中国科学技术大学2016级

信息科学与技术学院 4班

多模态场景下的自监督学习算法研究

答辩人：付哲仁　　导师：毛震东 特任研究员

目录

CONTENTS

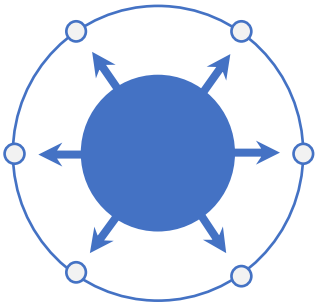
Part 1

绪论

选题背景

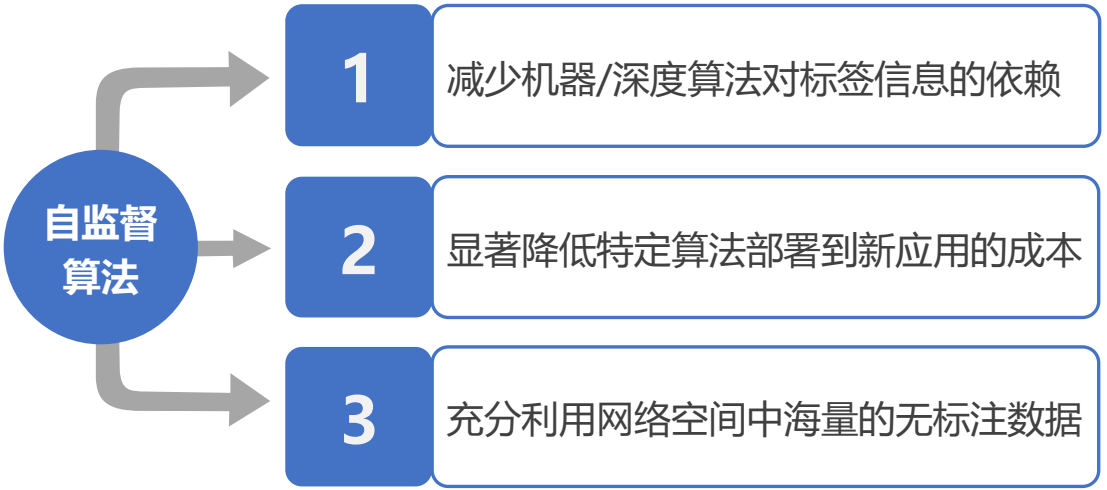
在计算机视觉领域，目前最成功的方法都是基于有监督的深度学习算法，背后极度依赖于庞大的人工标注数据。

- 人工标签的获取**耗时费力**，经济成本很高，难以大量获取。



- **自监督学习**算法可从未标注的数据中学习有用信息。

研究意义


自监督
算法
1
减少机器/深度算法对标签信息的依赖
2
显著降低特定算法部署到新应用的成本
3
充分利用网络空间中海量的无标注数据

问题定义

定义

自监督学习，就是从未标注的大量数据自动地抽取出“**监督信息**”，从而指导算法学习有用的特征。

思路

如何从未标注的数据挖掘**有效**的监督信号，即从未标注的数据中挖掘什么样的信息来**指导**网络的训练。

研究进展与挑战

研究主要集中在设计合适的**学习任务**：如预测图像块的位置关系，图像着色，旋转角度预测、特征聚类等。

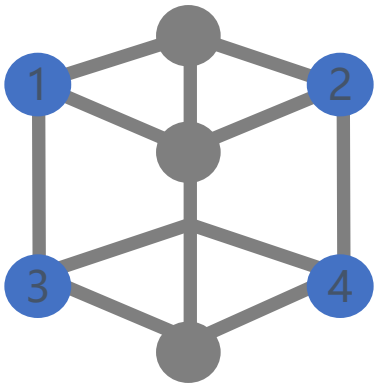
挑战一: 绝大部分工作都是基于ImageNet数据集完成的，仅限于**视觉模态**的信息被利用，其他模态未被考虑。

挑战二: 如何进一步挖掘图像间的关联性，寻找更**深层次**的潜在约束，作为监督信息的工作越来越难进行更深。

算法创新与优势

1. 基于多模态数据

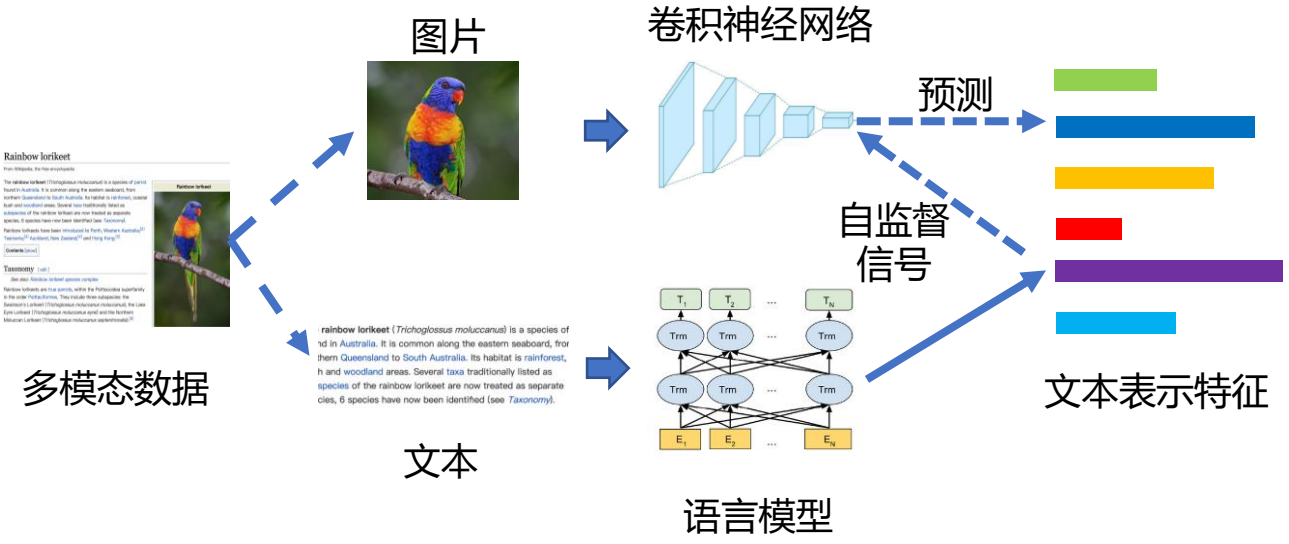
2. 利用数据内容之间的相关性



3. 文本的语义信息作为监督信号

4. 算法框架简单，易实现，可扩展性强

算法框架


图片
卷积神经网络
预测
自监督
信号
多模态数据
文本
语言模型
文本表示特征
Rainbow lorikeet
Taxonomy
rainbow lorikeet (Trichoglossus moluccanus) is a species of
Trm
T1
T2
TN
E1
E2
EN

主要贡献

A

提出了一种**多模态场景**下，利用图像与文本间的语义相关性，完成对**视觉特征**的自监督表示学习方法。

B

尝试了**不同种类**的多模态数据集、文本编码器、损失函数，综合比较了不同组合间的**差异**。

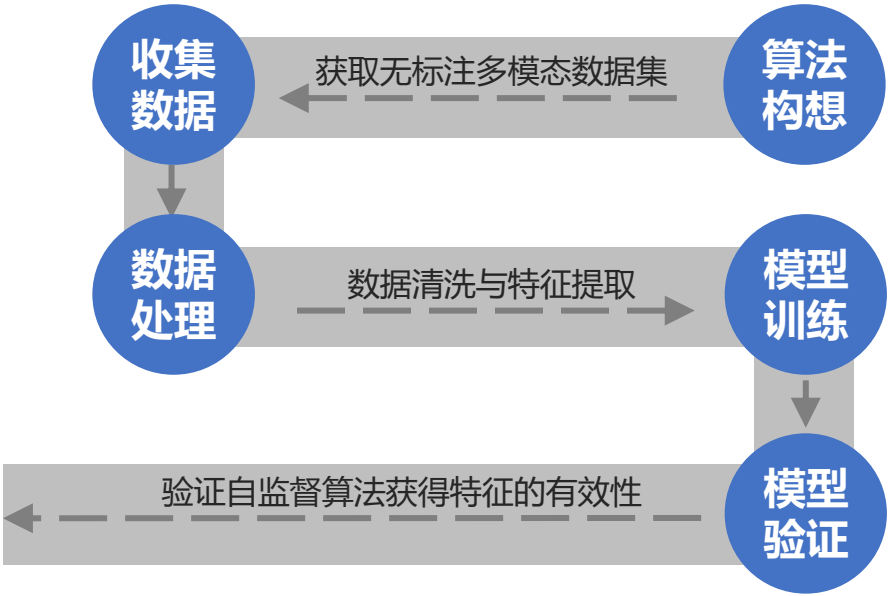
C

在多种**下游任务**上进行验证实验，证明该自监督算法学习的视觉特征，具有表示能力与**语义特性**。

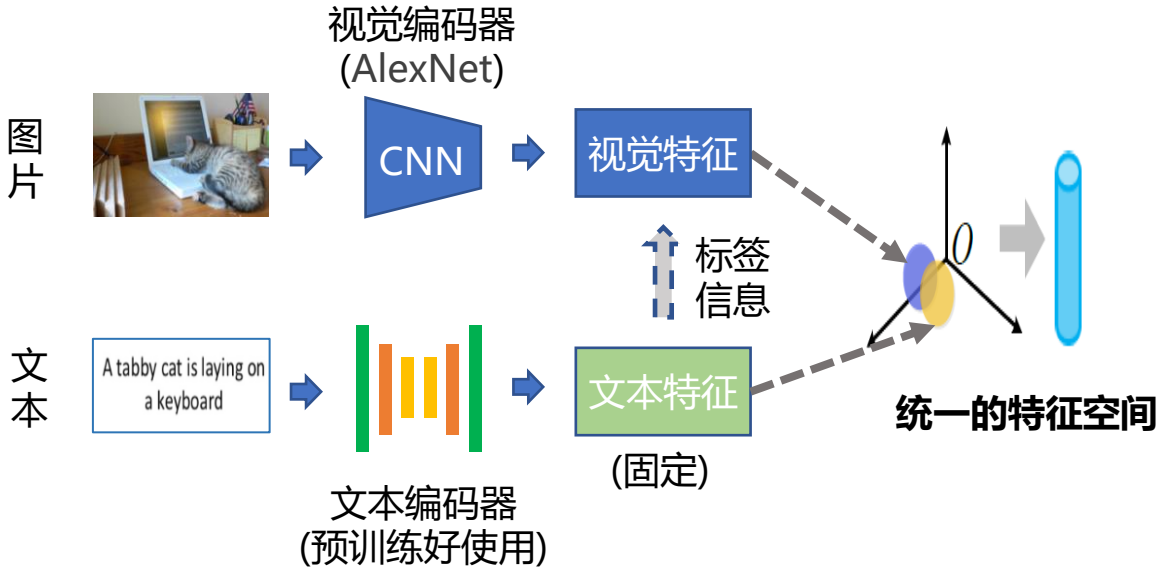
Part 2

思路与方法

研究思路


收集数据
获取无标注多模态数据集
算法构想
数据处理
数据清洗与特征提取
模型训练
模型验证
验证自监督算法获得特征的有效性

算法框架



希望语义相关的图片-文本对特征尽可能相似(靠近)

数据收集

维基百科网站

V S

社交媒体平台

Cat

From Wikipedia, the free encyclopedia

This article is about the species that is commonly kept as a pet. For the cat family, see Felidae. For other uses, see Cat (disambiguation) and Cats (disambiguation).
For technical reasons, "Cat #1" redirects here. For the album, see Cat 1 (album).

The **cat** (*Felis catus*) is a domestic species of small carnivorous mammal.[1][2] It is the only domesticated species in the family Felidae and is often referred to as the **domestic cat** to distinguish it from the wild members of the family.[4] A cat can either be a **house cat**, a **farm cat** or a **feral cat**; the latter ranges freely and avoids human contact.[5] Domestic cats are valued by humans for companionship and their ability to hunt pests such as rodents. About 60 cat breeds are recognized by various cat registries.[6]

The cat is similar in anatomy to the other felid species: it has a strong flexible body, quick reflexes, sharp teeth and retractable claws adapted to killing small prey. Its night vision and sense of smell are well developed. Cat communication includes vocalizations like meowing, purring, trilling, hissing, growling and grunting as well as cat-specific body language. It is a solitary hunter but a social species. It can hear sounds too faint or too high in frequency for human ears, such as those made by mice and other small mammals. It is a predator that is most active at dawn and dusk.[7] It secretes and perceives

Domestic cat

Various types of the domestic cat

Conservation status

Domesticated

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia

Bobby is only super obedient like this when there's food on offer. cocker spaniel bobby dog k-9 pet canon 5d 70-200mm f2.8 IS	These two vultures were part of a group of 6 sitting on this roof waiting for their turn with a dead squirrel on the road. bird vulture

成对出现的无标注图片-文本数据

文本编码器

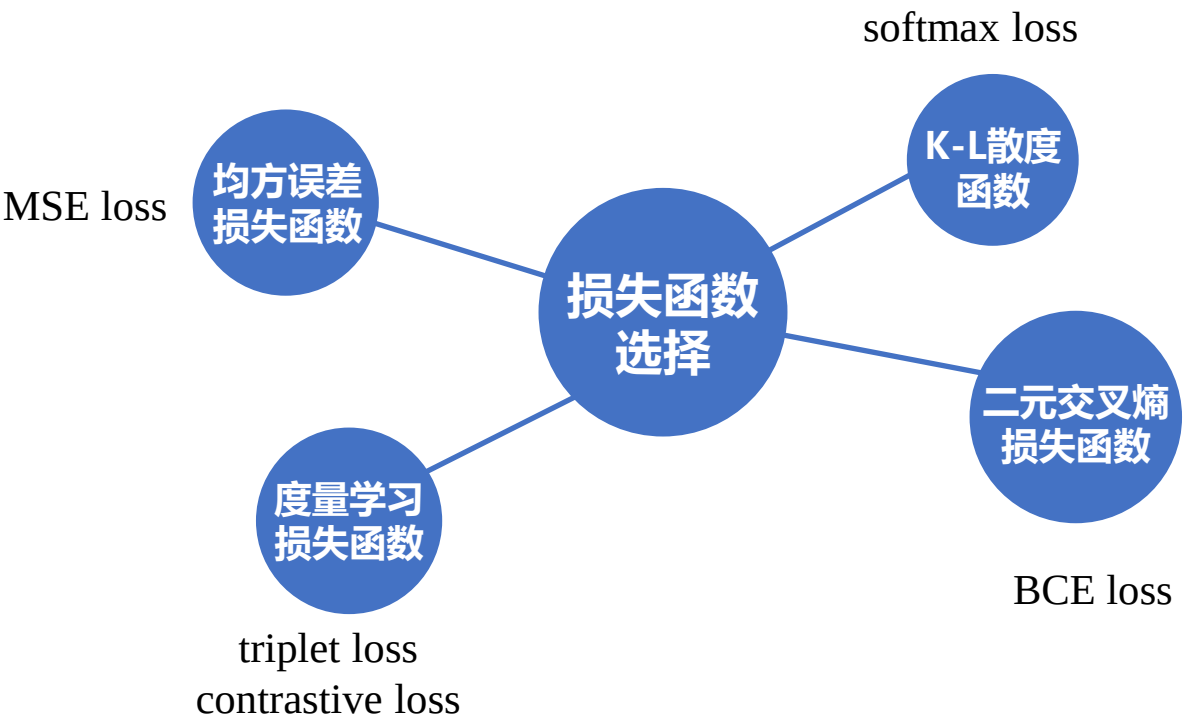
主题生成模型
(LDA)

文档嵌入向量
(Doc2Vec)

单词嵌入向量
(Word2Vec)

大型预训练模型
(BERT)

目标函数


softmax loss
K-L散度
函数
MSE loss
均方误差
损失函数
损失函数
选择
二元交叉熵
损失函数
BCE loss
度量学习
损失函数
triplet loss
contrastive loss

Part 3

实验结果

实验设计

数据集

名称	图片数	文本数	文本平均长度	语言	实际使用率
ImageCLEF	100k	35k	1200+	英文	100%
English-Wikipedia	420M	170M	1200+	英文	30%
Webvision-1.0	240M	240M	20+	多语种	40%
InstaCities1M	100M	100M	20+	多语种	100%

多模态无标注数据集的基本信息

消融实验

文本编码器	数据集	mAP (conv5)
LDA	ImageCLEF	47.1
Word2Vec	ImageCLEF	43.5
Doc2Vec	ImageCLEF	42.6
LDA	InstaCities1M	36.5
Word2Vec	InstaCities1M	40.2
Doc2Vec	InstaCities1M	32.4
BERT	InstaCities1M	35.1
LDA	Wikipedia	**50.8**
Word2Vec	Webvision	41.3

固定网络模型，提取中间层特征(conv5)，
在VOC07 数据集上训练SVM分类器，测试集计算mAP
LDA损失函数为softmax loss, 其余为BCE loss

消融实验

损失函数	特征归一化类型	mAP (conv5)
MSE loss	None	30.1
Softmax loss	Softmax	38.5
BCE loss	sigmoid	40.2
triplet loss	L2-norm	41.4
contrastive loss	L2-norm	**42.7**

InstaCities1M数据集 + Word2vec文本编码器

特征分辨能力

Method	conv1	conv2	conv3	conv4	conv5
ImageNet Supervised	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Jigsaw(2016)	18.2	28.8	34.0	33.9	27.1
Colorization(2016)	12.5	24.5	30.4	31.5	30.3
SplitBrain[76](2017)	17.7	29.3	35.4	35.2	32.8
Rotation[13](2018)	18.8	31.7	**38.7**	38.2	**36.5**
DeepCluster*[3] (2018)	12.9	29.2	38.2	39.8	36.1
Ours(Ins1M+Word2Vec)	16.4	26.5	31.0	29.3	26.3
Ours(ImageCLEF+LDA)	18.1	29.5	35.1	32.0	29.2
Ours(Wikipedia+LDA)	**20.0**	**32.1**	37.3	**38.6**	33.4

固定网络模型，提取中间层特征(conv1～5)，
在ImageNet数据集上训练线性分类器，并测试集计算准确率

特征泛化能力

Method	Classification (%mAP)	Detection (%mAP)	Segmentation (%mIoU)
ImageNet Supervised	79.9	59.1	48.0
Random	53.3	43.4	19.8
Jigsaw(2016)	67.6	53.2	37.6
Colorization(2016)	65.9	46.9	35.6
SplitBrain(2017)	67.1	46.7	36.1
Rotation(2018)	**73.0**	54.4	**39.1**
DeepCluster*(2018)	74.7	55.4	45.1
Ours(ImageCLEF+LDA)	66.9	50.3	36.7
Ours(Wikipedia+LDA)	70.3	**54.6**	38.5

在PASCAL VOC 数据集上进行迁移学习的结果
(图像分类、目标检测、语义分割任务)

跨模态检索应用

Method	Image Query	Text Query	Average
CCA[24](2010)	19.7	17.8	18.8
PLS[55](2006)	30.6	28.0	29.3
JFSSL*[65](2016)	**42.8**	39.6	**41.2**
CCA-3V*[18](2014)	40.5	36.5	38.5
LCFS*[66](2013)	41.9	38.5	39.9
Ours(Ins1M+Word2Vec)	32.2	30.6	31.4
Ours(ImageCLEF+LDA)	36.8	36.2	36.5
Ours(Wikipedia+LDA)	37.4	**39.8**	38.6

在Multimodal-Wikipedia 数据集上跨模态检索mAP，
带* 表示有监督方法(使用了类别信息)

Part 4

总结

研究总结

总结观点

- 提出了一种基于多模态场景的自监督学习算法。
- 通过文本与图像之间的内容相关性，用文本的语义信息作为监督信号，指导卷积神经网络训练。
- 尝试了不同类型的数据集、文本编码器、损失函数，找到了效果最好的组合。
- 在下游任务上验证了自监督算法学习的特征的分辨能力与泛化性能，且可用于跨模态检索。

附录：参考文献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

[2] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152, 1992.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. Lecture Notes in Computer Science, page 139–156, 2018.

[4] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In Proceedings of the IEEE International Conference on Computer Vision, pages 2959–2968, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Selfsupervised gans via auxiliary rotation loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12154–12163, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pages 1422–1430, 2015.

[10] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2051–2060, 2017.

附录：参考文献

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010.

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems, pages 2121–2129, 2013.

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations, 2018.

[14] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

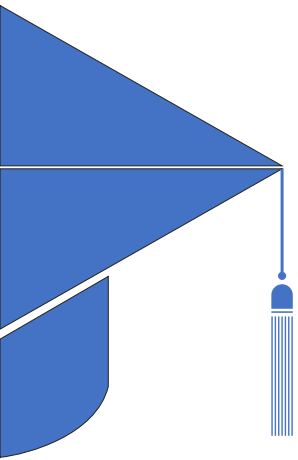
[15] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4230–4239, 2017.

[16] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. Selfsupervised learning from web data for multimodal retrieval, 2019.

[17] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. International journal of computer vision, 106(2):210–233, 2014.

[18] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In European conference on computer vision, pages 529–545. Springer, 2014.

[19] Albert Gordo, Jon Almazán, Naila Murray, and Florent Perronin. Lewis: latent embeddings for word images and their semantics. In Proceedings of the IEEE International Conference on Computer Vision, pages 1242–1250, 2015.



中国科学技术大学2016级

信息科学与技术学院 4班

演示完毕 请多指点

答辩人：付哲仁　　导师：毛震东 特任研究员