基于语境学习的图像生成

陈文虎

滑铁卢大学-助理教授

Google Deepmind

DataFunSummit # 2023

Google DeepMind

DataFun.

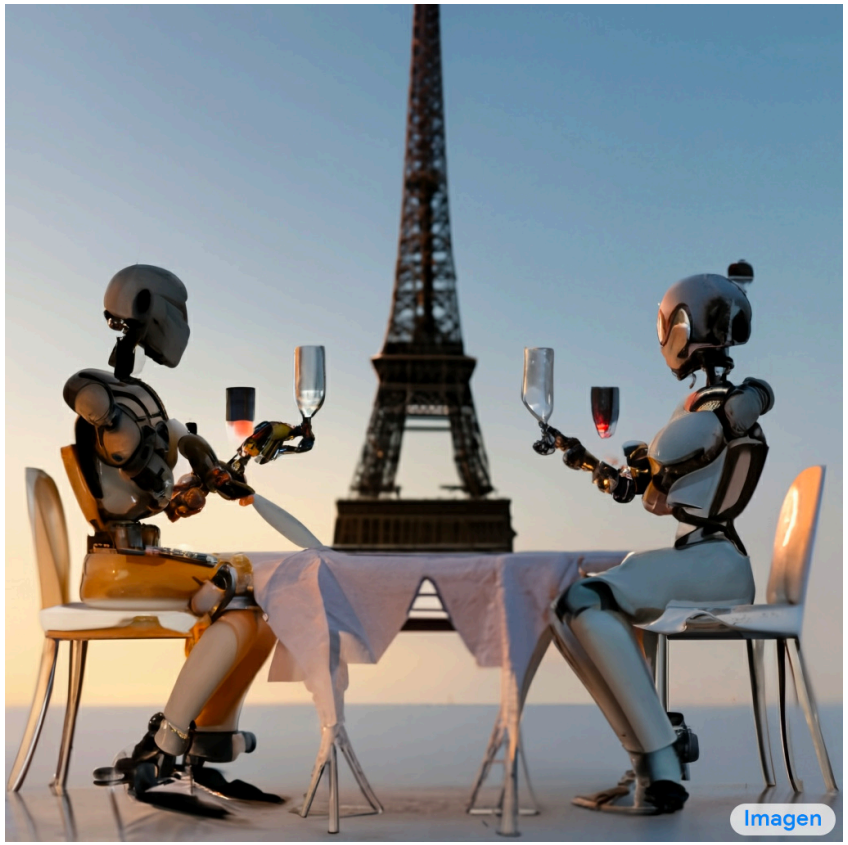
目录 CONTENT

01

背景知识

文本图像生成模型 (Imagen, Dalle2, …)

- 现有的文本图像生成模型已经取得令人骄傲的成绩
  - 生成的图片很符合文本
  - 极具想象能力
  - 图片高清晰度
- 然而，目前的图像生成模型的可控性较差
  - 通过文字无法描述视觉的信息
    - 位置/角度/姿势
  - 如何让生成模型个性化
    - 生成指定的物品
    - 生成指定的场景

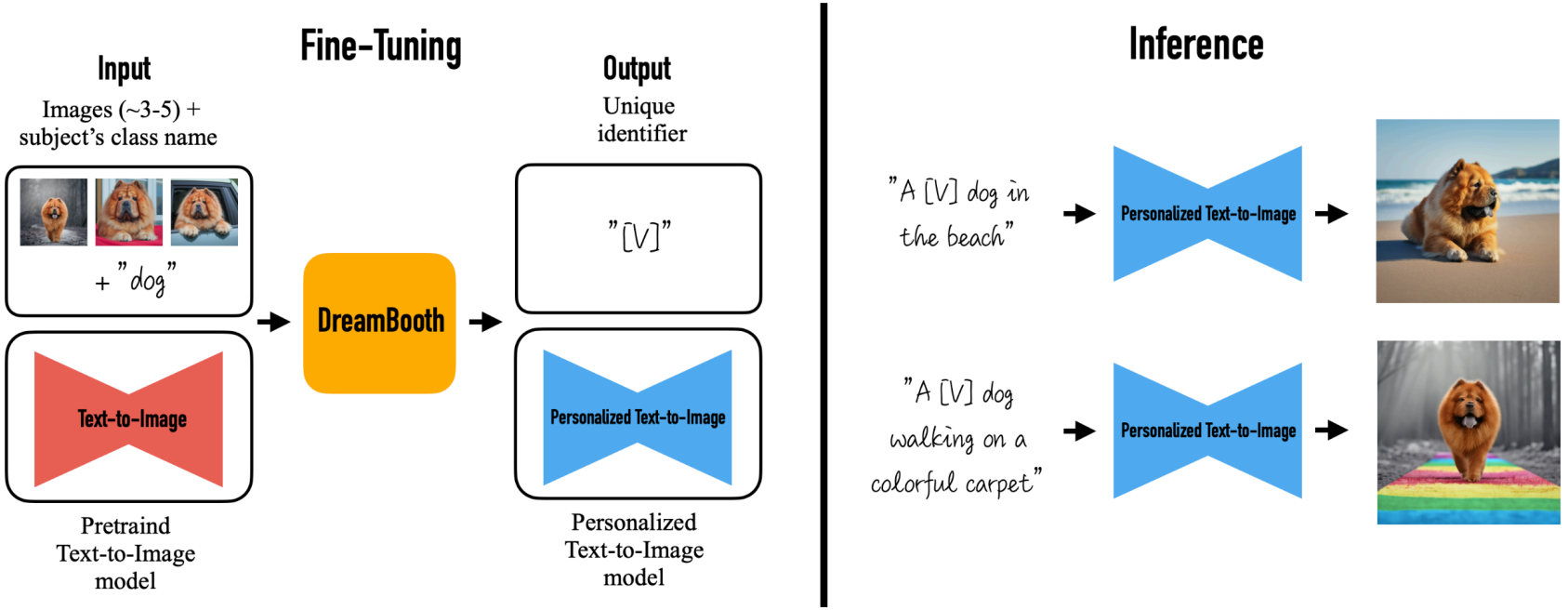


生成模型的个性化

- 如何让生成模型能够生成更加个性化的内容
  - 基于主体的图像生成模型
  - DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

DreamBooth

- 如何让生成模型能够生成更加个性化的内容
  - 先微调整个模型在3-5张关于制定物品的图片1000步，然后把模型存下来。
  - 用微调过的模型进行新的图片生成。

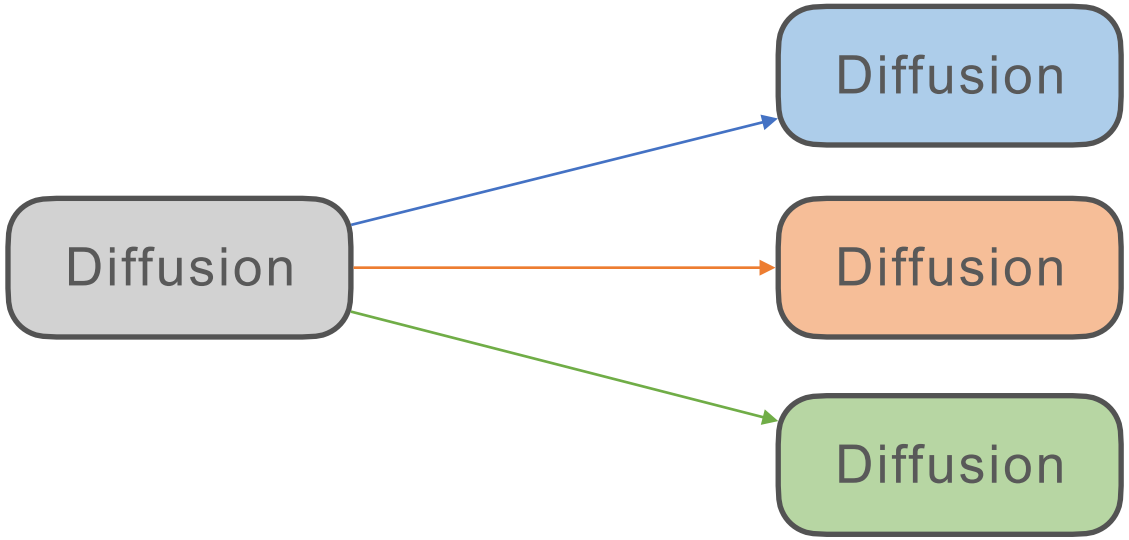


Textual Inversion

- 如何让生成模型能够生成更加个性化的内容
  - 先微调某个embedding在3-5张关于制定物品的图片1000步，把embedding存下来。
  - 用微调过的模型加上新的embedding进行新的图片生成。
  - memory的要求降低很多，但是效果明显差于DreamBooth。

当前模型的问题

- 需要微调整个模型才可以得到比较满意的效果
  - 1000 step的微调非常耗费时间和GPU内存
  - 存储额外的checkpoint需要大量CPU内存
  - 目前的方法扩展性比较弱



论文

Subject-driven Text-to-Image Generation via Apprenticeship Learning
Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, William W. Cohen
Manuscript

Website: https://open-vision-language.github.io/suti/

Re-Imagen: Retrieval-Augmented Text-to-Image Generator
Wenhu Chen,Hexiang Hu, Chitwan Saharia, William W. Cohen
Proceedings of ICLR 2023, Kigali, Rwanda

02

动机

语境学习

- 当前的个性化图像生成通过传统的微调学习
  - 微调学习仍然需要做梯度下降

- 自然语言处理的语境学习
  - 语境学习仅仅需要一些示例样本
  - 无需梯度下降，单个模型可以不断自适应到新的环境

语境学习

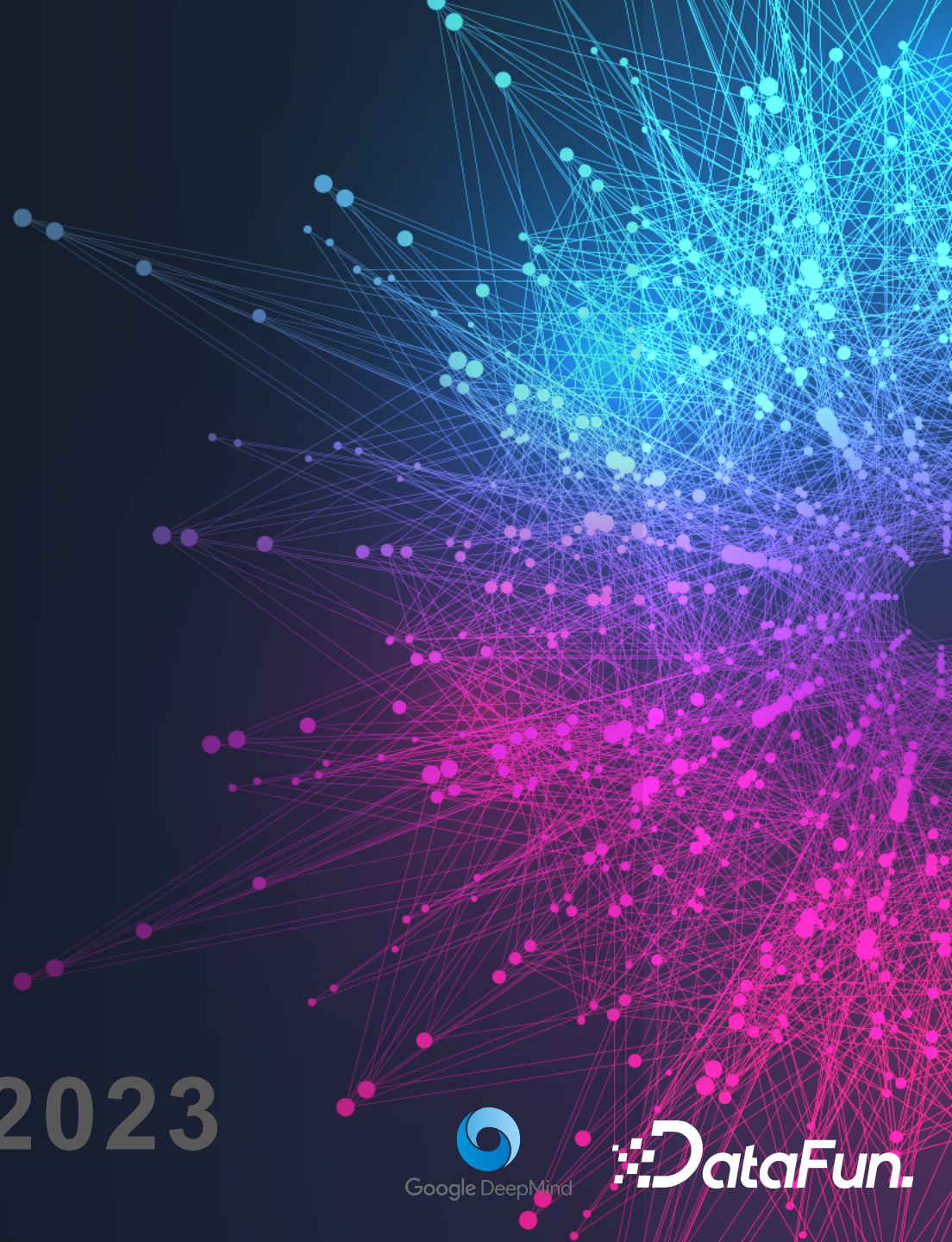
- 自然语言大模型的语境学习来源于预训练
  - 通过大量文本的next work prediction
  - 语言大模型可以自动获取语境学习的能力

- 图像生成模型目前的预训练是仅仅基于单个 (文本，图像) 的pair
  - 图像生成模型并不是连续多个图像文本一起训练
  - 预训练的图像生成模型并不具有任何语境学习的能力

- 因此，我们需要专门适配图像生成模型来获取这种能力
  - 网络架构需要获取示例文本图像信息
  - 训练数据也需要通过多个相似图像文本一起放置

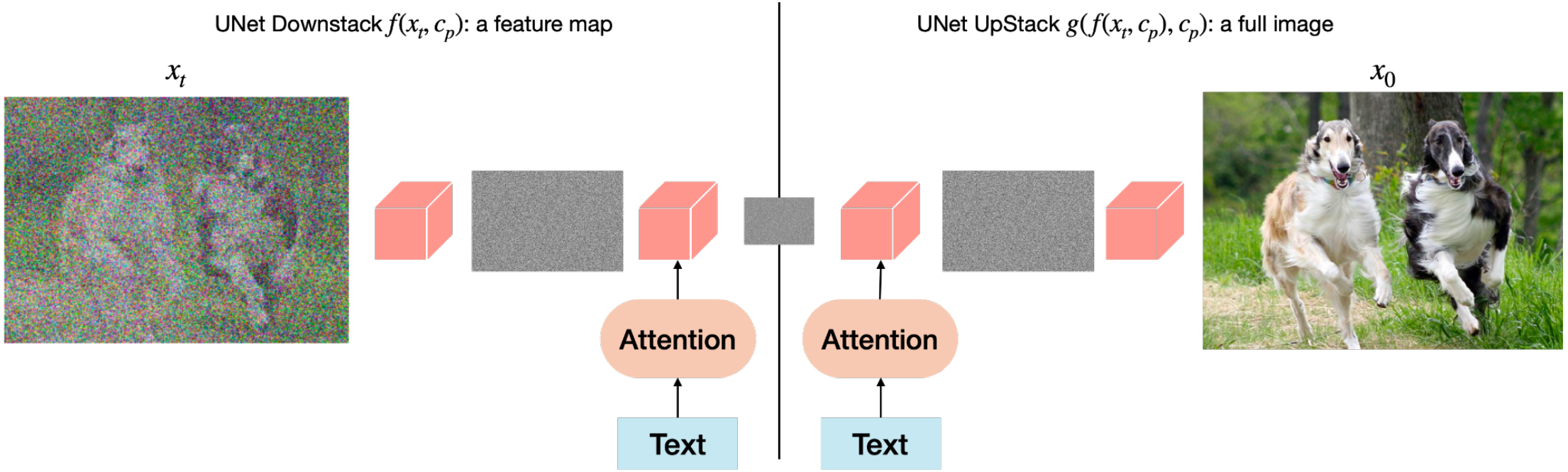
03

设计-网络架构



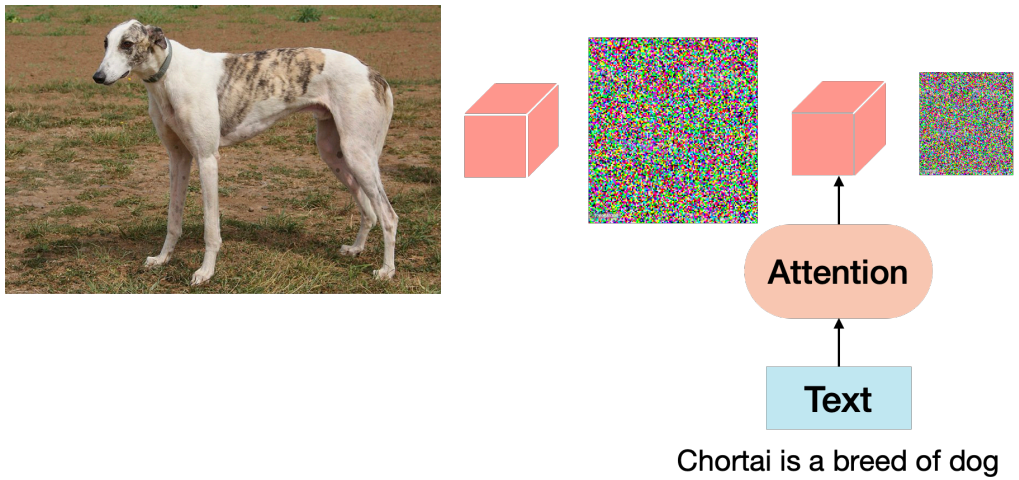
模型架构UNet

- UNet的图像生成架构



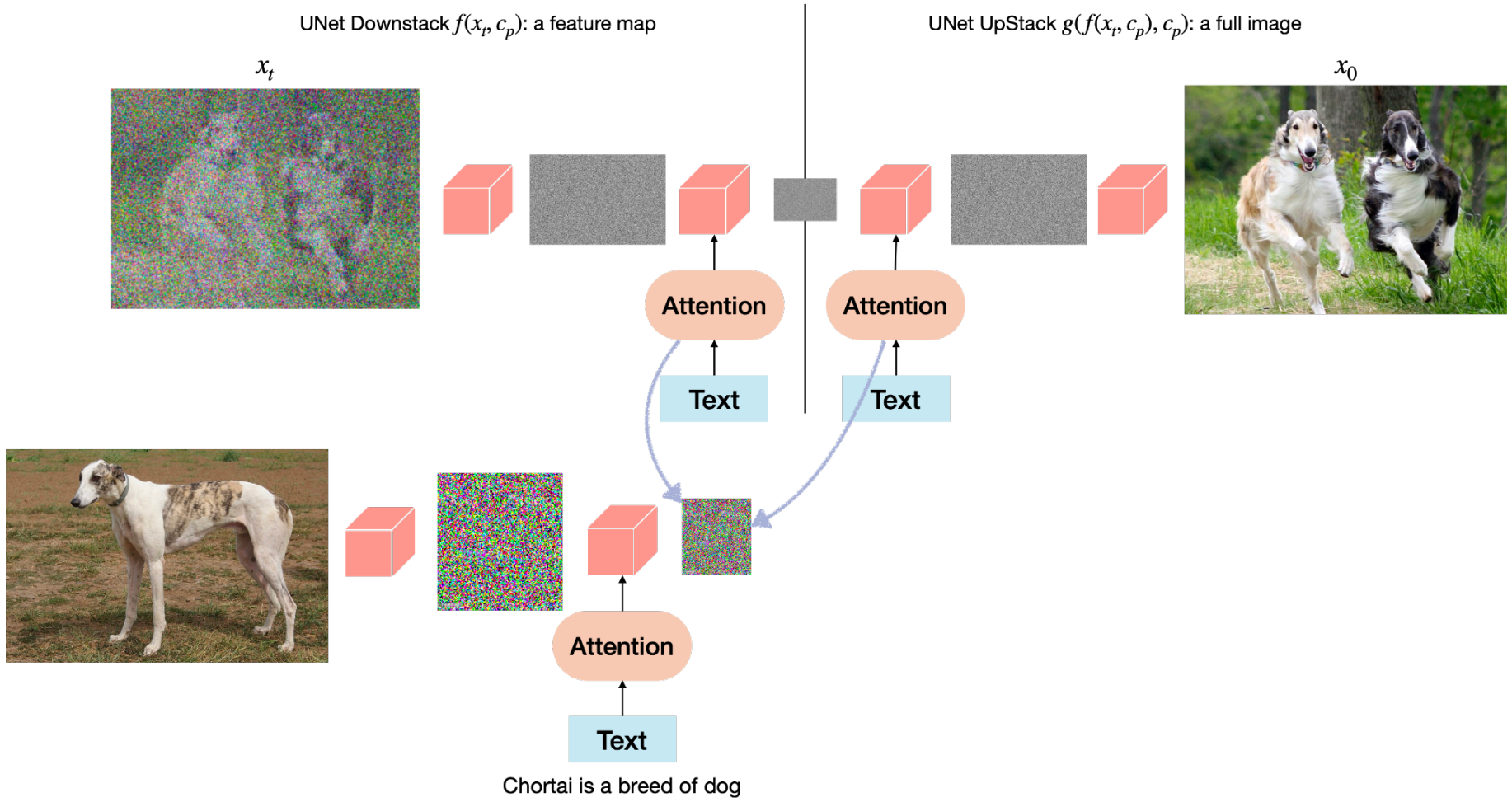
复用模型架构UNet的Encoder

- 如何最高效的把demonstration的信息输入?
  - 复用UNet的Downstack Encoder
  - demonstration的信息被编码到同一个空间



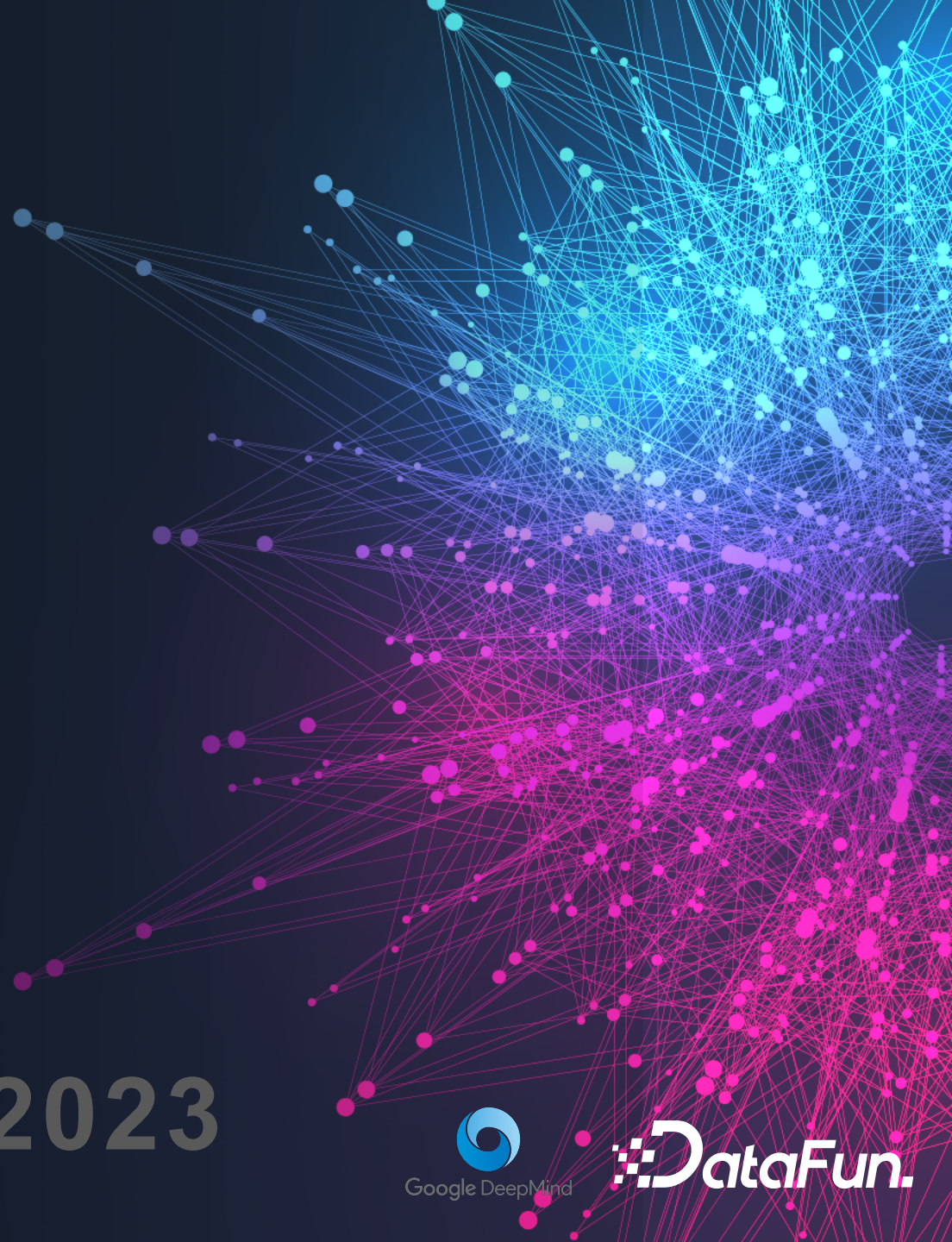
Exemplar: (image, text) pairs

额外的Attention Layer



03

设计-训练数据



语境学习的图文数据

- 比较理想的图文ICL数据应该长的比较像
  - (text_1, Image_1), (text_2, image_2), … (text_t, image_t)
  - 这些图文数据相互比较类似
  - 模型能够通过In-Context的exemplar理解如何生成text_t的图片

- 然而，目前public和internal都不存在这样的数据
  - 现有的图文数据都是从网上爬下来的
  - 都是独立的 (text, image) pair

语境学习的图文数据

- 聚类
  - 我们将这些(image, text)按照URL进行聚类
  - 来自于同一个URL的图片相关性很大
  - 我们通过inter-cluster similarity来过滤比较差的group

- 重新标注text
  - 我们发现这些cluster里面的text噪声比较大，没有反应这个cluster不同image的共性
  - 我们利用Google PaLI模型重新标注caption
  - 然后使用PaLM来找到不同caption直接的相关性进行整合

图文ICL数据集

Example 1

A teapot

A teapot on table

A teapot on table

Example 2

A dog

A cute dog

A dog front view

图文ICL数据集

- 目前我们已经拥有了连续的(image, text) pair的数据
  - 我们可以用其中的k-1作为exemplar，然后用第k个作为target训练模型
  - 然而效果却很差，最终模型变成了直接copy-paste，不管输入的text
  - 原因主要是group起来的cluster的无论是文本还是图像都过于相似，以至于模型陷入了一个local optimal

- 我们需要target和exemplar非常不一样
  - 我们可以用LLM产生很不一样的text
  - 通过DreamBooth作为target图像生成器

图文ICL数据集v2

A teapot

A teapot on table

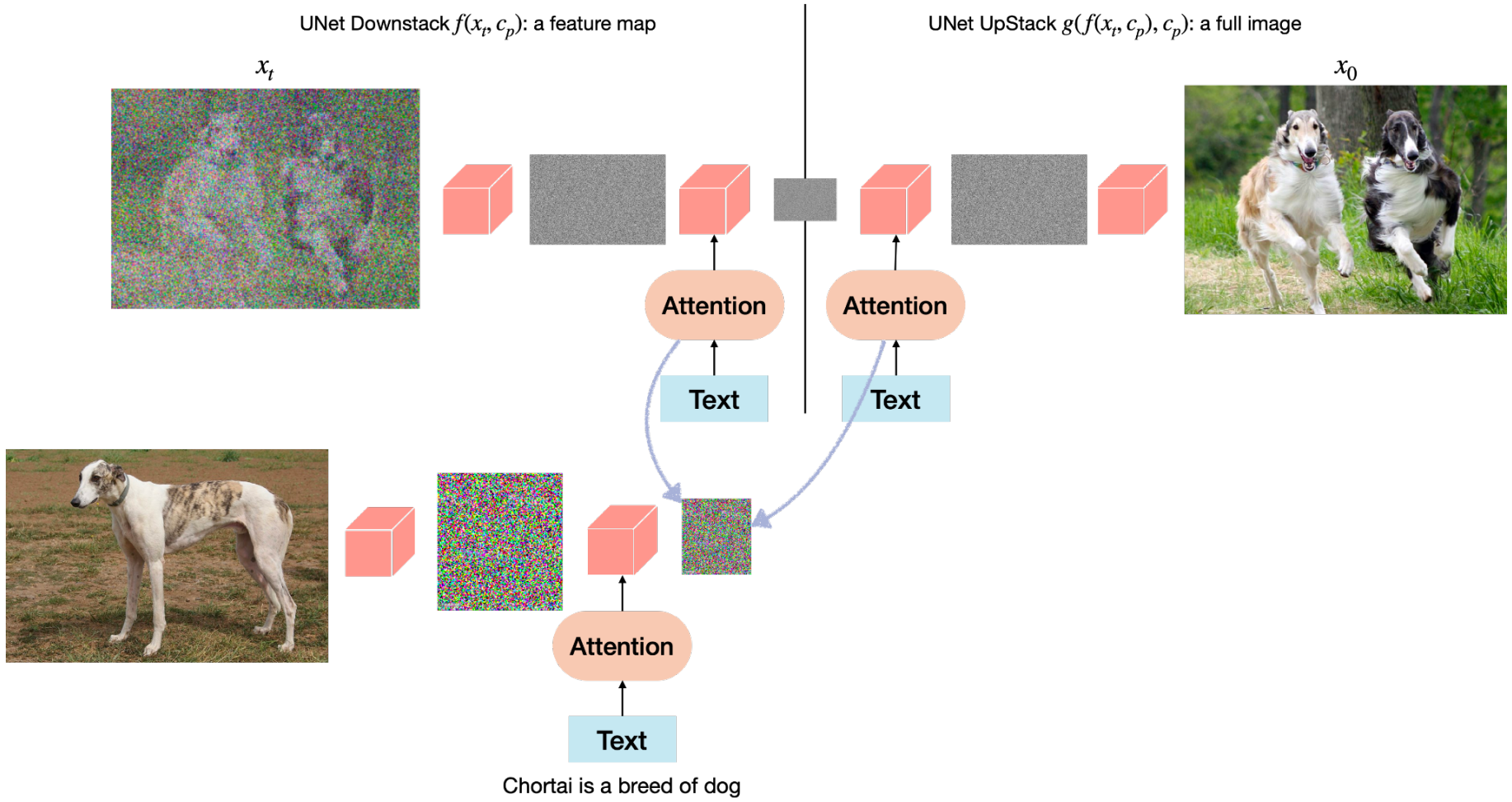
A teapot on table

DreamBooth

PaLM

A person holding a teapot

SuTI训练 (图文ICL数据集 v2)



04

结果和展望

实验参数

- 我们用ICL-v2的数据集训了我们的模型SuTI
  - 500K的训练数据，大约训练1天
  - 我们把三个exemplar的图文pair用encoder feed给Imagen
  - 然后Imagen通过attend到这三个exemplar就可以理解subject的appearance

- SuTI模型拥有五个技能
  - Stylization: 给物体增加艺术风格
  - Recontextualization: 把物体放在不同的环境下
  - Multi-View Synthesis: 从不同视角看物体
  - Attribute Modification: 改变物体的属性
  - Accessorization: 给物体增加不同的配件

输出的样例 [1]

A duck toy

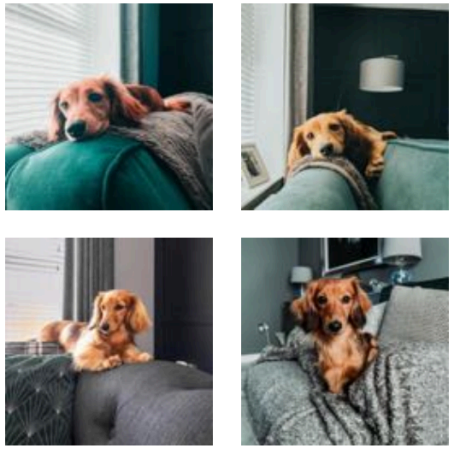
Pablo Picasso

Rembrandt

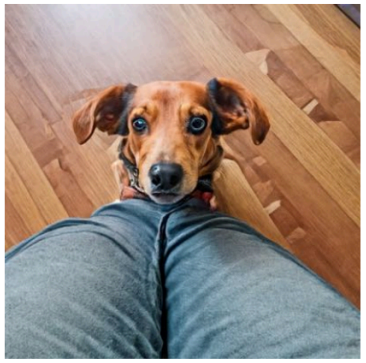
Rene Magritte

Vincent van Gogh

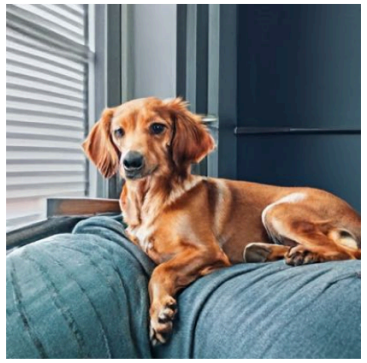
A dog



Top-down view

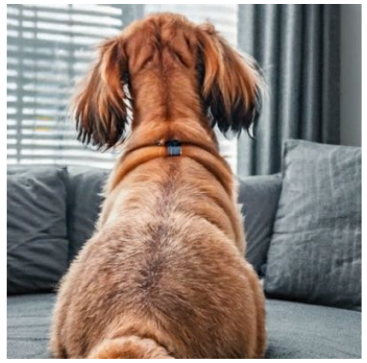


Side view



Bottom view

Back view



输出的样例 [2]

A dog

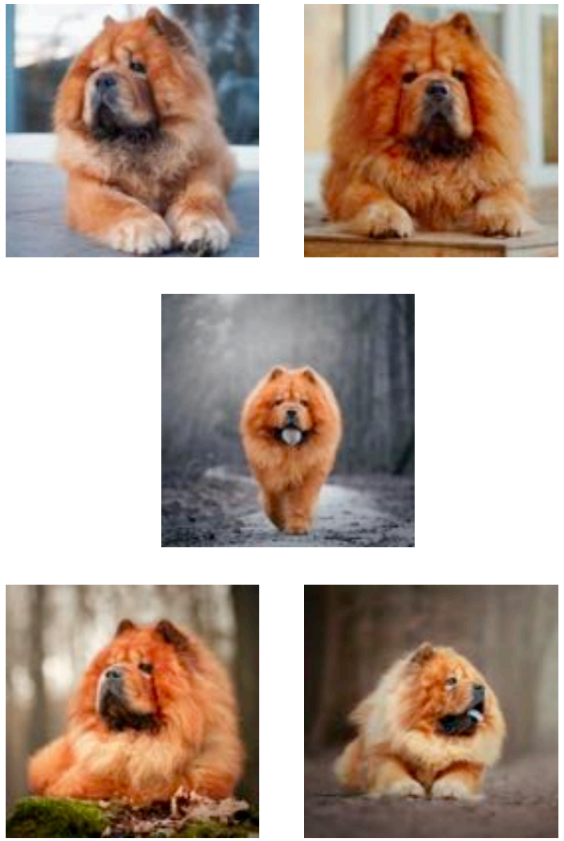
Depressed Joyous Sleepy Screaming

A monster toy

Blue Green Purple Pink

输出的样例 [3]

A dog



Chef outfit

Police outfit

Nurse outfit

Fire-Fighter outfit

Ironman outfit

Witch outfit

Superman outfit

Angel outfit

Human Evaluation

- 我们做了非常详尽的Human Evaluation来评测模型对Subject, Text的Alignment，还有生成的图片是否包含Artifacts

Methods	Backbone	Space	Time	Subject ↑	Text ↑	Photorealism ↑	Overall ↑
Models requiring test-time tuning							
Textual Inversion [10]	SD [25]	$	30 mins	0.22	0.64	0.90	0.14
Null-Text Inversion [19]	Imagen [28]	$$	5 mins	0.20	0.46	0.70	0.10
Imagic [15]	Imagen [28]	$$$$	70 mins	0.78	0.34	0.68	0.28
DreamBooth [27]	SD [25]	$$$	6 mins	0.74	0.53	0.85	0.47
DreamBooth [27]	Imagen [28]	$$$	10 mins	0.88	0.82	**0.98**	0.77
InstructPix2Pix [4]	SD [25]	-	10 secs	0.14	0.46	0.42	0.10
Re-Imagen [6]	Imagen [28]	-	20 secs	0.70	0.65	0.64	0.42
Ours: SuTI	Imagen [28]	-	30 secs	**0.90**	**0.90**	0.92	**0.82**

局限性和展望

- 目前的SuTI模型的输出还是存在不少artifacts
  - 尤其是人脸和文字之类的细节表达
  - 我们在尝试scale up模型到更大的size来解决这些问题

- 目前SuTI的技能还比较少，没办法像ControlNet一样给不同的signal
  - 我们目前在训练SuTI2可以把各种signal都feed给模型去生成output
  - 我们把所有技能都准备打包为一种instruction-tuning的format
  - 在未来的几个月即将launch进Google Cloud面世


感谢观看