南京郵電大學
Nanjing University of Posts and Telecommunications

面向特征融合与知识蒸馏的恶意软件分类

汇报人：王刚

引言

恶意软件分类旨在提取恶意软件的关键特征,并用其来构建分类器，从而对该恶意软件的类别进行判断，本质上是一个多分类任务。基于深度学习的恶意软件分类方法则因其良好的扩展性和稳定性受到广泛关注，目前，依据恶意软件特征提取方法的不同，其主要可以分为如下两类：

1.基于恶意软件特征的分类方法。这类方法需要人工制定特征工程来提取恶意软件特征，分类器的构建具有较高的成本。此外，在新型恶意软件数量激增的情况下，特征工程难以保持良好的特征抽取能力，导致模型性能难以提升。

2.基于恶意软件图像的分类方法。这类方法仅通过单一图像语义表示来抽取恶意软件特征，导致模型无法捕获恶意软件序列特征和分布特征之间的依赖关系，且模型的规模较大，对计算资源较为依赖。

针对上述问题，本研究提出面向特征融合与知识蒸馏的恶意软件分类方法，基于注意力机制，模型对不同特征之间的关联性和依赖关系进行挖掘与捕获。同时，利用知识蒸馏来降低模型的规模，以期减少模 型对计算资源的消耗.

目录
CONTENTS

南京郵電大學
Nanjing University of Posts and Telecommunications

模型方法

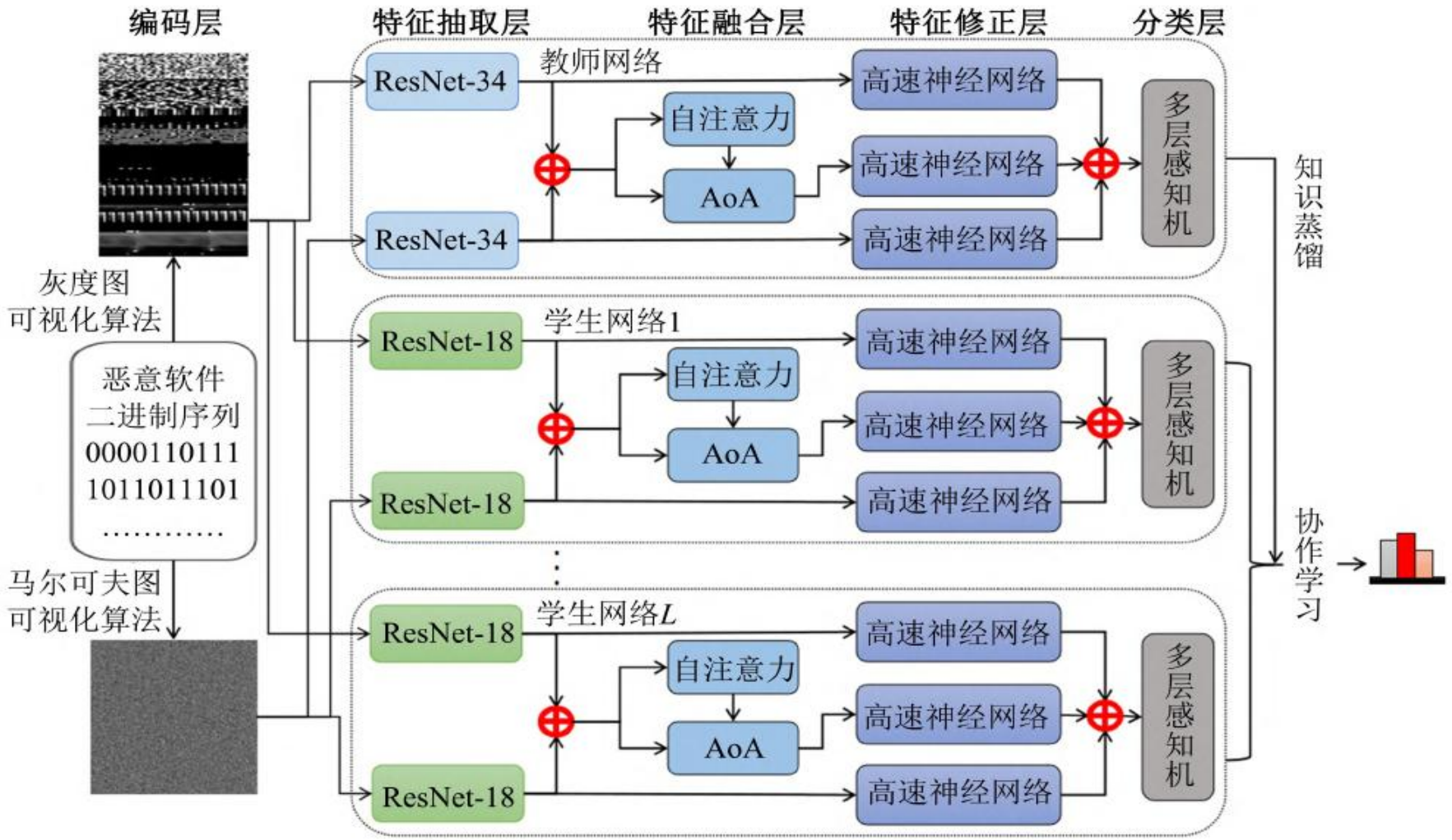


图 1　模型框架图

模型方法

- *编码层*

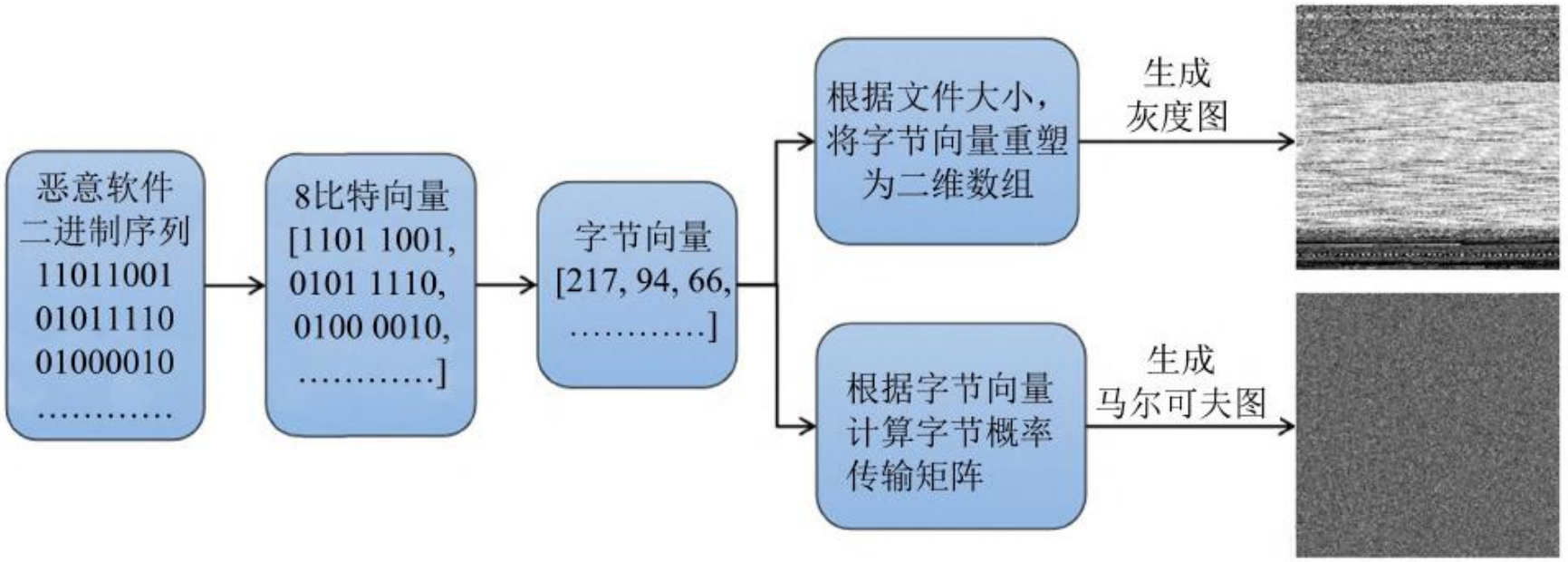


图 2　恶意软件可视化过程

模型方法

特征融合层

自注意力模块对恶意软件的序列特征和分布特征进行自注意力计算，捕捉不同特征之间的依赖关系，该模块将序列特征向量和分布特征向量进行拼接，得到输入向量 $f_{f_{\text{att\_in}}}$ 分别使用 W^{Q}、W^{K}、W^{V} 这三个矩阵作为Query、Key、Value，将其映射到相应的特征空间；然后再进行注意力计算，得到融合恶意软件的序列特征和分布特征的融合向量 $f_{\text{att\_out}}$，融合向量的生成过程可用公式表示为：

$$f_{\text{att\_out}} = \text{Att}(W^{Q}f_{\text{att\_in}}, W^{K}f_{\text{att\_in}}, W^{V}f_{\text{att\_in}})$$

$f_{\text{att\_out}}$ 可能存在结果向量与查询向量不匹配的情况，导致误导信息的产生，这会影响模型的性能。因此将 $f_{\text{att\_in}}$ 和 $f_{\text{att\_out}}$ 作为输入，利用AoA进行二次注意力计算，可以过滤掉无关的注意力结果．最终得到特征融合层的输出向量

$$f_{\text{AoA\_out}} = \text{Dropout}((f_{\text{att\_out}} \oplus f_{\text{att\_in}})W_1 + b_1) \otimes \delta((f_{\text{att\_out}} \oplus f_{\text{att\_in}}) \times W_2 + b_2)$$

模型方法

- *特征修正层*

为充分利用恶意软件的序列特征和分布特征，同时增强融合模块的泛化性，利用高速神经网络的门机制，可以对不同的特征向量进行修正，其输出向量定义为

$$y = \begin{cases} x, \text{ if } T(x, W_T) = 0 \\ H(x, W_H), \text{ if } T(x, W_T) = 1 \end{cases}$$

式中： x 为特征抽取层得到的恶意软件的序列特征向量、分布特征向量或特征融合层得到的融合特征向量；函数H(x)表示的是一个线性层和非线性层的组合；矩阵 W_H、W_T 为线性层的可学习参数；非线性层使用Sigmoid激活函数.函数T(x)的表达式为 $T(x) = \delta((W_T)^T x + b_T)$

式中：b_T为-1或者-3.由定义可知，高速神经网络将会直接保留输入特征向量中的有效信息，并对其中的无效信息进行映射，进一步增强特征向量的有效性

模型方法

- *模型压缩*

模型压缩的过程可分为４步：
１）预训练特征抽取层为ResNet-34的大模型，对其进行非结构化剪枝后，将其作为教师网络；
２）固定教师网络参数，对多个特征抽取层为ResNet-18的学生网络进行知识迁移；
３）多个不同的学生网络互相之间进行协作学习；
４）选择性能最优的学生网络作为最终的分类模型，用于输出分类结果。

知识蒸馏是一种深度学习中的模型优化技术，主要目标是让小型模型能够模仿大型模型的行为，并且在学习过程中，将大型模型的"知识"蒸馏（传递）给小型模型，使得小型模型在表现上能够接近或者甚至超过大型模型。

知识蒸馏的基本思想是通过在训练过程中，使用大型模型的预测结果（通常是类别概率）来指导小型模型的学习过程。一般来说，知识蒸馏会在损失函数中添加一个额外的项，该项用于衡量小型模型的预测与大型模型的预测之间的相似性。这个相似性度量通常使用交叉熵损失函数或其他类似的距离度量。

模型方法

知识蒸馏和协作学习的核心，都是用于更新学生网络参数的损失计算. 当对第 i 个学生网络进行参数更新时，对应的损失(L_{TL})将由散度值(L_1)和混合损失值(L_2)进行加权计算得到，其表达式为

$$L_1 = D_{KL}(\boldsymbol{p}_i \| \boldsymbol{p}_k) \tag{6}$$

$$L_2 = \text{Mix Loss}(\boldsymbol{p}_i, y) \tag{7}$$

$$L_{TL} = \alpha L_1 + \beta L_2 \tag{8}$$

式中：$D_{KL}(\cdot)$代表 KL 散度函数；$\boldsymbol{p}_i$ 为第 i 个学生网络的输出向量；在知识蒸馏阶段和协作学习阶段，$\boldsymbol{p}_k$ 分别为教师网络和性能最优学生网络的输出向量；y 为恶意软件类别对应的数值；$\alpha+\beta=1$，且 $\alpha \gg \beta$. 混合损失函数的定义为

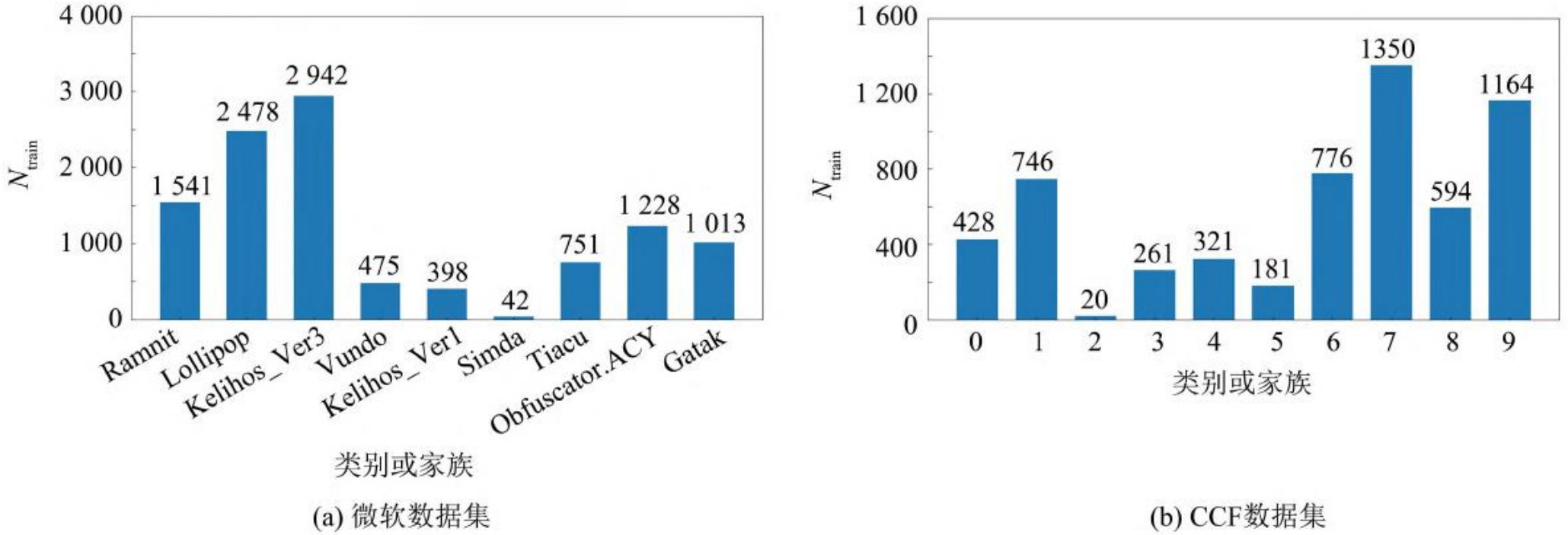
$$\text{Mix Loss}(\boldsymbol{p}_i, y) = L_{CE}(\boldsymbol{p}_i, y) + L_{FL}(\boldsymbol{p}_i, y) + L_{OHEM}(\boldsymbol{p}_i, y) \tag{9}$$

式中：$L_{CE}(\cdot)$、$L_{FL}(\cdot)$和 $L_{OHEM}(\cdot)$分别代表交叉熵损失函数 Cross Entropy、聚焦损失函数 Focal Loss[17] 和难样本损失函数 OHEM[18].

实验分析

- *实验设计*

数据集描述



(a) 微软数据集　　(b) CCF数据集

图 3　数据集训练样本分布

实验分析

评价指标

对于微软数据集，可以将测试集的测试结果上传到kaggle，官方将以下式计算多分类对数损失，即

$$L_{\mathrm{TL}} = -\frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}}\sum_{j=1}^{M} y_{ij}\ln p_{ij}$$

对于CCF公开数据集，只能在训练集上进行两折交叉验证，采用精确度 P_{ACC} 来衡量模型的性能．其计算式为

$$P_{\mathrm{ACC}} = \frac{n_{\mathrm{true}}}{n}$$

实验分析

基准模型

选取 5 个模型作为基准模型.

1）M-CNN模型．将灰度图作为输入，对VGG16模型进行训练.

2）IMCFN模型．将灰度图映射为彩图，并对图像进行数据增强，用于微调VGG16模型.

3）DCNM模型． 将马尔可夫图作为输入，用于训练改良的VGG16模型.

4）MNCD模型．将灰度图作为输入，训练所提出的一种基于深度学习的新的混合模型.

5）S-DCMN模型．将灰度图作为输入数据，训练所提出的由 3 个不同卷积神经网络组成的模型

参数设定

本研究模型需要50个迭代轮次才能完成训练，且每训练20个迭代轮次，都需要将学习率降低10倍.批大小、学习率、权重衰退、动量、非结构化剪枝率、学生网络数据的最优参数分别为8、0.005、0.0005、0.9、40%、4.

实验分析

- *有效性实验*

表 1　各模型在微软和 CCF 数据集上的实验结果对比

Tab.1　Comparison of experimental results of various models in Microsoft and CCF datasets

模型	参数量/$\times 10^6$	浮点运算数/$\times 10^9$	L_{TL}	P_{ACC}/%
M-CNN	134.30	15.50	0.062 64	93.18
DCNN	40.42	20.11	0.064 36	97.36
IMCFN	134.30	15.50	0.085 15	97.15
NMCD	155.92	4.85	0.062 32	97.87
S-DCNN	68.20	8.75	0.063 28	97.50
本研究模型	39.20	4.22	0.020 30	98.93

实验分析

- *单输入实验*

表 2　单输入实验结果

Tab.2　Experimental results of single input

模型	输入数据	L_{TL}	P_{ACC}
M-CNN	灰度图	0.062 64	93.18
VGG19	灰度图	0.101 43	92.10
ResNet-34	灰度图	0.059 63	95.83
VGG16	马尔可夫图	0.075 07	97.24
DCNN	马尔可夫图	0.064 36	97.36
VGG19	马尔可夫图	1.904 95	97.18
ResNet-34	马尔可夫图	0.040 55	97.58

实验分析

- *消融实验*

表 3　特征融合层和特征修正层消融实验结果

Tab.3　Experimental results of feature fusion layer and feature correction layer ablation

特征融合层	特征修正层	L_{TL}	P_{ACC}/%
—	—	0.034 09	98.15
√	—	0.023 56	98.45
—	√	0.029 10	98.36
√	√	0.020 30	98.93

注：“√”表示用到对应模块；“—”表示未用到对应模块.

实验分析

- *模型压缩实验*

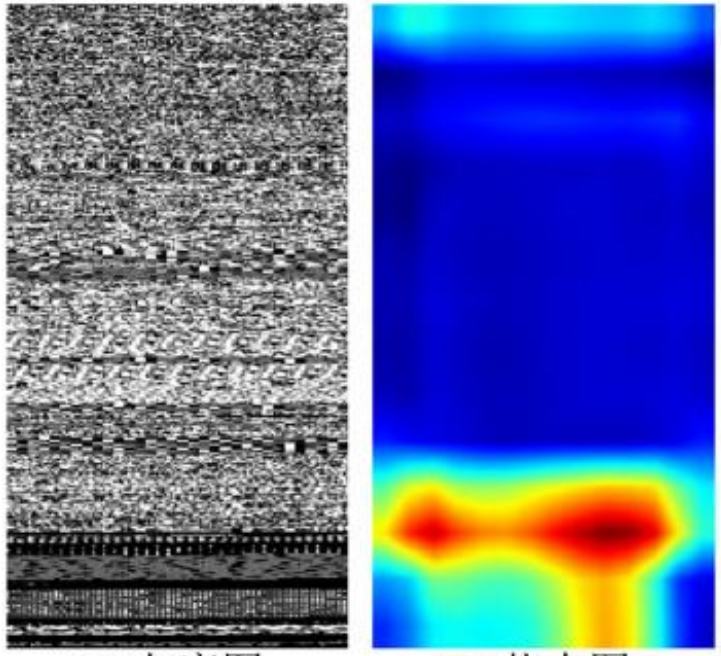
表 4　模型压缩实验结果

Tab.4　Experimental results of model compression

模型	参数量/$\times 10^6$	浮点运算数/$\times 10^9$	t/ms	L_{TL}	P_{ACC}/%
教师网络	59.4	8 509	29.76	0.023 14	98.52
学生网络(直接训练)	39.2	4 228	19.96	0.029 24	98.30
学生网络(基于本研究)	39.2	4 228	19.96	0.020 30	98.93

实验分析

- *可解释性实验*



(a) 灰度图　　(b) 热力图

图 4　“熊猫烧香”的灰度图和热力图

```
CODE:0040D1D8 asc_40D1D8      db '***武*汉*男*生*感*染*下**载*者***',0
CODE:0040D1D8                                          ; DATA XREF: start+57↑o
CODE:0040D1F9                 db    0
CODE:0040D1FA                 db    0
CODE:0040D1FB                 db    0
CODE:0040D1FC                 dd 0FFFFFFFFh, 29h
CODE:0040D204 aMopery         db '感谢艾玛,mopery,海色の月,对此木马的关注!~',0
CODE:0040D204                                          ; DATA XREF: start+66↑o
CODE:0040D22E                 db    0
CODE:0040D22F                 db    0
CODE:0040D230                 db 0FFh
CODE:0040D231                 db 0FFh
CODE:0040D232                 db 0FFh
CODE:0040D233                 db 0FFh
CODE:0040D234                 db  1Ch
CODE:0040D235                 db    0
CODE:0040D236                 db    0
CODE:0040D237                 db    0
CODE:0040D238 aPs             db 'PS: 服了。。。艾玛。。。  =,=',0
```

图 5　“熊猫烧香”的专有字符串

总结

本文提出一种面向特征融合与知识蒸馏的恶意软件分类方法，基于注意力机制融合恶意软件的序列特征 和分布特征，对不同特征之间的关联性进行挖掘．通过知识蒸馏和协作学习技术，降低分类模型的参数量和计算量．在微软和CCF两个公开数据集上的实验结果表明，与基准模型相比，本研究模型具有更加良好的性能和更小的规模．此外，实验分析证明，卷积神经网络可以从恶意软件图像中抽取关键特征，并将其用于模型学习和分类．后续研究将在对恶意软件进行可视化之前，对恶意软件中无用的垃圾字节进行消除，以降低作为模型输入的恶意软件图像噪声．

THANKS!

感谢您的观看与聆听

汇报人：王刚