



山东大学
SHANDONG UNIVERSITY

面向纠错的知识蒸馏研究

Research on Error Correction-Oriented Knowledge Distillation

指导老师：宋雪萌

答 辩 人：刘子鑫

答辩时间：2022. 05. 22

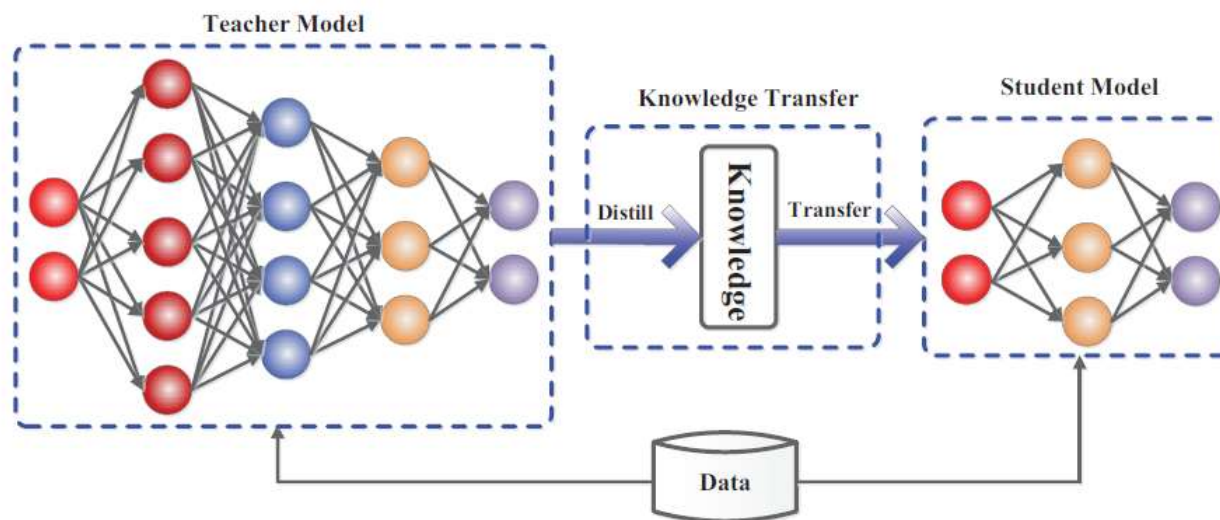
山东大学计算机科学与技术学院2019级硕士研究生毕业答辩

目录

- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

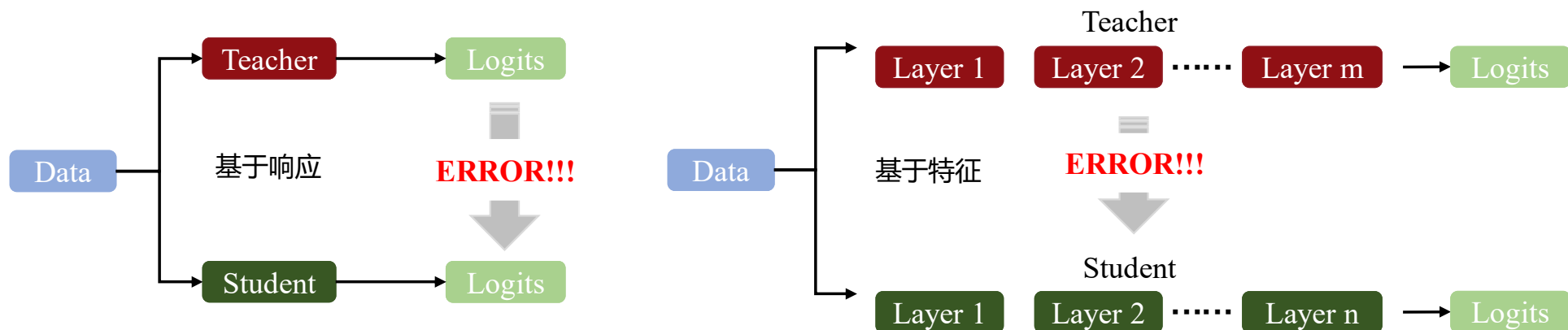
研究背景

- 复杂模型可以显著提升深度学习任务的最终效果，但却会带来高额的资源消耗问题，而知识蒸馏正是解决这一问题的方法之一；

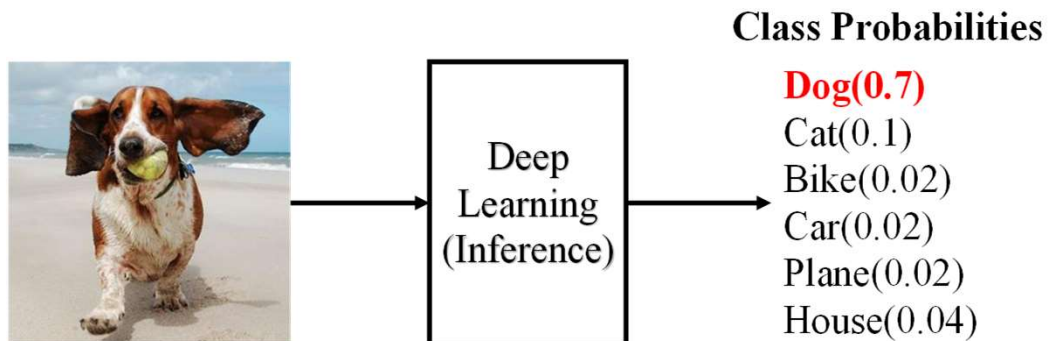


研究背景

- 已有的知识蒸馏方法，对知识的挖掘并不充分，传递的知识中包含错误信息；



- 知识蒸馏技术广泛应用于CV领域，本研究以图像分类为任务需求展开。



目录

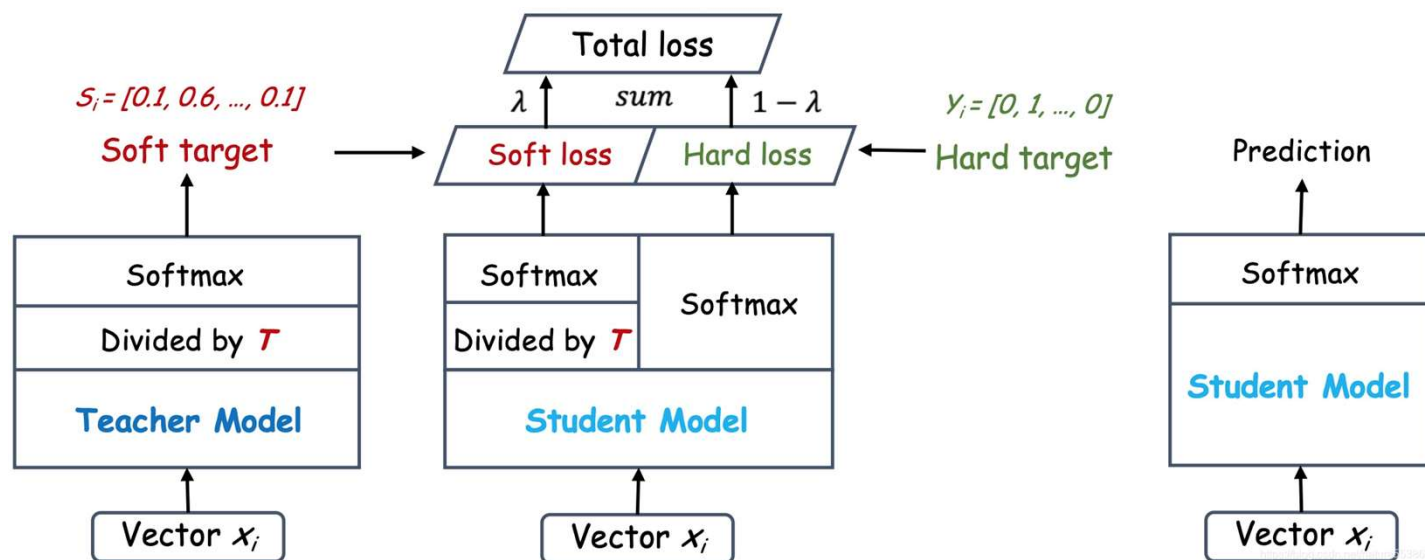
- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

相关工作

相关工作	方法分类
Distilling the Knowledge in a Neural Network, In Computer Science 2015.	离线蒸馏
Supervised contrastive learning, In NIPS 2020.	
Deep Mutual Learning, In CVPR 2018.	在线蒸馏
Feature fusion for online mutual knowledge distillation, In ICPR 2021.	
Revisit Knowledge Distillation: A Teacher-free Framework, In arXiv 2019.	自蒸馏
Regularizing class-wise predictions via self-knowledge distillation, In CVPR 2020.	

相关工作

➤ 离线蒸馏



引入“温度”：

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

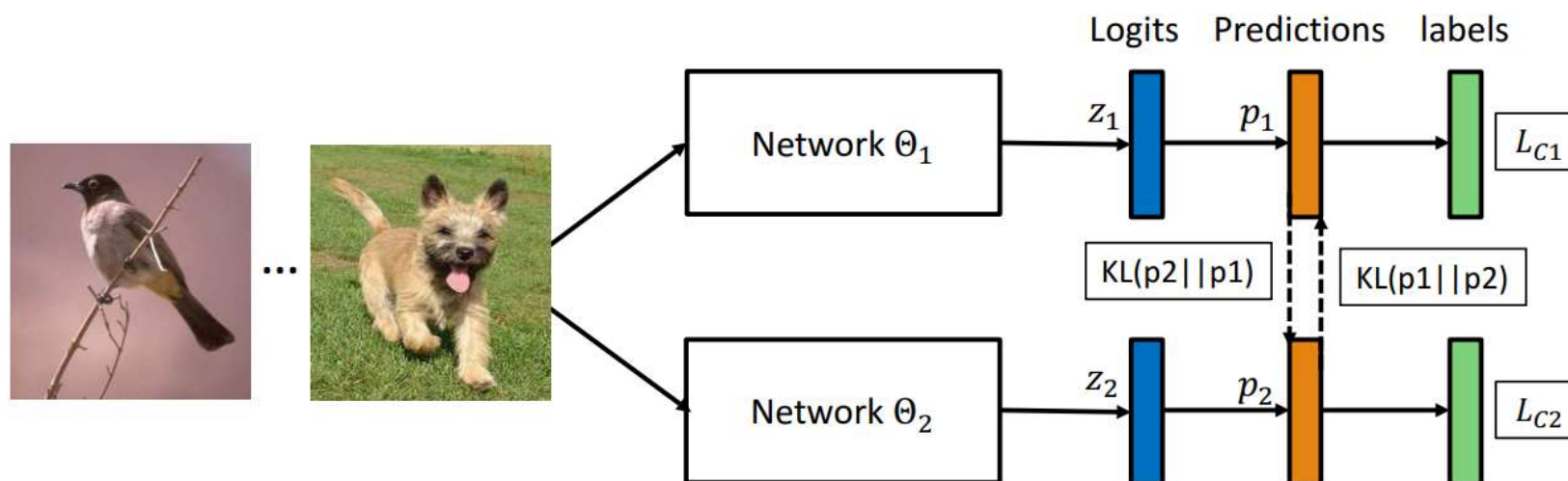
优点：简单、有效

缺点：蒸馏方式单一、资源消耗等

Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015, 2(7).

相关工作

➤ 在线蒸馏

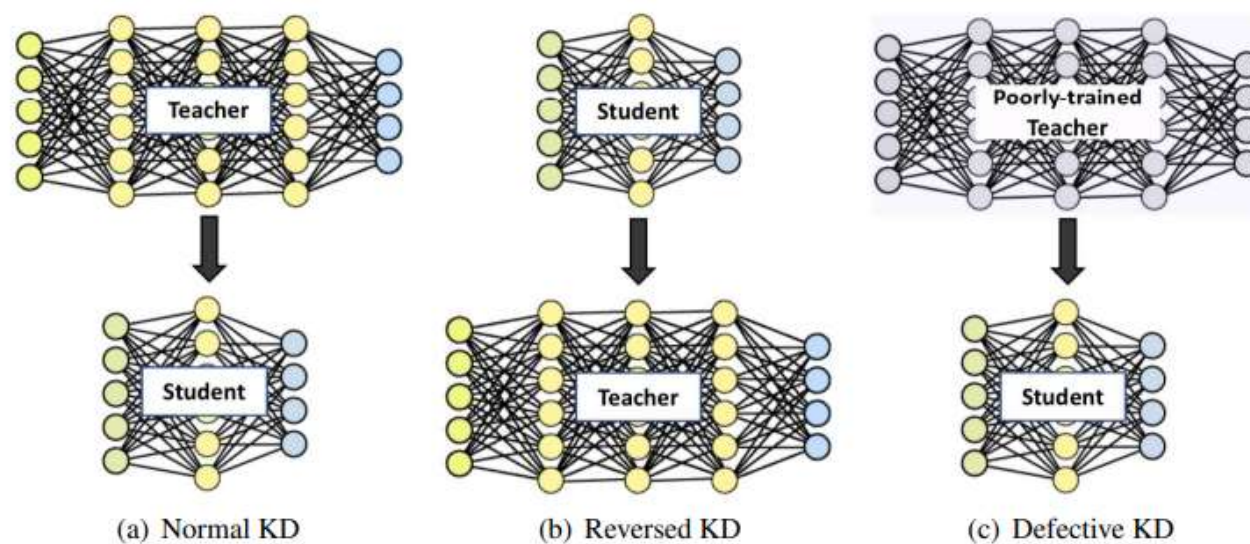


Zhang等人提出了一种深度相互学习策略，在此策略中，一组学生网络在整个训练过程中相互学习、相互指导，而不是静态的预先定义好教师和学生之间的单向转换通路。

Zhang Y, Xiang T, Hospedales T M, et al. Deep mutual learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4320-4328.

相关工作

➤ 自蒸馏




教师模型与学生模型使用相同的网络，可以被看作是一种特殊的在线蒸馏方法。

Yuan L, Tay F E H, Li G, et al. Revisit Knowledge Distillation: A Teacher-free Framework. arXiv 2019[J]. arXiv preprint arXiv:1909.11723.

相关工作

小结:



相关工作	是否使用层间知识	蒸馏方式	是否关注错误知识
KD	×	离线	×
FitNet	√	离线	×
DML	×	在线	×
Re-KD	×	自蒸馏	×

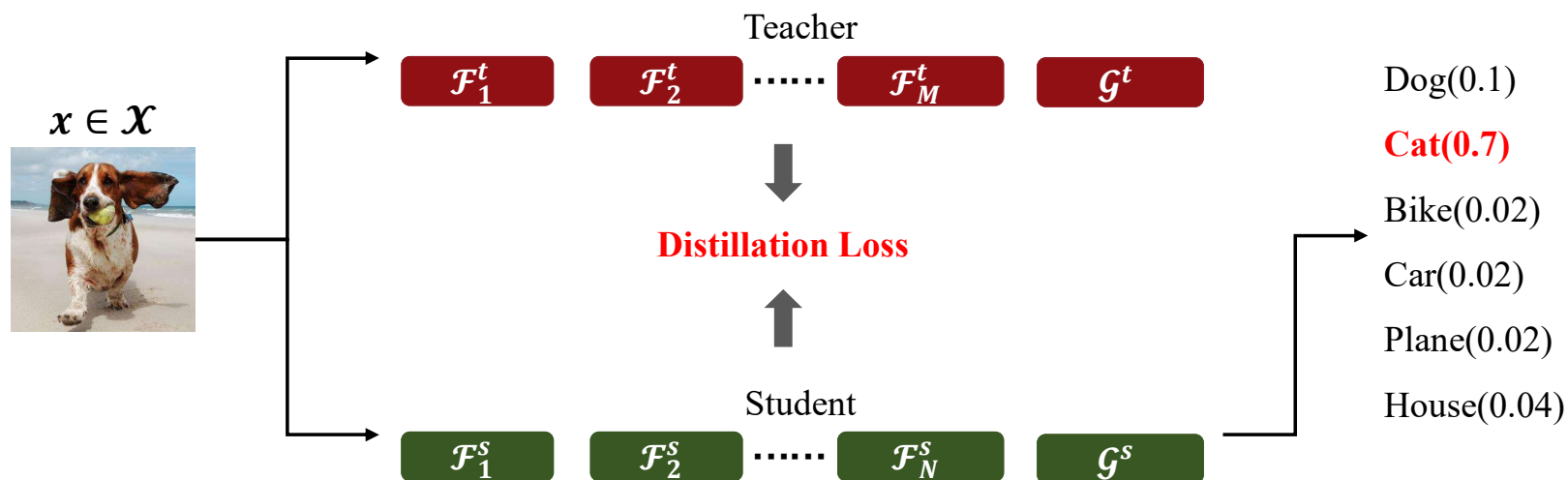
- 已有方法并未关注蒸馏过程中传递的错误知识，Re-KD虽指出此问题，但却并没有给出相应的处理方案
- 以往的工作大多围绕教师模型如何指导学生，很少提出教师模型指导学生的同时很可能也需要学生的帮助

目录

- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

研究思路

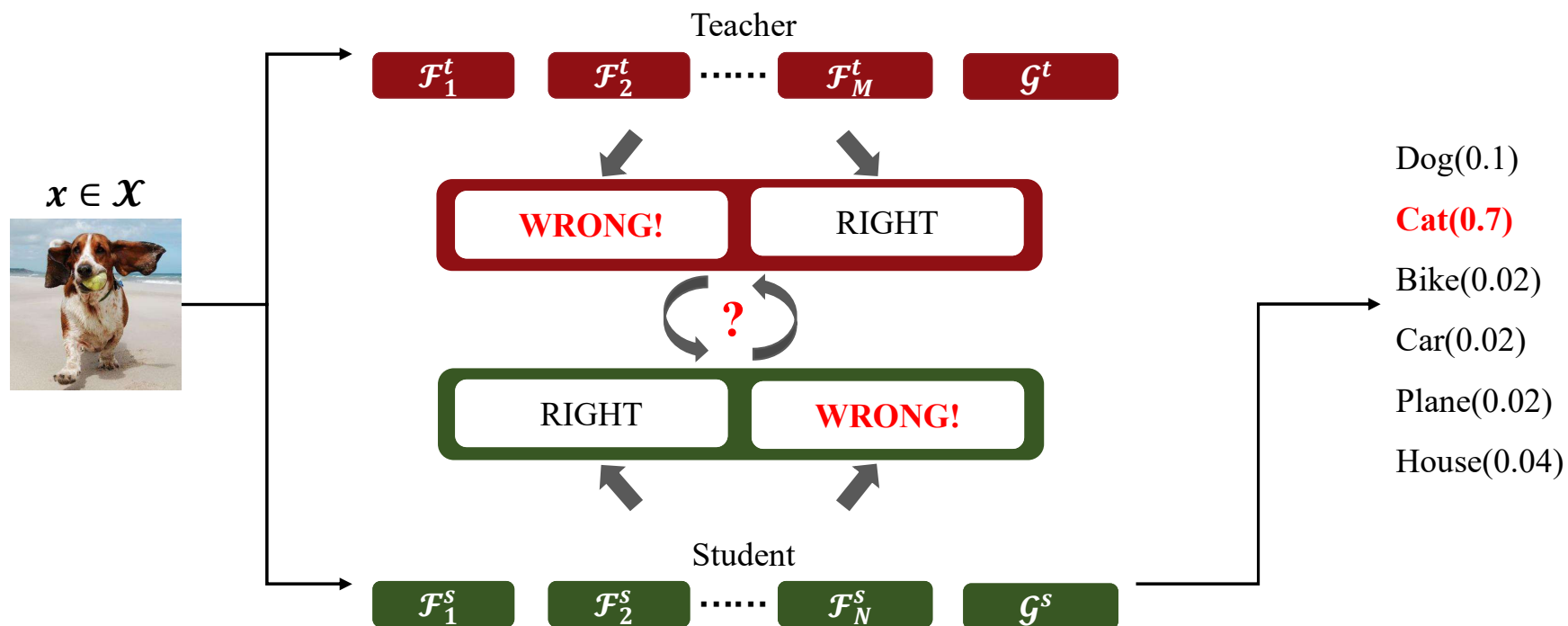
- 挑战一：如何捕捉模型在**整个学习过程中**的待纠正信息？



模型出错不仅体现在最终的分类阶段，也体现在中间的特征学习阶段。

研究思路

- 挑战二：如何设计有效的纠错机制？

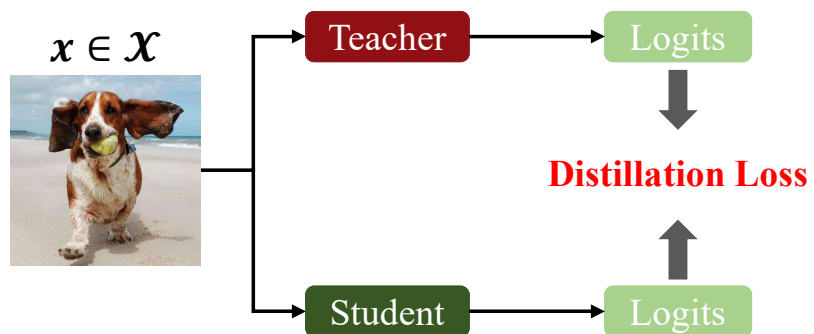


目录

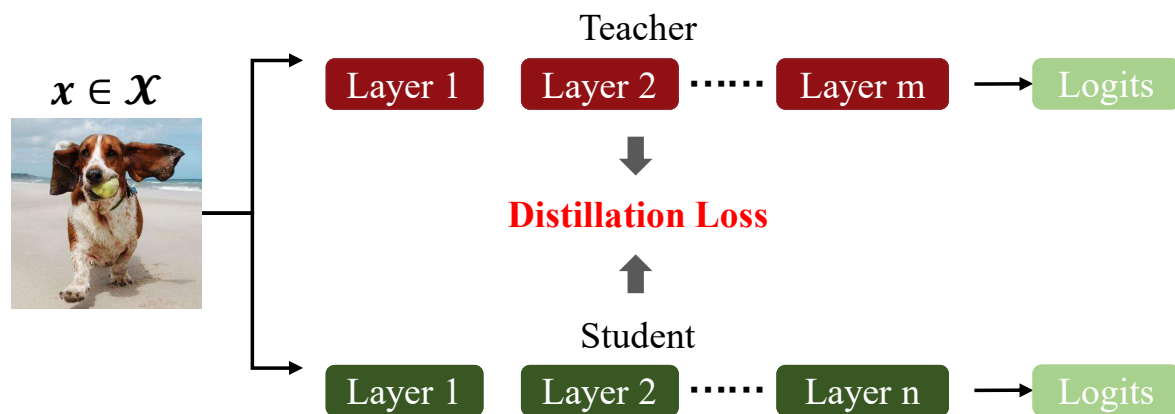
- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

模型框架

➤ 基于响应的知识蒸馏

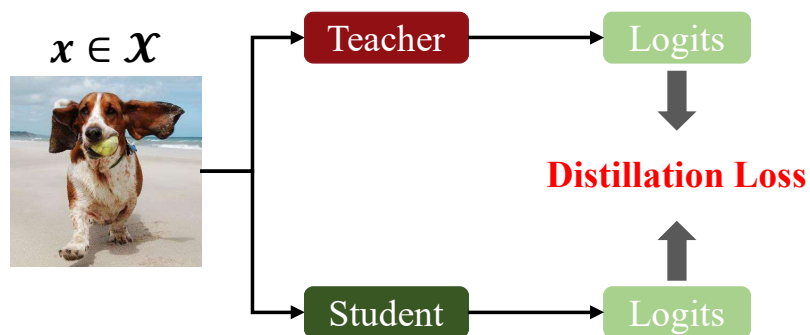


➤ 基于特征的知识蒸馏



模型框架

- 基于响应的知识蒸馏
 - 挑战一：如何捕捉模型在整个学习过程中的待纠正信息？
 - 挑战二：如何设计有效的纠错机制，增强教师模型和学生模型之间的知识蒸馏效果？



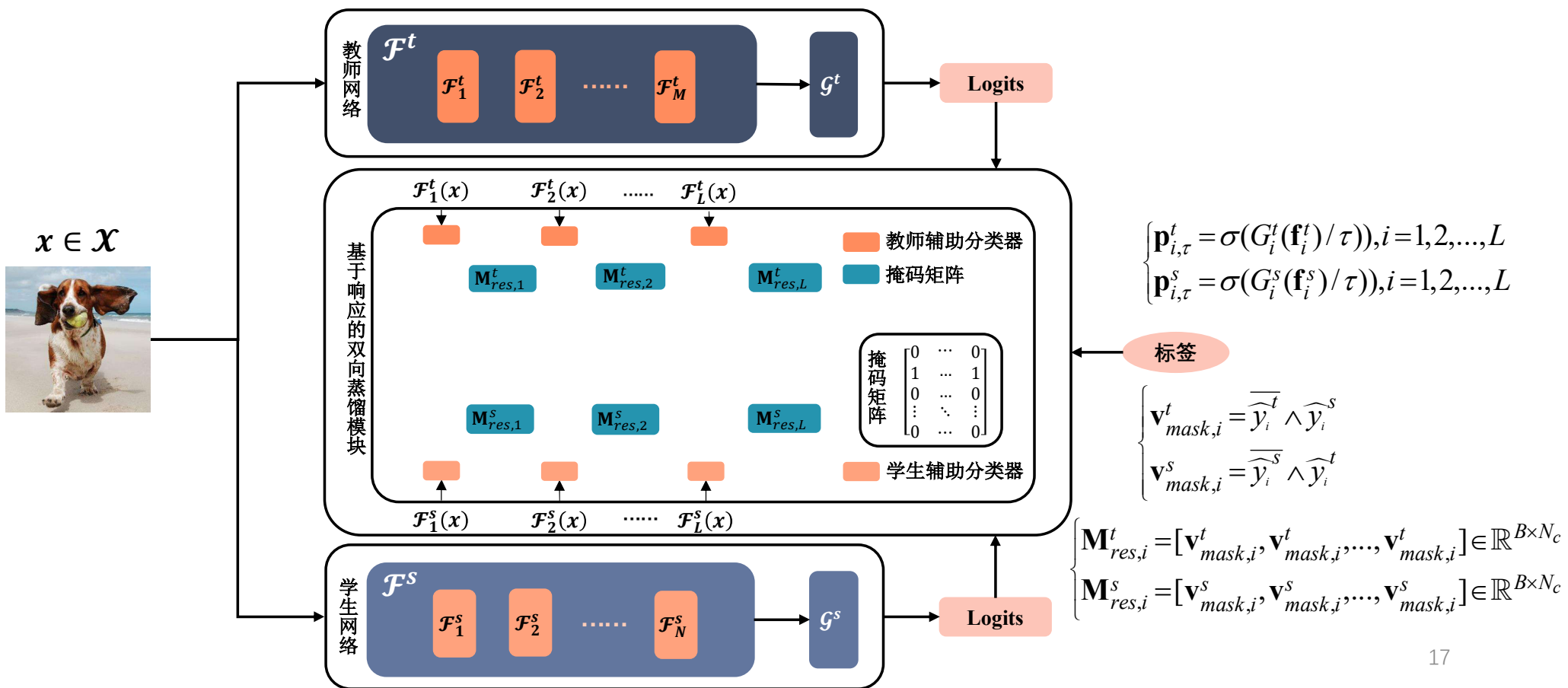
中间层信息可以有效的帮助教师模型
训练出性能更优异的学生模型

模型框架

➤ 基于响应的知识蒸馏

➤ 挑战一：如何捕捉模型在整个学习过程中的待纠正信息？

➤ 挑战二：如何设计有效的纠错机制，增强教师模型和学生模型之间的知识蒸馏效果？



研究思路

➤ 掩码矩阵

输入：学生模型的预测结果 $\hat{y}_i^s = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \dots \\ 1 \end{bmatrix}_{batchsize \times 1}$ ， 教师模型的预测结果 $\hat{y}_i^t = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 1 \end{bmatrix}_{batchsize \times 1}$

1: 学生模型预测正确的样本

1: 教师模型预测正确的样本

0: 学生模型预测错误的样本

0: 教师模型预测错误的样本

$$\mathbf{v}_{mask,i}^s = \overline{\hat{y}_i^s} \wedge \hat{y}_i^t = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \wedge \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}_{batchsize \times 1} \quad \longrightarrow \quad \mathbf{M}_{res,i}^s = [\mathbf{v}_{mask,i}^s, \mathbf{v}_{mask,i}^s, \dots, \mathbf{v}_{mask,i}^s] = \begin{bmatrix} 0 & \dots & 0 \\ 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}_{batchsize \times N_c}$$

$$\hat{\mathbf{p}}_{i,\tau}^s = \mathbf{M}_{res,i}^s \circ \mathbf{P}_{i,\tau}^s(x) = \begin{bmatrix} 0 & \dots & 0 \\ 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \circ \begin{bmatrix} 0.6543 & \dots & 0.2532 \\ 0.0562 & \dots & 0.7587 \\ \vdots & \ddots & \vdots \\ 0.1536 & \dots & 0.1124 \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 \\ 0.0562 & \dots & 0.7587 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}_{batchsize \times N_c}$$

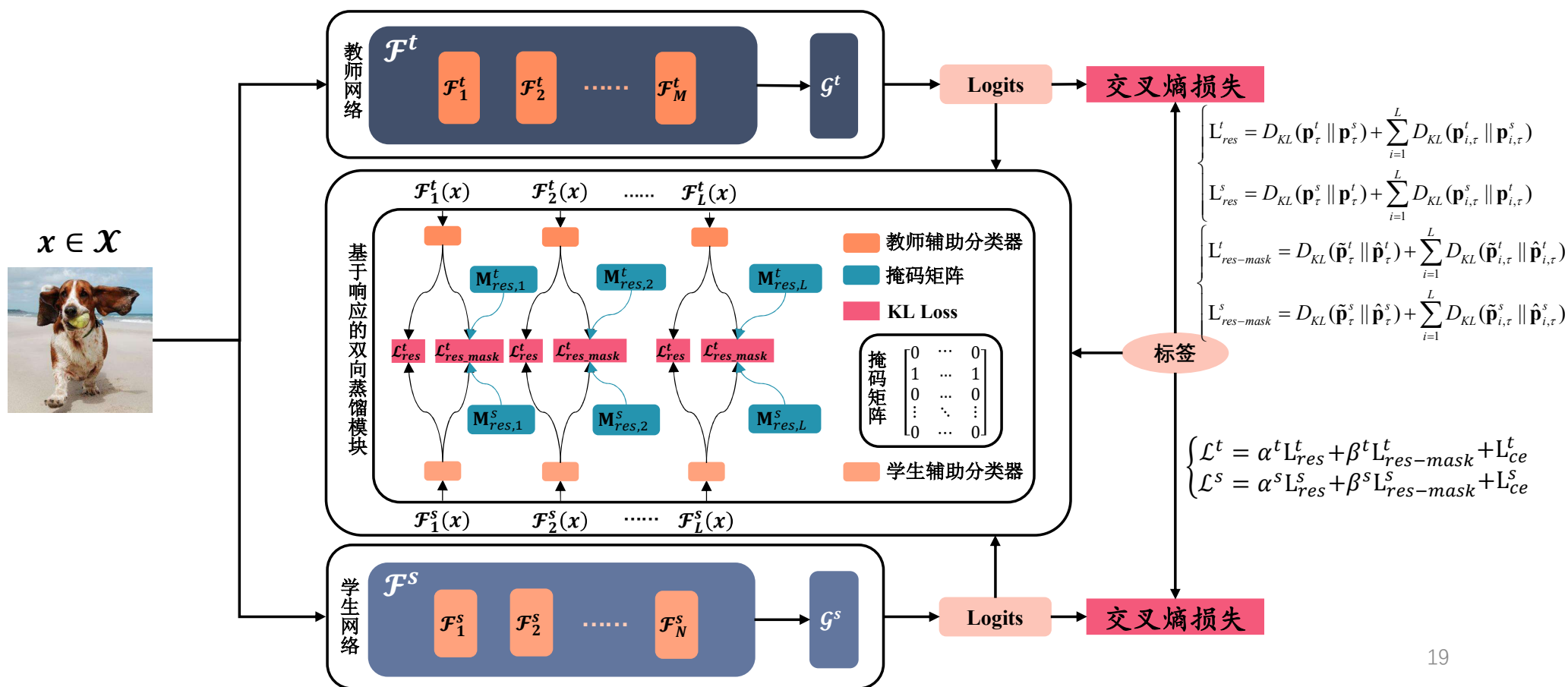
$$\tilde{\mathbf{p}}_{i,\tau}^s = \mathbf{M}_{res,i}^s \circ \mathbf{P}_{i,\tau}^t(x) = \begin{bmatrix} 0 & \dots & 0 \\ 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix} \circ \begin{bmatrix} 0.0403 & \dots & 0.5245 \\ 0.8192 & \dots & 0.1087 \\ \vdots & \ddots & \vdots \\ 0.1016 & \dots & 0.0143 \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 \\ 0.8192 & \dots & 0.1087 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}_{batchsize \times N_c}$$

模型框架

➤ 基于响应的知识蒸馏

➤ 挑战一：如何捕捉模型在整个学习过程中的待纠正信息？

➤ 挑战二：如何设计有效的纠错机制，增强教师模型和学生模型之间的知识蒸馏效果？

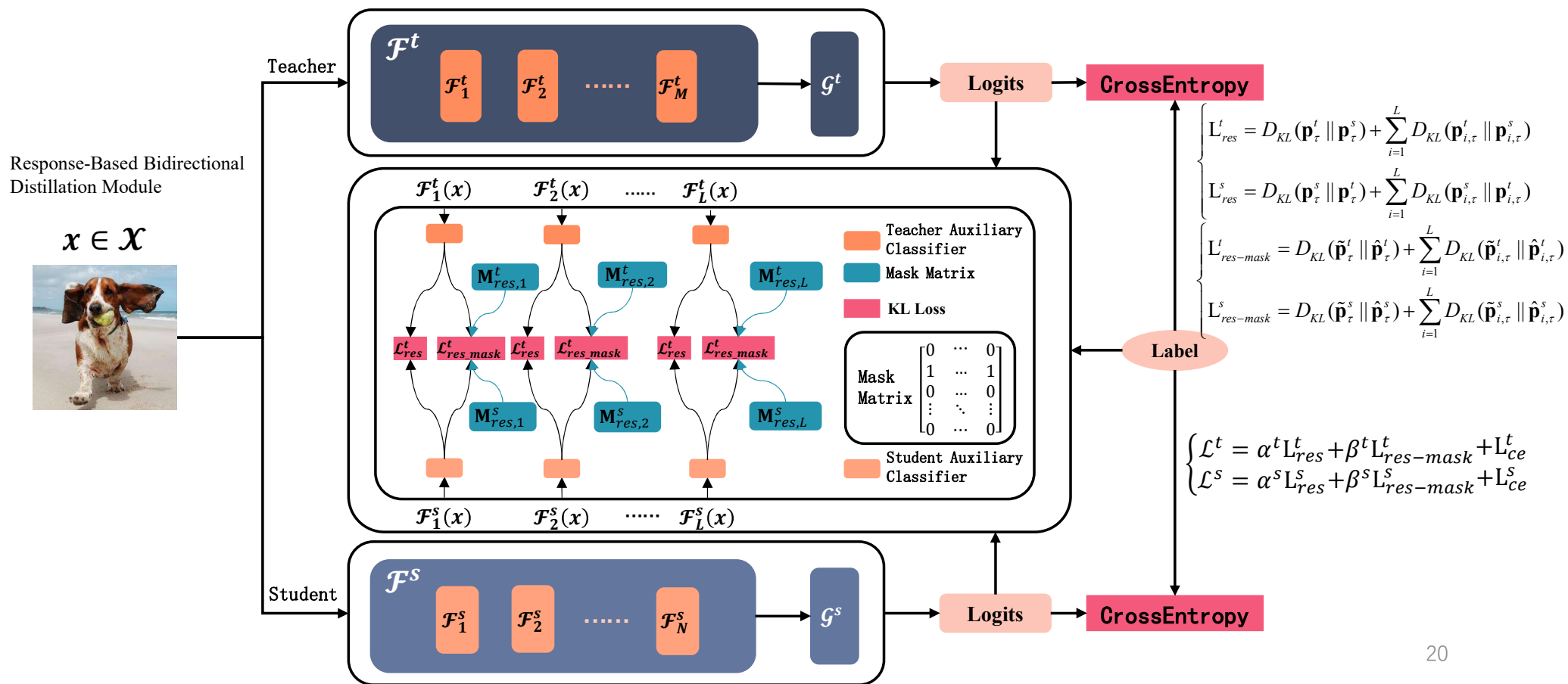


模型框架

➤ 基于响应的知识蒸馏

➤ 挑战一：如何捕捉模型在整个学习过程中的待纠正信息？

➤ 挑战二：如何设计有效的纠错机制，增强教师模型和学生模型之间的知识蒸馏效果？

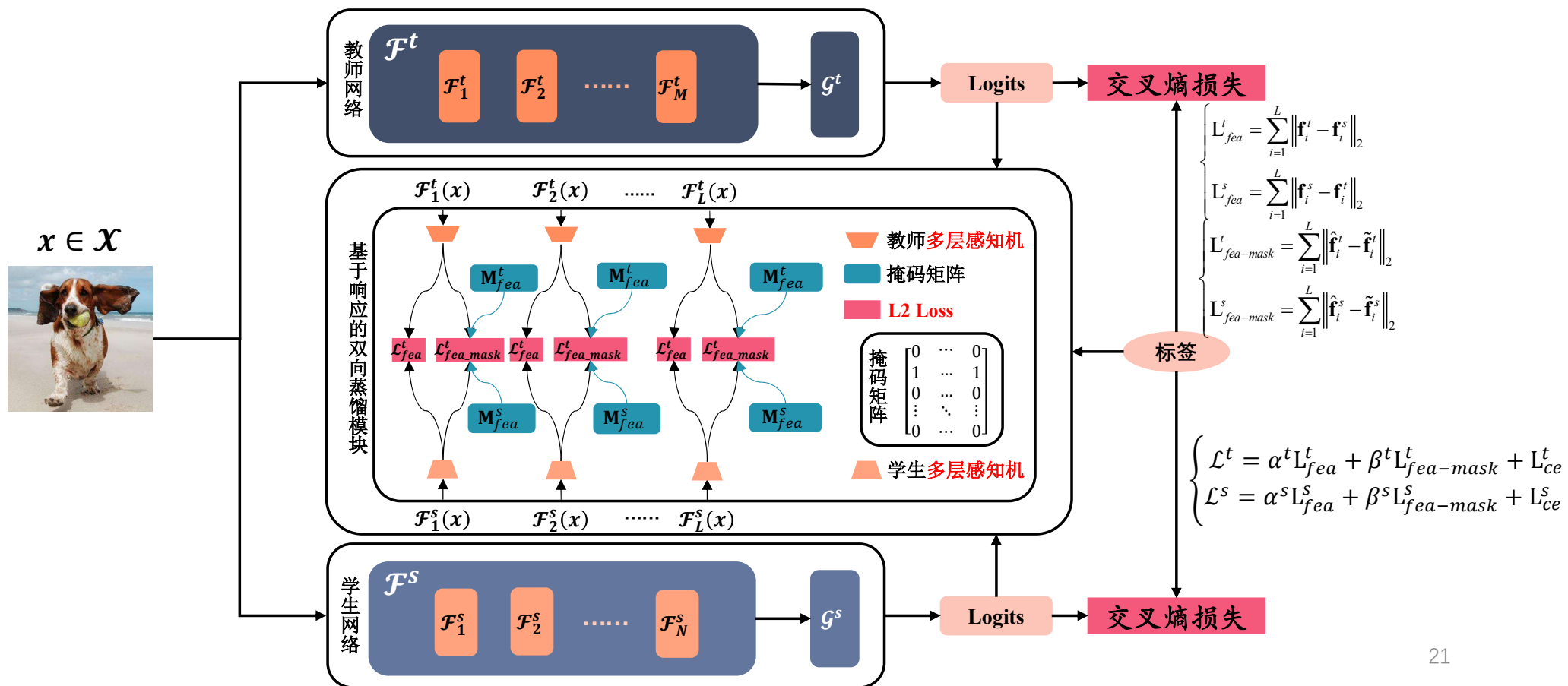


模型框架

➤ 基于特征的知识蒸馏

➤ 挑战一：如何捕捉模型在整个学习过程中的待纠正信息？

➤ 挑战二：如何设计有效的纠错机制，增强教师模型和学生模型之间的知识蒸馏效果？

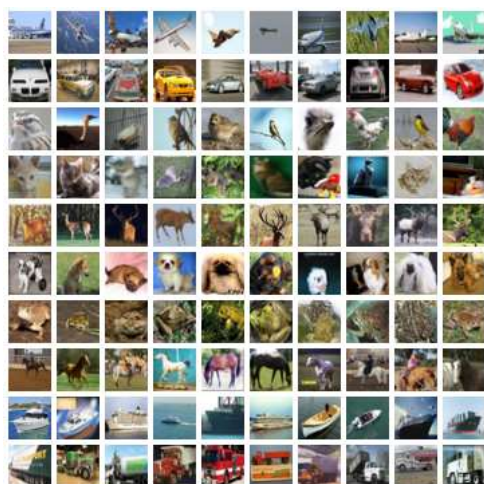


目录

- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

实验设置

➤ 数据集：CIFAR100



CIFAR100 数据集示例

图片数量：60000

类别数目：100

图片格式：32x32 RGB

超类	类别	超类	类别
水生哺乳动物	海狸, 海豚, 水獭, 海豹, 鲸鱼	大自然的户外场景	云, 森林, 山, 平原, 海
鱼	水族馆的鱼, 比目鱼, 射线, 鲨鱼, 鳕鱼	大杂食动物和食草动物	骆驼, 牛, 黑猩猩, 大象, 袋鼠
花卉	兰花, 罂粟花, 玫瑰, 向日葵, 郁金香	中型哺乳动物	狐狸, 豪猪, 负鼠, 浣熊, 臭鼬
食品容器	瓶子, 碗, 罐子, 杯子, 盘子	非昆虫无脊椎动物	螃蟹, 龙虾, 蜗牛, 蜘蛛, 蠕虫
水果和蔬菜	苹果, 蘑菇, 橘子, 梨, 甜椒	人	宝贝, 男孩, 女孩, 男人, 女人
家用电器	时钟, 电脑键盘, 台灯, 电话机, 电视机	爬行动物	鳄鱼, 恐龙, 蜥蜴, 蛇, 乌龟
家用家具	床, 椅子, 沙发, 桌子, 衣柜	小型哺乳动物	仓鼠, 老鼠, 兔子, 母老虎, 松鼠
昆虫	蜜蜂, 甲虫, 蝴蝶, 毛虫, 蟑螂	树木	枫树, 橡树, 棕榈, 松树, 柳树
大型食肉动物	熊, 豹, 狮子, 老虎, 狼	车辆1	自行车, 公共汽车, 摩托车, 皮卡车, 火车
大型人造户外用品	桥, 城堡, 房子, 路, 摩天大楼	车辆2	割草机, 火箭, 有轨电车, 坦克, 拖拉机

➤ 评价指标

$$Accuracy = \frac{TrueSamples}{TotalSamples}$$

TrueSamples 表示被正确分类的样本的总数

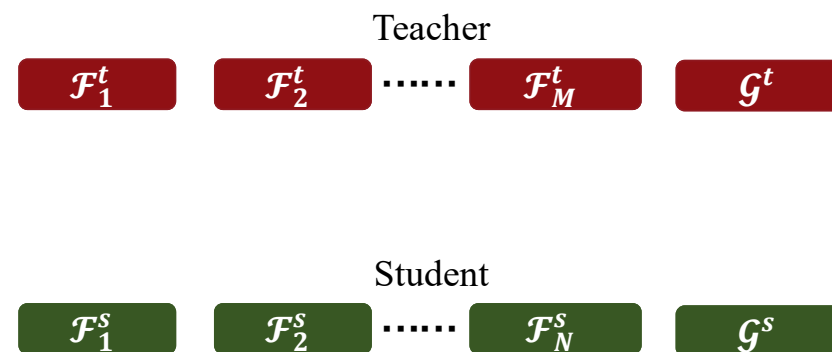
TotalSamples 表示所有样本的总数

实验设置

➤ 模型选择

表4-1 教师模型和学生模型的搭配

教师网络	输出选择 (M/FC)	学生网络	输出选择 (N/FC)
WRN-40-2	M=1/M=2/M=3/FC	WRN-16-2	N=1/N=2/N=3/FC
WRN-40-2	M=1/M=2/M=3/FC	WRN-40-1	N=1/N=2/N=3/FC
ResNet56	M=1/M=2/M=3/FC	ResNet20	N=1/N=2/N=3/FC
ResNet32x4	M=1/M=2/M=3/FC	ResNet8x4	N=1/N=2/N=3/FC
VGG13	M=1/M=2/M=3/M=4	MobileNetV2	N=2/N=3/N=5/N=7
ResNet50	M=1/M=2/M=3/M=4/FC	MobileNetV2	N=2/N=3/N=5/N=7/FC
WRN-40-2	M=1/M=2/M=3/FC	ShuffleNetV1	N=1/N=2/N=3/FC
ResNet32x4	M=1/M=2/M=3/FC	ShuffleNetV2	M=1/M=2/M=3/FC



实验结果

➤ 模型对比

表4-2 在CIFAR100数据集上的模型对比结果

教师网络	WRN-40-2	WRN-40-2	ResNet56	ResNet32x4	VGG13	ResNet50	WRN-40-2	ResNet32x4
学生网络	WRN-16-2	WRN-40-1	ResNet20	ResNet8x4	MobileNetV2	MobileNetV2	ShuffleNetV1	ShuffleNetV2
KD	74.31%	73.90%	70.97%	73.49%	75.19%	74.87%	75.83%	75.43%
FitNet	75.30%	74.30%	71.21%	75.37%	75.42%	75.41%	76.27%	76.91%
AT	75.64%	74.32%	71.35%	75.06%	74.08%	76.57%	76.51%	76.32%
AB	71.26%	74.55%	71.56%	74.31%	74.98%	75.87%	76.43%	76.40%
VID	75.31%	74.23%	71.35%	75.07%	75.67%	75.97%	76.24%	75.98%
RKD	75.33%	73.90%	71.67%	74.17%	75.54%	76.20%	75.74%	75.42%
SP	74.35%	72.91%	71.45%	75.44%	75.68%	76.35%	76.40%	76.43%
CC	75.30%	74.46%	71.44%	74.40%	75.66%	76.05%	75.63%	75.74%
CRD	75.81%	74.76%	71.83%	75.77%	76.13%	76.89%	76.37%	76.51%
SSKD	76.16%	75.84%	70.80%	75.83%	76.21%	78.21%	76.71%	77.64%
ECKD-R	76.94%	76.76%	72.17%	76.89%	76.81%	78.91%	77.23%	78.37%
ECKD-F	76.00%	75.79%	71.23%	76.01%	75.47%	76.12%	76.58%	77.02%

实验结果

➤ 消融实验

表4-3 ECKD-R在CIFAR100数据集中的消融研究结果

消融方法	正确率 (Acc)
w/o EC-R	76.56%
w/o Bi-R	76.17%
w/o Layer-R	76.23%
w/o Data-Aug-R	76.42%
ECKD-R	76.94%

结论：证明了ECKD的双向蒸馏模块设计的合理性。
同时，纠错机制对相互学习策略和层间知识有较强的依赖。

1. w/o EC-R: 移除纠错机制
2. w/o Bi-R: 移除双向蒸馏
3. w/o Layer-R: 移除层间知识
4. w/o Data-Aug-R: 移除数据增强

实验结果

➤ 消融实验

表4-4 KD与纠错机制结合前后结果对比

教师模型-学生模型	相同卷积结构		不同卷积结构	
	WRN-40-2	WRN-16-2	ResNet50	MobileNetV2
KD	75.44%	74.31%	77.82%	74.87%
KD+纠错	76.68%	75.64%	78.43%	74.97%

可以看出，无论教师模型与学生模型的卷积结构是否相同，在经过纠错导向机制后，教师模型与学生模型的Acc均有不同程度的提升。

实验结果

➤ 消融实验

表4-5 纠错前后KD预测样本分布变化

教师模型 学生模型	相同卷积结构			不同卷积结构		
	WRN-40-2			ResNet50		
	WRN-16-2			MobileNetV2		
方法	KD	KD+纠错	Δ	KD	KD+纠错	Δ
TRSR	67.27%	68.87%(↑)	+1.60%	69.23%	69.93%(↑)	+0.70%
TWSW	17.52%	16.55%(↓)	-0.97%	16.54%	16.53%(↓)	-0.01%
TRSW	8.17%	7.81%(↓)	-0.36%	8.59%	8.5%(↓)	-0.09%
TWSR	7.04%	6.77%(↓)	-0.27%	5.64%	5.04%(↓)	-0.60%

TRSR: 教师预测正确且学生预测正确的样本

TRSW: 教师预测正确且学生预测错误的样本

TWSR: 教师预测错误且学生预测正确的样本

TWSW: 教师预测错误且学生预测错误的样本

结论:

1) 教师模型和学生模型受益于它们可以通过相互学习**纠正自身错误**, 这也进一步的验证了我们的方法**可行性及先进性**。

2) 纠错机制会受制于教师模型和学生模型之间的差异化程度, 其差异性越大纠错效果越弱。

实验结果

➤ 消融实验

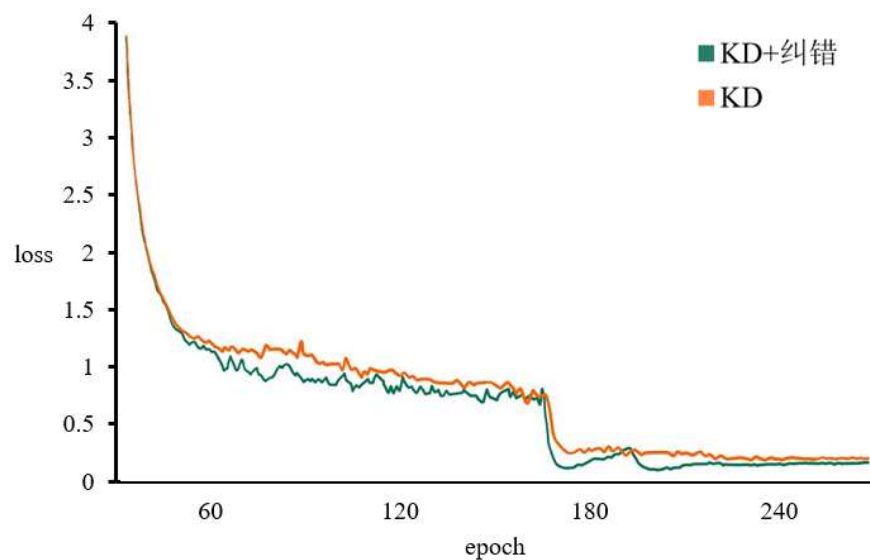


图4-1 纠错前后模型收敛对比

结论：在添加纠错机制后，学生模型的收敛效果明显提升。这表明，纠错机制可以帮助模型更好地收敛，从而提升蒸馏的效果。

目录

- 研究背景
- 相关工作
- 研究思路
- 模型框架
- 实验结果
- 总结与展望

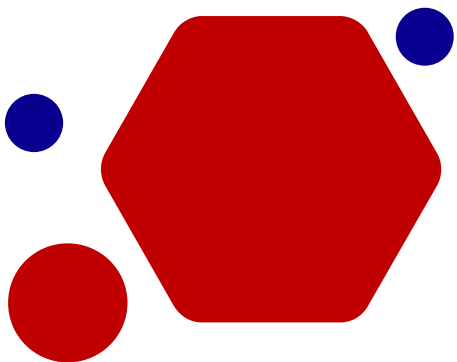
总结

1. 提出了以纠错为导向的知识蒸馏方法，有效缓解了蒸馏过程中传递错误知识的问题。
2. 从基于响应和基于特征两个角度出发，设计了两种以纠错为导向的知识蒸馏正则项，有效提升了知识蒸馏的效果。
3. 在基准数据集上的实验证明了方法的可行性与有效性。

未来工作

1. **提升模型的鲁棒性。**目前本文所提的方法使用了与以往工作相同的数据增强技术，但并未对此进行过多的研究。为了进一步提升模型的鲁棒性，我们将探索适用于知识蒸馏领域的数据增强技术。
2. **拓展模型的纠错方式。**本文所提方法的纠错方式依赖于计算得到的掩码矩阵，该矩阵是由独热编码变换而来，被用于提取模型的待纠正信息。然而，这种只用0和1提取知识的方式存在丢失信息的情况。为此，我们将拓展模型的纠错方式，探索一种自适应生成的解决方案，以取代只包含1和0的掩码矩阵。
3. **增强模型的通用性。**本文仅探讨了图像分类领域的知识蒸馏技术，事实上目前很多领域都会涉及知识蒸馏技术，如自动驾驶领域[4]、目标检测领域[3]等，未来我们将拓展我们模型的其他适用领域。
4. **提升模型的效率。**本文设计的以纠错为导向的知识蒸馏模型，在未增加模型体量的前提下，进一步提升了知识蒸馏效果。未来我们将设计更优的方案，进一步减少参数量，以提升模型的计算效率，加速其在嵌入式端和移动端的落地应用。

问答



感谢各位老师
指导！