LORA: Low-Rank Adaptation of Large Language Models

Nguyn Th Ngc Uyển Phm Th c

Khoa Toòn - C - Tin hc
Trng i hc Khoa hc T nhiển, HQGHN
Ging viển hng dn: PGS.TS Trn Trng Hiu

Hị Ni, 2025

Mc lc

Mc lc

Gii thiu

- Còc mũ hơnh ngũn ng ln (LLMs) nh GPT, BERT, LLaMA, T5 t nhiu thịnh tu vt tri.
- Tuy nhiổn, vic tinh chnh toịn b mũ hơnh (Full Fine-tuning) ùi hi:
  - Tịi nguyổn tờnh toòn ln (GPU, b nh)
  - Thi gian hun luyn dịi
  - Nguy c *overfitting*
- Gii phòp: **Parameter-Efficient Fine-Tuning (PEFT)** ch tinh chnh mt phn nh tham s.

ng c nghiổn cu

- **Thòch thc:**
  - Full fine-tuning tiổu tn hịng trm GB b nh GPU.
  - Mi tòc v cn mt bn tinh chnh riổng bit.
- **Gii phòp:** PEFT cho phỗp:
  - Gi nguyổn mũ hơnh gc, ch hc vịi tham s mi.
  - D chia s, tòi s dng mũ hơnh.
- **LoRA (Low-Rank Adaptation)**: gim ti **10.000 ln s tham s** mị vn gi hiu nng tng ng.

Kin thc nn tng

- **Transformer**: kin trữc da trổn c ch *self-attention*, giữp mũ hnh hc quan h gia còc t song song.
- **Full Fine-tuning**: cp nht toịn b trng s mũ hnh hiu nng cao nhng chi phờ ln.
- **PEFT**: ch tinh chnh còc mũ-un nh, giữp gim chi phờ hun luyn mị vn duy trơ hiu nng.

Còc phng phòp PEFT ph bin

- **Adapter Tuning**: thổm tng ph nh gia còc tng Transformer.
- **Prefix/Prompt Tuning**: hc thổm chui tham s c chốn vịo u vịo.
- **BitFit**: ch tinh chnh còc tham s bias.
- **Compacter**: biu din hng thp cho adapter.
- **LoRA**: òp dng rịng buc hng thp lổn ma trn trng s.

Mc lc

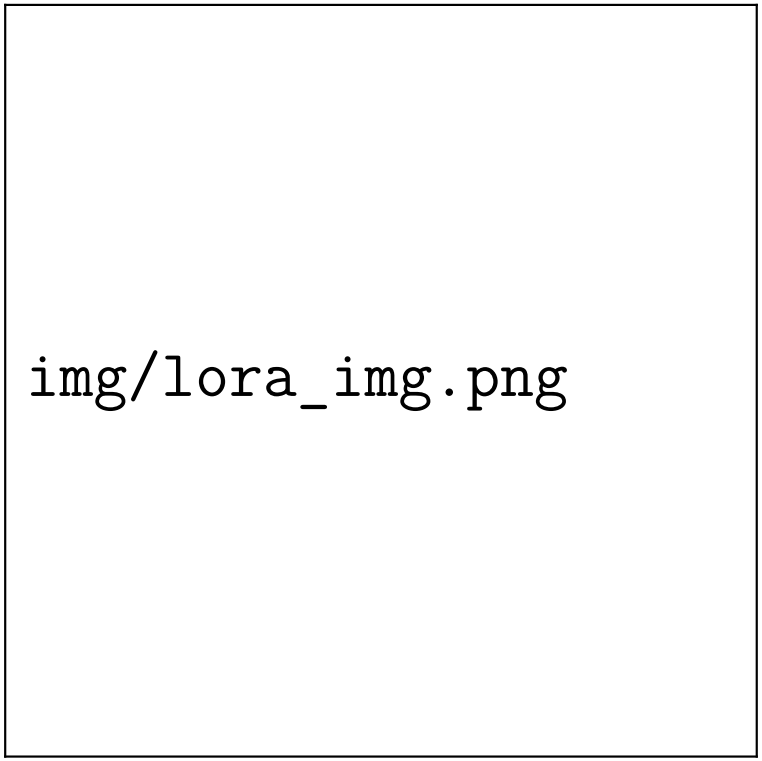
Nguyổn lỳ hot ng ca LoRA

- Quan sòt: thay i trng s khi fine-tuning ch nm trong khũng gian con hng thp.
- LoRA tòi tham s húa ma trn trng s:

$$W = W_0 + BA, \quad B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}, \ r \ll \min(d,k)$$

- Gim tham s hun luyn t $d \times k$ xung $r(d+k)$.
- Vi $r = 8$, gim hn 500 ln s tham s cn tinh chnh.

Minh ha c ch LoRA



Hơnh minh ha c ch tòi tham s húa trng s trong LoRA.

- W_0: trng s gc c gi nguyổn.
- A, B: hai ma trn hng thp c hun luyn.
- Kt qu: ch cn hun luyn rt ờt tham s.
- Khi hun luyn xong, cú th gp $(W_0 + BA)$ dứng cho suy lun mị khũng tng tr.

So sònh LoRA vi còc phng phòp khòc

Phng phòp	**Thay i kin trữc**	**T l tham s**	**tr**	**c im**
Adapter Tuning	Cú	13%	Cú	Hiu nng cao nhng tng sóu mũ hơnh
Prefix Tuning	Khũng	<1%	Khũng	n gin, ph thuc dịi prefix
BitFit	Khũng	0.1%	Khũng	Rt nh, hiu nng gii hn
Compacter	Cú	0.5%	Nh	Adapter nỗn, hiu qu cao
LoRA	Khũng	<1%	Khũng	Rịng buc hng thp, d tờch hp

u vị nhc im ca LoRA

u im:

- Hiu qu tham s cao ($<1\%$ tham s hun luyn).
- Khũng tng tr suy lun.
- Hun luyn nhanh hn, tit kim b nh.
- D kt hp vi Adapter hoc Prefix Tuning.

Hn ch:

- Gi nh khũng gian hng thp cú th khũng ững cho mi tòc v.
- Giò tr r cn c chn phứ hp.
- Vi mũ hơnh nh (nh BERT-base), li ờch khũng òng k.

Mc lc

Mc tiổu thc nghim

- ònh giò LoRA so vi Full Fine-Tuning trổn bịi toòn sinh ngũn ng t nhiổn.
- D liu: **E2E NLG Challenge** 50.000 cp (MR, cóu mũ t).
- Cóu hi chờnh:
  - LoRA cú t hiu nng tng ng Full Fine-Tuning?
  - Mc tit kim tịi nguyổn ra sao?

Thit lp thc nghim

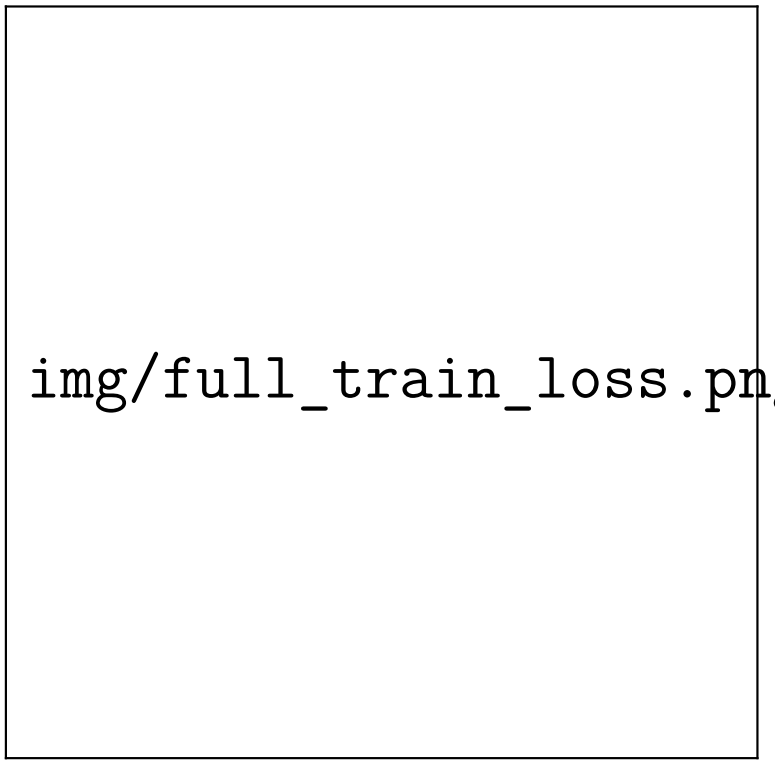
Mũ hơnh: GPT-2 Medium (355M tham s).

Cu hơnh:

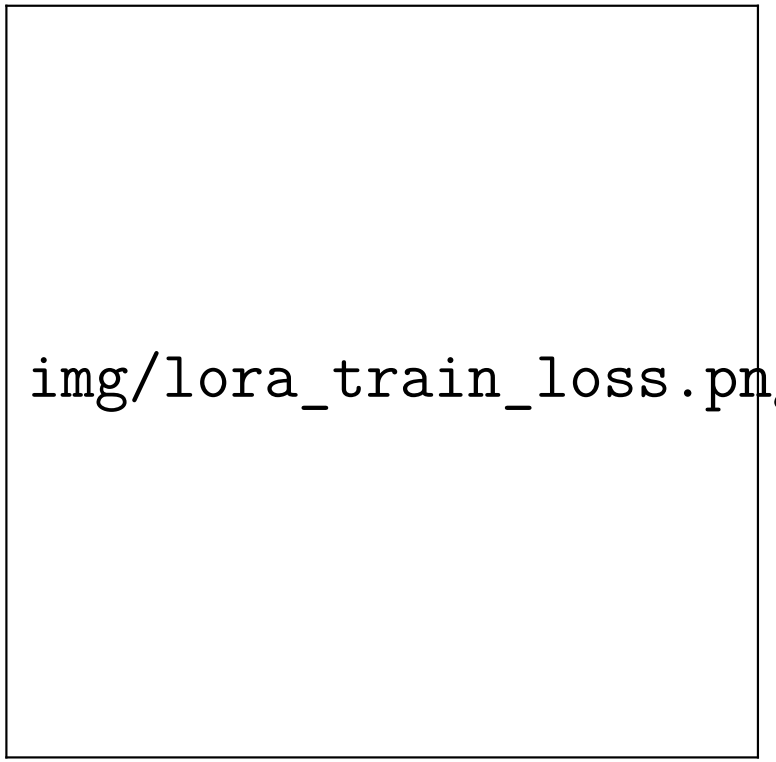
- Optimizer: AdamW, lr = 2e-4, batch size = 4
- Epochs = 5, dropout = 0.1
- LoRA rank = 4, scaling factor = 32
- Ch tinh chnh **0.3% tham s mũ hơnh**

ònh giò: BLEU, NIST, METEOR, ROUGE-L, CIDEr.

So sònh Train Loss gia còc mũ hơnh



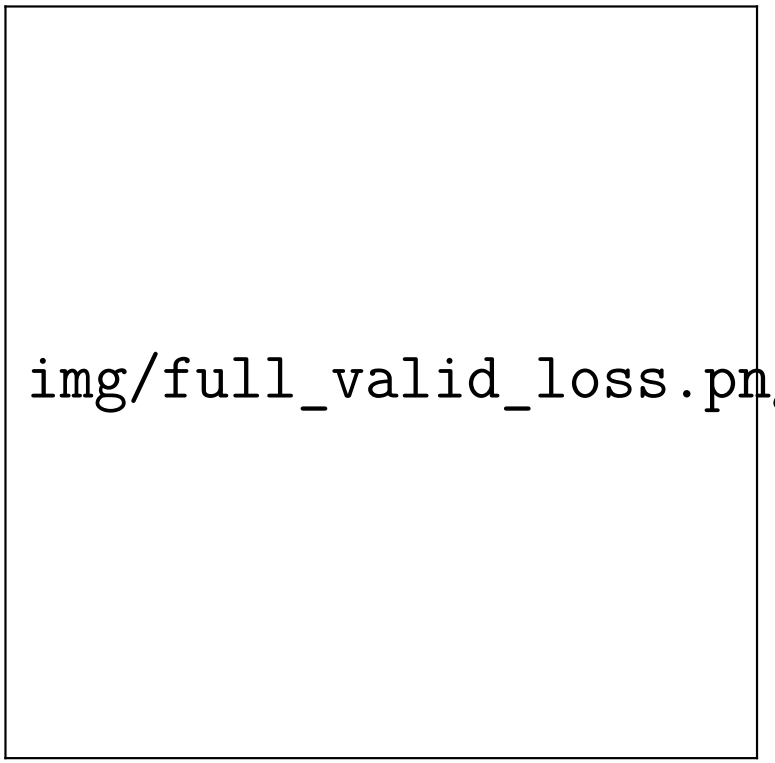
Loss ca mũ hơnh Full Fine-tuning



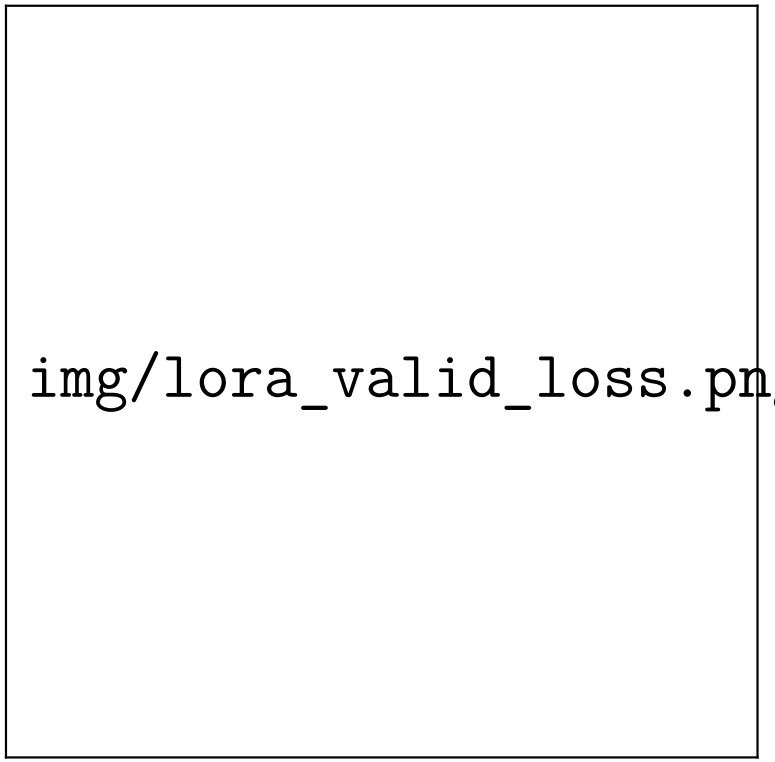
Loss ca mũ hơnh LoRA

- C hai mũ hơnh u hi t sau khong 34 epoch.
- LoRA cú ng loss mt hn, gim nhanh hn giai on u.
- iu niy cho thy vic gii hn khũng gian hng thp giữp quò trơnh ti u n nh hn

So sònh Validation Loss gia còc mũ hơnh



Validation Loss ca mũ hơnh Full Fine-tuning



Validation Loss ca mũ hơnh LoRA

- Validation Loss ca LoRA gim nhanh vị n nh hn.
- Full fine-tuning cú xu hng dao ng vị dng sm hn du hiu overfitting.
- Kt lun: LoRA giữp mũ hơnh tng quòt húa tt hn trổn tp kim th.

Kt qu nh lng

Tiểu chờ	Full Fine-tuning	LoRA
BLEU	0.6734	**0.6803**
NIST	8.5452	**8.6419**
METEOR	0.4622	0.4556
ROUGE-L	0.7033	0.7018
CIDEr	2.3843	**2.4465**

Nhn xỗt: LoRA t hiu nng tng ng hoc cao hn Full Fine-Tuning dứ ch hun luyn 0.3% tham s.

Phón tờch vị nhn xỗt

- LoRA hi t nhanh hn, gim dao ng gradient.
- Giữp trònh overfitting vị n nh trổn tp validation.
- Kt qu chng minh: thay i trng s quan trng ch nm trong khũng gian hng thp.

Mc lc

Kt lun vị hng phòt trin

- LoRA lị phng phòp tinh chnh hiu qu, tit kim tịi nguyổn mị vn gi hiu nng cao.
- Phứ hp vi h thng cú hn ch phn cng (GPU nh).
- Hng m:
  - ng dng trổn tòc v túm tt, i thoi, truy xut thũng tin.
  - Kt hp LoRA vi Adapter hoc Prefix Tuning.
  - Ti u giò tr rank r vị h s α.

Xin chón thịnh cm n!

Cm n thy vị còc anh/ch õ lng nghe.