格格尔加诺夫 / **美洲驼.cpp** 公共

法典 问题 230 **拉取请求** 52 讨论 行动 项目 4 维基

Code ▾ Jump to bottom

[研究] 转向矢量 #1472

草案 **狡猾的回声**想要将 4 个提交合并到 主人 从 转向 译

谈话 29 提交 4 检查 21 文件已更改 5

SlyEcho 评论 yesterday • 编辑 ▾ 合作者

对于#1460

```
./bin/main --model ../models/llama-7b-q4_0.bin -n 32 \
   --seed 123 \
  --prompt "I want to kill you because you're such a" \
  --steering-add "I love you so much" \
  --steering-sub "I hate you so much" \
  --steering-source 1 \
  --steering-layer 20 \
  --steering-mul 2
```

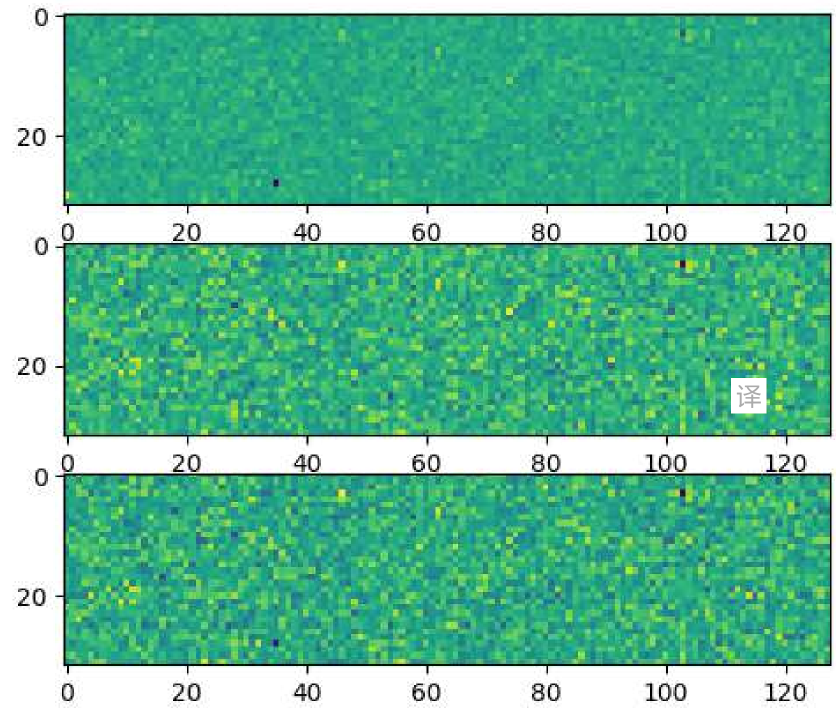
待办事项：使用新参数进行新测试。

我还想看看向量是什么样子的，所以我将它们导入 Numpy 并绘制它们：

```
import numpy as np
from matplotlib import pyplot as plt

steer = np.fromfile("~/src/llama.cpp/build/steering.bin", dtype=np.float32).reshape((512, -1))

fig, ax = plt.subplots(3)
for i in range(0, len(ax)):
    ax[i].imshow(steer[3+i, :].reshape((32, -1)))
```



☺ 🚀 7 👀 1

Steering ✓ 021e6d9

SlyEcho补充说 研究 🔬 标签 yesterday

此评论已被隐藏。 显示评论

此评论已被隐藏。 显示评论

此评论已被隐藏。 显示评论

此评论已被隐藏。 显示评论

SlyEcho 提到了这个拉取请求 yesterday

Investigate and play with "steering vectors" post (paper) #1460

○ Open

j-f1 reviewed yesterday

View reviewed changes

examples/common.cpp Outdated Show resolved

cleanup and stuff ✕ 8388aaa

SlyEcho force-pushed the `steering` branch from **`e63aa89`** to **`8388aaa`** 19 hours ago Compare

× Repository owner deleted a comment from **github-actions** bot 19 hours ago

Azeirah commented 17 hours ago

It's good to note that the authors of the post said they were going to try this out with vicuna-13B as well, so we can see how it generalizes accross different models

👍 1

Azeirah commented 17 hours ago • edited

Also, from a quick glance through your code I saw that the steering vector retrieval layer is always the same as the steering vector add layer.

They also allow steering vectors sourced from earlier layers to be used at later layers, which might be necessary to get good behavior.

> However, the norm of early-layer residual streams is significantly smaller than at later layers (like 20). In particular, we've found a large jump between layers 0 and 2. Let's try sourcing a steering vector from the residual stream just before layer 2, and then adding that layer-2 vector to layer 20.

> I also didn't get much response from the higher layer numbers (like 20 in the paper).

Did you source the steering vector from a lower layer? That's what they do.

source = layer 2

Add = layer 20

Not

source = layer 20

add = layer 20

SlyEcho commented 16 hours ago Collaborator Author

> Did you source the steering vector from a lower layer? That's what they do.

I didn't notice that in the article, all mentions of layers are about only one layer and where they inserted it.

But it should be easy to test.

`separate source layer for steering vector.` × `c90059f`

extradosages reviewed 14 hours ago

View reviewed changes

`examples/common.cpp` Outdated Show resolved

`Update examples/common.cpp` … ✕ `1b0ff2c`

译

评论家

额外剂量

j-f1

github-actions

合并此拉取请求至少需要 1 次批准审核。

受让人

没有人分配

标签

研究 🔬

项目

暂无

里程碑

无里程碑

发展

成功合并此拉取请求可能会解决这些问题。

暂无

4 名参与者