MultiModal-Training

This model is the concatenation of the Text modal, i.e. Bert-base-cased and Image modal BeIT from Microsoft. The initial four epochs of the training were nice as validation and training losses decreased, but losses started increasing in the next ten epochs.

Samarth Garg

Draft autosaved 2 minutes ago

Save

the image the same and asking different questions doesn't change the output probability much. This implies multimodal is weighing the text modal less than the image modal.

Another interesting thing is the gradients' distribution, which suggests they might be prone to vanishing gradients issues.

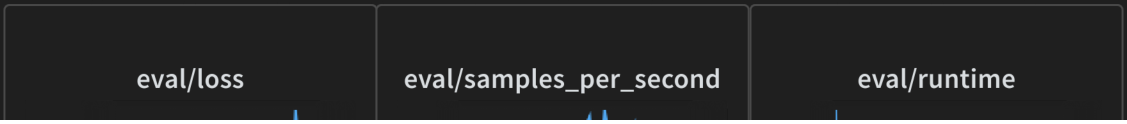
Although the training dataset was huge and the loss kept on decreasing, it could either be due to overfitting, or it could be improving.

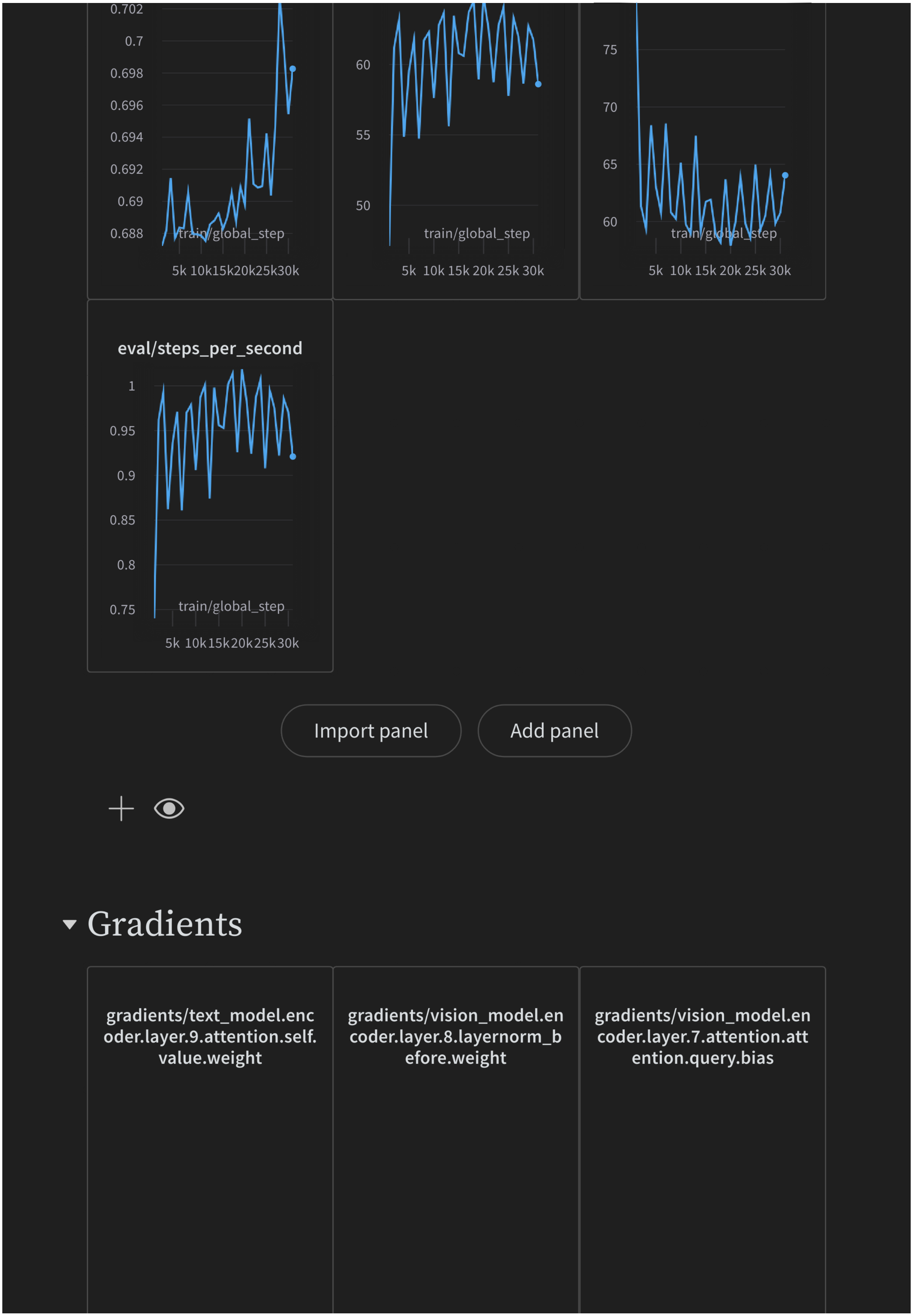
Improvements

I suggest increasing the validation size and training for 10-15 more epochs.

If overfitting is not the issue, we can introduce more complexity in a model, like adding an LSTM layer.

Validation





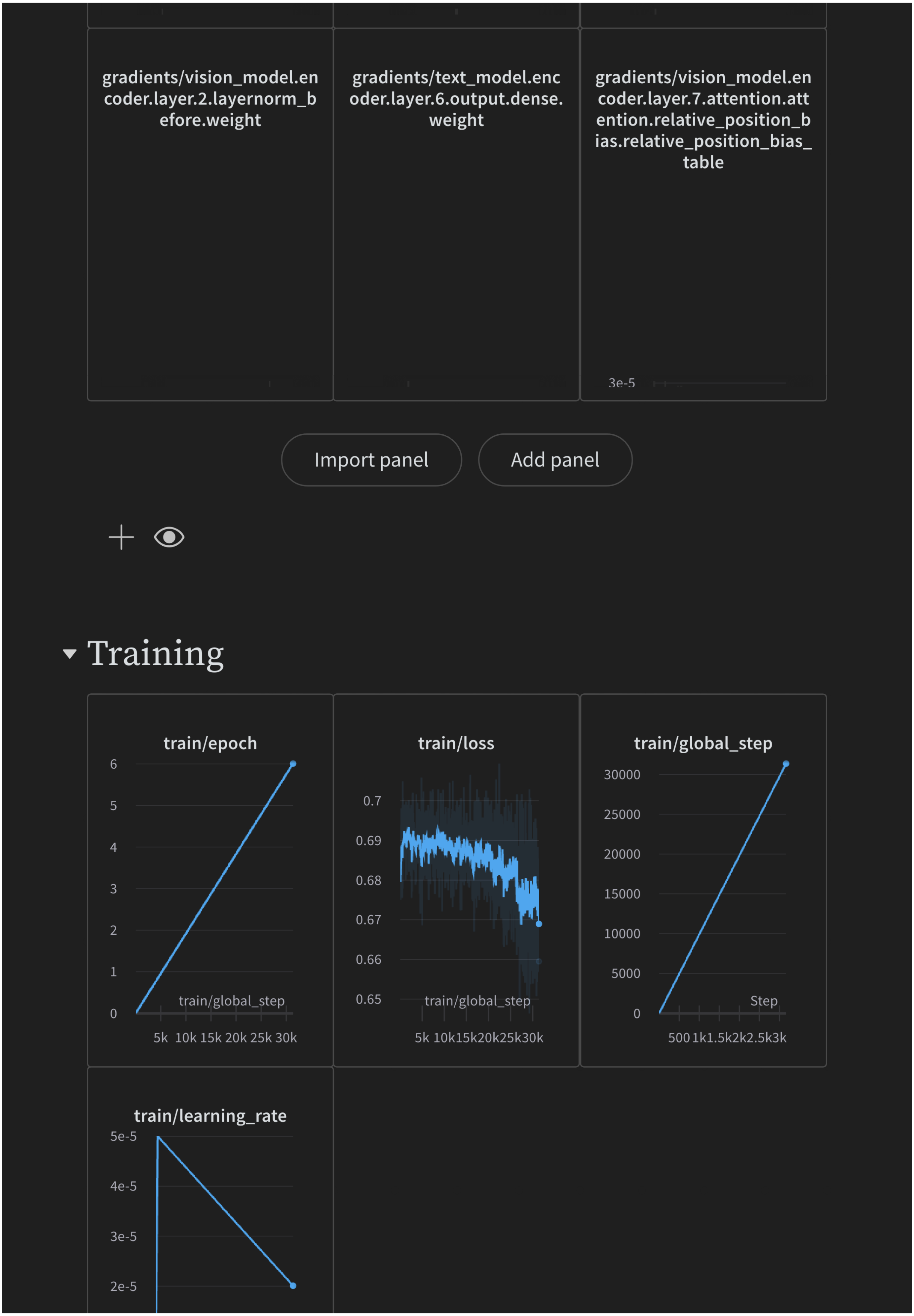
eval/steps_per_second

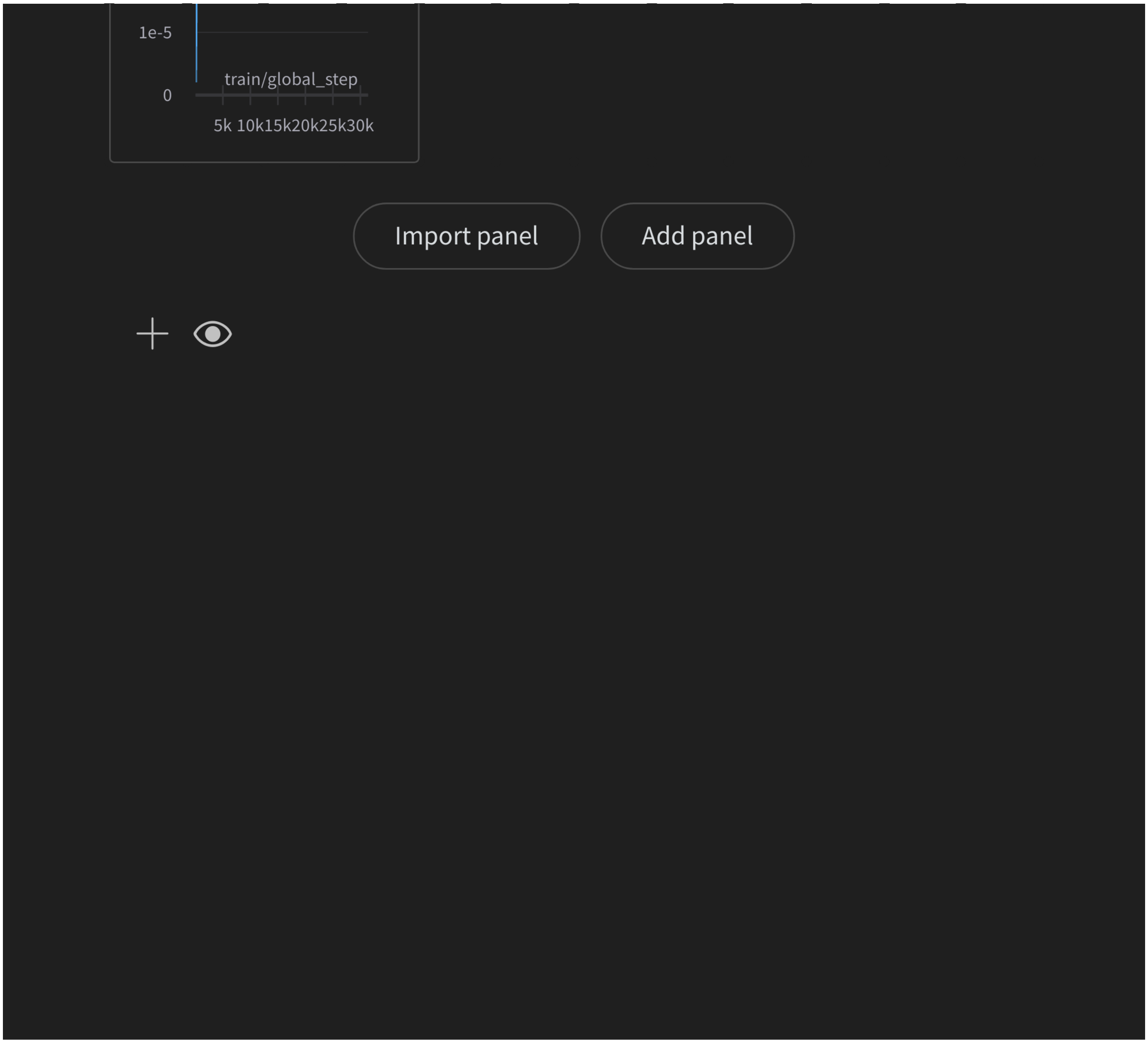
▾ Gradients

gradients/text_model.encoder.layer.9.attention.self.value.weight

gradients/vision_model.encoder.layer.8.layernorm_before.weight

gradients/vision_model.encoder.layer.7.attention.attention.query.bias


gradients/vision_model.encoder.layer.2.layernorm_before.weight
gradients/text_model.encoder.layer.6.output.dense.weight
gradients/vision_model.encoder.layer.7.attention.attention.relative_position_bias.relative_position_bias_table
3e-5
Import panel
Add panel
Training
train/epoch
6
5
4
3
2
1
0
train/global_step
5k 10k 15k 20k 25k 30k
train/loss
0.7
0.69
0.68
0.67
0.66
0.65
train/global_step
5k 10k15k20k25k30k
train/global_step
30000
25000
20000
15000
10000
5000
0
Step
500 1k1.5k2k2.5k3k
train/learning_rate
5e-5
4e-5
3e-5
2e-5


1e-5
0
train/global_step
5k 10k15k20k25k30k
Import panel
Add panel