分类号　　密级　　UDC

北京林业大学

学术型硕士学位论文

（中文题目）

（英文题目）

（作者姓名）

指　导　教　师　导师姓名 职称

学　　院　理学院

学　科　专　业　数学

研　究　方　向　计算数学

二〇 二四 年　五 月二十五 日

独 创 性 声 明

本人声明所呈交的论文是本人在导师指导下独立进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京林业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____　日　期：_____

关于学位论文使用授权的声明

本人完全了解北京林业大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京林业大学；学校有权保留并向国家有关部门或机构送交论文的纸质版和电子版，允许学位论文被查阅、借阅和复印；学校可以将学位论文的全部或部分内容公开或编入有关数据库进行检索，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。

签　名：_____导师签名：_____日　期：_____

答 辩 委 员 会 成 员 信 息

姓　名		职　称	工　作　单　位
主席		教授	北京林业大学工学院
委员		教授	北京林业大学理学院
		教授	北京林业大学理学院

摘要

分布式训练主要的应用场景就是例如当训练数据量较大时，单卡训练耗时过长，需要分布式训练技术提高训练效率；或者当单卡无法支持训练时，即单卡显存无法放下全量模型参数时，可使用分布式训练技术将模型参数等模型信息切分到多张卡或多台设备上，以保证模型可训练。简单来说无外乎两点：

- 加快训练速度；
- 让单卡无法训练的模型可训练。

当下分布式训练主要有两种训练模型，第一种是集合通信模型，第二种是参数服务器模型。集合通信模型适用于 CV/NLP 相关的任务，参数服务器模型适用于 CTR 等相关的任务。这是因为一般来说 CV/NLP 相关的任务模型的参数都比较稠密，而搜索/推荐相关的任务模型的参数都比较稀疏。稠密参数（Dense_Var）是指每个 step 都会更新的参数，比如 FC 的 Weight 参数。稀疏参数（Sparse_Var）是指每个 step 不是必须更新的参数，如 Embedding 参数，只会更新当前 step 中涉及到的 key 的相应 value。

关键词：计算机视觉，自然语言处理，语言/视觉大模型，分布式计算，高性能计算

Please text your english title here

Master Candidate: Zhang San
(Mathematics)
Directed by Li Si

ABSTRACT

Please text your english abstract.

Key Words: Computer Vision, Natural Language Processing, Large Language/Vision Models, Distributed Computing, High Performance Computing

目录

1 章标题

1.1 节标题

1.1.1 二级节标题

2 常用 LaTeX 命令参考样式

2.1 参考文献插入的参考样式

使用 BibTeX 格式参考文献，使用命令`\cite`即可引用，如：自然语言处理中的 Attention is All You Need[1]，分布式计算中 NVIDIA 的 Megatron 系列论文[2-4]....

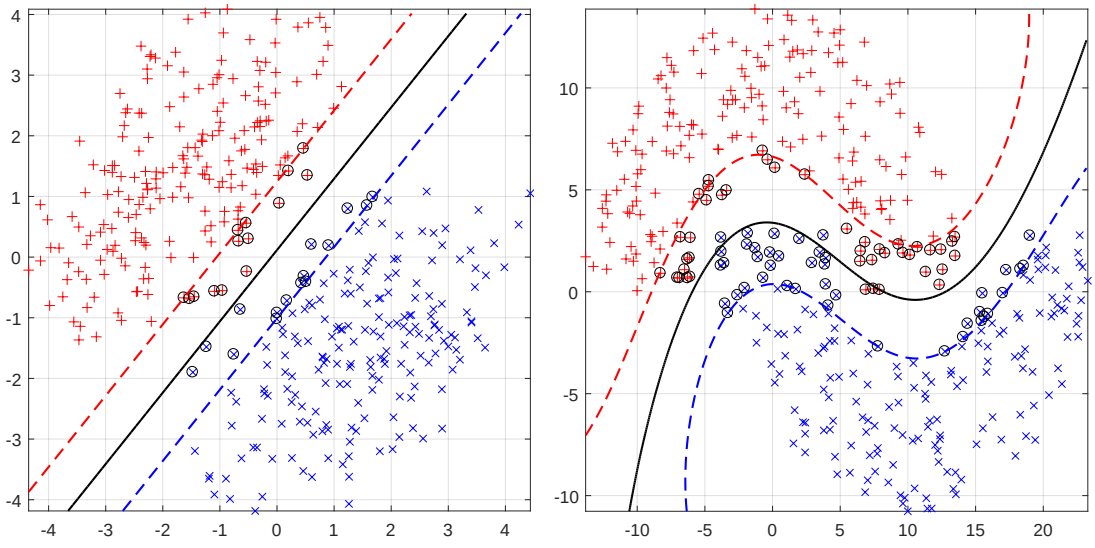
2.2 罗列要点的参考样式

本研究的主要创新点和贡献可总结为以下五点：

- 也可以通过；
- itemize 实现；
- 罗列；

2.3 图片插入的参考样式

SVM 有着优良的分类性能，从图 2.1中可以看出，SVM 能够准确地识别出不同类别的数据，并给出合理的的决策边界 (图中黑色实线)。



(a) 传统软间隔 SVM(线性核)　　(b) 非线性软间隔 SVM(高斯核)

图 2.1 线性 SVM 和非线性 SVM 在不同数据类型下的决策边界

Figure 2.1 Classification boundaries of linear SVM and nonlinear SVM under different data distributions

2.4 表格插入的参考样式

为了简便起见，表 2.1总结了本论文出现的主要数学符号。

表 2.1 符号对照说明表

Table 2.1 Table of notations

符号	说明
(x_i, y_i)	第 i 个单视角训练样本及其标签 y_i
(x_i^A, x_i^B, y_i)	第 i 个多视角训练样本及其标签 y_i
l^+, l^-	正类 (负类) 训练样本的个数
$\boldsymbol{c}_{\pm}^A,\ \boldsymbol{c}_{\pm}^B$	分别对应于视角 A 和视角 B 的正类 (负类) 超球球心
$R_{\pm}^A,\ R_{\pm}^B$	分别对应于视角 A 和视角 B 的正类 (负类) 超球半径
$\langle x_i,\ x_j \rangle$	向量 x_i 和 x_j 的内积，或写为 $x_i^\top x_j$
$\varphi(\cdot),\ \varphi_A(\cdot),\ \varphi_B(\cdot)$	将原空间映射到高维特征空间的正定核映射
$k(x_i, x_j)$	传统单视角中的核函数 $\langle \varphi(x_i),\ \varphi(x_j) \rangle$
$k^A(x_i^A, x_j^A)$	A 视角核函数 $\langle \varphi_A(x_i^A),\ \varphi_A(x_j^A) \rangle$
$k^B(x_i^B, x_j^B)$	B 视角核函数 $\langle \varphi_B(x_i^B),\ \varphi_B(x_j^B) \rangle$
$\|\cdot\|_p$	向量的 l_p-范数 (如果 p 省略，则默认取 l_2-范数)
$\boldsymbol{Q}_{(\cdot\times\cdot)}$	大写黑体西文字符表示矩阵，其规模可见下标
$\boldsymbol{q}_{(\cdot\times\cdot)}$	小写黑体西文字符表示列向量，其长度可见下标
$diag(\cdot)$	表示取矩阵对角线元素的运算符

2.5 公式输入的参考样式

支持向量机 (SVM) 是机器学习的经典算法，它的优化学习目标是在两类样本点之间寻求最大间隔和最小误差的权衡。令 $\mathscr{X} \in R^d$ 为 d 维的样本属性空间，$\mathscr{Y} = \{-1, +1\}$ 为样本标签空间。考虑一个有监督二分类问题，其训练样本集表示为 $T = \{(x_i, y_i)\,|x_i \in \mathscr{X}, y_i \in \mathscr{Y}\}_{i=1}^l$，SVM 将在样本空间中寻找能将两类样本点以最大间隔分开的决策超平面：

$$\langle \boldsymbol{\omega}, \varphi(x) \rangle + b = 0, \tag{2-1}$$

其中，$\boldsymbol{\omega}$ 和 b 分别表示样本属性空间中决策超平面的法向量与偏置项。

SVM 的原问题是：

$$\begin{aligned} \min_{\boldsymbol{\omega}, b, \xi} \quad & \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i\left(\langle \boldsymbol{\omega}, \varphi(x_i) \rangle + b\right) \le 1 - \xi_i, \\ & \xi_i \ge 0,\ i = 1, 2, ..., l, \end{aligned} \tag{2-2}$$

其中，$\varphi(.)$ 表示任意正定核映射，C 是松弛变量的惩罚参数，且 $C > 0$。

在 SVM 的目标函数中，正则化项 $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ 旨在最大化超平面 $\langle \boldsymbol{\omega}, \varphi(x) \rangle + b = +1$ 和 $\langle \boldsymbol{\omega}, \varphi(x) \rangle + b = -1$ 之间的距离，起到结构风险最小化的作用；而约束条件要求样本点在两条超平面之外，若落入超平面之间，则引入松弛变量 $\{\xi_i\}_{i=1}^l$ 来使得约束成立；目标函数的第二项是最小化松弛变量之和，旨在尽可能地避免样本落入间隔内，

起到经验风险最小化的作用；惩罚因子 C 用于权衡目标函数的结构风险和经验风险。

SVM 的对偶问题是：

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i,x_j) - \sum_{i=1}^{l}\alpha_i \\ s.t. \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \le \alpha_i \le C,\ i = 1,2,\ldots,l. \end{aligned} \tag{2-3}$$

通过对偶问题求解出 $\boldsymbol{\omega}$ 和 b 后，可得到 SVM 对未知样本 x 的决策函数：

$$f(x) = sign\left(\langle \boldsymbol{\omega}, \varphi(x)\rangle + b\right). \tag{2-4}$$

2.6 普通表格的插入样式

表 2.2 注意力类型研究

Table 2.2 Ablation study of Attention Type.

Attention Type	Information Exchange	mIoU (%)
Window Attention	Local	69.4
Window Attention†	Local	70.1
Region-Local Attention	Regional-Local	71.0
Broadcast-and-Mixing	Global-Regional-Local	**71.9**

2.7 长表格插入的参考样式

表 2.3 AwA 数据集性能对比 (Accu. ± std.)

Table 2.3 Benchmark results on AwA data sets (Accu. ± std.)

数据集名称	**SVM-2K** Accu. (%)	**MVMED** Accu. (%)	**MvTwSVMs** Accu. (%)	**MvNPSVM** Accu. (%)	**Ours** **Accu. (%)**	**Ours-2C** **Accu. (%)**
chi vs. pan	68.15±4.81	47.99±4.48	68.42±2.31	85.57±0.83	86.56±0.71	**88.70±0.68**
chi vs. leo	75.13±4.40	47.40±2.91	76.40±0.99	84.39±1.64	85.09±0.77	**87.03±0.62**
chi vs. per	65.43±2.60	83.74±0.72	75.70±1.03	**85.10±0.78**	84.13±0.71	84.29±0.61
chi vs. pig	78.33±2.47	47.18±3.31	66.94±2.02	82.34±0.56	**83.81±1.04**	83.03±0.74
chi vs. hip	45.58±1.81	49.57±4.41	71.71±1.69	81.70±1.12	80.88±0.66	**84.07±1.12**
chi vs. hum	90.73±2.14	94.81±0.56	88.10±0.53	94.88±0.54	93.82±0.65	**95.94±1.03**
chi vs. rac	83.43±1.33	46.86±2.81	63.65±1.65	83.31±0.75	84.69±0.99	**84.95±1.11**
chi vs. rat	64.90±1.59	76.18±1.01	66.41±1.95	79.48±1.08	**81.90±1.10**	79.73±0.96

续 表 2.3

chi vs. seal	82.05±1.37	84.04±0.72	74.51±1.25	47.68±3.64	86.49±0.76	**88.14±0.58**
pan vs. leo	82.00±2.36	82.46±0.66	69.17±1.83	**88.97±0.97**	87.07±0.89	88.09±0.64
pan vs. per	56.90±0.76	87.73±0.76	80.56±1.10	89.27±0.74	**89.45±0.71**	88.17±0.35
pan vs. pig	63.21±1.29	77.91±1.66	61.43±2.13	82.53±0.99	**82.65±0.74**	80.79±1.24
pan vs. hip	75.56±0.78	83.65±0.75	68.94±1.79	88.51±1.03	85.72±0.98	**89.04±0.66**
pan vs.hum	49.22±4.23	96.03±0.49	89.00±1.14	96.26±0.33	96.14±0.34	**97.07±0.47**
pan vs. rac	76.75±2.50	84.56±1.14	61.44±1.00	87.26±1.19	**87.90±1.00**	87.31±0.91
pan vs. rat	64.26±1.06	82.62±1.22	66.83±2.66	85.40±0.99	**88.00±0.93**	86.21±1.05
pan vs. sea	80.50±1.40	87.06±0.64	74.77±2.19	88.59±0.84	88.66±0.66	**90.23±0.70**
leo vs. per	52.16±4.36	50.66±7.14	85.70±0.76	**90.91±0.69**	90.56±0.31	90.07±0.45
leo vs. pig	65.01±1.10	47.10±3.70	72.45±1.28	78.19±0.87	**78.99±1.11**	78.42±1.21
leo vs. hip	73.44±0.86	82.42±0.87	78.38±1.03	82.94±1.12	84.14±1.06	**84.37±0.58**
leo vs. hum	85.05±3.77	95.57±0.61	91.65±0.29	95.25±0.40	95.13±0.19	**95.94±0.25**
leo vs. rac	56.29±5.19	48.31±3.56	60.82±3.52	**77.32±1.46**	75.29±1.11	76.39±1.46
leo vs. rat	59.85±0.84	83.91±0.64	76.71±1.40	75.15±6.60	83.13±0.56	**83.96±0.89**
leo vs. sea	62.94±4.96	87.26±0.67	84.37±0.71	**88.46±0.52**	88.24±0.84	87.72±1.14
per vs. pig	66.00±6.37	78.24±0.84	73.93±0.96	79.06±0.99	**79.50±1.03**	79.24±1.63
per vs. hip	72.66±3.78	83.85±0.79	78.37±1.00	85.66±0.50	86.02±1.21	**86.54±1.15**
per vs. hum	91.89±0.88	88.46±0.86	83.42±0.78	92.19±0.42	91.41±0.29	**95.91±0.42**
per vs. rac	47.42±5.11	47.78±3.58	73.39±1.40	85.84±0.69	84.06±0.75	**87.43±0.83**
per vs. rat	66.52±3.26	68.18±1.34	61.35±1.70	68.27±1.69	69.29±0.78	**69.76±1.28**
per vs. sea	78.52±1.82	84.72±1.08	75.11±0.82	**85.30±1.17**	84.01±0.76	83.35±0.68
pig vs. hip	71.36±2.02	70.94±1.44	64.53±1.19	71.87±0.82	72.29±1.14	**72.43±1.29**
pig vs. hum	90.11±1.29	91.26±0.67	**94.15±0.40**	93.01±0.42	91.50±0.40	92.76±0.51
pig vs. rac	72.32±1.94	73.88±1.49	58.46±0.81	80.98±1.36	**80.96±0.85**	80.07±1.20
pig vs. rat	57.73±2.40	48.59±4.62	55.90±2.55	68.89±1.96	71.46±1.39	**72.22±0.85**
pig vs. sea	75.08±0.83	78.03±1.02	67.10±2.11	**78.97±0.91**	77.71±1.14	78.70±0.72
hip vs. hum	81.40±1.05	88.15±1.07	75.67±1.54	**88.45±0.87**	88.04±0.47	88.36±0.74
hip vs. rac	79.20±1.11	79.61±0.83	69.54±1.23	**81.60±1.12**	81.57±1.03	80.53±1.31
hip vs. rat	59.95±0.85	47.85±2.67	67.42±2.82	75.01±1.07	**79.08±1.21**	77.19±0.97
hip vs. sea	54.37±4.15	70.24±1.09	60.20±2.99	**70.95±1.35**	69.85±0.90	69.96±1.26
hum vs. rac	63.09±4.41	90.90±0.66	86.24±1.17	92.34±0.50	92.69±0.56	**93.81±0.57**
hum vs. rat	84.48±0.67	88.22±0.67	81.62±0.81	**89.94±0.55**	87.81±0.76	89.27±0.56
hum vs. sea	84.91±1.54	83.99±0.63	74.72±1.43	85.08±0.72	85.28±0.86	**85.37±0.59**
rac vs. rat	57.80±1.62	71.74±0.91	65.97±1.47	46.78±2.09	**76.12±1.13**	76.10±1.10
rac vs. sea	52.77±5.45	85.72±0.79	76.25±1.34	**86.29±0.50**	85.58±0.82	85.66±0.55
rat vs. sea	66.59±2.30	74.66±0.98	66.22±1.77	75.65±1.07	76.26±1.04	**76.58±0.99**
平均准确率	69.58	74.45	72.97	82.26	83.98	**84.46**
平均耗时	3.03	63.64	0.98	5.32	**0.20**	2.07
平均排名	5.18	4.18	5.07	2.42	2.38	**1.78**

2.8 定义、定理、引理与证明环境的参考样式

引理 1 ($\boldsymbol{\pi}$ 的稀疏性). 假设(**??**)的解为 $\boldsymbol{c}_+^{A*}, \boldsymbol{c}_+^{B*}, R_+^{A^2*}, R_+^{B^2*}$，其对偶问题 (**??**)的最优解为 $\boldsymbol{\pi}^* = \left(\boldsymbol{\alpha}_+^{A\top*}, \boldsymbol{\alpha}_+^{B\top*}, \boldsymbol{\beta}^{+\top*}, \boldsymbol{\beta}^{-\top*}\right)^\top$，则 $\forall i \in I^+$，多视角训练样本 $\boldsymbol{x}_i = \left\{x_i^A, x_i^B\right\}$ 和 $\boldsymbol{\pi}^*$ 满足如下关系：

样本的 A 视角特征 x_i^A 和 $\boldsymbol{\alpha}_+^{A*}$ 之间满足:

$$x_i^A \in \mathcal{R}_+^A = \left\{x^A \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 < R_+^{A^2*} \right.\right\} \Rightarrow \alpha_i^{A*} = 0, \tag{2-5}$$

$$x_i^A \in \mathcal{E}_+^A = \left\{x^A \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 = R_+^{A^2*} \right.\right\} \Rightarrow \alpha_i^{A*} \in \left[0, \frac{c_1^A}{l^+}\right], \tag{2-6}$$

$$x_i^A \in \mathcal{L}_+^A = \left\{x^A \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 > R_+^{A^2*} \right.\right\} \Rightarrow \alpha_i^{A*} = \frac{c_1^A}{l^+}, \tag{2-7}$$

样本的 B 视角特征 x_i^B 和 $\boldsymbol{\alpha}_+^{B*}$ 之间满足:

$$x_i^B \in \mathcal{R}_+^B = \left\{x^B \left| \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 < R_+^{B^2*} \right.\right\} \Rightarrow \alpha_i^{B*} = 0, \tag{2-8}$$

$$x_i^B \in \mathcal{E}_+^B = \left\{x^B \left| \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 = R_+^{B^2*} \right.\right\} \Rightarrow \alpha_i^{B*} \in \left[0, \frac{c_1^B}{l^+}\right], \tag{2-9}$$

$$x_i^B \in \mathcal{L}_+^B = \left\{x^B \left| \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 > R_+^{B^2*} \right.\right\} \Rightarrow \alpha_i^{B*} = \frac{c_1^B}{l^+}, \tag{2-10}$$

多视角样本 $\boldsymbol{x}_i = \left\{x_i^A, x_i^B\right\}$ 和 $\boldsymbol{\beta}^{+*}, \boldsymbol{\beta}^{-*}$ 之间满足：

$$\begin{aligned}
&\boldsymbol{x}_i \in \mathcal{B}^+ = \left\{\boldsymbol{x} \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 - R_+^{A^2*} - \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 + R_+^{B^2*} > \epsilon \right.\right\} \\
&\quad \Rightarrow \beta_i^{+*} = D_1, \beta_i^{-*} = 0, \qquad (2\text{-}11) \\
&\boldsymbol{x}_i \in \mathcal{E}^+ = \left\{\boldsymbol{x} \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 - R_+^{A^2*} - \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 + R_+^{B^2*} = \epsilon \right.\right\} \\
&\quad \Rightarrow \beta_i^{+*} = D_1, \beta_i^{-*} = 0, \qquad (2\text{-}12) \\
&\boldsymbol{x}_i \in \mathcal{E} = \left\{\boldsymbol{x} \left| \left| \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 - R_+^{A^2*} - \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 + R_+^{B^2*} \right| < \epsilon \right.\right\} \\
&\quad \Rightarrow \beta_i^{+*} = 0, \beta_i^{-*} = 0, \qquad (2\text{-}13) \\
&\boldsymbol{x}_i \in \mathcal{E}^- = \left\{\boldsymbol{x} \left| \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 - R_+^{B^2*} - \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 + R_+^{A^2*} = \epsilon \right.\right\} \\
&\quad \Rightarrow \beta_i^{+*} = 0, \beta_i^{-*} = [0, D_1], \qquad (2\text{-}14) \\
&\boldsymbol{x}_i \in \mathcal{B}^- = \left\{\boldsymbol{x} \left| \left\|\varphi_B(x^B) - \boldsymbol{c}_+^{B*}\right\|^2 - R_+^{B^2*} - \left\|\varphi_A(x^A) - \boldsymbol{c}_+^{A*}\right\|^2 + R_+^{A^2*} > \epsilon \right.\right\} \\
&\quad \Rightarrow \beta_i^{+*} = 0, \beta_i^{-*} = D_1, \qquad (2\text{-}15)
\end{aligned}$$

进一步地，在**命题**和**引理**1成立的条件下，我们有关于 $\boldsymbol{\beta}^{+*}, \boldsymbol{\beta}^{-*}$ 的进一步推论：

推论 1. 当参数 $\epsilon > 0$ 时，令 $\boldsymbol{u}^{-*} = \boldsymbol{\beta}^{+*} - \boldsymbol{\beta}^{-*}$，多视角样本 $\boldsymbol{x}_i = \left\{x_i^A, x_i^B\right\}$ 和 $\boldsymbol{u}^{-*}$ 之间满足：

定义 1 (多视角边界误差点 (multi-view margin errors, MEs))**.** 多视角边界误差点指的是，该样本点在所有视角上，其特征均在超球外，即 $\boldsymbol{x}_i = \left\{x_i^A, x_i^B\right\} \in \mathcal{L}_+^A \cap \mathcal{L}_+^B$。该类样本点的集合 *(MEs)* 可表述为：

$$MEs = \left\{ \boldsymbol{x}_i = \left\{x_i^A, x_i^B\right\} \middle| \|\varphi(x_i^A) - \boldsymbol{c}_+^{A*}\| > R_+^{A^2*} \ \& \ \|\varphi(x_i^B) - \boldsymbol{c}_+^{B*}\| > R_+^{B^2*} \ , i \in I^+ \right\}. \tag{2-16}$$

证明.

$$\begin{aligned}
&\sum_{i \in MEs} \left(\alpha_i^A + \alpha_i^B\right) \leq \sum_{i \in I^+} \left(\alpha_i^A + \alpha_i^B\right) = \sum_{i \in CVs} \left(\alpha_i^A + \alpha_i^B\right) = 2 \\
\Rightarrow &|MEs| \frac{c_1^A + c_1^B}{l^+} \leq 2 \leq |CVs| \frac{c_1^A + c_1^B}{l^+} \\
\Rightarrow &\frac{|MEs|}{l^+} \leq \frac{2}{c_1^A + c_1^B} \leq \frac{|CVs|}{l^+} \Rightarrow \ |MEs| \leq \frac{2l^+}{c_1^A + c_1^B} \leq |CVs|.
\end{aligned}$$

□

2.9 算法伪代码的参考样式

算法环境可以使用 algorithm2e 宏包 (输入注释等非算法包命令可支持中文)。如下为添加了自适应梯度裁剪 (Adaptive Gradient Clipping) 方法[5]的 AdamW[6]优化器伪代码：

2.10 超链接的参考样式

示例：

- 国产深度学习框架 Paddle 仓库地址为 https://github.com/PaddlePaddle/Paddle
- 基于 Paddle 的大模型分布式训练框架 PaddleNLP 的仓库地址为：https://github.com/PaddlePaddle/PaddleNLP

Algorithm 1: AdamW Optimizer with Adaptive Gradient Clipping

Input: Parameter vector to optimize w; objective function $\mathcal{L}$; learning rate schedule η_t; initial moving averages $v_0 = 0, u_0 = 0$; exponential moving average parameters β_1, β_2; weight decay λ; regularization constants ϵ.

Input: AGC hyperparameters: clipping threshold λ.

while w_t *not converge* **do**

  Compute $g_t = \frac{\partial \mathcal{L}(w_t)}{\partial w_t} \rightarrow$ obtain stochastic gradient.

  apply adaptive gradient clipping

  $h_{t,i} = \min\{\frac{\lambda \|w_{t,i}\|}{g_{t,i}}, 1.0\}, \quad i \in |w|$

  $g_{t,i} = h_{t,i} \cdot g_{t,i}, \quad i \in |w|$

  apply correction term to debias moving averages.

  $\hat{\beta}_1 = \beta_1 \cdot \frac{1-\beta_1^{t-1}}{1-\beta_1^t}$

  $\hat{\beta}_2 = \beta_2 \cdot \frac{1-\beta_2^{t-1}}{1-\beta_2^t}$

  update moving averages

  $v_t = \hat{\beta}_1 v_{t-1} + (1 - \hat{\beta}_1) g_t$

  $u_t = \hat{\beta}_2 u_{t-1} + (1 - \hat{\beta}_2) g_t^2$

  compute updates

  $r_t = \frac{v_t}{\sqrt{u_t} + \epsilon}$

  $w_{t+1} = w_t - \eta_t \lambda w_t - \eta_t r_t$

end

Output: w_T

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] SHOEYBI M, PATWARY M, PURI R, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism[A]. 2019.

[3] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient large-scale language model training on gpu clusters using megatron-lm[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2021: 1-15.

[4] KORTHIKANTI V A, CASPER J, LYM S, et al. Reducing activation recomputation in large transformer models[J]. Proceedings of Machine Learning and Systems, 2023, 5.

[5] BROCK A, DE S, SMITH S L, et al. High-performance large-scale image recognition without normalization[C]//International Conference on Machine Learning. PMLR, 2021: 1059-1071.

[6] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[A]. 2017.

附录

请填写你的附录内容

个人简介

请填写你的个人简介

导师简介

请填写导师简介

获得成果目录清单

请填写成果清单

致谢

请写上真诚的致谢！