# 会议纪要

**会议主题：** 生成式AI与大模型的前沿与思考——2024年中国生成式AI大会内容分享

**会议时间：** 2024年 4月29日19：00至21：00

**会议地点：** 西安电子科技大学北校区会议中心303-1会议室

**线下人员：**承楠孙瑞锦 尹志胜 贺靖超 王葳 沈京龙 王秀程 杨浩 张茹倩 马龙飞 周新阳 权赟昊 张玉洁 黄蕾 杨杰 胡陆莹 齐阁 傅连浩 侯毓真 邓川 李青壮 朱煜朋 祝馨平 郑佩林 孙兴栋 方忠盛

**线上人员：** 温瑶 王兆薇 邱子仪 韩松明 尚佳瑶 贾宏刚 郑雯馨 刘苏 解思舀 万佳林 陈梦豪 王天宇 刘永红 李昊坤 尉家豪

**记 录 人：万佳林**

**会议内容：**1、沈京龙介绍了具生智能的感知和规划以及相关研究进展 2、郑培林介绍AI Agent智能体的概念构成和多智能体Agent的应用以及agent 在通信中应用思考 3、傅连浩介绍了算力网络、大模型发展进程、大模型架构及其应用

* **承楠：**什么是具生智能，举个例子？

**沈京龙：**具生智能就是你给他一个多模态的输入他会直接输出一个动作action ，就比如说最直接的——机器人

* **承楠：**（在介绍Embodied Generalist Agent in 3D World时）2D Encoder是什么？

**沈京龙：**这个就是第一人称的图像，因为他那个机器人上面是会有一些摄像头之类的。

* **承楠**：（在介绍Scene Representation时）他这个模型的架构是类似于Transformer的架构么？

**沈京龙：** 对，他这个整体上就是一个Transformer，老师你可以把这一块就把它看成一个我们transformer的decoder ,跟Tranformer那个decoder结构是一样的。他其实他改动的是什么地方呢，就是在Spatial Self-Attention这一块,在每一层的detention那层。

**承楠：**那么还有一个问题,它那个object就是object tokens,比如说左边那个三object proposals,它这个是什么样的一个数据形式。

**沈京龙：**点云形式，然后通过Point Net++输出向量，是一个token的形式。

**承楠：**那Point Net++作用是啥？

**沈京龙：**它是提特征的，然后你可以把它理解成对图像处理那种。

**孙瑞锦：**他这个是处理序列之间的相关性还是空间的相关性

**沈京龙：**其实我觉得他本质上还是在处理序列之间的相关性,只不过是啥呢，只不过他是在这个序列每个token里面,他去把那个空间信息给加到那个token里面去了。

* **承楠：**具生智能训练数据是哪来的？

**沈京龙：**数据也是他自己采集的，就自己有一个数据集，它里面包含了这些 就是输入输出的内容，就是一个样本。

**承楠：**比如说我现在假设说我这个lama他训练的时候,映射的是输入和输出之间的这个文字,是语言上一种关系对吧，那我现在把它微调一下,然后我这个输入变成空间上的东西。那这个是让输出要变成机械臂的一些控制,我感觉它这个这种映射的关系。和你刚开始的时候,训练的映射的关系它差的很远,你怎么微调才能把它微调回来？

**承楠：**就他如果有这样的能力,那我觉得确实挺厉害。如果这样的化,我们可以用这个大模型做自己事情,我可以做通信。

**沈京龙：**他这个反正他最终的结果，那个结果还是挺好的，会有一个视频可以放一下。

* **承楠：**（放机械臂视频时）这个看上去像是假的？

**沈京龙：**这个是仿真的，他这个实际上这个演示效果是从仿真软件里面去做的。那工业机器人里面这种仿真软件很多。

**承楠**：他这个如果不用大模型做？还是说只有这种才能做，其实我想问的就是他到底带来了什么意义？

**沈京龙：**就是不用他的话,就是很难做。就是像如果你直接去用,比如说就用GTP4去怎么去做。你只能输入一个图片,输入图片是他其实是没有这个三维场景的信息的。你就很难去告诉他这个物体或者他也是很难去识别出来。

**承楠：** 那是因为他处理不了点云数据。我假如说我用点云数据拿来,然后用一些就比如用你那个什么point net++,然后用一些不是用那么大的模型,用一些比如说cv的东西把它处理一下,不能实现这个？

**沈京龙：**是这样的，就是如果只用point net++的话，它是只包含这一个物的语义信息的，它没有空信信息。

**承楠：**那你要微调的时候,就每个任务你都要收集一些数据,然后来训练,放到这个大模型里面,啥都可以去做了。相当于把llm这样的一种就是在自然语言处理的能力泛化到工业，但是实际上他的核心能力还是为他这个llm能处理很多, 那我还是不知道这个里面是怎么实现它自然语言的通用,然后它的任务是通用的？

**付连浩：**首先用这个大模型的话,那肯定能投一些数据进去。比如说就是假如说我要一个端到端的东西出来的话,我肯定需要大量的数据,然后他这部分他本身就有这个能力了。

**承楠：**话也不能说太满，现在这个lim到底有啥能力，我对这个还是要打一个问号，另外就是，他这个微调是加一个小网络么？

**沈京龙：**不是,比如说像是那个cell atention里面它有很多的全连接的网络,它是一个个矩阵。然后他是把这些矩阵去做了一个低质分解,就可以想象成是把一个比如说一个正方形,它做低质分解就变成两个类似于一个哑一形的,其实就是降低了参数量。

**郑佩林：**像那个控制机械臂的话,其实你可以把那个llm想象成是一个就是什么都懂一点的人。就是你要怎么让他就是学会去控制那个机械臂,你就可以在那个llm那个输入的前端,你可以加一个那种专用的encode。然后把一些控制机械臂相关的命令，或者是一些模态之类的,相关的那个数据,送到那个encoder里面,然后，在那个大语言模型的前面,你可以加一些引导,比如说我现在给你这一个桌面上的一些东西的信息,然后请你告诉我要怎么控制机械力。使得我可以把比如说从把这个球从这个点推到那边一个点,然后同时训练数据里,你那个输入就是你输入的那个图片信息,然后输出就是那一个控制机械臂的参数。然后通过这个方法,来对整大语言模型微调。

或者比如说一个普通人,他不懂得怎么打字,那样就给他灌输一些打字相关的知识,然后让他一直去练。

**承楠：**GTP或者说LAMA,已经到了这样的一种程度了么？就是说我可以把它形容成一个未被某些专业技能训练过的人,但是已经达到了普通人的这种知识水平,或者说是有认知能力的这样的一种状态？

**郑佩林：**或者说他至少对语言这方面的信息理解的是很透彻的。

**王秀程：**但是，我就觉得llm会有一个问题,比如说你教它做这个小数乘法,除法,他会很固执。就我曾经试过百以内乘除法,就死活教不好。

**郑佩林：**呃,这个其实就是那个大语言中的那个幻觉问题,他会很固执。在那胡说八道。所以我认为一般的话有解决方案,就是说你不直接教他怎么算,你去教他用计算器。

**承楠：**对，不会直接算，但是如果有什么方式来对齐他能去算，但他的能力是不会算的。

**王秀程：**就它得驱动一个工具，就是你直接用它不行。

**承楠：**那他（机械臂）这块是驱动什么呢？

**沈京龙：**其实这一块的话,这块直接输出一个数值的结控制结果方面，可能我其实我也觉得就是他每一次输出的结果可能不会特别的精确。因为他是本身是一个控制的问题,其实你在运动的过程中可能他也不需要有特别精确。

**王秀程：**他这个示例选的比较好。因为导航的话,他就只需要走输出一个前后左右的轴距。如果说是机械臂的话,就大概的跟他说往前往右往上往下。就不需要那么的精确。

**承楠：**他这个控制没有那么准确，所以在通信里边你只能做一些这个语义的，或者做一些vui的东西，不需要那么准确的东西。

**王秀程：**觉得是我觉得语义的东西其实更需要解释。我就一直在想,你看nlp为什么能够成功，是因为它的word embedding做的很好,它能够把具有相似语义的word embedding到一个相近的向量空间当中。为什么图像到现在分割,就图像到现在没有出现一个很好的东西。,我觉得他这种分割现在做的难度很大。就是你想segmentation做到眼睛和嘴的分割还是没有那么的好的。就是而且尤其是如果你只是一个人脸,你不具备任何的动作,你光看他的眼神来判断他是开心还是伤心的话,他这个对于提取分割的精度是很高的。

**承楠：**对,他和文字不一样,尤其是对表情理解很不客观,你知道吧，有的时候那种就是虽然在笑里,但他其实很悲伤，以用笑容来掩饰悲伤。

**王秀程：**我觉得现在其实语义通信他们做这种图片的不太合理。我觉得应该做的是文字的,因为文字可以把它映附到一个相近隐域空间。你两个图片,你怎么表示隐域空间到底近还是不近。

* **承楠：**（在讲open six door时）这个VLLM是？

**沈京龙：**视觉语言模型。

* **承楠：**（在谈到抓娃娃的仿真物理引擎时，随机播撒娃娃，然后问GTP4,看哪一个娃娃朝前，最终不断的选出一个用户满意的位姿，然后这个位姿通过物理引擎是可以提前得到的，所以机械臂可以根据这个位姿的旋转矩阵做控制）这物引擎它只能针对这个娃娃还是可以针对其他的？

**沈京龙：**可以，这块他把图像提取出去做了一个类似于三个重建的那种感觉,去把它三个模型面做出来 ,就是说换其他的,比如说换那个碗跟其他的一些。都是可以的。

**承楠：**但是问题是这个东西的材质和重量等,它表现的出来么？否则你物理它需要的,其实不单单是一个外形。

**沈京龙：**其实他这个工作只考虑了这个。

**承楠：**就是说现在的网络,你说它是一种具生的智能,其实也没太大的问题。他现在也要感知，同感一体。=他自身可以做决策。比如分类资源,它也可以控制，比如说你控制它这个下游的一些这个自动驾驶,或者说是无人机调度。

**承楠：**我其实还有问题,具生智能这个词，国外认可还是说只有国内人用这个词儿,就比如说空间是一体化,国外就没有这个词儿,国外就没有这个概念。

**沈京龙：**应该有吧。

**承楠：**当然了,咱们抛开他这个现象,看本质的话,实际上它是一个Agent,就是只是我们把它定义成了一个叫具生智能的词，际上它这样的一个范式在做这样的一个任务。你叫不叫他这个他也实际存在。

**承楠：**那比如说我们放到网络里面很多任务他存在显著解法，那我们是造一个新任务呢还是？因为你要用这个东西解决问题，可能大家会问你这个东西到底怎么做?那这是我的一些思考,好。

* **承楠：**（谈到多智能体交流时）大模型驱动的智能体间的信息交流有什么特点？交流的是什么东西？以何种形式？何时交流？对于这种交流应该采用什么样的支持？或者说LLM驱动的一个能执行任务的单元，那么他到底通信的是什么?协议又是什么样?这个至少我觉得没有人人真正去做这件事情或者它不存在什么特特别问题,可能就不需要做了，用传统的就行。如果它是存在什么特别的，就必须要解决，但是这需要挖掘。
* **承楠：**LAMA有多模态能力么？

**傅连浩：**LAMA3有。

**承楠：**可以把GPT4和GPT4V是两个不同的模型？不是说GPT4V是GPT4加上了一些东西？

**傅连浩：**它极大概率是先训练好GPT4，然后我再把这个预训练好的这个框架给拿过来，然后我再做一些，比如说我再加一个中间加一个分支，然后进去再继续在这个里面做。

**承楠：**那有没有人说闭源的SORA有多少参数？

**傅连浩：**暂时没有看到。

* **王秀程：**（NIVDIA架构）我有个问题,就当时我记得就是李木木的,他那篇很有名的那个文章,他当时就是说做种模型分割,但我记得他说cos是没法做这种分割,只有tf能行,他说这个也能搞cos的吗?

**傅连浩：**这个可以，我不知道你那个是什么版本，但是它现在的话主流的框架都能支持，他有一个很大的优势就是支持很多。

**总结：**

**承楠：**挺好的，我感觉就了这个报告跟咱们那个去年年会还是好很多的，就是从技技术细节角度来说的话。