智源学者成果展示——自然语言处理

作者　何晓冬（京东人工智能研究院）

2020年6月

项目负责人：何晓冬

单位 / 团队：京东人工智能研究院

成果名称：

智能疫情助理（Intelligent Assistant for COVID-19 Epidemic）

成果简介：

为及时助力抗疫，产品基于京东人工智能研究院与人机交互部自研的通用智能对话平台，快速完成搭建。利用“通用深度语义模型”、“超长上下文理解模型”、”智能数据蒸馏”等一系列具有自主知识产权的前沿技术，深度理解用户的问题和潜在意图；结合权威媒体 / 机构发布的最新疫情信息，挖掘生成答案，确保应答质量；基于业内领先的 ASR、TTS 技术，提供多种高拟人音色，与用户进行拟人化的多轮语音交互。上线后，基于日志与用户反馈，机器自主学习持续优化，并结合实时热点挖掘机制，快速联动运营，利用语音语义调优工具快速对热点进行干预，从而实现机器与人工双向结合调优应答，持续保障应答效果。规模商用的情感智能机器人，具有良好的情绪感知能力。整体准确率可达 95% 以上，对于用户高频问题基本可以完全准确识别，可以保证为用户提供比较良好的咨询体验。该项目提供技术支撑的智联云自研通用智能对话平台，在全链路集成、智能算法、行业认知、安全、操作体验等方面，具备显著的先进性与创新性。项目上线以来，已有武汉市长专线、华西医院、蚌埠市人民政府官网、中信银行、南京银行、壳牌中国、北京人寿、首旅集团、中信集团、中国联通、联想等在内的 1000+ 政府 / 企业接入，覆盖政府、医疗、金融等多个行业，为上千万用户提供抗疫帮助，助力抗疫的表现得到了人民网、新浪科技等知名媒体的密切关注和报道。

报道链接：http://finance.ifeng.com/c/7tofl8Oyski。

http://www.chinaweekly.cn/43644.html。

http://epaper.cqna.com.cn/web/na/2020-02/17/content_17725.html。

项目负责人：何晓冬

单位 / 团队：京东人工智能研究院

成果名称：

The BAAI-JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service

成果简介：

我们基于京东的客服真实脱敏对话日志，构建了大规模多轮中文对话数据集 -JDDC 数据集。该数据集包含了 100 万对话日志，2,000 万条句子，和 1.5 亿词；覆盖了任

务型对话、知识问答和开放领域闲聊等多种对话类型。此外，我们还提供了句子级别的意图信息和 3 个人工标注的评测集。JDDC 数据集为学术界和工业界进行复杂场景下多轮对话技术研究提供了良好的基础，并且成功支持了 2018、2019 年京东对话大赛，累计参赛队伍超过 600 支，在业界积累了较高的知名度。

论文和数据集下载地址：https://www.aclweb.org/anthology/2020.lrec-1.58.pdf

对话大赛的网址：https://jddc.jd.com/

项目负责人：何晓冬

单位 / 团队：京东人工智能研究院

成果名称：

IEEE JSTSP Special Issue：Deep Learning for Multimodal Intelligence across Speech, Language, Vision, and Heterogeneous Signals, April 2020

成果简介：

在信号处理和模式识别领域内的顶级刊物 IEEE JSTSP（IF 6.8）编辑 April 2020 特刊：跨语音、语言、视觉和异构信号的多模态智能深度学习 (Deep Learning for Multimodal Intelligence across Speech, Language, Vision, and Heterogeneous Signals)，在多模态人工智能这个前沿领域从五十多篇投稿中通过严格审稿收录了 10 篇高质量论文，涉及跨文本、图像、视频、语音等多模态的各种多元互补的机器学习算法及重要应用。论文作者包括高文院士、邓力院士等本领域国内外著名学者，作者机构包括北京大学、中国科学院、卡耐基梅隆大学、约翰霍普金斯大学、剑桥大学、伦敦帝国学院、哥本哈根大学、新加坡国立大学等世界著名学府及京东、腾讯、字节跳动、CitaDel 等中美企业界实验室。

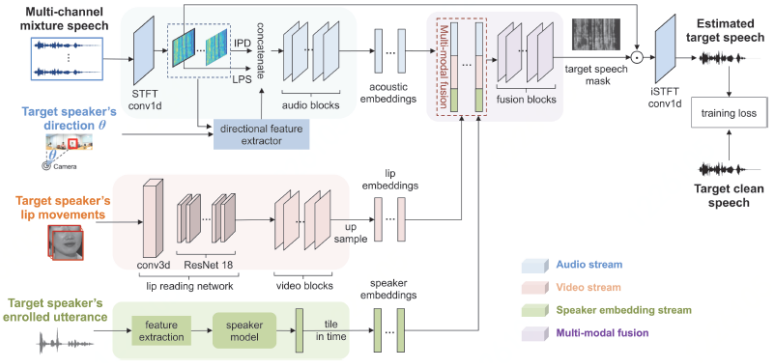
IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society

https://signalprocessingsociety.org/

APRIL 2020 VOLUME 14 NUMBER 4 IJSTGY (ISSN 1932-4553)

ISSUE ON DEEP LEARNING FOR MULTI-MODAL INTELLIGENCE



The diagram of proposed multi-modal target speech separation framework. For more, see "Multi-Modal Multi-Channel Target Speech Separation," by Gu *et al.*, p. xxxx.

1. J-STSP-DLMMI-00258-2019
Title: Where is the Model Looking At? --Concentrate and Explain the Network Attention.
Authors: Wenjia Xu, Jiuniu Wang, Yang Wang, Guangluan Xu, Wei Dai, Yirong Wu
Institutions: Institute of Electronics, Chinese Academy of Sciences

2. J-STSP-DLMMI-00263-2019
Title: Audio-Visual Speech Separation and Dereverberation with a Two-Stage Multimodal Network
Authors: Ke Tan, Yong Xu, Shi-Xiong Zhang, Meng Yu, Dong Yu
Institutions: Ohio State University, Tencent AI

3. J-STSP-DLMMI-00265-2019
Title: An Efficient Threshold-Driven Aggregate-Label Learning Algorithm for Multimodal Information Processing
Authors: Malu Zhang, Xiaoling Luo, Jibin Wu, Yi Chen, Zihan Pan, Hong Qu, Haizhou Li
Institutions: University of Electronic Science and Technology of China, National University of Singapore

4. J-STSP-DLMMI-00267-2019
Title: Speech-to-Image Translation without Text
Authors: Jiguo Li, X. Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, Wen Gao
Institutions: Institute of Computing Technology Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Peking University, Bytedance Inc.

5. J-STSP-DLMMI-00271-2019
Title: Learning to Recognize Visual Concepts for Visual Question Answering with Structural Label Space
Authors: Difei Gao, Ruiping Wang, Shiguang Shan, Xilin Chen
Institutions: Institute of Computing Technology Chinese Academy of Sciences

6. J-STSP-DLMMI-00272-2019
Title: Grounded Sequence-to-Sequence Transduction
Authors: Lucia Specia, Raman Arora, Loic Barrault, Ozan Caglayan, Amanda Duarte, Desmond Elliott, Spandana Gella, Nils Holzenberger, Chiraag Lala, Sun Jae Lee, Jindrich Libovicky, Pranava Madhyastha, Florian Metze, Karl Mulligan, A. Ostapenka, S. Palaskar, R. Sanabria, Josiah Wang
Institutions: Imperial College London, Johns Hopkins University, Universitat Politecnica de Catalunya, University of Copenhagen, University of Edinburgh, University of Sheffield, University of Pennsylvania, Charles University, Carnegie Mellon University, Worcester Polytechnic Institute

7. J-STSP-DLMMI-00276-2019
Title: Perfect Match: Self-Supervised Embeddings for Cross-modal Retrieval
Authors: Soo-Whan Chung, Joon Son Chung, Hong Goo Kang
Institutions: Yonsei University, Naver Corporation

8. J-STSP-DLMMI-00282-2019
Title:A Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions
Authors: Lingyue Zhang, Richard J. Radke
Institutions: Rensselaer Polytechnic Institute

9. J-STSP-DLMMI-00284-2019
Title: Multi-modal Multi-channel Target Speech Separation
Authors: Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, Dong Yu
Institutions: Peking University Shenzhen Graduate School, Tencent AI Lab

10. J-STSP-DLMMI-00301-2019
Title: Multimodal Intelligence: Representation Learning, Information Fusion, and Applications
Authors: Chao Zhang, Zichao Yang, Xiaodong He, Li Deng
Institutions: JD AI Research, University of Cambridge, Citadel LLC.

项目负责人：何晓冬

单位 / 团队：京东人工智能研究院

成果名称：

多模态智能：表征学习、信息融合、典型应用（Multimodal Intelligence：Representation Learning, Information Fusion, and Applications）

成果简介：

虽然深度学习推动了语音、语言处理和计算机视觉等单一模态领域的巨大进步，但更多的人工智能应用场景其实同时涉及到多种模态的输入特征。本文主要关注于结合文本和图像的多模态任务，尤其是近年来一些侧重数学模型和训练方法的相关研究工作。论文主要选取了表征学习、信息融合和具体应用三个角度来分析多模态视觉与语言信息处理领域的核心问题和应用场景，具体来说：（1）学习输入特征的更好的表征是深度学习的核心内容。（2）对不同模态表征的融合也是任何多模态任务的关键内容。（3）在具体应用方面，论文主要综述了三种不同任务，包括：图像字幕生成、基于文字的图像生成，以及 VQA。虽然多模态智能研究已经取得了重大进展，并成为了人工智能发展的一个重要分支，但如果以构建能够感知多模态信息并利用不同模态之间的联系来提高其认知能力的智能体为最终目标，关于多模态智能的研究仍处于起步阶段，其中既面临着巨大的挑战，也存在着巨大的机遇。

（本文获 BAAI 资助，发表于 IEEE JSTSP April 2020。）

论文链接：https://arxiv.org/abs/1911.03977

Multimodal Intelligence: Representation Learning, Information Fusion, and Applications

Chao Zhang, Zichao Yang, Xiaodong He, *Fellow, IEEE*, and Li Deng, *Fellow, IEEE*

***Abstract*—Deep learning methods haverevolutionized speech recognition, image recognition, and natural language processing since 2010. Each of these tasks involves a single modality in their input signals. However, many applications in the artificial intelligence field involve multiple modalities. Therefore, it is of broad interest to study the more difficult and complex problem of modeling and learning across multiple modalities. In this paper, we provide a technical review of available models and learning methods for multimodal intelligence. The main focus of this review is the combination of vision and natural language modalities, which has become an important topic in both the computer vision and natural language processing research communities. This review provides a comprehensive analysis of recent works on multimodal deep learning from three perspectives: learning multimodal representations, fusing multimodal signals at various levels, and multimodal applications. Regarding multimodal representation learning, we review the key concepts of embedding, which unify multimodal signals into a single vector space and thereby enable cross-modality signal processing. We also review the properties of many types of embeddings that are constructed and learned for general downstream tasks. Regarding multimodal fusion, this review focuses on special architectures for the integration of representations of unimodal signals for a particular task. Regarding applications, selected areas of a broad interest in the current literature are covered, including image-to-text caption generation, text-to-image generation, and visual question answering. We believe that this review will facilitate future studies in the emerging field of multimodal intelligence for related communities.**

***Index Terms*—Multimodality, representation, multimodal fusion, deep learning, embedding, speech, vision, natural language, caption generation, text-to-image generation, visual question answering, visual reasoning.**

I. Introduction

SIGNIFICANT progress has been made in the field of machine learning in recent years based on the rapid development of deep learning algorithms [1]–[6]. The first major milestone was a significant increase in the accuracy of large-scale automatic speech recognition based. on the use of fully connected deep neural networks (DNNs) and deep autoencoders around 2010 [7]–[17]. Shortly thereafter, a series of breakthroughs was achieved in computer vision (CV) using deep convolutional neural network (CNN) models [18] for large-scale image classification around 2012 [19]–[22] and large-scale object detection around 2014 [23]–[25]. All of these milestones have been achieved for pattern recognition with a single input modality. In natural language processing (NLP), recurrent neural network (RNN) based semantic slot filling methods [26] have achieved state-of-the-art for spoken language understanding. RNN-encoder-decoder models with attention mechanisms [27], which are also referred to as sequence-to-sequence models [28], have achieved superior performance for machine translation in an end-to-end fashion [29], [30]. For additional NLP tasks with small amounts of training data, such as question answering (QA) and machine reading comprehension, generative pre-training has achieved state-of-the-art results [31]–[33]. This method transfers parameters from a language model (LM) pre-trained on a large out-of-domain dataset using unsupervised training or self-training, which is followed by fine-tuning on small in-domain datasets.

Although there have been significant advances in vision, speech, and language processing, many problems in the artificial intelligence field involve more than one input modality, such as intelligent personal assistant systems that must understand human communication based on spoken words, body language, and pictorial languages [34]. Therefore, it is of broad interest to study modeling and training approaches across multiple modalities [35]. Based on advances in image processing and language understanding [36], tasks combining images and text have attracted significant attention, including visual-based referred expression understanding and phrase localization [37]–[39], as well as image and video captioning [40]–[45], visual QA (VQA) [46]–[48], text-to-image generation [49]–[51], and visual-and-language navigation [52]. In these tasks, natural language plays a key role in helping machines in "understanding" the content of images, where "understanding" means capturing the underlying correlations between the semantics embedded in languages and the visual features obtained from images. In addition to text, vision can also be combined with speech to perform audio-visual speech recognition [53]–[55], speaker recognition [56]–[58], speaker diarization, [59], [60], as well as speech separation [61], [62] and enhancement [63].

This paper provides a technical review of the models and training methods used for multimodal intelligence. Our main focus

Manuscript received November 10, 2019; revised March 23, 2020 and April 3, 2020; accepted April 7, 2020. This work was supported by the Beijing Academy of Artificial Intelligence. The guest editor coordinating the review of this paper and approving it for publication was Dr. Isabel Trancoso. *(Corresponding author: Xiaodong He.)*

Chao Zhang was with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K., and also with the JD AI Research, JD.com Inc, Beijing 100101, China (e-mail: cz277@cam.ac.uk).

Zichao Yang was with the Citadel LLC, Chicago, IL 60603 USA (e-mail: yangtze2301@gmail.com).

Li Deng was with the Citadel America, Seattle, WA, 98121 USA (e-mail: deng629@gmail.com).

Xiaodong He was with the JD AI Research, JD.com Inc, Beijing 100101, China (e-mail: xiaodong.he@jd.com).

Digital Object Identifier 10.1109/JSTSP.2020.2987728

项目负责人：何晓冬

单位 / 团队：京东人工智能研究院

成果名称：

基于异构图推理的跨文档多跳阅读理解（Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs）

成果简介：

相比单文档机器阅读理解，多跳机器阅读理解需具备跨文档推理能力才能更好的回答问题。对此，我们提出了一种基于 Heterogeneous Document-Entity (HDE) 异构文档 – 实体图和 GNN（Graph Neural Networks）的推理模型。具体来说、HDE 图由候选答案 (Candidate)、文档 (Support Document)、实体 (Entity) 三种类型的节点组成，包含多颗粒度信息，并使用 Co-attention 和 Self-attention 的上下文编码方法初始化 HDE 图表示，最后通过基于 GNN 的消息传递算法收集证据实现推理。我们的模型在 WikiHop 盲测试集上达到了单模型精度 70.9%，Ensemble 模型精度 74.3% 的效果，并获得 WikiHop 2019 多文档机器阅读和问答大赛的第一名。

论文链接：https://www.aclweb.org/anthology/P19-1260.pdf

#	Model / Reference	Affiliation	Date	Accuracy[%]
1	JDReader (ensemble)	JD AI Research	March 2019	74.3
2	DynSAN (ensemble)	Samsung Research (SRC-B)	March 2019	73.8
3	DynSAN basic (single)	Samsung Research (SRC-B)	February 2019	71.4
4	Entity-GCN v2 (ensemble)	University of Amsterdam && University of Edinburgh	November 2018	71.2
5	HDEGraph	JD AI Research	February 2019	70.9
6	CFC	Salesforce Research	September 2018	70.6
7	[anonymized]	[anonymized]	November 2018	69.6
8	[anonymized]	[anonymized]	February 2019	69.1
9	BAG	University of Sydney	March 2019	69.0
10	[anonymized]	[anonymized]	September 2018	67.6

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院