GradCraft: Elevating Multi-task Recommendations through Holistic Gradient Crafting

Yimeng Bai[1], Yang Zhang[1], Fuli Feng[1], Jing Lu[2], Xiaoxue Zang[2], Chenyi Lei[2], Yang Song[2]

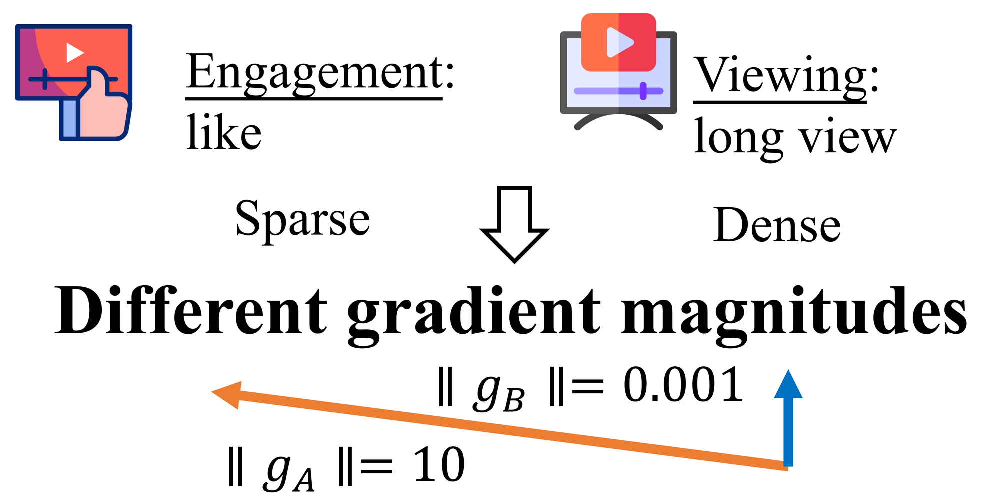
[1]University of Science and Technology of China [2]Kuaishou Technology

KDD2024 BARCELONA, SPAIN

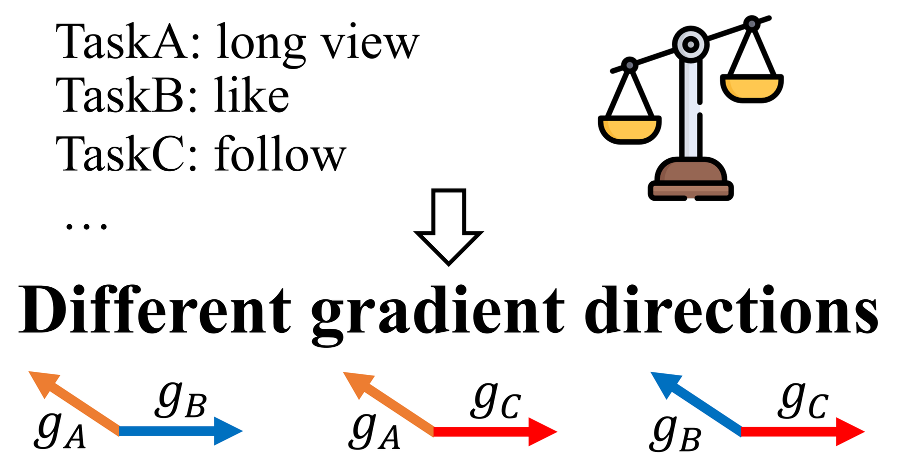
Codes for GradCraft: https://github.com/baiyimeng/GradCraft

Multi-task Recommendations: Challenges

Task Heterogeneity

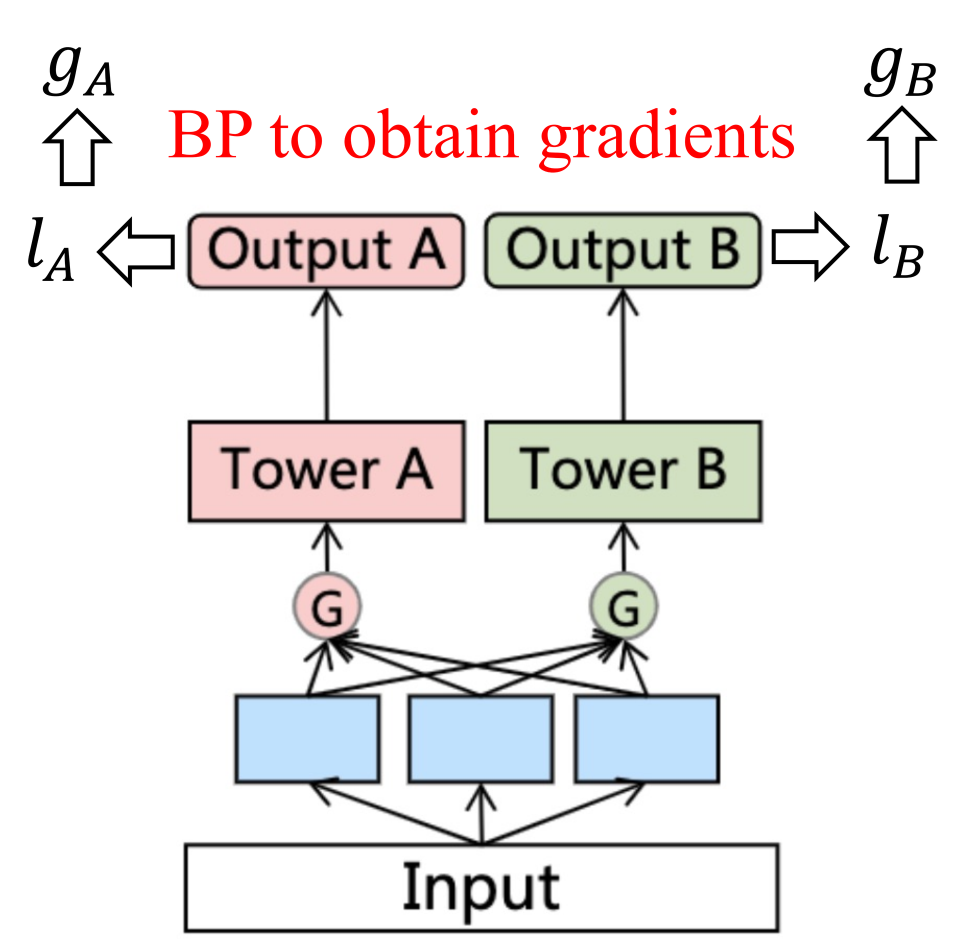


Task Cardinality



Conflicting task pairs (inner product < 0)

Problem Formulation: Gradient Perspective



Gradient Balance

Magnitude Balance

- ☐ Bad case: $\|g_i\| \gg \|g_j\|$
- ☐ Consequence: optimization dominance
- ☐ Solution: ensure similar magnitudes

Direction Balance

- ☐ Bad case: $\langle g_i, g_j \rangle < 0$
- ☐ Consequence: optimization conflicts
- ☐ Solution: ensure similar directions

Methodology: GradCraft

Overall Framework

- ☐ 1. Compute all task losses, $[l_1, \dots, l_T]$
- ☐ 2. Obtain all task gradients, $[g_1, \dots, g_T]$
- ☐ 3. Gradient magnitude adjustment, $[\hat{g}_1, \dots, \hat{g}_T]$
- ☐ 4. Gradient direction deconfliction, $[\tilde{g}_1, \dots, \tilde{g}_T]$
- ☐ 5. Gradient combination, just average as $\frac{1}{T}\sum_{i=1}^{T} \tilde{g}_i$
- ☐ 6. Gradient update by the optimizer (shared parameters)

Gradient Magnitude Adjustment

- ☐ **Key**: ensure an appropriate level of magnitude balance
- ☐ **How**: align gradient norm with the maximum norm

For i in range(T): $\hat{g}_i = \tau \frac{\max_j \|g_j\|}{\|g_i\|} g_i + (1-\tau) g_i,$

Number of tasks

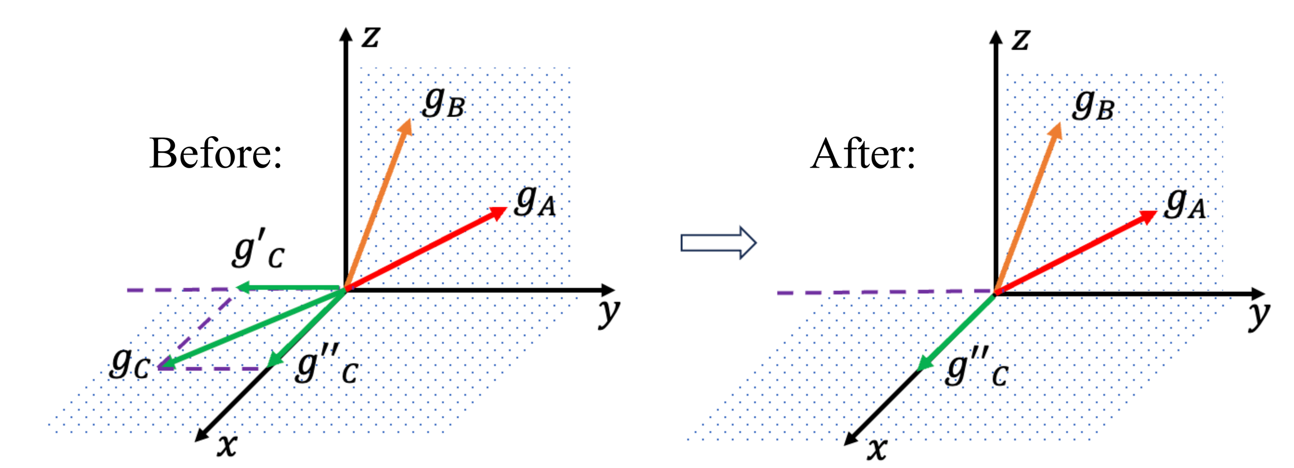
Hyper-parameter: control the proximity to the maximum norm

After adjustment: $\max_j \|g_j\| \,/\, \|\hat{g}_i\| < 1/\tau$

- ☐ **Effect**: mitigate interference from magnitudes for subsequent manipulation

Gradient Direction Deconfliction

- ☐ **Key:** ensure task gradient does not conflict with others
- ☐ **How:** global gradient projection



$g_C g_A^T < 0$, $g_C g_B^T < 0$ ⟶ $g''_C g_A^T = 0$, $g''_C g_B^T = 0$

- ☐ **Formulation:** global gradient projection
  - ✓ 1. Stack all conflicting gradients $G_i = [\hat{g}_{i_1}, \dots, \hat{g}_{i_n}] \in \mathbb{R}^{n \times d}$,
  - ✓ 2. Set the projection target $G_i \tilde{g}_i^\top = \boldsymbol{z}$, $\boldsymbol{z} = [\epsilon \|\hat{g}_i\| \|\hat{g}_{i_1}\|, \dots, \epsilon \|\hat{g}_i\| \|\hat{g}_{i_n}\|]$,

    Hyper-parameter: allow the positive inner product
  - ✓ 3. Give the projection to the space $\tilde{g}_i = \hat{g}_i + \sum_{k=1}^{n} w_k \hat{g}_{i_k} = \hat{g}_i + \boldsymbol{w}^\top G_i$,
  - ✓ 4. Solve the weight vector $\boldsymbol{w} = (G_i G_i^\top)^{-1}(-G_i \hat{g}_i^\top + \boldsymbol{z})$.

  For i in range(T): Input: $\hat{g}_i$ ⟶ Output: $\tilde{g}_i$
- ☐ **Effect :** simultaneously addresses all conflicting tasks

Experiments & Results

Offline Experiment

➢ **Setting**

- ☐ **Dataset:** Kuaishou (private), Wechat (public)
- ☐ **Task (Binary for simplicity):**
  - ☐ Viewing behavior: EffectiveView, LongView, CompleteView
  - ☐ Engagement behavior: Like, Follow, Forward
- ☐ **Evaluation:** average value of all tasks' AUC and GAUC
- ☐ **Baseline:**
  - ☐ Simple: Single, EqualWeighting
  - ☐ Other multi-task learning methods like Uncertainty and PCGrad

➢ **Results on two datasets**

Method	Wechat		Kuaishou	
	AUC	Rea.Imp.	GAUC	Rea.Imp.
Single	0.8361	0.000%	0.6770	0.000%
EW	0.8369	0.091%	0.6798	0.413%
UC	0.8358	-0.038%	0.6824	0.791%
DWA	0.8367	0.078%	0.6809	0.559%
MGDA	0.8309	-0.639%	0.6718	-0.809%
PCGrad	0.8371	0.118%	0.6830	0.887%
PCGrad+	0.8373	0.135%	0.6802	0.472%
GradVac	0.8366	0.065%	0.6790	0.288%
CAGrad	0.8369	0.099%	0.6815	0.653%
IMTL	0.8367	0.056%	0.6810	0.589%
DBMTL	0.8350	-0.129%	0.6796	0.380%
GradCraft	**0.8385**	**0.278%**	**0.6842**	**1.056%**

Online Experiment

➢ **Setting**

- ☐ **Traffic**: 1 week, 15 million users
- ☐ **Backbone**: QIN [1]
- ☐ **Label**: EffectiveView, LongView, Click
- ☐ **Evaluation**:
  - ☐ the average time users spend watching videos (WT)
  - ☐ the number of effective video viewing records (VV)
  - ☐ the instances of video sharing (Share)

➢ **Results on our video search system**

	WT	VV	Share
Base	-	-	-
GradCraft	**+0.505%**	**+0.950%**	**+1.746%**

Base: the SOTA multi-task learning method on our platform

Significance: improvements > 0.1% for WT and VV, and >1.0% for Share

[1] Guo tong et al. "Query-dominant User Interest Network for Large-Scale Search Ranking." CIKM2023