\cdot)$ objective predictive function.

  $\boldsymbol{f}(\cdot)$ can be learned from training data $\{(x_i, y_i) | i \in \{1, 2, \dots, N\}\}$, where $\boldsymbol{x_i} \in \boldsymbol{\mathcal{X}}$ & $\boldsymbol{y_i} \in \boldsymbol{\mathcal{Y}}$

From probabilistic view, $\boldsymbol{f}(\boldsymbol{x_i})$ can be written as $\boldsymbol{P}(\boldsymbol{y_i} | \boldsymbol{x_i})$, and the task as $\boldsymbol{\mathcal{T}} = \{\boldsymbol{\mathcal{Y}}, \boldsymbol{P}(Y | \boldsymbol{X})\}$

In general, if two tasks are different, they may have different;

label spaces $\boldsymbol{\mathcal{Y}_s} \neq \boldsymbol{\mathcal{Y}_T}$, or $\boldsymbol{P}(\boldsymbol{Y_s} | \boldsymbol{X_s}) \neq \boldsymbol{P}(\boldsymbol{Y_T} | \boldsymbol{X_T})$

Pan & Yang: A survey on Transfer Learning. IEEE 2010
Weiss et al., A survey of transfer Learning. Journal of Big Data 2016

Definition of Transfer Learning

Given $\boldsymbol{D_s}$ and $\boldsymbol{T_s}$, $\boldsymbol{D_T}$ and $\boldsymbol{T_T}$,

Transfer learning aims to improve the learning of the target predictive function $\boldsymbol{f_T(\cdot) \sim P(Y_T|X_T)}$ in $\boldsymbol{D_T}$ using the knowledge from $\boldsymbol{D_s}$ & $\boldsymbol{T_s}$, where $\boldsymbol{D_s \neq D_T}$, or $\boldsymbol{T_s \neq T_T}$.

1. A domain is a pair $\boldsymbol{D} = \{\boldsymbol{\mathcal{X}}, \boldsymbol{P(X)}\}$ thus the condition;

   $\boldsymbol{D_s \neq D_T}$ implies that either $\mathcal{X}_s \neq \mathcal{X}_T$ or $P(X_s) \neq P(X_T)$

2. A task is a pair $\boldsymbol{\mathcal{T}} = \{\boldsymbol{\mathcal{Y}}, \boldsymbol{P(Y|X)}\}$ thus the condition;

   $\boldsymbol{T_s \neq T_T}$ implies that either $y_s \neq y_T$ or $P(Y_s|X_s) \neq P(Y_T|X_T)$

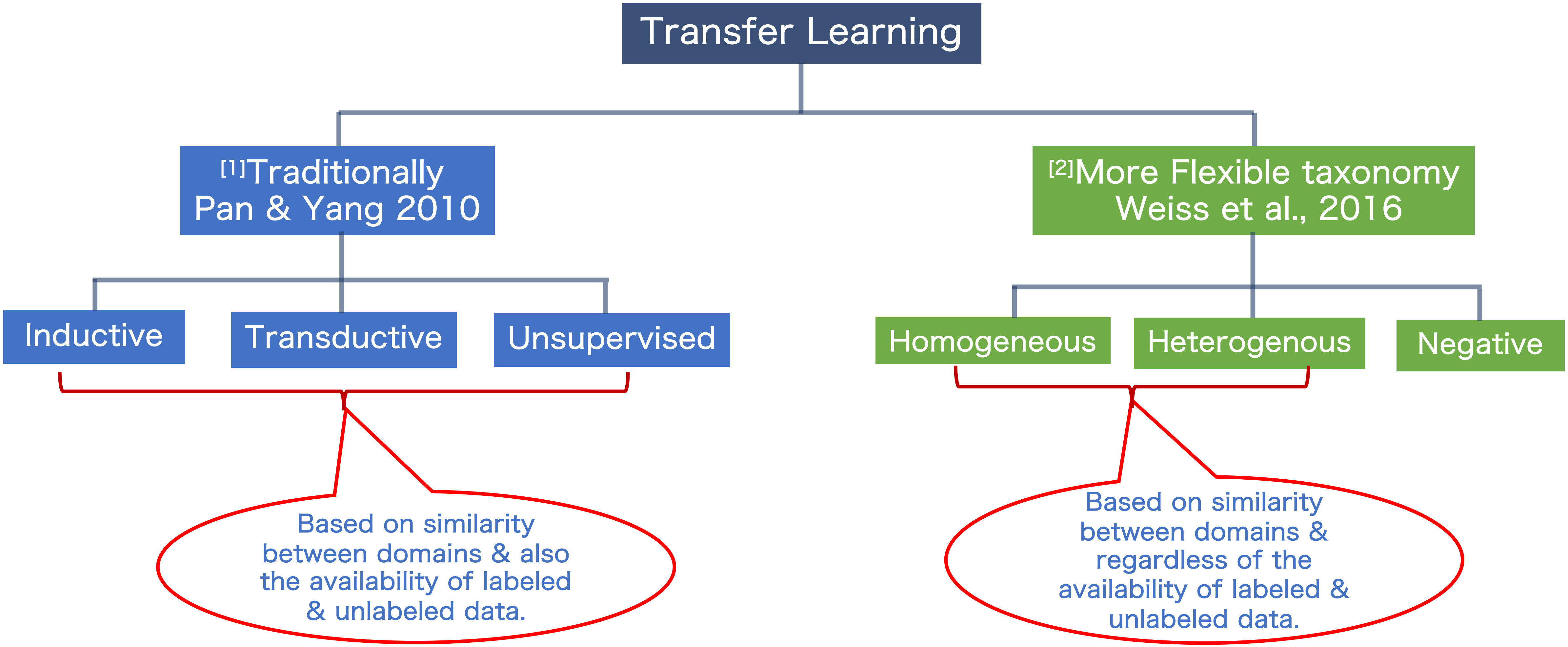
Definition of Transfer Learning

Table 1: All possible combinations for domain and task pair

Scenario	**Example in Sentiment classification**
$X_s \neq X_t$	The source domain could be English and the target could be Arabic.
$P(X_s) \neq P(X_t)$	The review could be written in the topic of hotels in the first domain while on restaurants on the target domain.
$y_s \neq y_t$	As an example, the reviews in the source task might be binary while in the target task is categorical.
$P(y_s \mid x_s) \neq P(y_t \mid x_t)$	For example, given a specific review in the domain task might have a label negative while in the target task it has a label neutral.

Table Source; Zaid et al., A survey on Transfer Learning in Natural Language Processing

Categorization of Transfer Learning Problems



Pan & Yang: A survey on Transfer Learning. IEEE 2010
Weiss et al., A survey of transfer Learning. Journal of Big Data 2016

Categorization of Transfer Learning

[1]Traditionally
Pan & Yang 2010

Inductive

Learn $f_T(\cdot)$ in D_T using knowledge in D_s, & T_s, where $T_s \neq T_T$

D_s, Labels: Available, Unavailable
D_T, Labels: Available

Transductive

Learn $f_T(\cdot)$ in D_T using knowledge in D_s, & T_s, where $D_s \neq D_T$ & $T_s = T_T$

D_s, Labels: Available
D_T, Labels: Unavailable

Unsupervised

Learn $f_T(\cdot)$ in D_T using knowledge in D_s, & T_s, where $T_s \neq T_T$ & y_s, & y_T are not observed

D_s, Labels: Unavailable
D_T, Labels: Unavailable

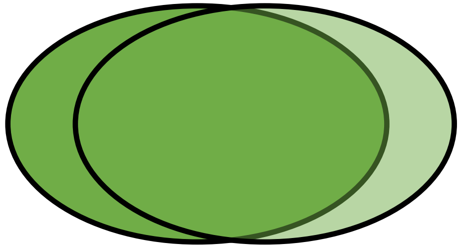
Dai et al., Boosting for Transfer Learning ICML' 07
Arnold et al., A Comprehensive study of Methods for Transductive Transfer Learning IEEE' 07
Dai et al., Self-taught Clustering ICML' 08

Categorization of Transfer Learning Problems

1. Homogenous Transfer Learning

$(X_s = X_T$ and $Y_s = Y_T)$

$X_s \approx X_T$



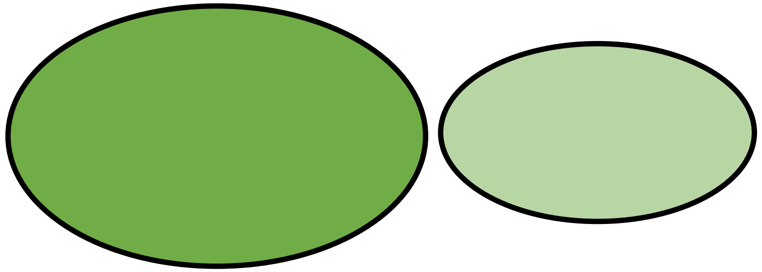
Task: Bridge the gap between the source and target data distributions i.e.,

$P(X_s) \neq P(X_T)$ and/or $P(Y_s|X_s) \neq P(Y_T|X_T)$

2. Heterogenous Transfer Learning

$(X_s \neq X_T$ and/or $Y_s \neq Y_T)$

$X_s \neq X_T$



Task: Bridge the gap between feature spaces and reduce the problem to homegenous

3. Negative Transfer: If the source domain is not very similar to the target domain, the information learned from the source can have a detrimental effect on a target learner.

Weiss et al., A survey of transfer Learning. Journal of Big Data 2016

Categorization of Transfer Learninng Solutions

Table 2. Homogenous Transfer learning Approaches

Transfer Learning Approaches	Description
Instance-transfer	Try to re-weight samples in the source domain for use in the target domain. *Sugiyama et al., 2008, Yao et al., 2010, Asgarian et al., 2018*
Feature-based-transfer	Aim to reduce gap between marginal and conditional distribution between source and target domains. *Long et al., 2014, Oquab et al., 2014, Pan et al., 2011.* Two transformation groups: 1. **Asymmetric:** Transforms one of the domain into the other *[Hoffman et al., 2014].* 2. **Symmetric:** Transforms both domains to a common latent space *[Ganin et al., 2014]*
Parameter-Transfer	Discover shared parameters or priors between the source domain and target domain models, which can benefit for transfer learning. *Duan et al., 2012, Yao et al., 2010,*
Relational-knowledge-transfer	Transfer knowledge through learning a common relationship between source and target domain. *Li et al., 2012, Yang et al., 2018*
Hybrid-based	Transfer through both instance and shared parameters *[Xia et al., 2013]*

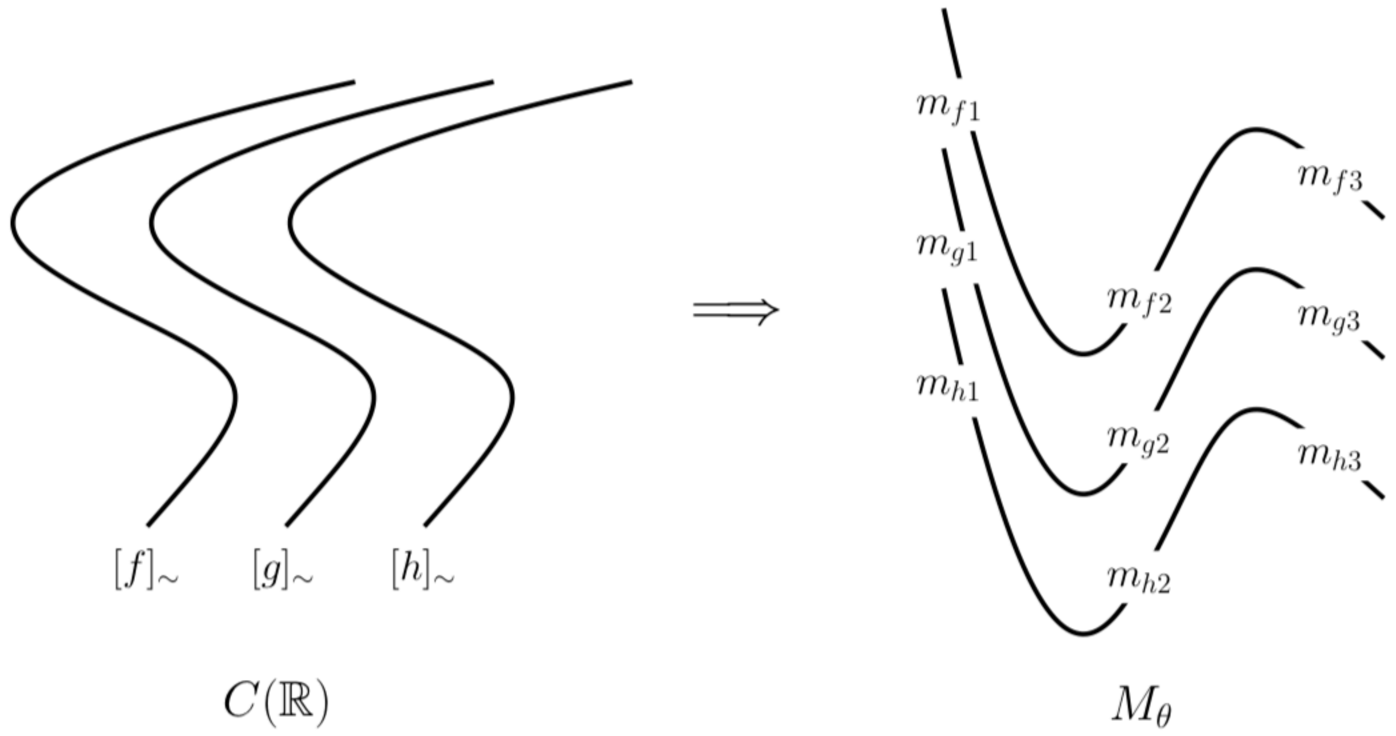
DISCUSSION ON TRANSFER LEARNING

Fariz Ikhwantri
ABI-RIKEN AIP, Tokyo Institute of Technology

Current Work on Transfer Learning

Formalization of Relation Between Task

- Relatedness
  - From several equivariance of Continuous function C(R)

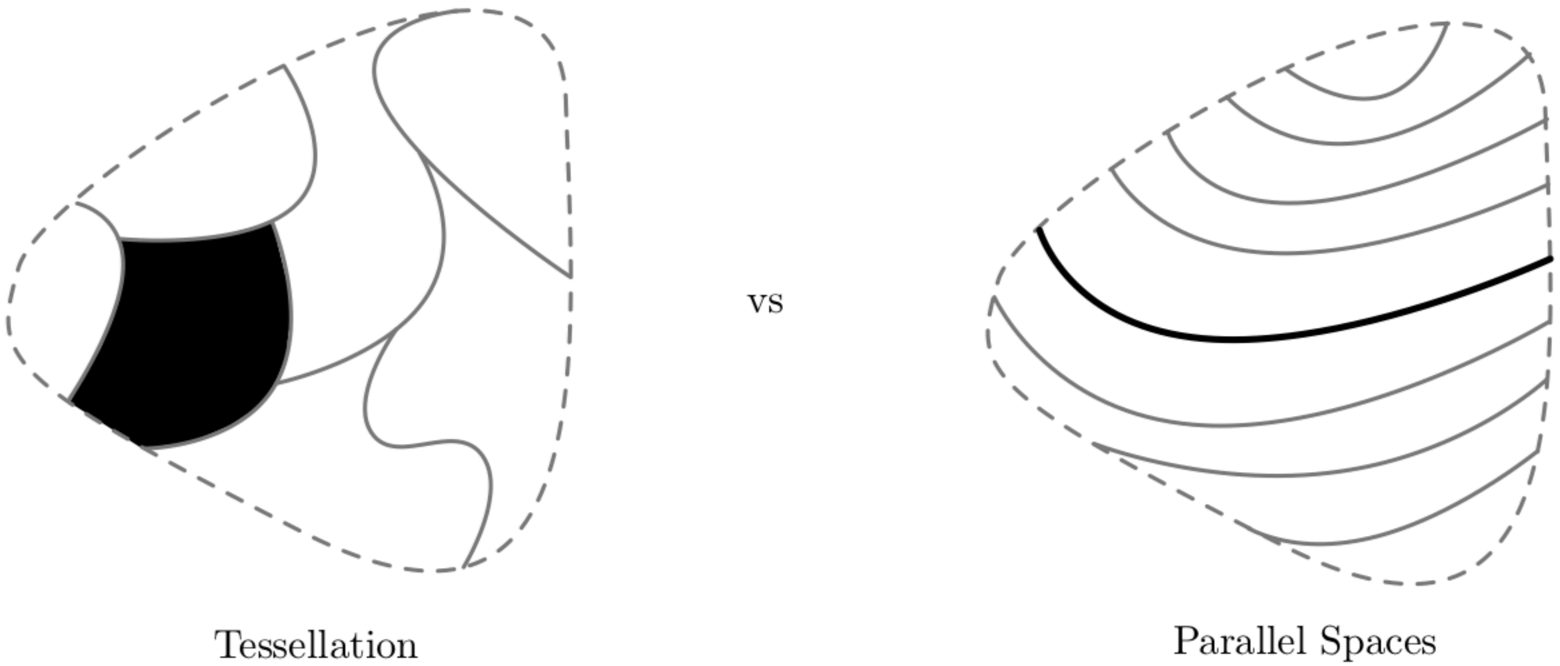


Petangoda, Janith C. et al. “A Foliated View of Transfer Learning.” 2020.

Current Work on Transfer Learning

Formalization of Relation Between Task

- Relatedness vs Similarity in a Topological

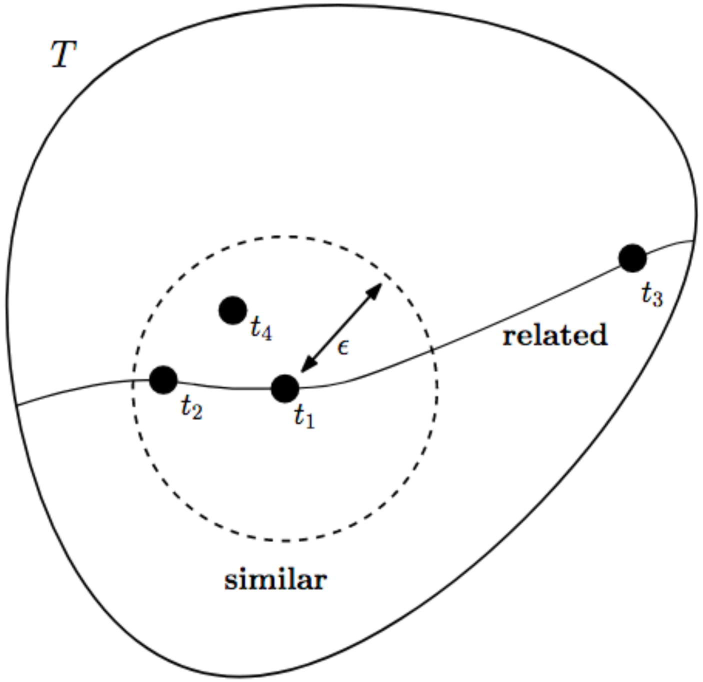


Petangoda, Janith C. et al. “A Foliated View of Transfer Learning.” 2020.

Current Work on Transfer Learning

Formalization of Relation Between Task

- Related
  - Defined in transformation of function f to g
- Similarity
  - Defined in geometric distance between two task f and g
  - f and g are similar iff $\rho(f, g) < \varepsilon$, where ε is to be chosen

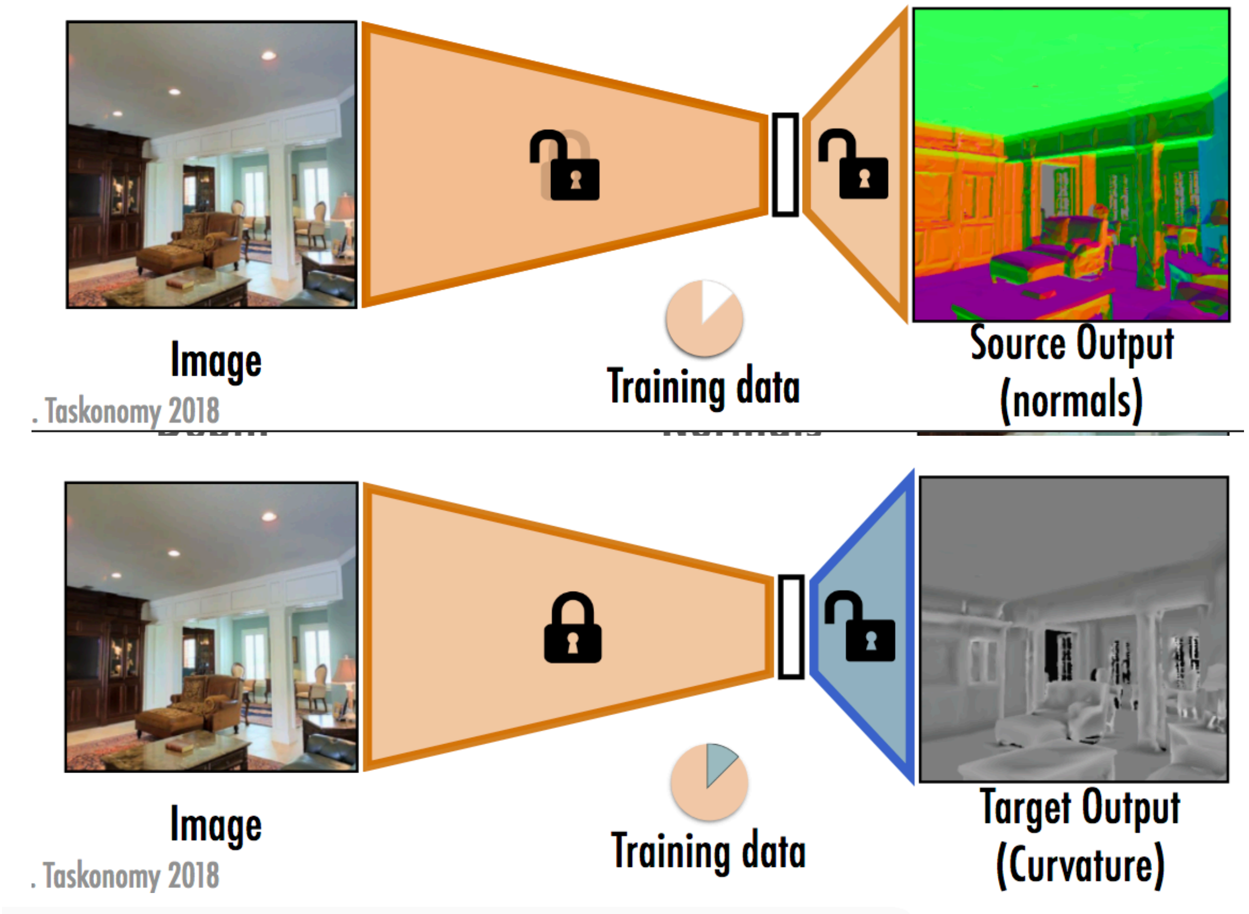


Petangoda, Janith C. et al. "A Foliated View of Transfer Learning." 2020.

Current Work on Transfer Learning

Relation Between Task

- Task taxonomy example in Computer Vision (Zamir et al., 2018)
  1. Specific Task Training
  2. Transfer Learning Model
     a. One to one task
     b. Many to one task

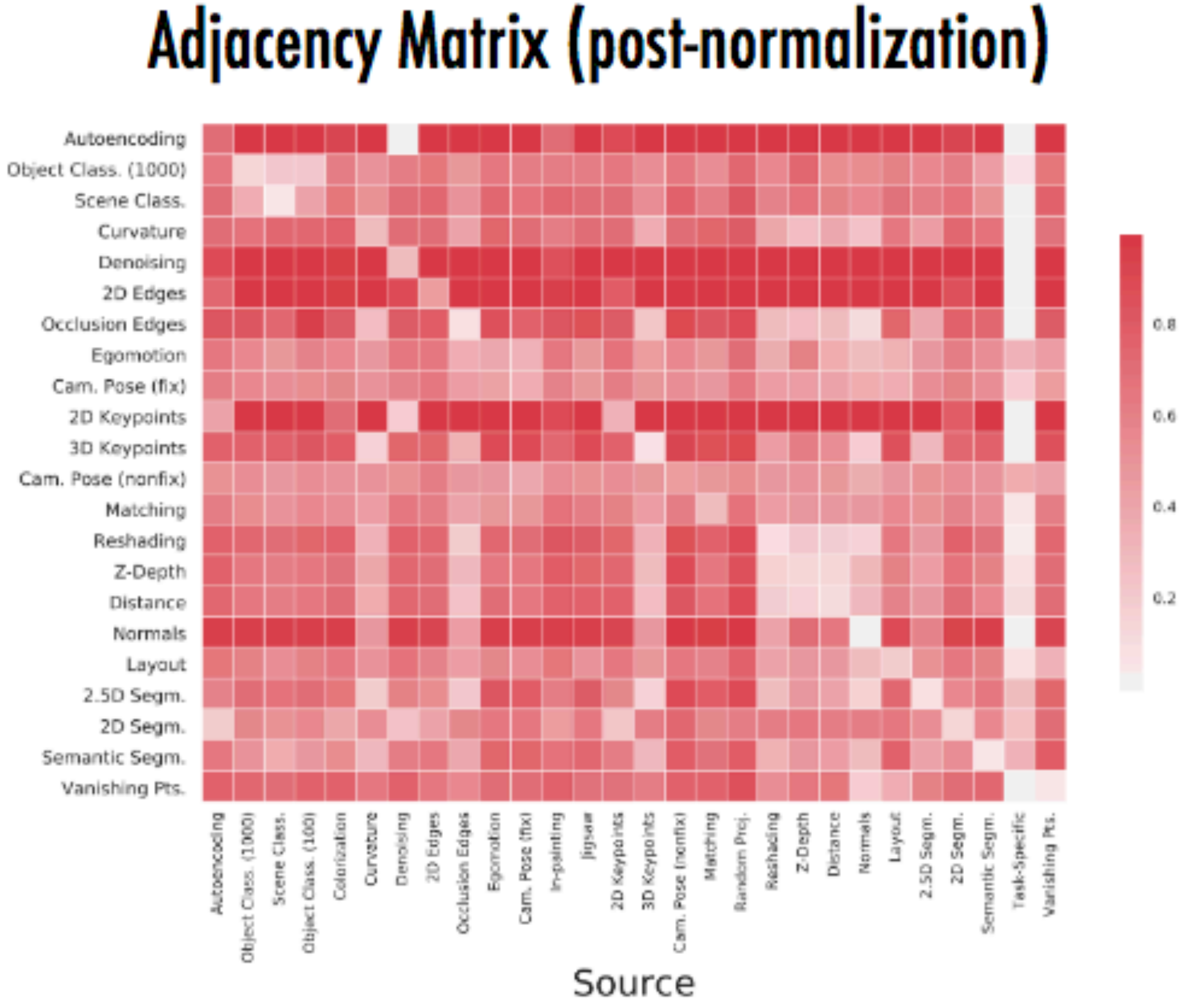


Zamir et al. "Taskonomy: Disentangling Task Transfer Learning." 2018 CVPR.

Current Work on Transfer Learning

Relation Between Task

- Task taxonomy example in Computer Vision (Zamir et al., 2018)
  1. Specific Task Training
  2. Transfer Learning Model
     a. One to one task
     b. Many to one task
  3. Aggregate normalized (ordinal) raw-loss/evaluation between transfer model in Pairwise Matrix

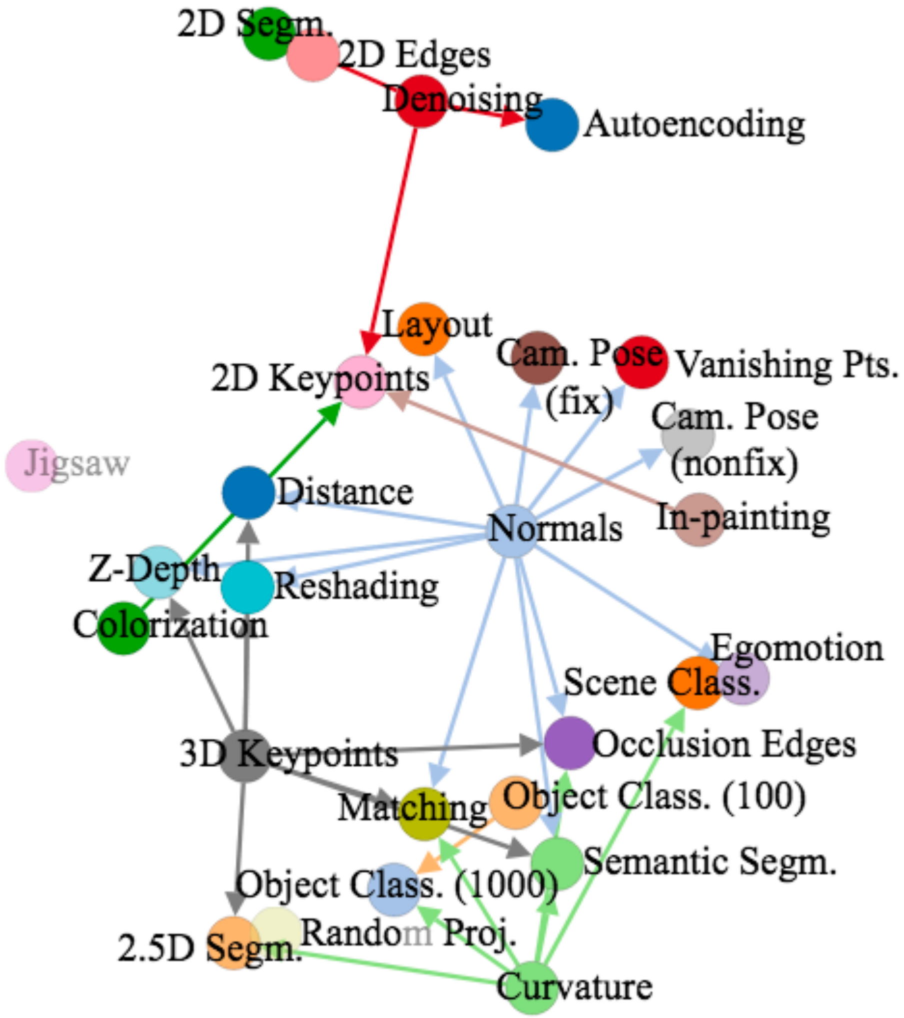


Zamir et al. "Taskonomy: Disentangling Task Transfer Learning." 2018 CVPR.

Current Work on Transfer Learning

Relation Between Task

- Task taxonomy example in Computer Vision (Zamir et al., 2018)
  1. Specific Task Training
  2. Transfer Learning Model
     a. One to one task
     b. Many to one task
  3. Aggregate normalized (ordinal) raw-loss/evaluation between transfer model in Pairwise Matrix
  4. Compute Global Taxonomy



Zamir et al. "Taskonomy: Disentangling Task Transfer Learning." 2018 CVPR.

Current Work on Transfer Learning

Analyzing Model Weight in Transfer Learning

- Inductive Bias
  - What information that source task contains ??
- Transferred Knowledge: weight Similarity between Random and Pre-train (**Neyshabur et al. 2020**)
  - Centered Kernel Alignment (**Kornblith et al., 2019**)

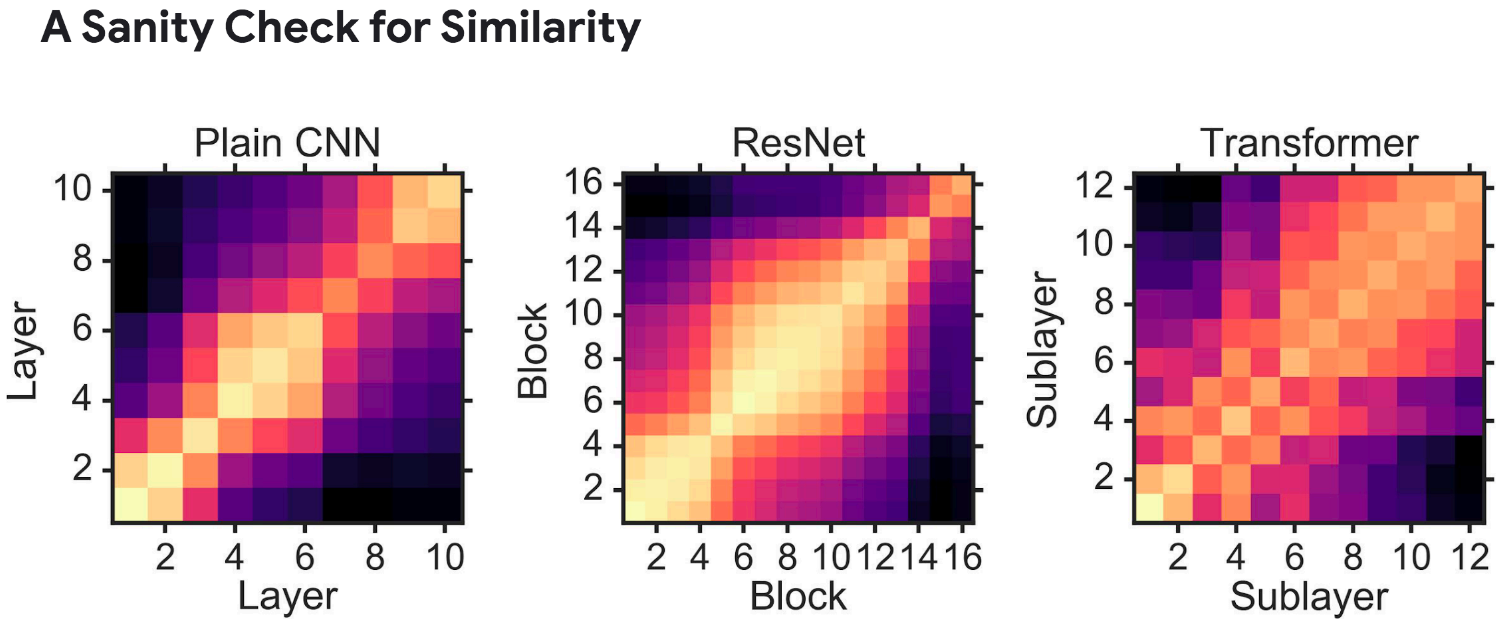
$$\mathrm{CKA}(XX^{\mathrm{T}}, YY^{\mathrm{T}}) = \frac{||Y^{\mathrm{T}}X||_{\mathrm{F}}^2}{||X^{\mathrm{T}}X||_{\mathrm{F}}||Y^{\mathrm{T}}Y||_{\mathrm{F}}} = \frac{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\lambda_X^i\lambda_Y^j\langle \mathbf{u}_X^i, \mathbf{u}_Y^j\rangle^2}{\sqrt{\sum_{i=1}^{p_1}(\lambda_X^i)^2}\sqrt{\sum_{j=1}^{p_2}(\lambda_Y^j)^2}}.$$

Neyshabur et al. "What is being transferred in transfer learning?" (2020).
Kornblith et al. "Similarity of Neural Network Representations Revisited." ICML (2019).

Current Work on Transfer Learning

Analyzing Model Weight in Transfer Learning

- Inductive Bias
  - What information that source task contains ??
- Transferred Knowledge: weight Similarity between Random and Pre-train (**Neyshabur et al. 2020**)
  - Centered Kernel Alignment (**Kornblith et al., 2019**)



Neyshabur et al. "What is being transferred in transfer learning?" (2020).
Kornblith et al. "Similarity of Neural Network Representations Revisited." ICML (2019).

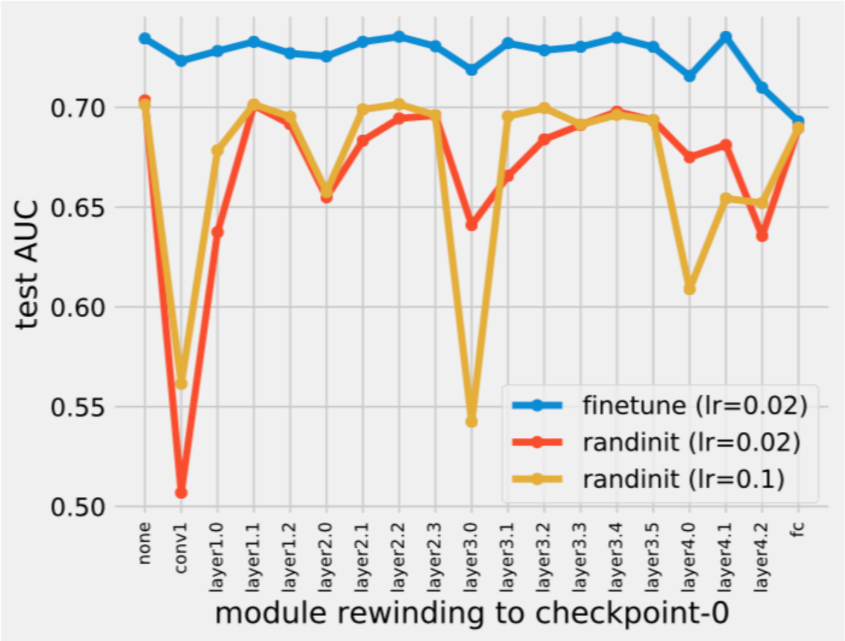
Current Work on Transfer Learning

Analyzing Model Weight in Transfer Learning

- Inductive Bias
  - What information source task contains ??
- Transferred Knowledge : weight Similarity between Random and Pre-train (**Neyshabur et al. 2020**)

Table 1: Feature similarity for different layers of ResNet-50, target domain CheXpert

models/layer	conv1	layer 1	layer 2	layer 3	layer 4
P-T & P	0.6225	0.4592	0.2896	0.1877	0.0453
P-T & P-T	0.6710	0.8230	0.6052	0.4089	0.1628
P-T & RI-T	0.0036	0.0011	0.0022	0.0003	0.0808
RI-T & RI-T	0.0016	0.0088	0.0004	0.0004	0.0424



RI (random initialization), P (pre-trained model),
RI-T (model trained on target domain from random initialization),
P-T (model trained/fine-tuned on target domain starting from pre-trained weights).

Neyshabur et al. "What is being transferred in transfer learning?" (2020).

Transfer Learning vs Continual Learning

Difference	Transfer learning	Continual Learning
Task Boundaries	Source → Target	No Boundaries
Learning Goal	Target Task	Past (Source) and Future (Target) Task
Access to Past Data/Task	Directly (Instance, Parameter re-use/sharing)	Indirectly (Memory in weight/function space)
??? (Comment welcome)	???	???
Similarity	Feature / Parameter Re-use ??? (Comment welcome)	

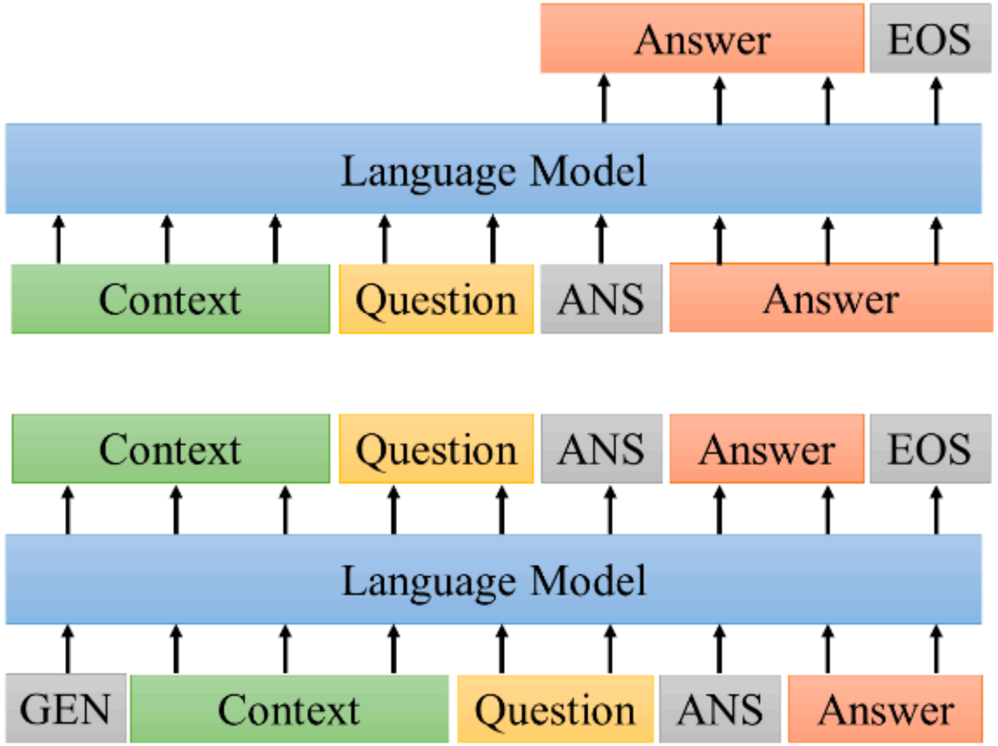
Example: Transfer Learning vs Continual Learning

Representation Across Task

- Cross-Entropy loss over large number of Classes (e.g. |Vocabulary|)
- Generative Model

Examples

Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive



McCann, B. et al. “The Natural Language Decathlon: Multitask Learning as Question Answering.” 2019
Sun, Fan-Keng et al. “LAMOL: LAnguage MOdeling for Lifelong Language Learning.” ICLR 2020

Example: Transfer Learning vs Continual Learning

Representation Across Task

Methods	SST SRL WOZ	SST WOZ SRL	SRL SST WOZ	SRL WOZ SST	WOZ SST SRL	WOZ SRL SST	Average	Std
Fine-tuned	50.2	24.7	62.9	31.3	32.8	33.9	39.3	12
EWC	50.6	48.4	64.7	35.5	43.9	39.0	47.0	8.7
MAS	36.5	45.3	56.6	31.0	49.7	30.8	41.6	8.9
GEM	50.4	29.8	63.3	32.6	44.1	36.3	42.8	11
$\text{LAMOL}^{0}_{\text{GEN}}$	46.5	36.6	56.6	38.6	44.9	45.2	44.8	6.0
$\text{LAMOL}^{0.05}_{\text{GEN}}$	79.6	78.9	73.1	73.7	68.6	75.7	74.9	3.4
$\text{LAMOL}^{0.2}_{\text{GEN}}$	80.0	80.7	79.6	78.7	78.4	80.5	**79.7**	0.8
$\text{LAMOL}^{0}_{\text{TASK}}$	41.0	33.5	50.1	41.9	49.3	41.5	42.9	5.2
$\text{LAMOL}^{0.05}_{\text{TASK}}$	77.3	76.9	78.1	74.7	73.4	75.8	76.0	1.5
$\text{LAMOL}^{0.2}_{\text{TASK}}$	79.4	79.9	80.1	78.7	79.8	79.0	79.5	**0.5**
$\text{LAMOL}^{0.05}_{\text{REAL}}$	81.0	78.9	80.1	80.9	77.7	78.0	79.4	1.2
$\text{LAMOL}^{0.02}_{\text{REAL}}$	81.8	80.6	81.6	81.2	80.4	80.5	81.0	0.5
Multitasked				81.5				

Sun, Fan-Keng et al. “LAMOL: LAnguage MOdeling for Lifelong Language Learning.” ICLR 2020

Possible Direction on Transfer for Continual Learning

Analyzing weight space in transfer learning

- On different architecture
- Continual Learning
  - Could useful for identifying general parameter across task
- On heterogenous transfer learning (output space is different)
  - e.g. classification vs structured prediction