ckkissane first commit 5bde1d5 · 4 months ago History

Code Blame 59 lines (49 loc) · 1.57 KB Raw

```python
# %%
from utils import *
from crosscoder import CrossCoder
torch.set_grad_enabled(False);
# %%
cross_coder = CrossCoder.load_from_hf()

# %%
norms = cross_coder.W_dec.norm(dim=-1)
norms.shape
# %%
relative_norms = norms[:, 1] / norms.sum(dim=-1)
relative_norms.shape
# %%

fig = px.histogram(
    relative_norms.detach().cpu().numpy(),
    title="Gemma 2 2B Base vs IT Model Diff",
    labels={"value": "Relative decoder norm strength"},
    nbins=200,
)

fig.update_layout(showlegend=False)
fig.update_yaxes(title_text="Number of Latents")

```

Symbols

Find definitions and references for functions and other symbols in this file by clicking a symbol below or in the code.

Filter symbols r

const cross_coder
const norms
const relative_norms
const fig
const shared_latent_mask
const cosine_sims
const fig

```
    # Update x-axis ticks
    fig.update_xaxes(
        tickvals=[0, 0.25, 0.5, 0.75, 1.0],
        ticktext=['0', '0.25', '0.5', '0.75', '1.0']
    )

    fig.show()

    # %%
    shared_latent_mask = (relative_norms < 0.7) & (relative_norms > 0.3)
    shared_latent_mask.shape
    # %%
    # Cosine similarity of recoder vectors between models

    cosine_sims = (cross_coder.W_dec[:, 0, :] * cross_coder.W_dec[:, 1, :]).sum(dim=-1)
    cosine_sims.shape
    # %%
    import plotly.express as px
    import torch

    fig = px.histogram(
        cosine_sims[shared_latent_mask].to(torch.float32).detach().cpu().numpy(),
        #title="Cosine similarity of decoder vectors between models",
        log_y=True,  # Sets the y-axis to log scale
        range_x=[-1, 1],  # Sets the x-axis range from -1 to 1
        nbins=100,  # Adjust this value to change the number of bins
        labels={"value": "Cosine similarity of decoder vectors between models"}
    )

    fig.update_layout(showlegend=False)
    fig.update_yaxes(title_text="Number of Latents (log scale)")

    fig.show()
    # %%
```