Emu2：新一代生成式多模态模型

2023年12月21日，智源研究院开源发布新一代多模态基础模型Emu2，通过大规模自回归生成式多模态预训练，显著推动多模态上下文学习能力的突破。Emu2在少样本多模态理解任务上大幅超越Flamingo-80B、IDEFICS-80B等主流多模态预训练大模型，在包括VQAv2、OKVQA、MSVD、MM-Vet、TouchStone在内的多项少样本理解、视觉问答、主体驱动图像生成等任务上取得最优性能。

根据少量演示和简单指令完成听、说、读、写、画等多模态任务是人类的基本能力。对于AI系统而言，如何利用多模态环境下的各种信息、实现少样本多模态理解与生成是有待攻克的「技术高地」。

Emu2是目前最大的开源生成式多模态模型，基于Emu2微调的Emu2-Chat和Emu2-Gen模型分别是目前开源的性能最强的视觉理解模型和能力最广的视觉生成模型。Emu2-Chat可以精准理解图文指令，实现更好的信息感知、意图理解和决策规划。Emu2-Gen可接受图像、文本、位置交错的序列作为输入，实现灵活、可控、高质量的图像和视频生成。

相较2023年7月发布的第一代「多模态to多模态」Emu模型，Emu2使用了更简单的建模框架，训练了从编码器语义空间重建图像的解码器、并把模型规模化到37B参数实现模型能力和通用性上的突破。此外仍延续采用大量图、文、视频的序列，建立基于统一自回归建模的多模态预训练框架，将图像、视频等模态的token序列直接和文本token序列交错在一起输入到模型中训练。

SOTA理解与生成

通过对多模态理解和生成能力的定量评测，Emu2在包括少样本理解、视觉问答、主体驱动图像生成在内的多个任务上取得最优性能。

在少样本评测上，Emu2在各个场景下显著超过Flamingo-80B，例如在16-shot TextVQA上较Flamingo-80B 超过12.7个点。

经过指令微调的Emu2可以对图像和视频输入进行自由问答，以统一模型在VQAv2、OKVQA、MSVD、MM-Vet、TouchStone等十余个图像和视频问答评测集上取得最优性能。

在零样本的DreamBench主体驱动图像生成测试上，较此前方法取得显著提升，例如比Salesforce的BLIP-Diffusion的CLIP-I分数高7.1%, 比微软的Kosmos-G的DINO分数高7.2%。

多模态上下文学习

生成式预训练完成后，Emu2 具备全面且强大的多模态上下文学习能力。基于几个例子，模型可以照猫画虎的完成对应理解和生成任务。例如在上下文中描述图像、在上下文中理解视觉提示（覆盖图像上的红圈）、在上下文中生成类似风格的图像、在上下文中生成对应主体的图像等。

强大的多模态理解

经过对话数据指令微调的Emu2-Chat，可以精准理解图文指令、更好的完成多模态理解任务。例如推理图像中的要素、读指示牌提供引导、按要求提取和估计指定属性、回答简单的专业学科问题等。

基于任意prompt序列的图像生成

经过高质量图像微调的Emu2-Gen，可以接受图像、文本、位置交错的序列作为输入，生成对应的高质量图像，这样的灵活性带来高可控性。