Introduction to Multi-task Learning

목충협

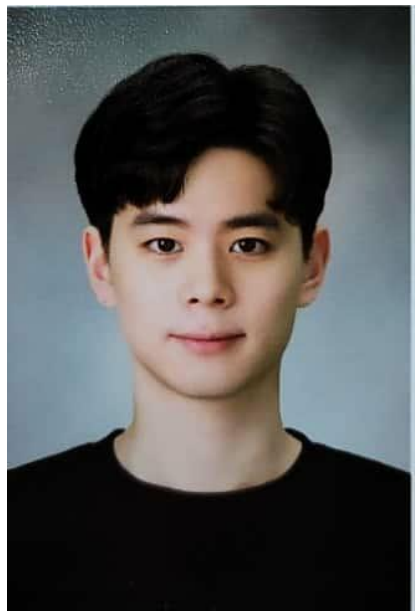
2021.10.22

Introduction to Multi-task Learning

목충협

2021.10.22

발표자 소개



- 목충협(Chunghyup Mok)
  - ✓ 고려대학교 산업경영공학
  - ✓ Data Mining & Quality Analytics Lab.
  - ✓ 석박통합과정(2019.03 ~)
- 관심분야
  - ✓ Machine learning / Deep learning
  - ✓ Meta-learning / Multi-task learning
- E-mail : mokch@korea.ac.kr

Contents

Introduction

What is mulit-task learning?

Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks.

-Wikipedia

여러 태스크들의 공통점과 차이점을 활용하면서

동시에 해결하는

머신러닝의 하위 분야

https://en.wikipedia.org/wiki/Multi-task_learning

Introduction

What is a task?

❖ Task

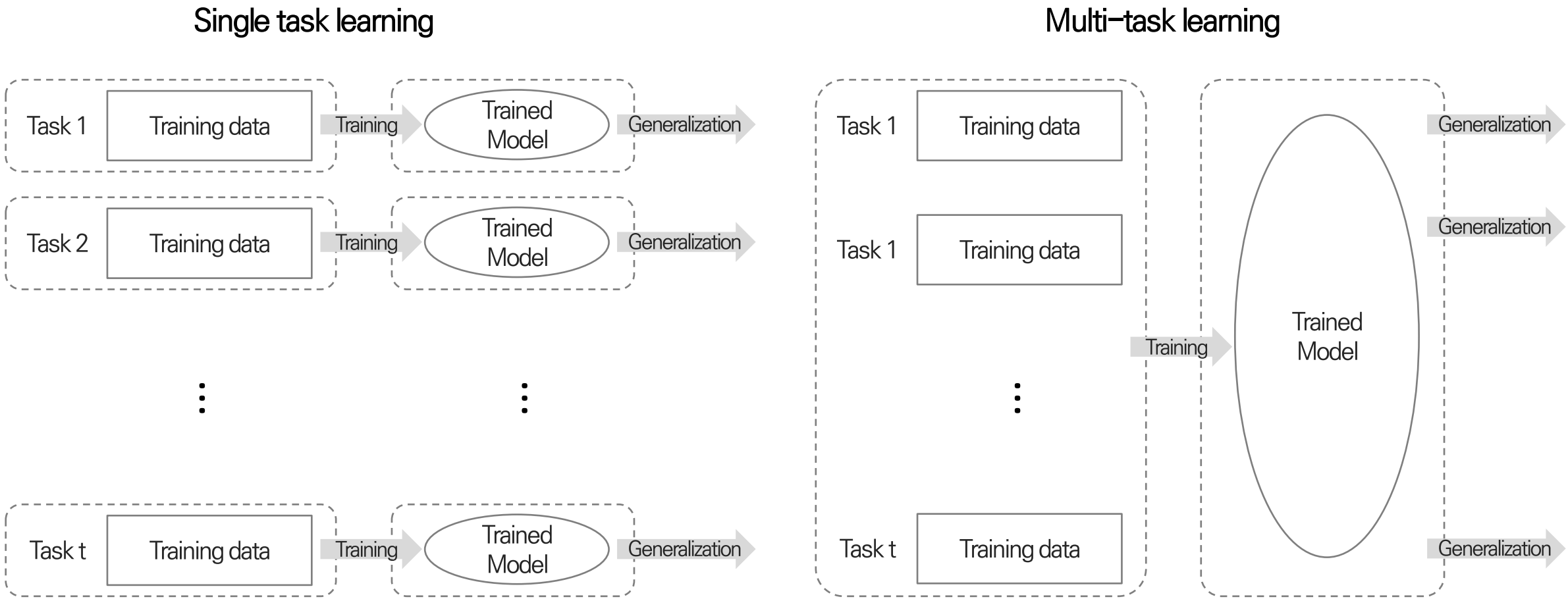
데이터셋 $D(X, y)$, 손실함수 L

X→y 를 일반화하는 모델 구축

$$T_i \triangleq \{p_i(X), p_i(y|X), L_i\}$$

Introduction

Single task learning vs Multi-task learning



Introduction

Learning methods

❖ Transfer learning

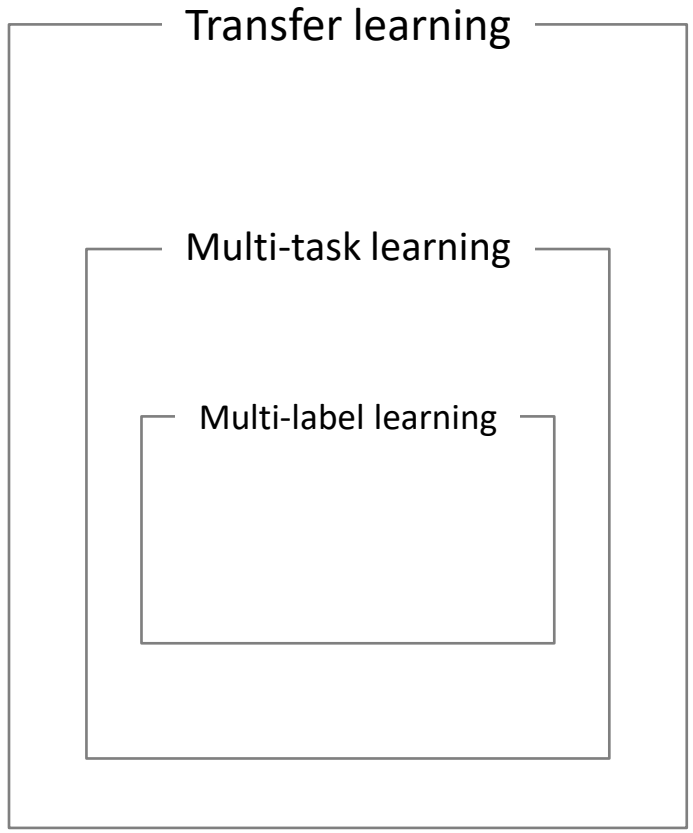
- 여러 Task를 사용하는 대부분의 방법론을 포함하는 개념
  - ✓ Self-supervised, meta-learning, …

❖ Multi-task learning

- 여러 Task를 동시에 학습하는 방법론

❖ Multi-label learning

- 하나의 데이터셋에 여러 레이블을 동시에 학습하는 방법론



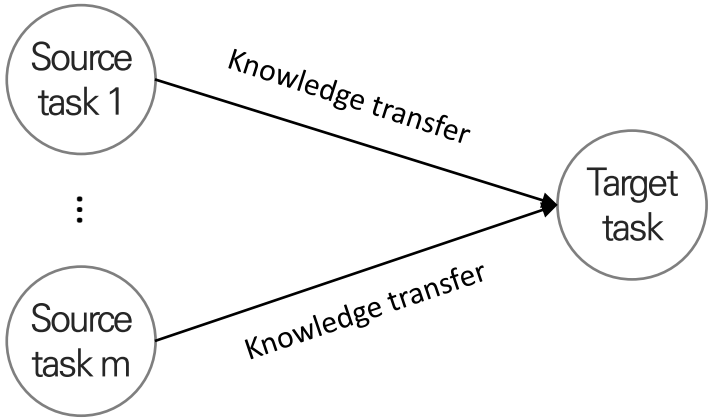
Introduction

Transfer learning vs multi-task learning

❖ Transfer learning

- Target task와 source task로 구분됨
- Target task를 잘 하는 것이 목표
- Source task에서 얻은 지식을 target task에서 활용

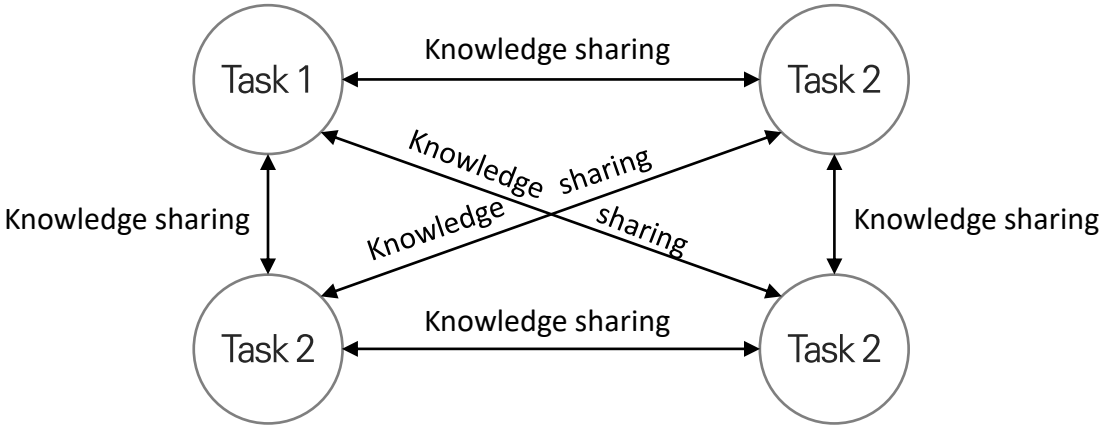
  → Sequential learning



❖ Multi-task learning

- Target task와 source task로 구분되지 않음
- 모든 task들을 잘 하는 것이 목표
- 각 task에서 얻은 지식들을 공유하여 각각 태스크에 활용

  → Parallel learning



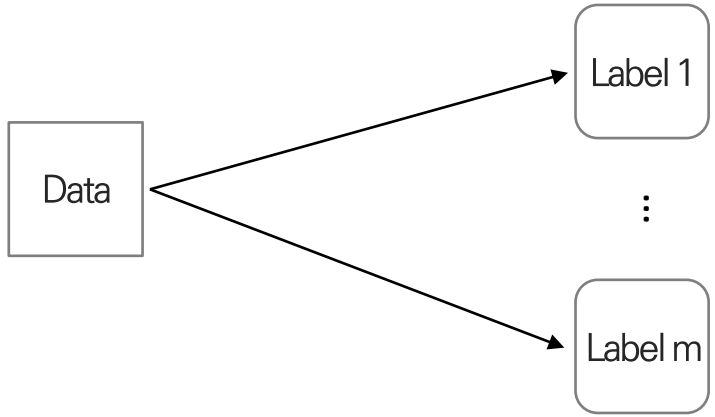
Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. National Science Review, 5(1), 30-43.

Introduction

Multi-label learning vs multi-task learning

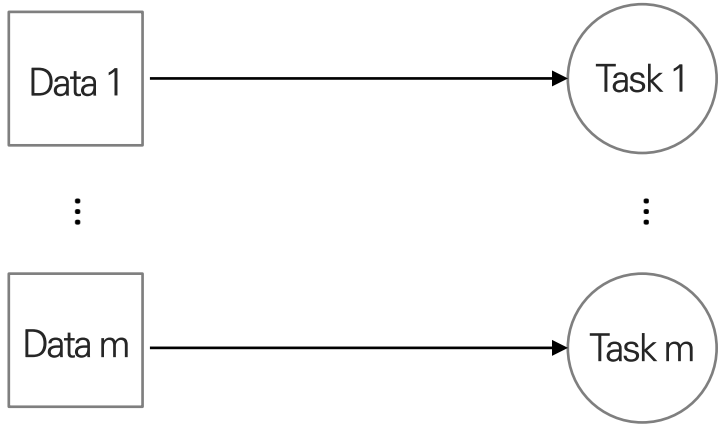
❖ Multi-label learning

- 하나의 데이터셋 사용
- 각 데이터 샘플별로 m개의 레이블이 존재
- 동시에 여러 레이블들을 학습



❖ Multi-task learning

- 여러 데이터셋으로 구성 가능
- 데이터별로 다른 태스크 수행 가능
- 동시에 여러 데이터셋을 학습



Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. National Science Review, 5(1), 30-43.

Multi-task model

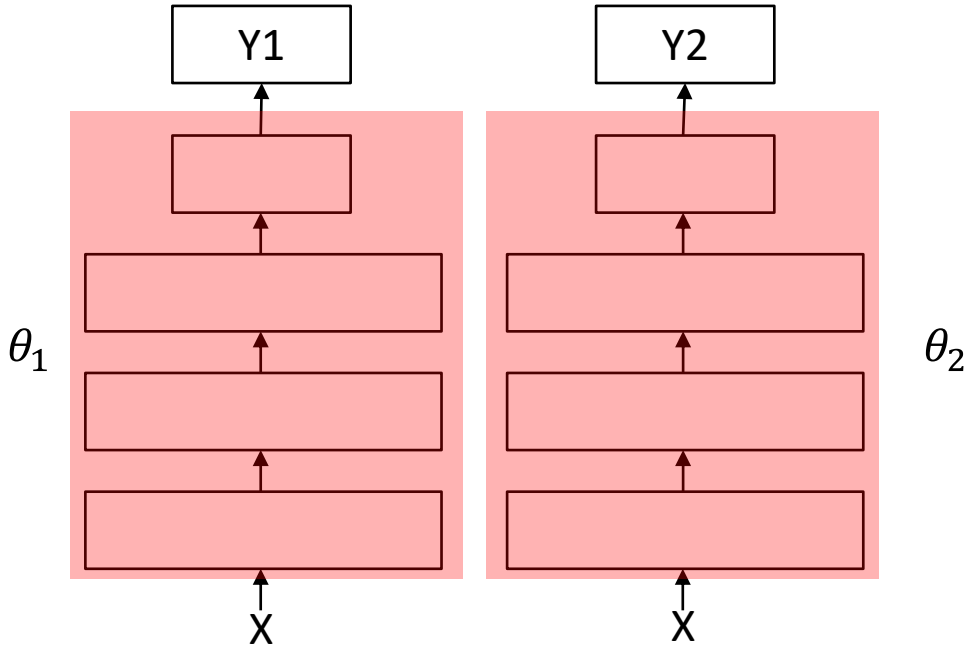
For multi-label dataset

❖ 하나의 데이터셋을 사용할 경우 (without sharing)

- 레이블별로 모델링이 가능
- 같은 데이터가 각각의 모델의 입력 값으로 사용
- 각 모델은 하나의 레이블을 예측하는데 최적화됨
- 손실함수 (Loss function)

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta_i, \mathcal{D})$$

θ_i별로 사용되는 Loss가 다름



Multi-task model

For multi-label dataset

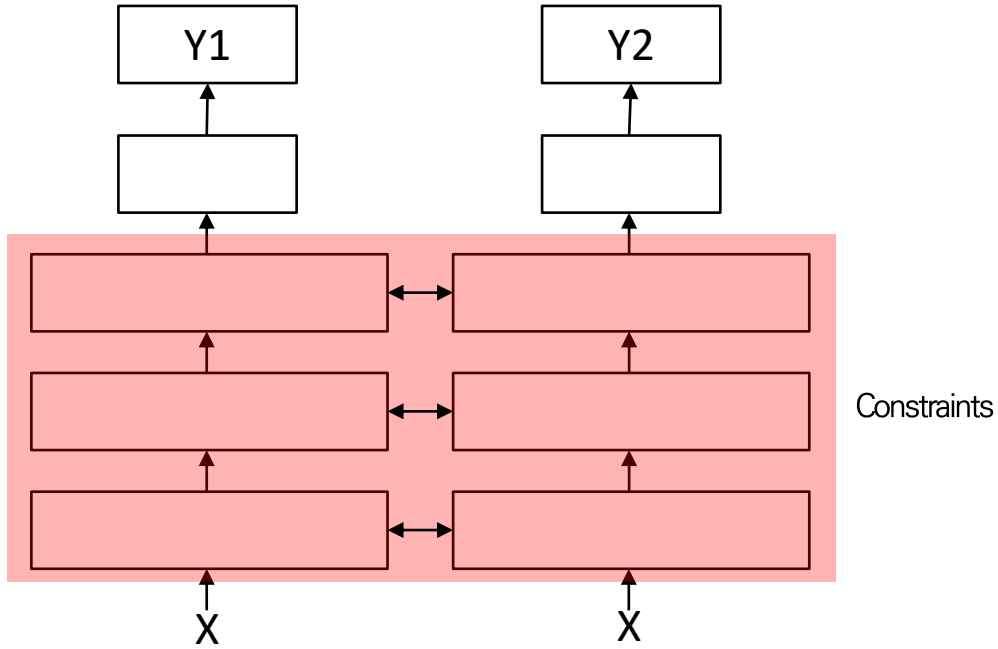
❖ 파라미터를 공유하는 경우 (Soft parameter sharing)

- 각 모델의 파라미터간 거리를 줄이는 제약식 추가됨

  →ex. L2 norm regularization
- 파라미터들이 비슷하게 학습되며 일반화(regularization) 효과
- 일반적으로 입력 데이터에 가까운 파라미터부터 공유
- 손실함수 (Loss function)

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta_i, \mathcal{D}) + \sum_{t'=1}^{T} \|\theta^t - \theta^{t'}\|$$

Knowledge sharing이 일어나는 부분



Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Multi-task model

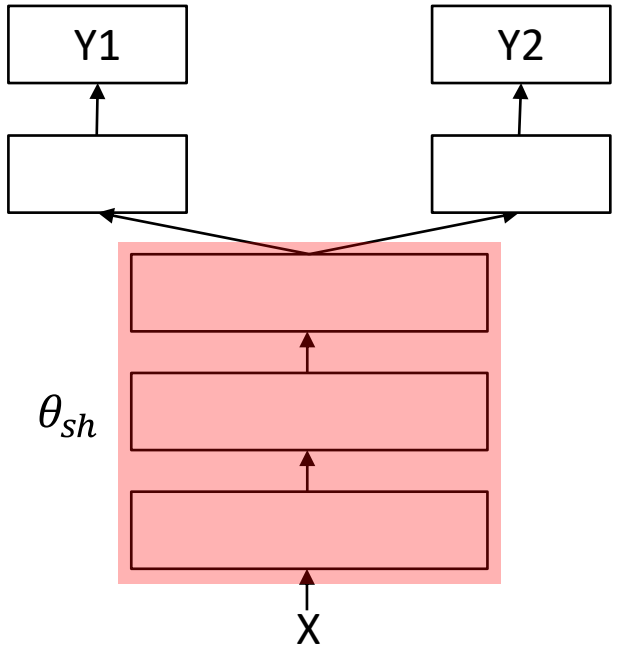
For multi-label dataset

❖ 파라미터를 공유하는 경우 (Hard parameter sharing)

- 일부 파라미터들을 완전히 공유하는 형태
- 하나의 모델로 여러 개의 레이블 예측
- 가장 많이 알려진 구조 (Multi-head)
- 공유되는 파라미터(θ_{sh})와 태스크별로 다른 파라미터(θ_i)로 나뉨
- 파라미터 수가 적으며, 가장 일반화(regularization) 효과가 큼
- 손실함수 (Loss function)

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\{\theta_{sh}, \theta_i\}, \mathcal{D})$$

Knowledge sharing이 일어나는 부분
공유되는 파라미터(θ_{sh})는 모든 Loss를 사용



Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

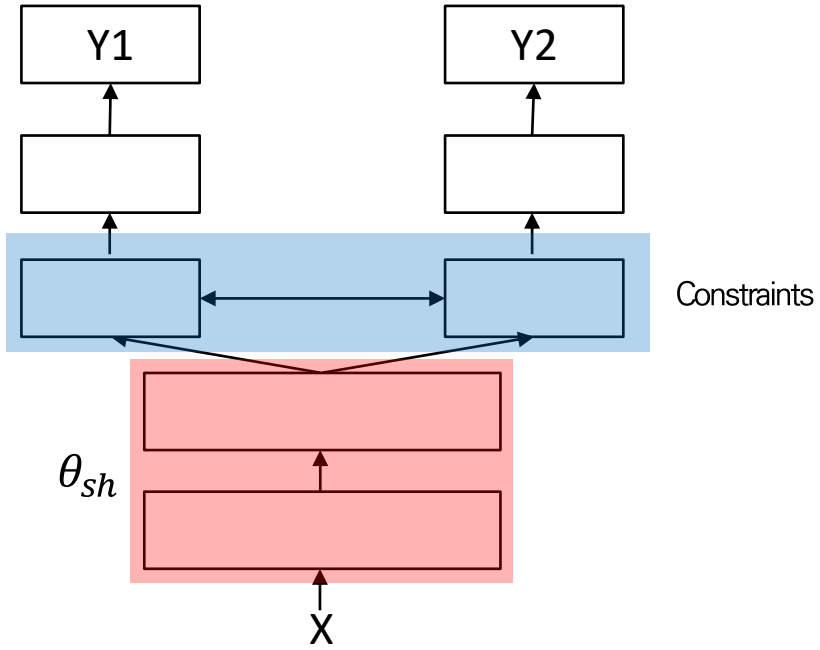
Multi-task model

For multi-label dataset

❖ 파라미터를 공유하는 경우

- 두 가지의 혼합 형태도 가능
- 데이터 특성에 따라 자유롭게 구성 가능

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\{\theta_{sh}, \theta_i\}, \mathcal{D}) + \sum_{t'=1}^{T} \left\| \theta^t - \theta^{t'} \right\|$$



Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

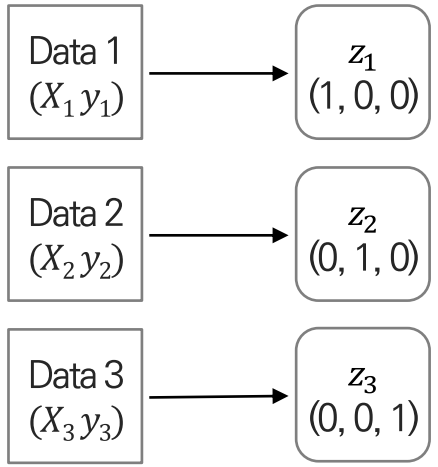
Multi-task model

For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector



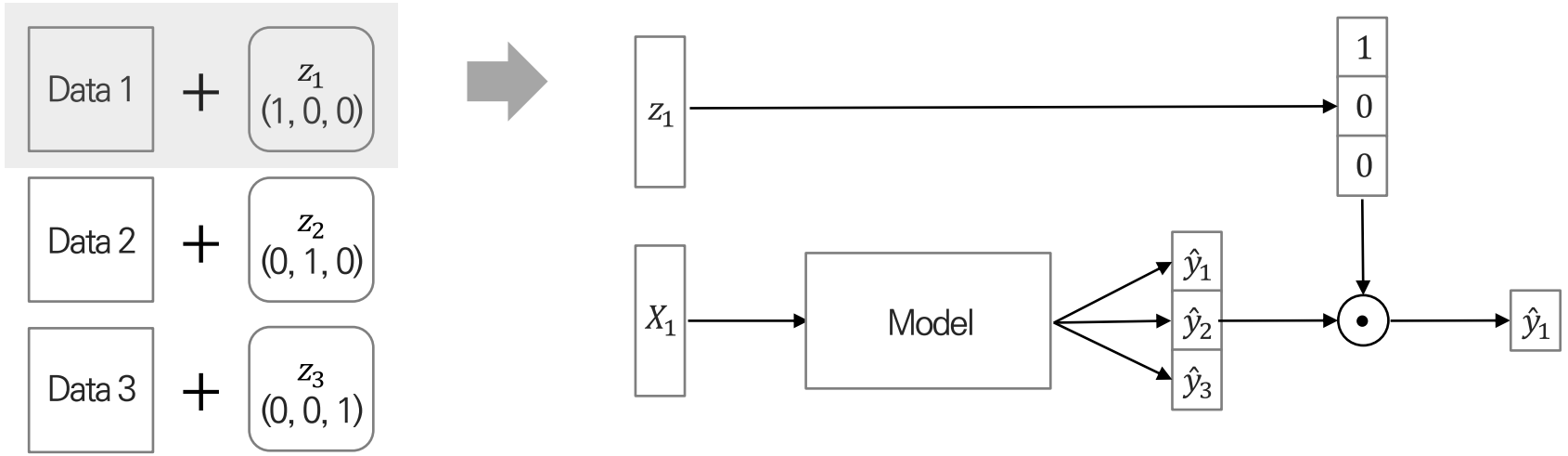
Multi-task model

For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector



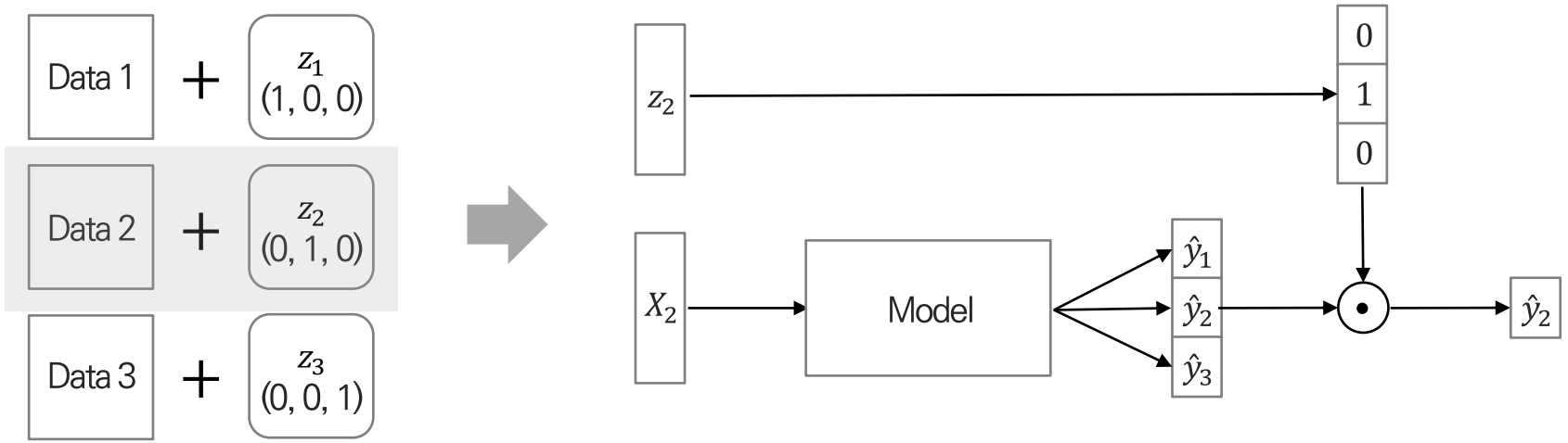
Multi-task model

For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector



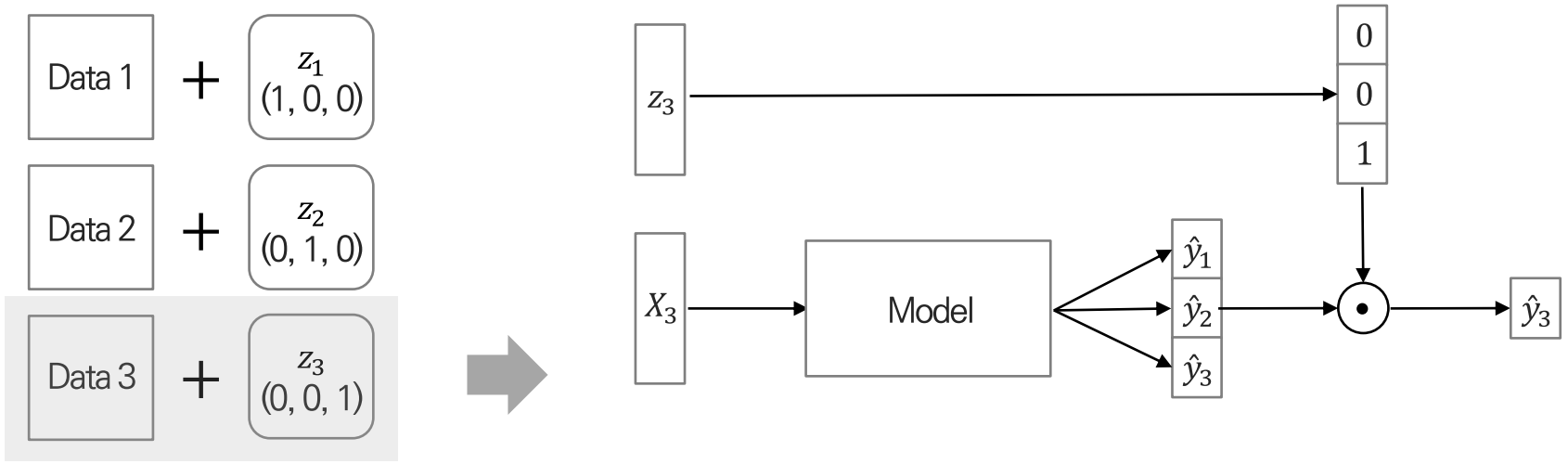
Multi-task model

For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector



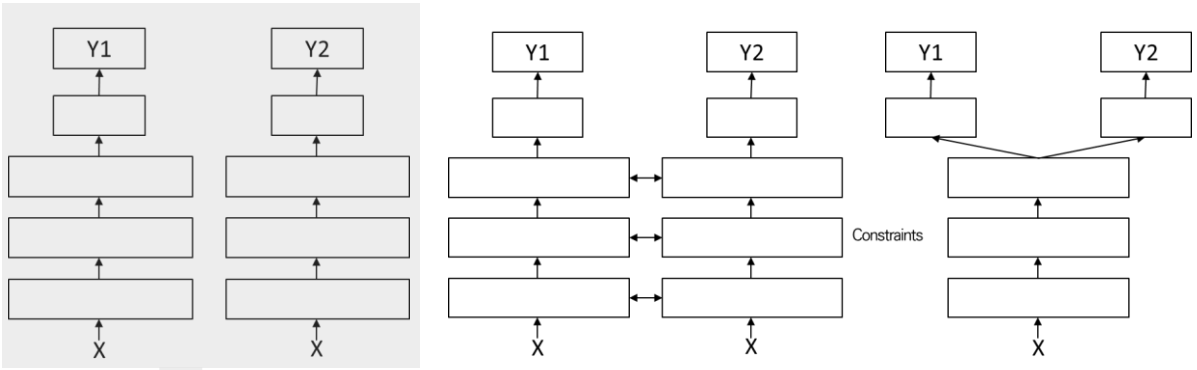
Multi-task model

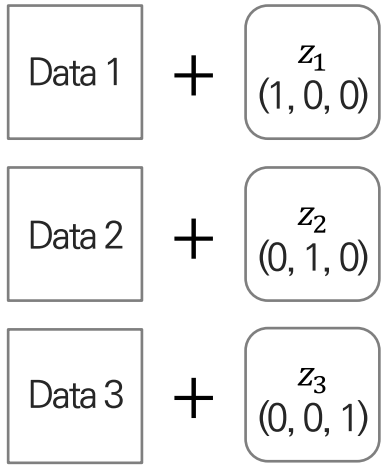
For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

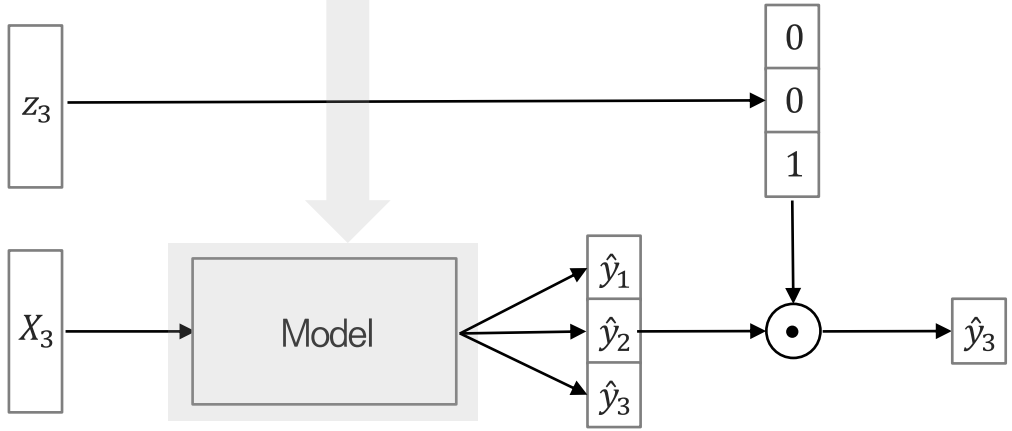
- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector



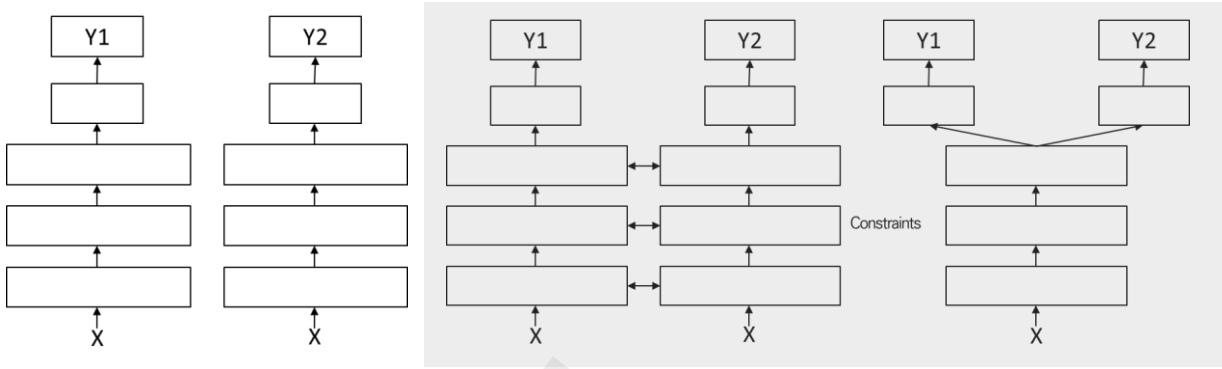


Expert model 여러 개를 사용하는
Multiplicative gating 효과



Multi-task model

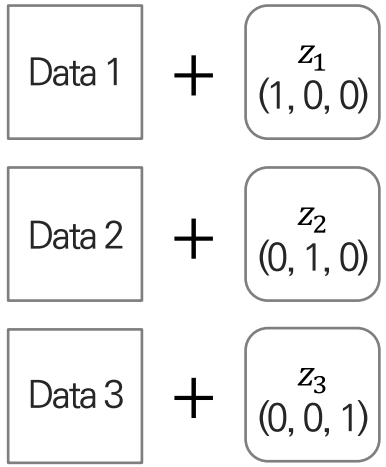
For multi dataset

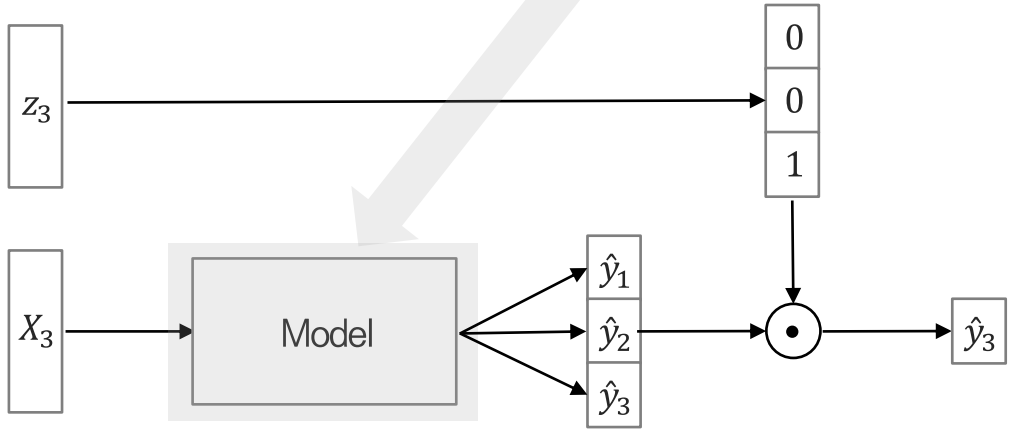


❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. One-hot vector





Multi-task model

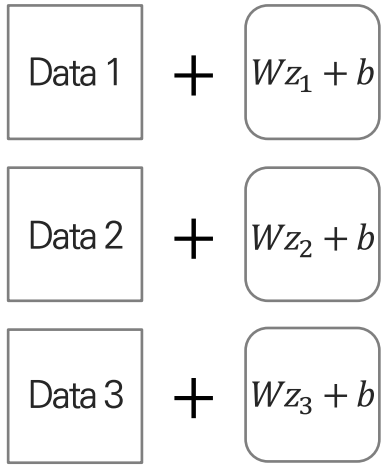
For multi dataset

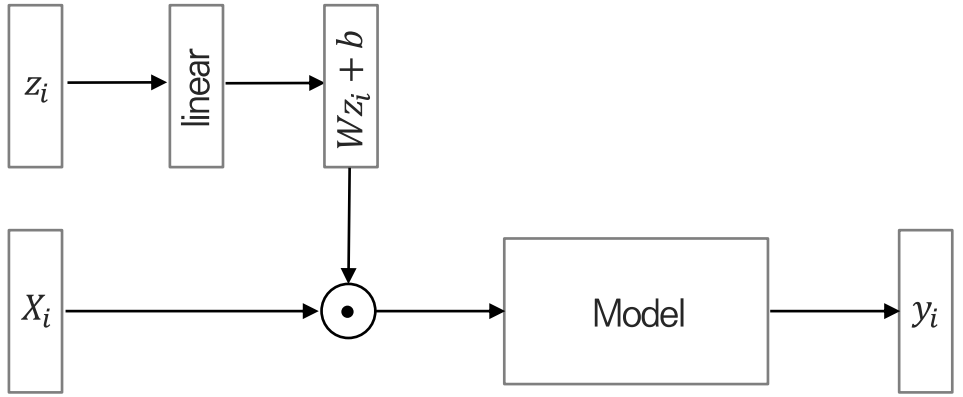
❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. Embedding vector

  → Multi-head 구조를 사용하지 않아도 됨





Multi-task model

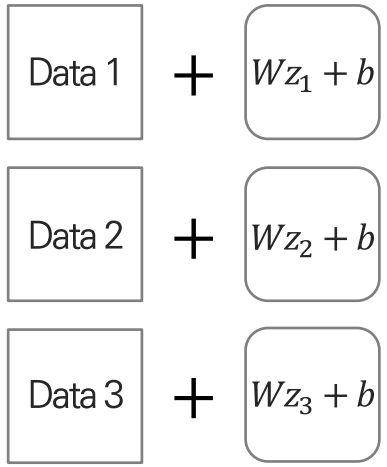
For multi dataset

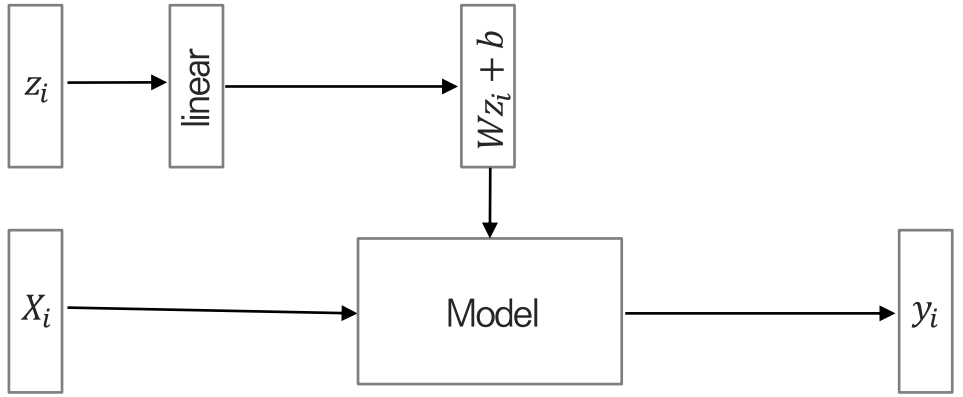
❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. Embedding vector

  → Multi-head 구조를 사용하지 않아도 됨





Multi-task model

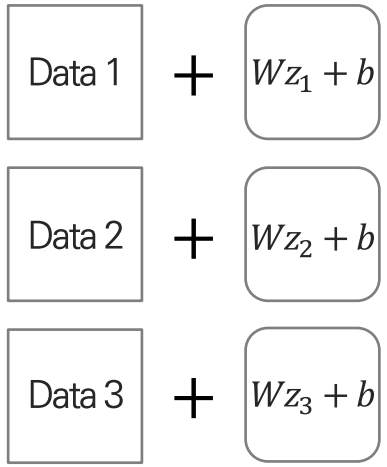
For multi dataset

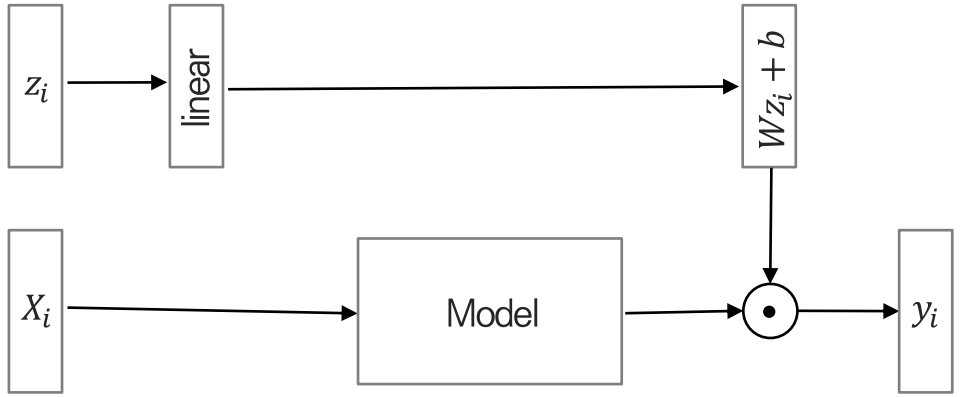
❖ 여러 데이터셋을 학습시킬 경우

- 입력 데이터(X)와 레이블(y)가 Task별로 다름
- 데이터의 태스크 상태를 알려주는 정보(z_i)가 필요

  → ex. Embedding vector

  → Multi-head 구조를 사용하지 않아도 됨



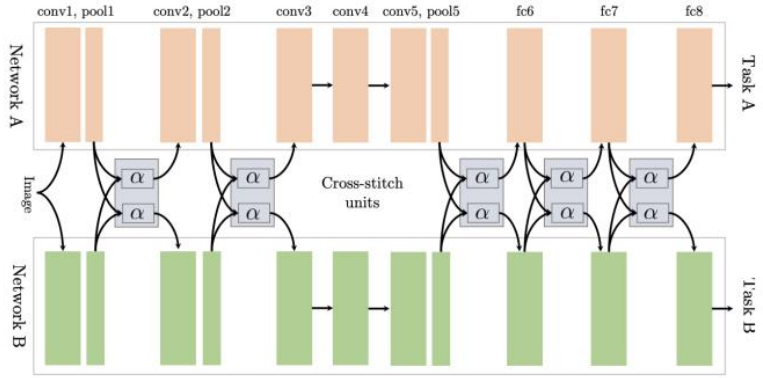


Multi-task model

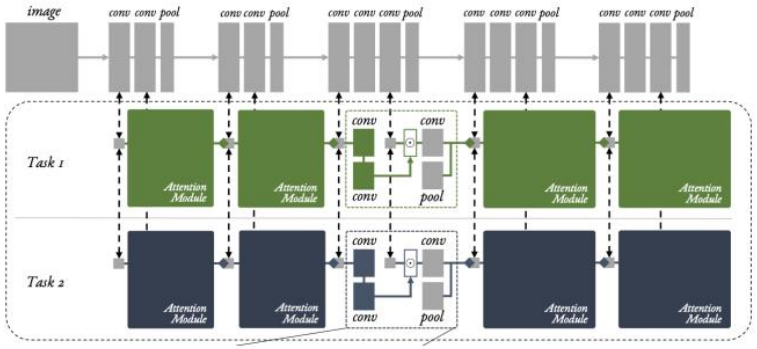
For multi dataset

❖ 여러 데이터셋을 학습시킬 경우

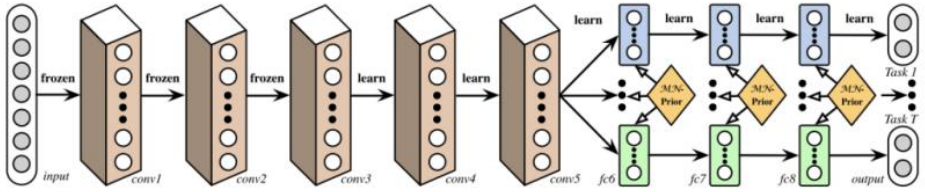
- Multi-task model은 정말 다양한 형태로 자유롭게 변형이 가능



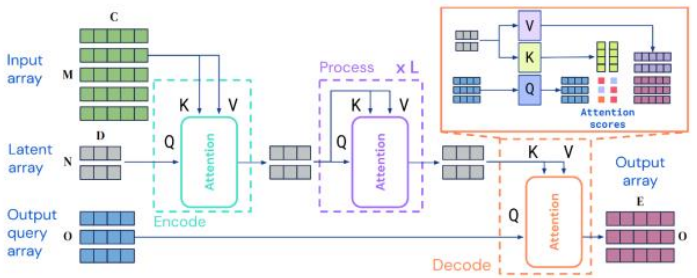
Cross-Stitch Networks. Misra, Shrivastava, Gupta, Hebert '16



Multi-Task Attention Network. Liu, Johns, Davison '18



Deep Relation Networks. Long, Wang '15



Perceiver IO. Jaegle et al. '21

Optimization

Objective

❖ Loss function

- 기본적인 형태

→ 각 태스크별 loss가 같은 비율로 더해지는 경우

→ Uniform weight

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)$$

Optimization

Objective

❖ Loss function

- 가중합 (weighted sum)

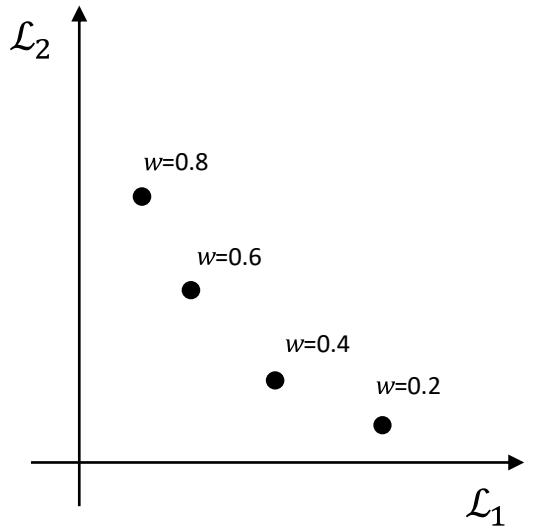
  → 각 태스크별 loss가 다른 비율로 더해지는 경우

  → 가중치는 task별 반영 정도(task balancing)를 의미함

  → 가중치에 따라 태스크별 성능의 trade-off가 일어남

$$\min_{\theta} \sum_{i=1}^{T} w_i \mathcal{L}_i(\theta, \mathcal{D}_i)$$

$$\min_{\theta} w\mathcal{L}_1 + (w-1)\mathcal{L}_2$$



Optimization

Objective

❖ Loss function

- 가중치를 정하는 방법 (task balancing)
  1. 태스크별 중요도/정확도에 따라 직접 설정
     - ✓ 어렵거나 중요한 태스크의 가중치 높게 설정
  2. Heuristics 방법론을 사용
     - ✓ Gradient의 크기가 유사해지도록 설정[1]
  3. Uncertainty를 이용한 가중치 탐색[2]
  4. 모델 학습시 동시에 가중치 학습을 진행[3]
  5. Loss가 높은 task만 학습

$$\min_{\theta} \max_{i} \mathcal{L}_1(\theta, D_i)$$

1. Chen, Z., Badrinarayanan, V., Lee, C. Y., & Rabinovich, A. (2018, July). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International Conference on Machine Learning (pp. 794-803). PMLR.
2. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7482-7491).
3. Cortes, C., Gonzalvo, J., Mohri, M., & Storcheus, D. (2020). Agnostic learning with multiple objectives. Advances in Neural Information Processing Systems, 2020.

Pros and Cons

Pros

❖ 장점

- 지식 공유 (Knowledge sharing)

  → Task 1을 학습하면서 얻은 정보가 다른 연관 task들에 좋은 영향을 줌

- 과적합 방지

  → 같은 모델에 여러 task들의 데이터를 학습하기 때문에 효율적

  → 여러 task들을 학습하면서 보다 일반화된 특징(generalized representation)을 학습

- 계산 효율성

  → 동시에 학습하기 때문에 계산 비용이 적다

- 현실 적용 가능성

  → 현실에서는 더욱 다양한 task들을 수행할 수 있는 인공지능을 요구함

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Pros and Cons

Cons

❖ 단점

- Negative transfer

  → 연관성이 부족한 Task들을 학습할 때 단일 모델보다 성능이 낮은 경우

- Task balancing이 어려움

  → task별 차이가 크면 학습하기 어려움

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Related paper

PCGrad: Project conflicting gradients

❖ Gradient Surgery for Multi-Task Learning

- Multi-task learning 관련 최신논문
- 34th Conference on Neural Information Processing Systems (NeurIPS 2020)
- 논문의 아이디어 소개

Gradient Surgery for Multi-Task Learning

Tianhe Yu[1], **Saurabh Kumar**[1], **Abhishek Gupta**[2], **Sergey Levine**[2],
Karol Hausman[3], **Chelsea Finn**[1]
Stanford University[1], UC Berkeley[2], Robotics at Google[3]
tianheyu@cs.stanford.edu

Abstract

While deep learning and deep reinforcement learning (RL) systems have demonstrated impressive results in domains such as image classification, game playing, and robotic control, data efficiency remains a major challenge. Multi-task learning has emerged as a promising approach for sharing structure across multiple tasks to enable more efficient learning. However, the multi-task setting presents a number

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

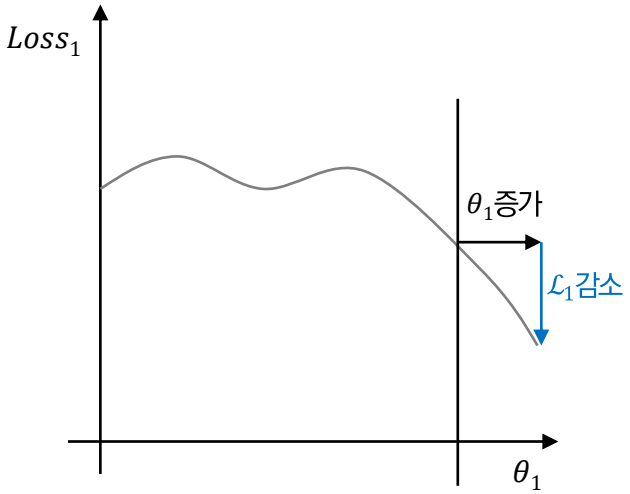
PCGrad: Project conflicting gradients

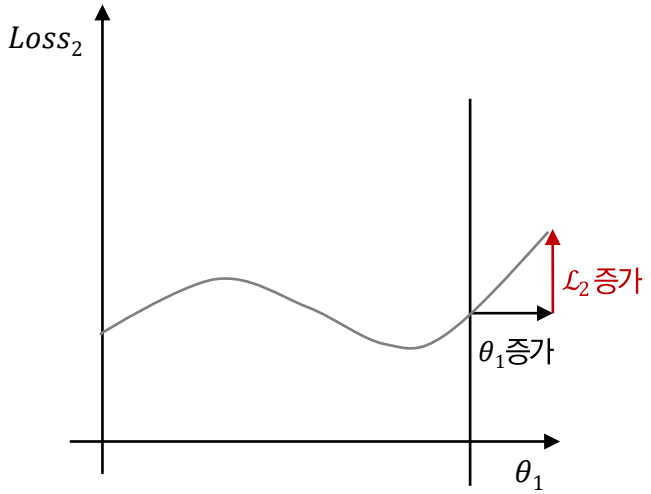
❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

✓ θ_1증가 → $\mathcal{L}_1$감소, $\mathcal{L}_2$증가





Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

PCGrad: Project conflicting gradients

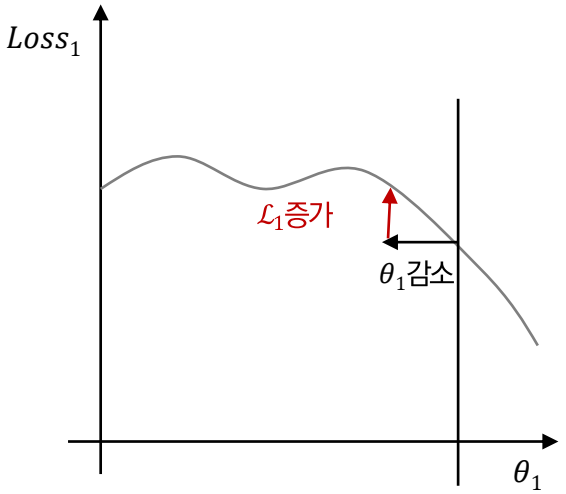
❖ Conflicting gradients

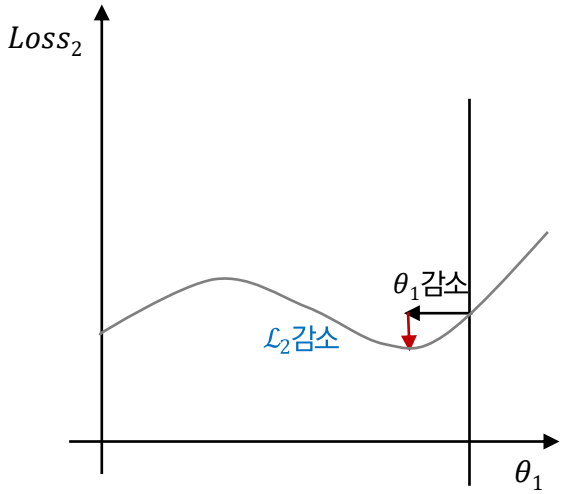
- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

✓ θ_1증가 → $\mathcal{L}_1$감소, $\mathcal{L}_2$증가

✓ θ_1감소 → $\mathcal{L}_1$증가, $\mathcal{L}_2$감소





Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

PCGrad: Project conflicting gradients

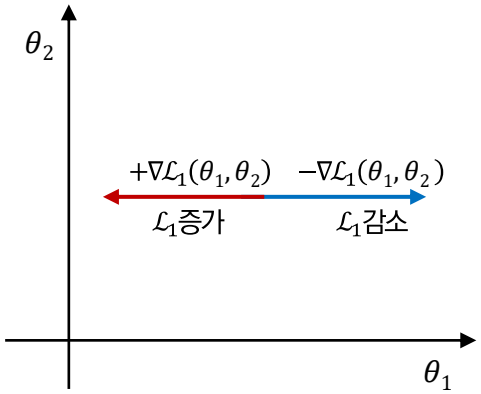
❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

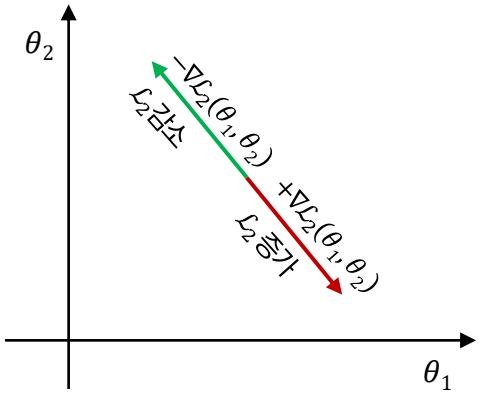
$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) \rightarrow \mathcal{L}_1$감소

$(\theta'_1, \theta'_2) + \nabla\mathcal{L}_1(\theta_1, \theta_2) \rightarrow \mathcal{L}_1$증가



$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2) \rightarrow \mathcal{L}_2$ 감소

$(\theta'_1, \theta'_2) + \nabla\mathcal{L}_2(\theta_1, \theta_2) \rightarrow \mathcal{L}_2$증가



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

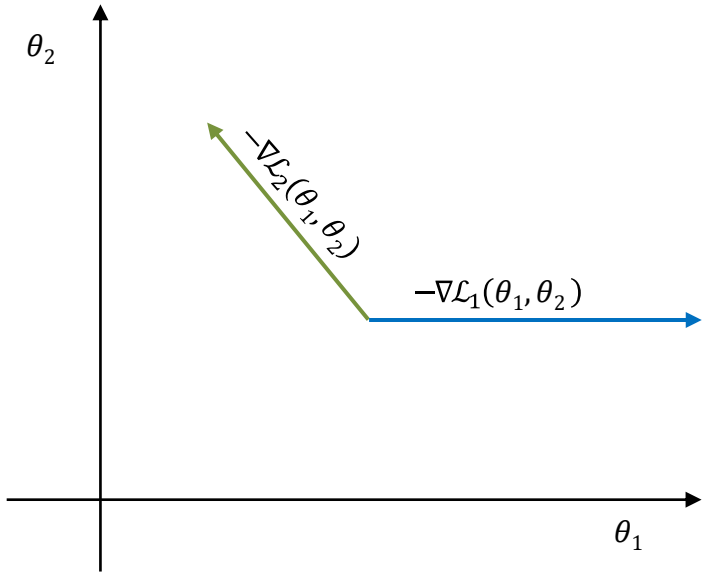
PCGrad: Project conflicting gradients

❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\boldsymbol{\theta'_1}, \boldsymbol{\theta'_2}) - \boldsymbol{\nabla \mathcal{L}_1(\theta_1, \theta_2)} - \boldsymbol{\nabla \mathcal{L}_2(\theta_1, \theta_2)}$$



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

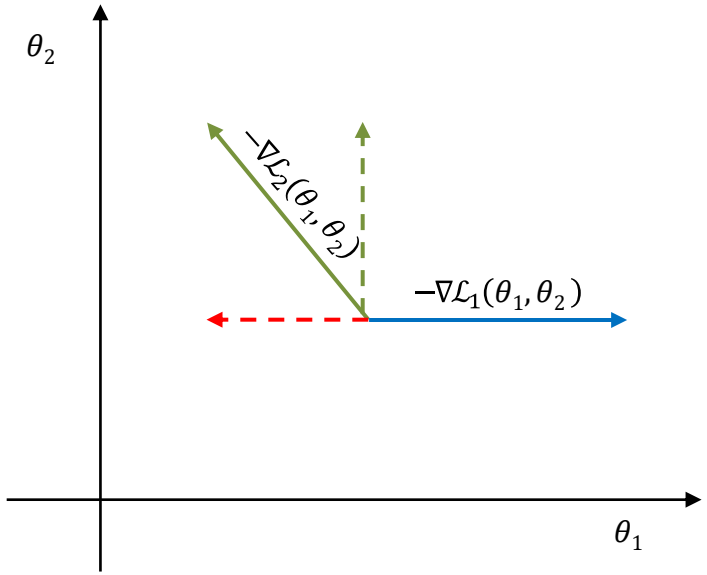
PCGrad: Project conflicting gradients

❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

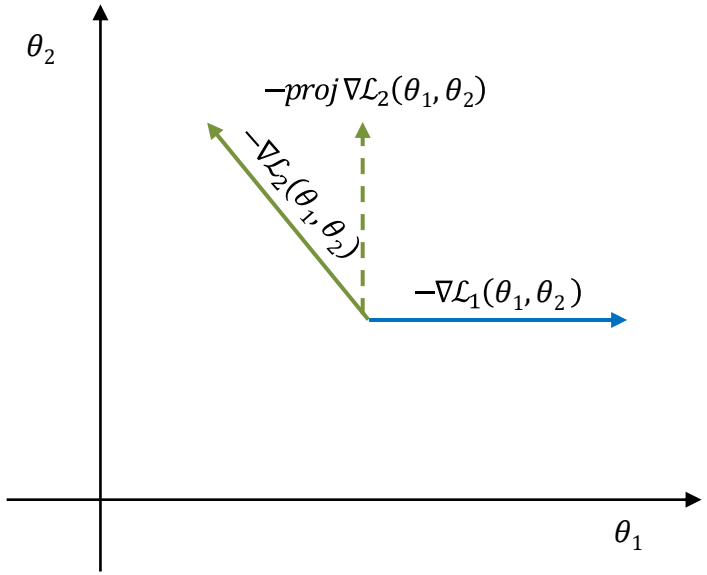
PCGrad: Project conflicting gradients

❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

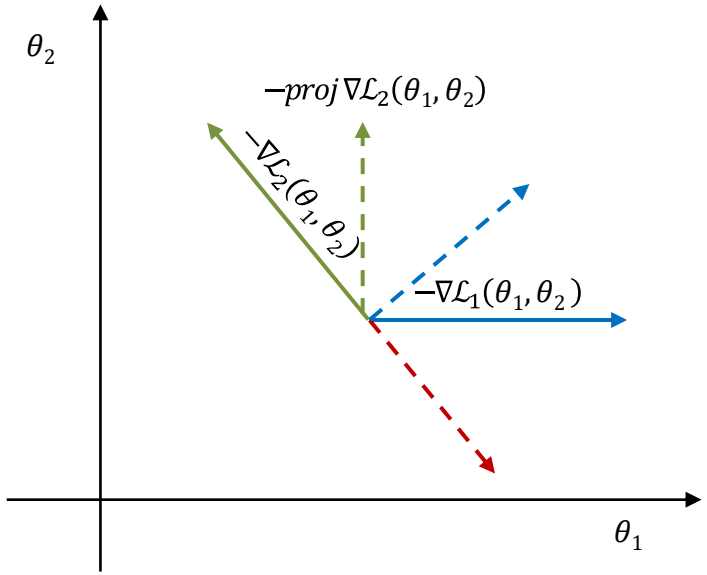
PCGrad: Project conflicting gradients

❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

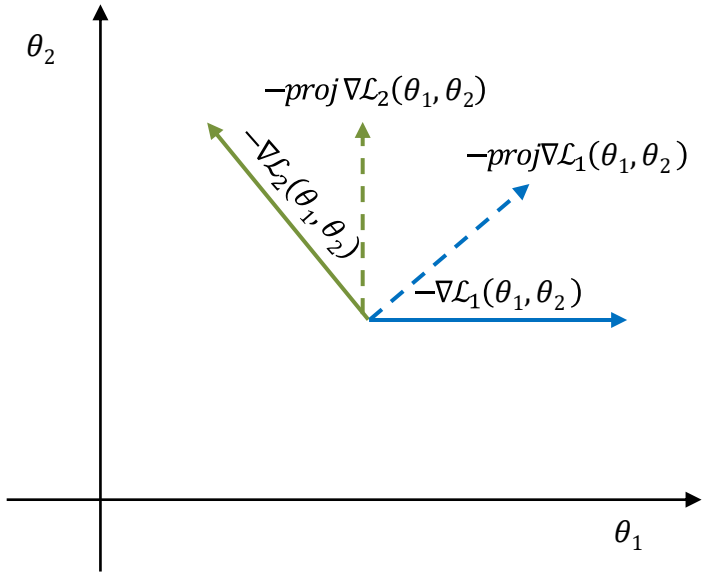
PCGrad: Project conflicting gradients

❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

PCGrad: Project conflicting gradients

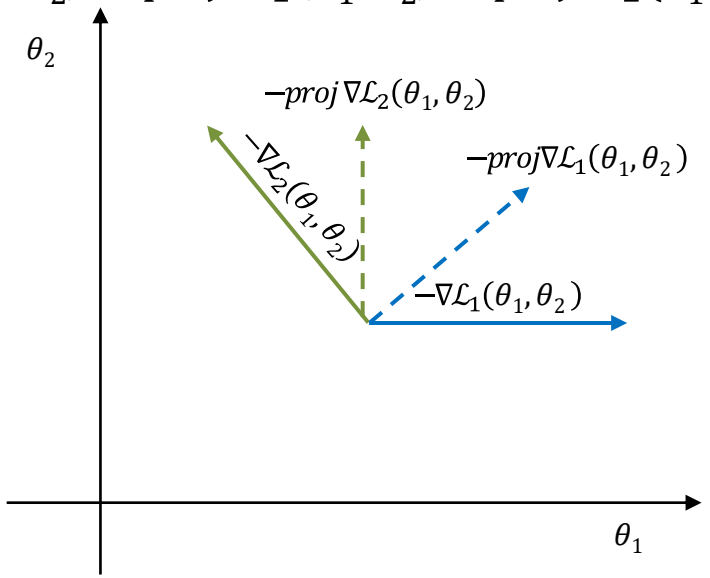
❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$

$$(\theta'_1, \theta'_2) - \text{proj}\nabla\mathcal{L}_1(\theta_1, \theta_2) - \text{proj}\nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Projection 된 gradient를 이용하여
task간 부정적인 영향 최소화

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Related paper

PCGrad: Project conflicting gradients

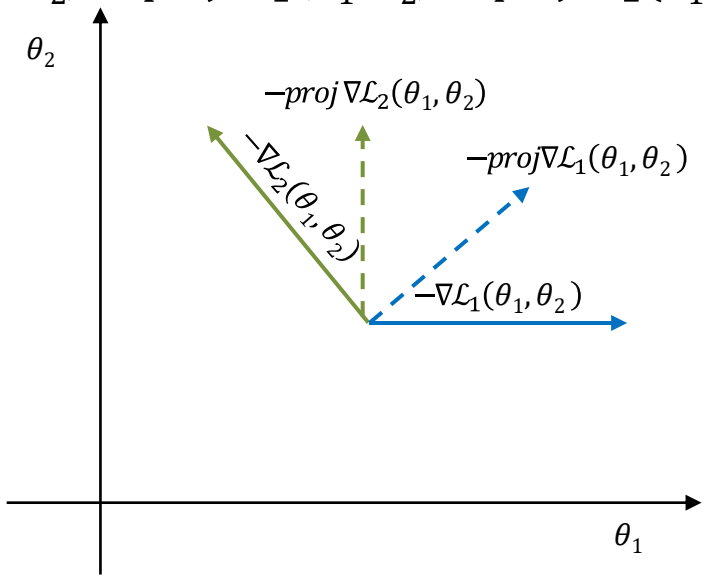
❖ Conflicting gradients

- 여러 개의 손실함수를 최적화할 때, 서로 반대 방향인 경우가 발생

$$\min_{\theta} \mathcal{L}_1(\theta, \mathcal{D}) + \mathcal{L}_2(\theta, \mathcal{D})$$

$$(\theta'_1, \theta'_2) - \nabla\mathcal{L}_1(\theta_1, \theta_2) - \nabla\mathcal{L}_2(\theta_1, \theta_2)$$

$$(\theta'_1, \theta'_2) - \text{proj}\nabla\mathcal{L}_1(\theta_1, \theta_2) - \text{proj}\nabla\mathcal{L}_2(\theta_1, \theta_2)$$



Projection 된 gradient를 이용하여
task간 부정적인 영향 최소화

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

Code – https://github.com/tianheyu927/PCGrad

Conclusions

- ❖ 현실에서는 더욱 다양한 task들을 수행할 수 있는 인공지능을 요구하며, 이에 맞는 방법론 중 하나
- ❖ Multi-task learning은 데이터를 효율적으로 사용할 수 있는 방법 중 하나
- ❖ 매우 다양한 형태의 모델들이 가능함
- ❖ Multi-task model은 입력 데이터(X)와 task 정보(z_i)가 필요하며, 마찬가지로 다양한 형태로 결합 가능
- ❖ 여러 손실함수를 최적화 해야하며, 이는 매우 어려운 문제이고 아직 명확한 해결책이 없음

References

❖ Review papers

- Caruana, R. (1997). Multitask learning. Machine learning, 28(1), 41–75.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.
- Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering.

❖ Multi-task models

- Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016). Cross-stitch networks for multi-task learning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3994-4003).
- Wang, Q., Han, T., Qin, Z., Gao, J., & Li, X. (2020). Multitask attention network for lane detection and fitting. IEEE transactions on neural networks and learning systems.
- Jaegle, A., Borgeaud, S., Alayrac, J. B., Doersch, C., Ionescu, C., Ding, D., ... & Carreira, J. (2021). Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795.

References

❖ Task balancing

- Chen, Z., Badrinarayanan, V., Lee, C. Y., & Rabinovich, A. (2018, July). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In International Conference on Machine Learning (pp. 794-803). PMLR.
- Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7482-7491).

❖ Optimization

- Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. arXiv preprint arXiv:1810.04650.
- Cortes, C., Gonzalvo, J., Mohri, M., & Storcheus, D. (2020). Agnostic learning with multiple objectives. Advances in Neural Information Processing Systems, 2020
- Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient surgery for multi-task learning. arXiv preprint arXiv:2001.06782.

감사합니다.