Multi-Task Learning via Generalized Tensor Trace Norm

Yu Zhang (张宇)
Southern University of Science and Technology

Contents

Yi Zhang, **Yu Zhang**, and Wei Wang. Multi-Task Learning via Generalized Tensor Trace Norm. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD**), pp. 2254-2262, 2021.

Multi-Task Learning via Generalized Tensor Trace Norm

01 Introduction

Introduction

Multi-Task Learning

Human Learning

Learn multiple tasks simultaneously

Use the knowledge learned in a task to help the learning of another task

play tennis

play squash

Introduction

Multi-Task Learning

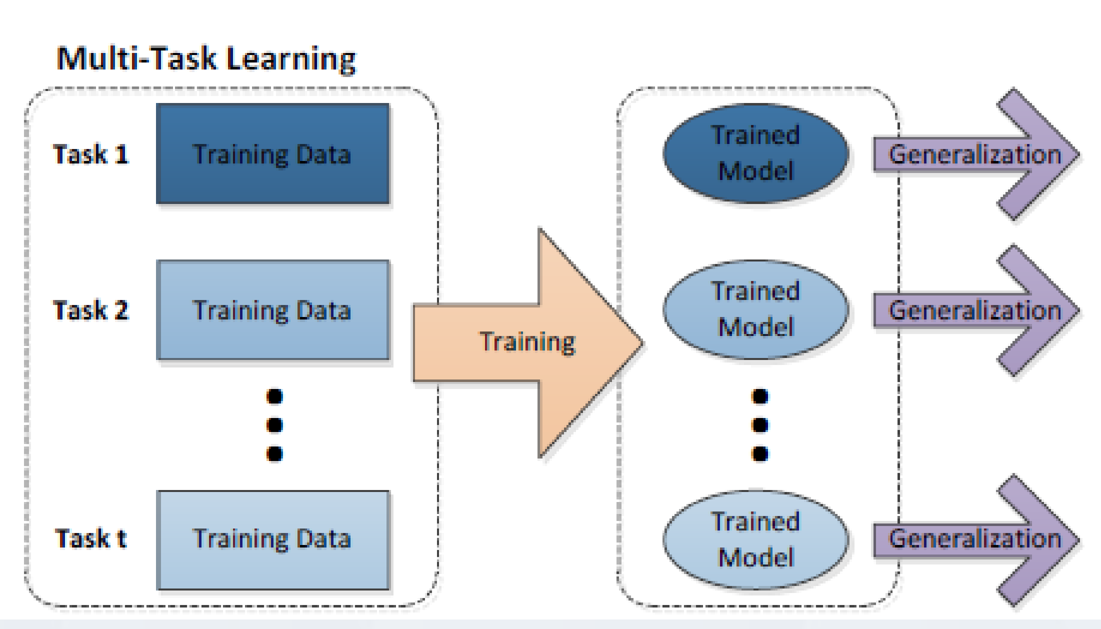
Learn multiple related tasks jointly

⬇

The knowledge contained in a task can be leveraged by other tasks

⬇

Improve the generalization performance of all the tasks



Introduction

Multi-Task Learning in Natural Language Processing

arXiv.org > cs > arXiv:2109.09138

Search...

Help | Advanced

Computer Science > Artificial Intelligence

[Submitted on 19 Sep 2021]

Multi-Task Learning in Natural Language Processing: An Overview

Shijie Chen, Yu Zhang, Qiang Yang

Deep learning approaches have achieved great success in the field of Natural Language Processing (NLP). However, deep neural models often suffer from overfitting and data scarcity problems that are pervasive in NLP tasks. In recent years, Multi-Task Learning (MTL), which can leverage useful information of related tasks to achieve simultaneous performance improvement on multiple related tasks, has been used to handle these problems. In this paper, we give an overview of the use of MTL in NLP tasks. We first review MTL architectures used in NLP tasks and categorize them into four classes, including the parallel architecture, hierarchical architecture, modular architecture, and generative adversarial architecture. Then we present optimization techniques on loss construction, data sampling, and task scheduling to properly train a multi-task model. After presenting applications of MTL in a variety of NLP tasks, we introduce some benchmark datasets. Finally, we make a conclusion and discuss several possible research directions in this field.

Subjects: **Artificial Intelligence (cs.AI)**

Cite as: arXiv:2109.09138 **[cs.AI]**

(or arXiv:2109.09138v1 **[cs.AI]** for this version)

Shijie Chen, **Yu Zhang**, Qiang Yang. Multi-Task Learning in Natural Language Processing: An Overview. arXiv:2109.09138, 2021.

Introduction

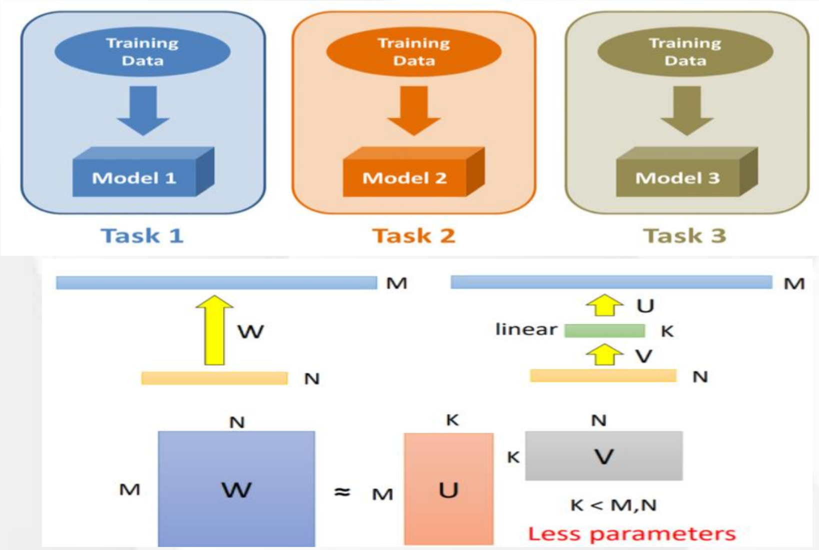
Multi-task learning models

- Feature transformation approach
- Feature selection approach
- Low-rank approach
- Task clustering approach
- Decomposition approach
- Task relation learning approach

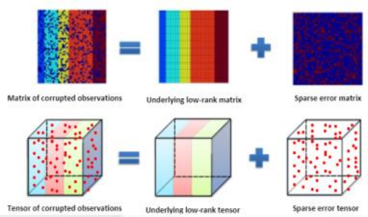
Yu Zhang and Qiang Yang, A Survey on Multi-Task Learning, IEEE TKDE 2021

Introduction

Low-rank approach

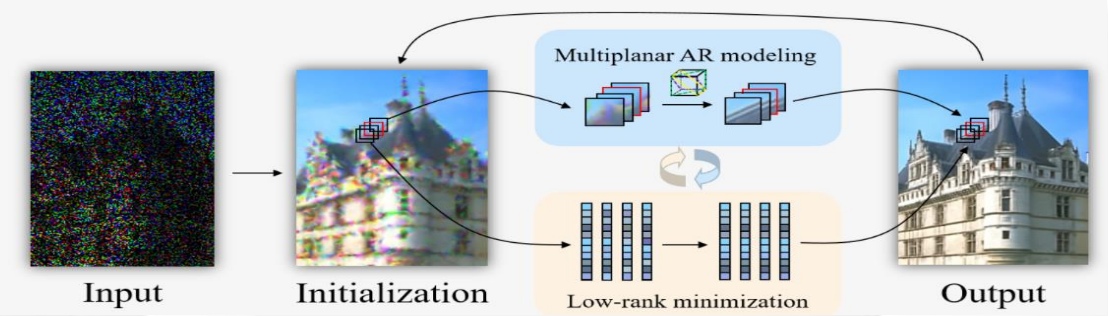


Relatedness among multiple tasks



Low-rank of parameters

Low-rank approach



Introduction

Low-rank approach

Matrix parameters → Matrix trace norm

Multi-task

Image

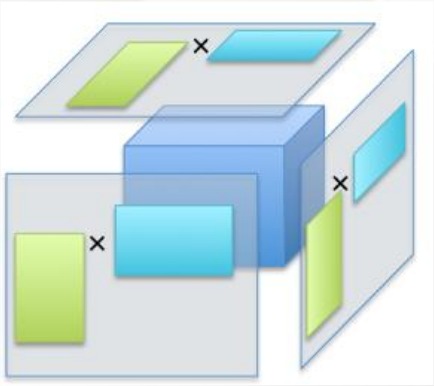
→ Tensor trace norm

Multi-class classification

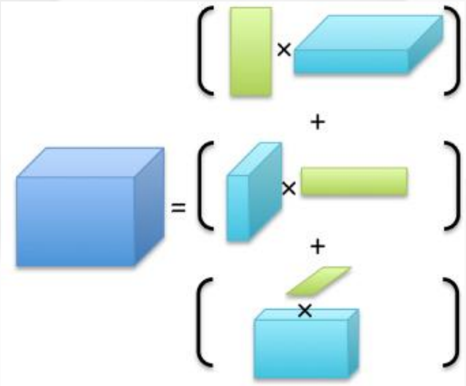
Introduction

Tensor trace norm

Overlapped tensor trace norms

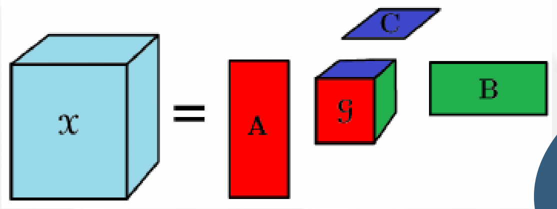


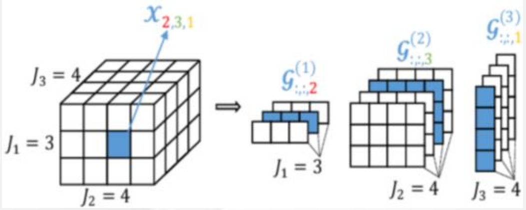
Latent tensor trace norms



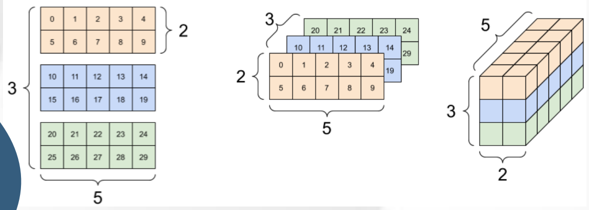
Introduction

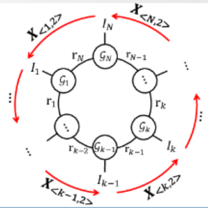
Tucker Trace Norm





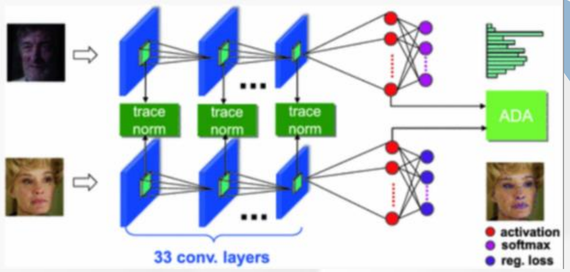
LAF Trace Norm





Overlapped tensor trace norms

Tensor-Train (TT) Trace Norm



Tensor-Ring (TR) Trace Norm

Multi-Task Learning via Generalized Tensor Trace Norm

02 Existing tensor trace norms

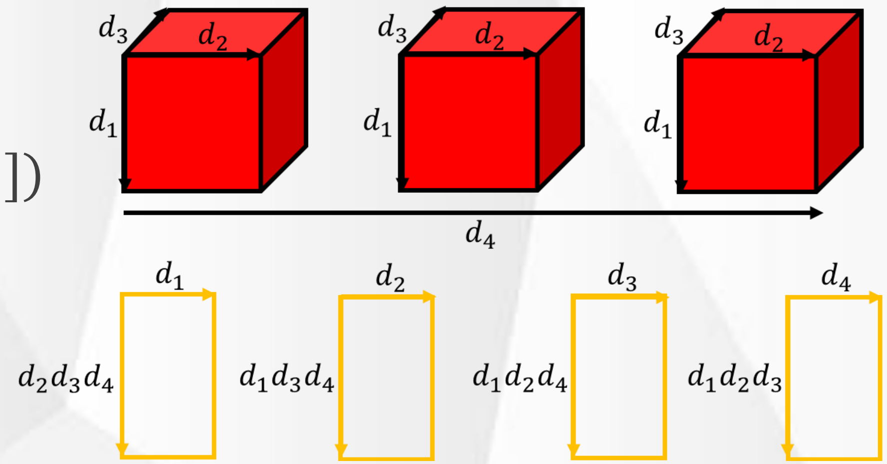
Existing tensor trace norms

Tucker Trace Norm

$$|||\mathcal{W}|||_* = \sum_{i=1}^{p} \alpha_i \, ||\mathcal{W}_{(i)}||_* \qquad \mathcal{W} \in \mathbb{R}^{d_1 \times \cdots \times d_\mathrm{p}}$$

$$\text{s.t.} \begin{cases} \boldsymbol{\mathcal{W}}_{(i)} \coloneqq \mathrm{reshape}(\mathrm{permute}(\boldsymbol{\mathcal{W}}, [i, 1, \ldots, i-1, \mathrm{i}+1, \ldots, p]), [d_i, \prod_{j \neq i} d_j]) \\ \alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1 \end{cases}$$

4-way tensor



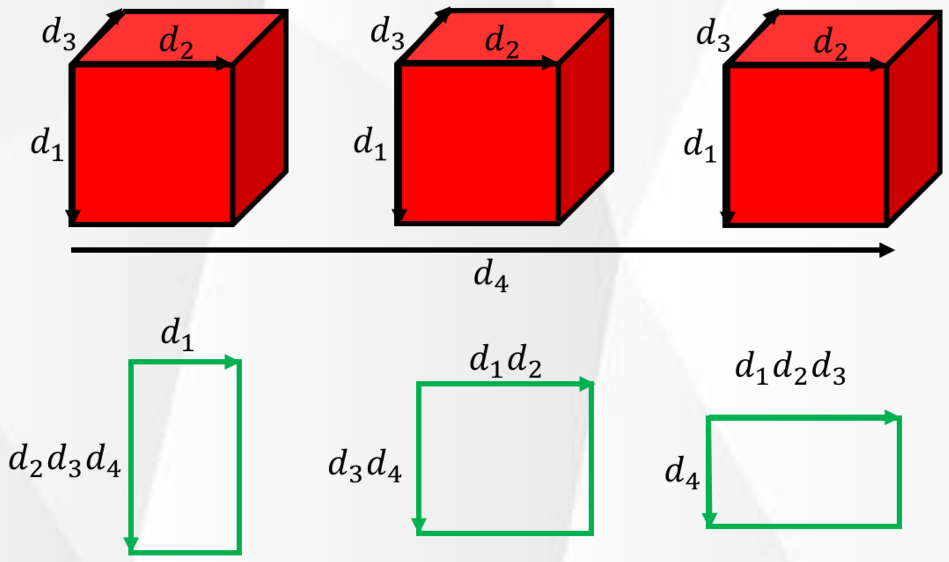
Existing tensor trace norms

TT Trace Norm

$$|||\mathcal{W}|||_* = \sum_{i=1}^{p-1} \alpha_i \, ||\mathcal{W}_{[i]}||_* \qquad \mathcal{W} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$$

$$\text{s.t.} \begin{cases} \mathcal{W}_{[i]} := \text{reshape}(\mathcal{W}, [\prod_{j=1}^{i} d_j, \prod_{j=i+1}^{p} d_j]) \\ \alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1 \end{cases}$$

4-way tensor



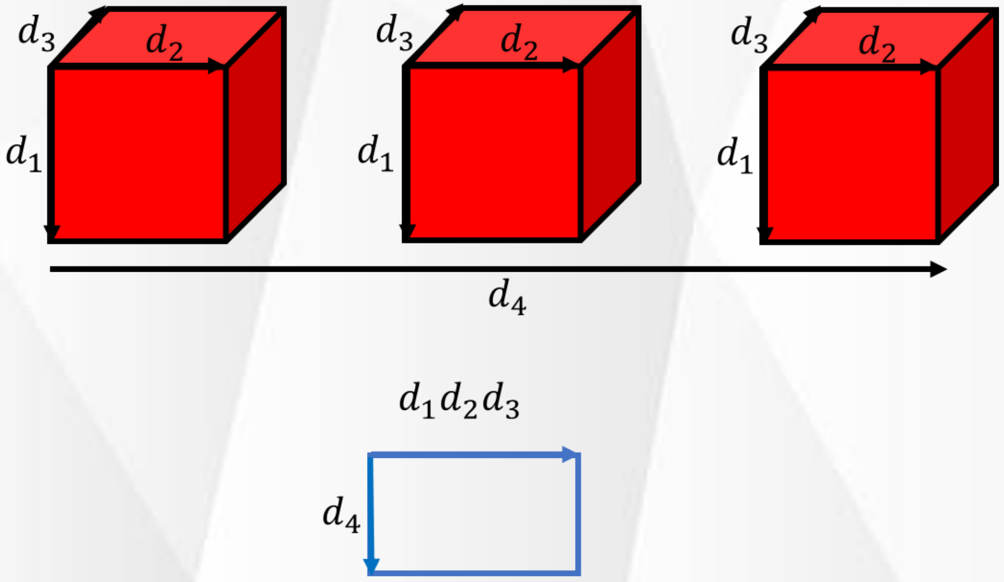
Existing tensor trace norms

LAF Trace Norm

$$|||\mathcal{W}|||_* = ||\mathcal{W}_{(p)}||_* \qquad \boldsymbol{\mathcal{W}} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$$

$$\text{s.t.} \quad \boldsymbol{\mathcal{W}}_{(p)} := \text{reshape}(\boldsymbol{\mathcal{W}}, [\prod_{j=1}^{p-1} d_j \, , d_p])$$

4-way tensor



Existing tensor trace norms

TR trace norm

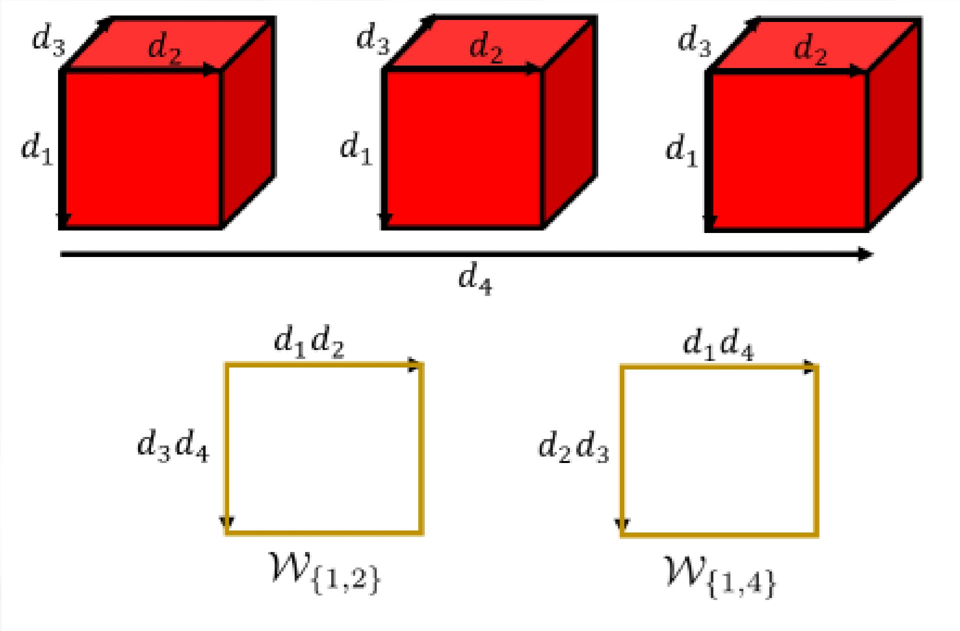
$$|||\mathcal{W}|||_* = \sum_{i=1}^{p} \alpha_i \, ||\mathcal{W}_{<i,d>}||_*$$

$$\mathcal{W} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$$

s.t.

$$\begin{cases} \mathcal{W}_{<i,d>} := reshape(permute(\mathcal{W}, [t, \dots, i, i+1, \dots, t-1]), [\prod_{j=t}^{i} d_j, \prod_{j=i+1}^{t+1} d_j]) \\ t = \begin{cases} i-d+1 & \text{if } d \le i \\ i-d+1+p & \text{otherwise} \end{cases} \qquad \alpha_i \ge 0, \sum_{i=1}^{p} \alpha_i = 1 \end{cases}$$

4-way tensor



Multi-Task Learning via Generalized Tensor Trace Norm

03 Generalized Tensor Trace Norm (GTTN)

Generalized Tensor Trace Norm (GTTN)

How to choose the way of tensor flattening?

- ◆ Try all possible ways of tensor flattening.

Analysis on Existing Tensor Trace Norms

Given the way of tensor flattening, how to determine the importance of resultant tensor flattenings?

Learn Weights:

- ◆ Variable
- ◆ Minimum
- ◆ Maximum
- ◆ Meta-learning

Generalized Tensor Trace Norm (GTTN)

How to choose the way of tensor flattening?

Try all possible ways of tensor flattening

$$\mathcal{W}_* = \sum_{\mathbf{s}} \alpha_{\mathbf{s}} \, ||\mathcal{W}_{\{\mathbf{s}\}}||_* \qquad \mathcal{W} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$$

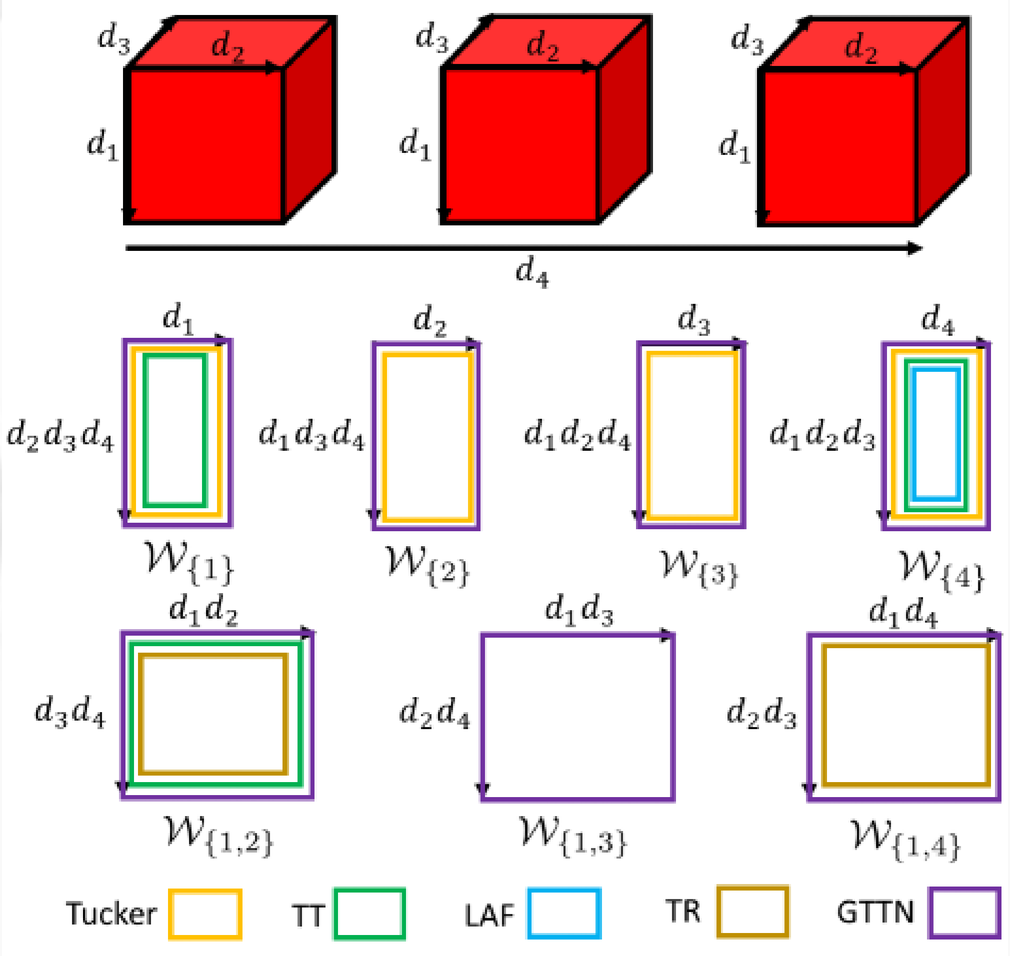
$$\text{s.t.} \begin{cases} \mathcal{W}_{\{\mathbf{s}\}} := \text{reshape}(\text{permute}(\mathcal{W}, [\mathbf{s}, \neg\mathbf{s}]), [\prod_{i \in \mathbf{s}} d_i, \prod_{j \in \neg\mathbf{s}} d_j]) \\ \alpha_{\mathbf{s}} \geq 0, \sum_{\mathbf{s}} \alpha_{\mathbf{s}} = 1 \end{cases}$$

Generalized Tensor Trace Norm (GTTN)

Comparison among the Tucker trace norm, TT trace norm, LAF trace norm, TR trace norm with *d*=2, and GTTN for a 4-way tensor.

How to choose the way of tensor flattening?

- Lemma 1: For a *p*-way tensor, There are $2^{p-1}-1$ distinct tensor flattening.
- $p \leq 5$
- # distinct tensor flattening$\leq$ 15



# distinct tensor flattening:	4	3	1	2	7

Generalized Tensor Trace Norm (GTTN)

How to determine α?

Learn Weights

- View α as variables to be optimized
- Optimize maximum matrix trace norm
- Optimize minimum matrix trace norm
- Meta-learning method

Generalized Tensor Trace Norm (GTTN)

Learning Weights

Viewing α as variables to be optimized

$$\min_{\Theta,\boldsymbol{\alpha}} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i=1}^{n_i} l\left(f_i\left(x_j^i;\Theta\right), y_j^i\right) + \lambda \sum_{i=1}^{N} \alpha_s \,||\mathcal{W}_{\{s\}}||_* \quad \mathrm{s.t.}\, \alpha_{\mathbf{s}} \geq 0, \sum\nolimits_{\mathbf{s}} \alpha_{\mathbf{s}} = 1$$

softmax function

$$\boldsymbol{\alpha}_s = \frac{exp\{\beta_s\}}{\sum_{t\in[p],t\neq\emptyset} \exp\{\beta_s\}}$$

β_s instead of $\alpha_{\mathbf{s}}$ is treated as a variable.

Generalized Tensor Trace Norm (GTTN)

Learning Weights

Optimizing minimum matrix trace norm

$$\min_{\Theta} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i=1}^{n_i} l\left(f_i\left(x_j^i;\Theta\right), y_j^i\right) + \lambda \min_{\substack{s \subset [p] \\ s \neq \emptyset}} ||\mathcal{W}_{\{s\}}||_*$$

Mathematically Equivalent

$$\min_{\Theta,\boldsymbol{\alpha}} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i=1}^{n_i} l\left(f_i\left(x_j^i;\Theta\right), y_j^i\right) + \lambda \sum_{i=1}^{N} \alpha_s ||\mathcal{W}_{\{s\}}||_* \ \text{s.t.}\ \alpha_{\mathbf{s}} \geq 0, \sum_{\mathbf{s}} \alpha_{\mathbf{s}} = 1$$

softmax function

cannot achieve 0 or 1 exactly

Generalized Tensor Trace Norm (GTTN)

Learning Weights

Optimizing maximum matrix trace norm

$$\min_{\Theta} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i=1}^{n_i} l\left(f_i\left(x_j^i; \Theta\right), y_j^i\right) + \lambda \max_{\substack{s \subset [p] \\ s \neq \emptyset}} ||\mathcal{W}_{\{s\}}||_*$$

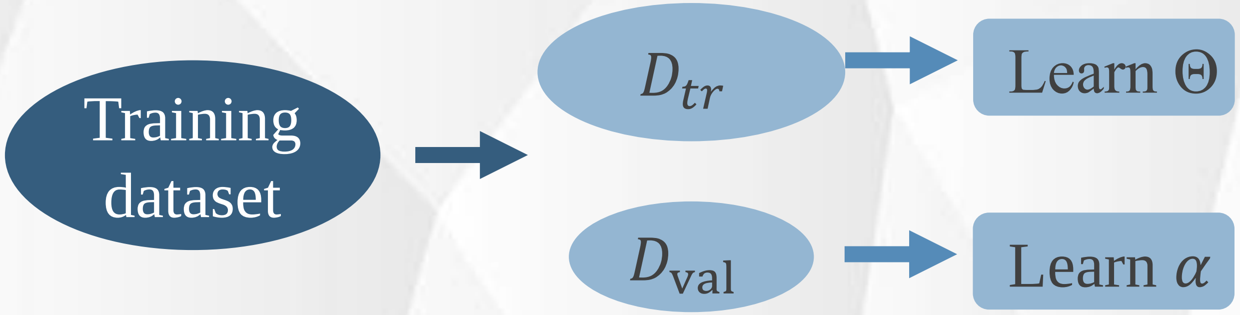
Pay attention to flattenings with the largest trace norm

Penalize all the matrix trace norm

Generalized Tensor Trace Norm (GTTN)

Learning Weights

Meta-learning method



$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \frac{1}{|D_{val}^{i}|} \sum_{(x,y)\epsilon D_{val}^{i}} l(f_i(\boldsymbol{x}; \boldsymbol{\Theta}^*), y)$$

$$\text{s.t. } \boldsymbol{\Theta}^* = argmin \sum_{i=1}^{m} \frac{1}{|D_{tr}^{i}|} \sum_{(x,y)\epsilon D_{tr}^{i}} l(f_i(\boldsymbol{x}; \boldsymbol{\Theta}), y) + \lambda \sum_{i=1}^{p} \alpha_s \, ||\mathcal{W}_{\{s\}}||_*$$

Generalized Tensor Trace Norm (GTTN)

Objective function

Given the Generalized Tensor Trace Norm (GTTN), the objective function of the deep multi-task model can be expressed as:

$$\min_{\Theta} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i=1}^{n_i} l\left(f_i\left(x_j^i; \Theta\right), y_j^i\right) + \lambda |||\boldsymbol{\mathcal{W}}|||_*$$

Empirical loss

Regularization term: GTTN

Generalized Tensor Trace Norm (GTTN)

Analysis

THEOREM 2. *For the solution $\hat{\mathcal{W}}$ of problem (8) and $\delta > 0$, with probability at least $1-\delta$, we have*

$$L(\hat{\mathcal{W}}) \leq \hat{L}(\hat{\mathcal{W}}) + \frac{2\rho\gamma C}{mn_0} \min_{\substack{s \neq \emptyset \\ s \subset [p]}} \left(\frac{\kappa m \sqrt{\ln d_s}}{\alpha_s n_0 d} + \frac{\ln d_s}{\alpha_s n_0} \right) + \sqrt{\frac{2}{m} \ln \frac{1}{\delta}}.$$

Multi-Task Learning via Generalized Tensor Trace Norm

04 Experiments

Experiments

Baselines

DMTL	Deep Multi-Task Learning method
Tucker	The Tucker trace norm method
TT	The TT trace norm method
LAF	The LAF trace norm method
LAF-i	The trace norm regularization method based on the i-th axis flattening
TR	TR trace norm method ($d = 2$)
LAF-TF	LAF tensor factorization method
Prod	The rank-product regularization method
GTTN-a	Setting the weights in GTTN to be same

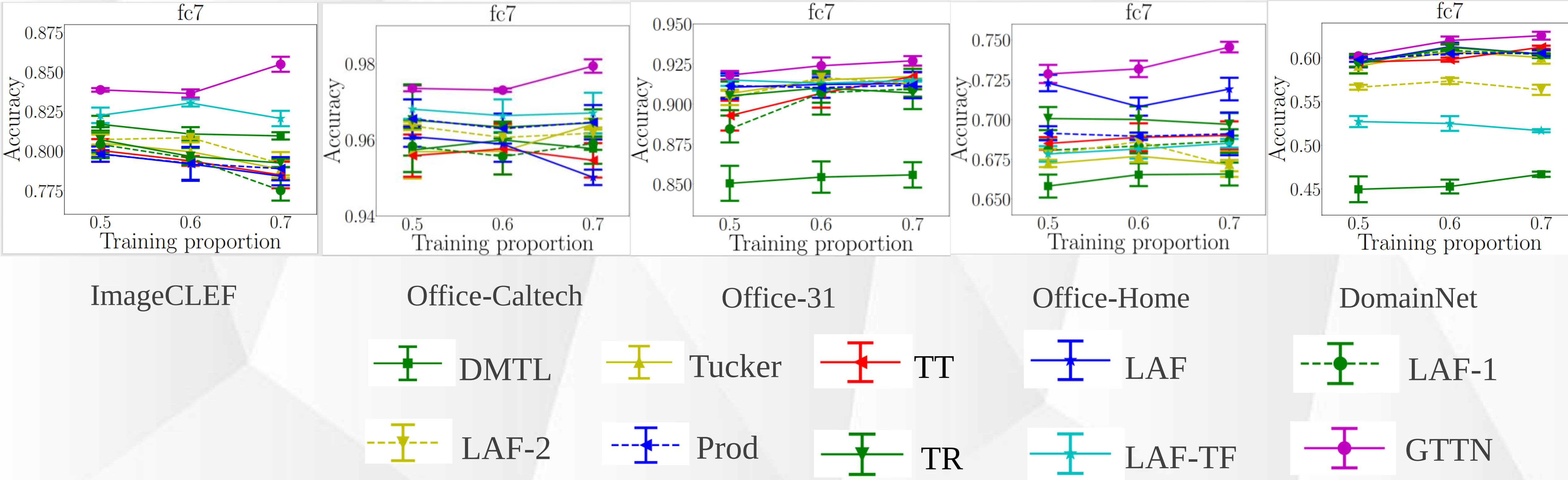
Experiments

Datasets

Dataset	#Images	#Classes	#Tasks
ImageCLEF	2,400	12	4
Office-Caltech	2,533	10	4
Office-31	4,110	31	3
Office-Home	15,500	65	4
DomainNet	600,000	345	6

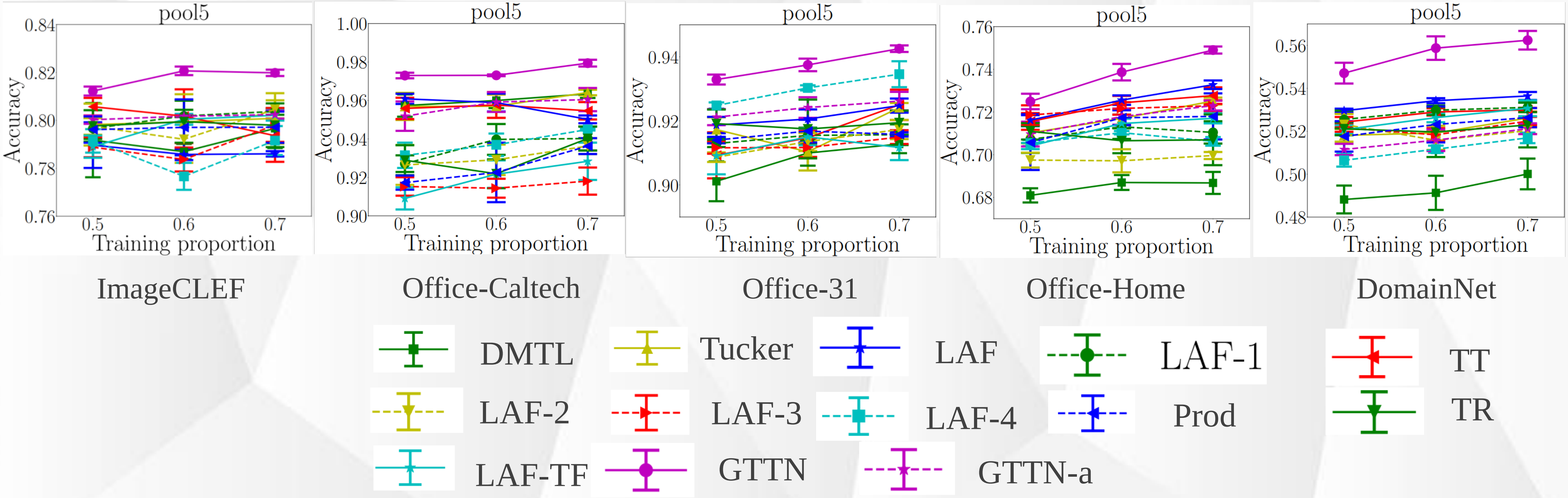
Experiments

Results: fc7 layer



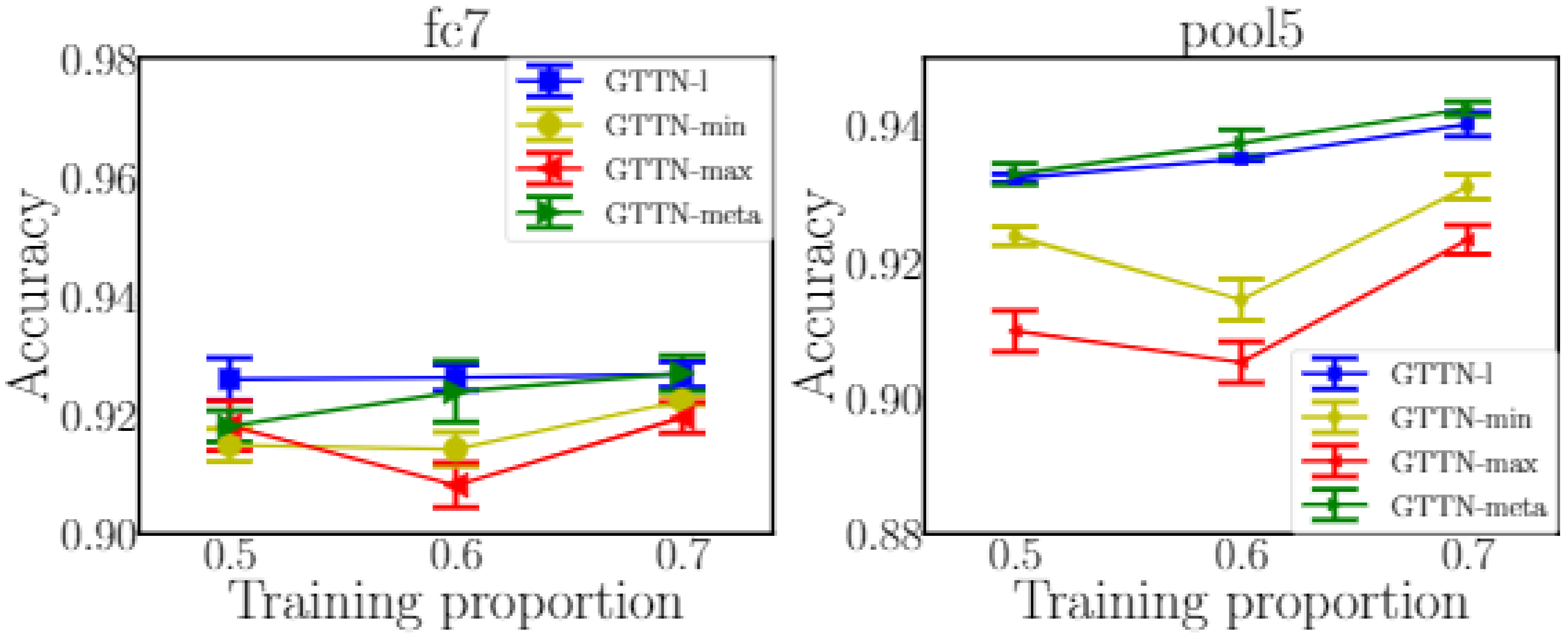
Experiments

Results: pool5 layer



Experiments

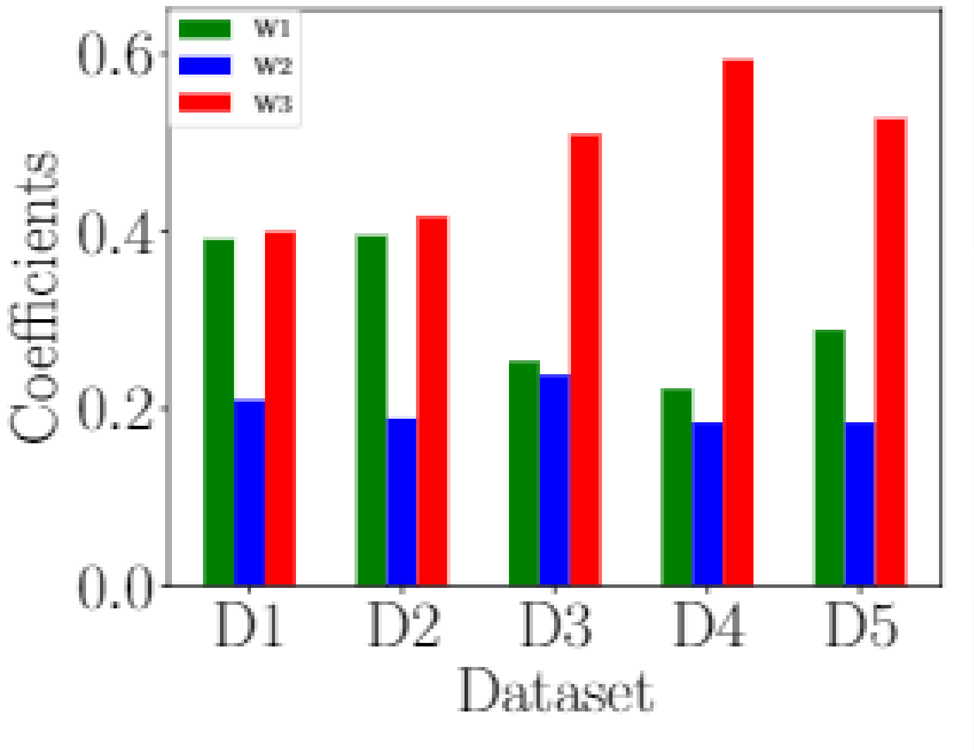
Comparison on Strategies to Learn Weights



(a) Comparison of GTTN (b) Comparison of GTTN

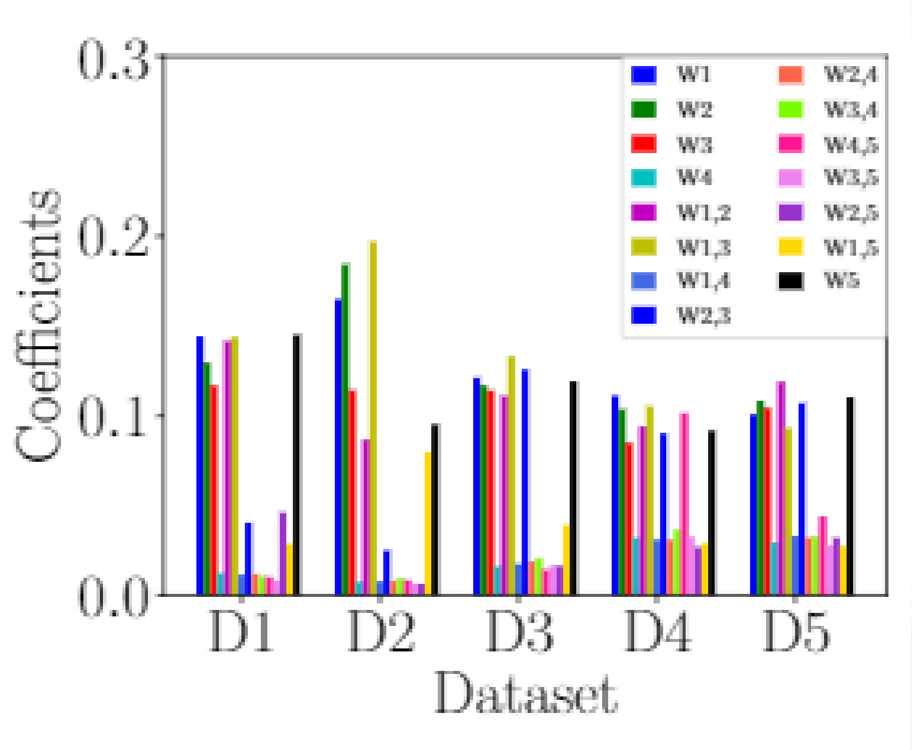
Experiments

Analysis on Learned Weights



(c) Learned α (fc7)

$\mathcal{W}_{\{2\}}$ is smaller

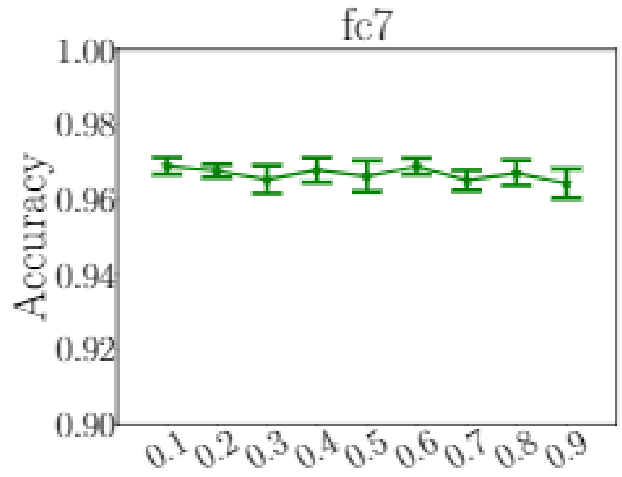


(d) Learned α (pool5)

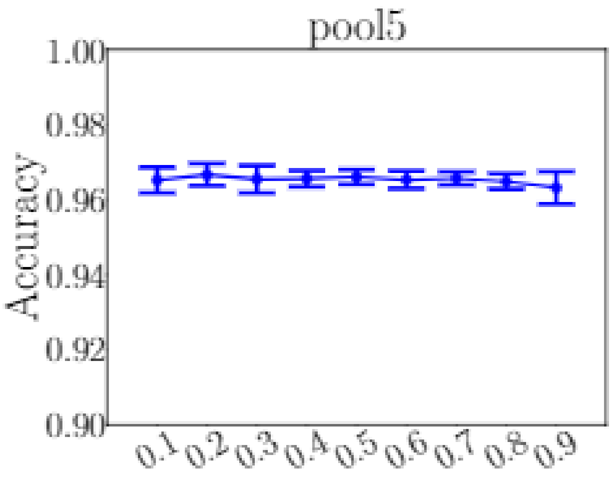
$\mathcal{W}_{\{1\}}, \mathcal{W}_{\{2\}}, \mathcal{W}_{\{3\}}, \mathcal{W}_{\{1,2\}}, \mathcal{W}_{\{1,3\}}, \mathcal{W}_{\{5\}}$ are larger

Experiments

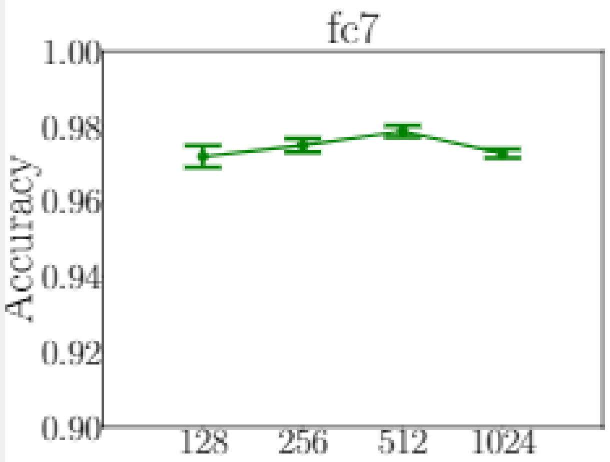
Sensitivity Analysis



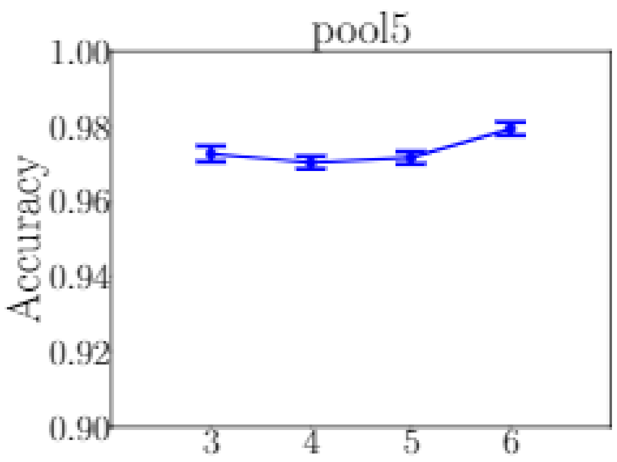
(a) λ (fc7)



(b) λ (pool5)



(c) Hidden units (fc7)



(d) d (pool5)

The performance is not sensitive to λ

number of hidden units = 512, d=6

Conclusion

- The generalized tensor trace norm (GTTN) to capture all the low-rank is effective.
- Learning weights of each tensor flattening to identify the importance of each structure is helpful.
- The GTTN method performs better than baseline methods.

Thank you !