DCN: The importance of the buffer size for shufling data

For the DCN model, the DThis report shows the importance of the buffer size used for shuffling data in datasets before training the model.
Specifically in our project, the different data samples are measures on objects detected in images. In one image, we can detect hudreds, thousands or even tens of thousands of stars, galaxies or other objects, all related to the same class
because the annotation holds on one image. So the risk of grouping data samples with the same annotations is strong.

Mike Fournigault

Created on April 17 | Last edited on April 17

This might lead to the appearance of specific patterns in the graph of metrics like batch/accuracy or batch/loss. And cause issues of the model exploration for global optimimums, leading to mean performances in the end.

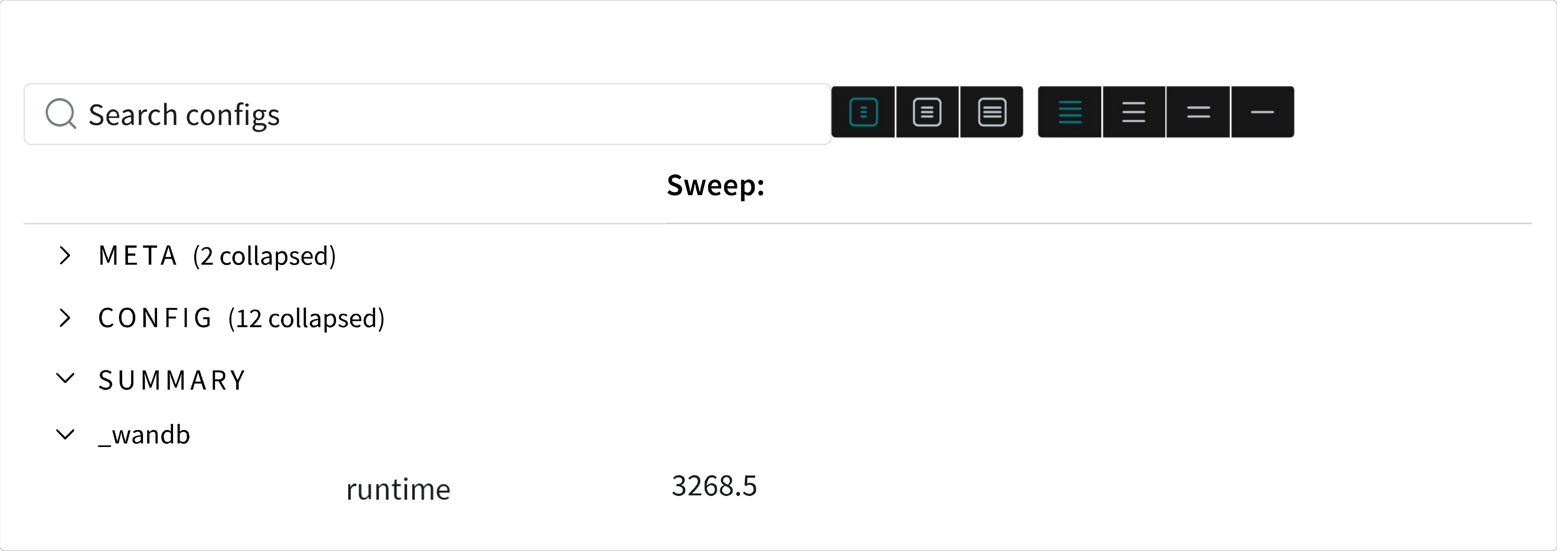
These phenomenons are observed for the run "stilted-salad-50", for which the buffer size of data shuffling was fixed to 10k samples.

In comparison, the buffer size was fixed to 100k samples for the run "kind-firefly-51", leading to much better performances with 97.8% epoch- accuracy and 99.9% of accuracy for epoch-validation (and test) ; against 65.9% of epoch-accuracy and 64.5% of epoch-val_accuracy for the run "stilted-salad-51".

The source datasets are catalogs loaded from several parquet files, with various sources. They are then concatenated in one set, shuffled with numpy during this operation. It seems that the numpy shuffling was not good enough, and adding a significant shuffling with Tensorflow improves greatly the model performances.

The observation made for the DCN model should be true for other models as well, the same experiment should be tested with other models.

Configuration Summary



Sweep:

META (2 collapsed)

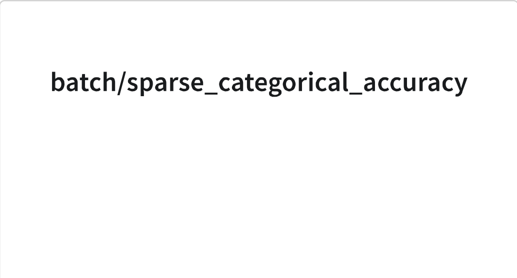
CONFIG (12 collapsed)

SUMMARY

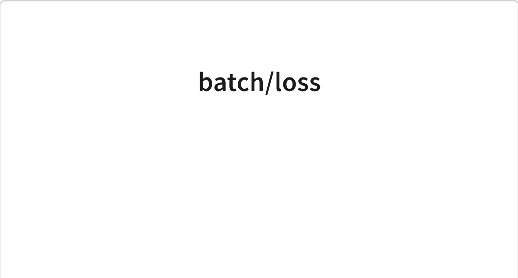
_wandb

runtime 3268.5

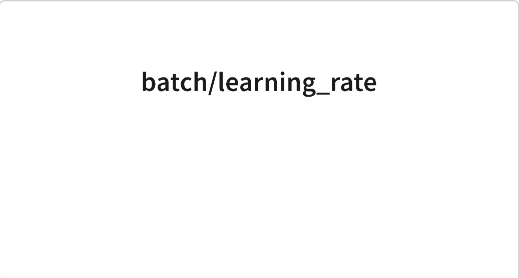
Batch analysis



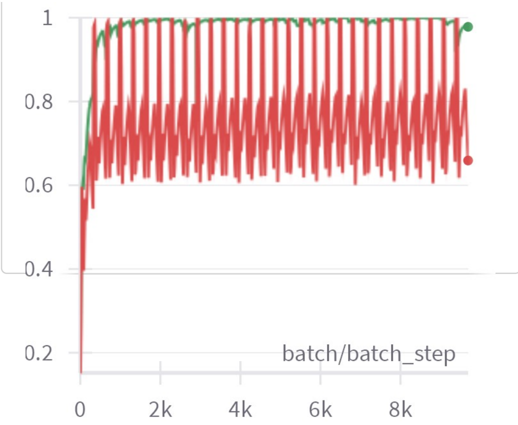
batch/sparse_categorical_accuracy

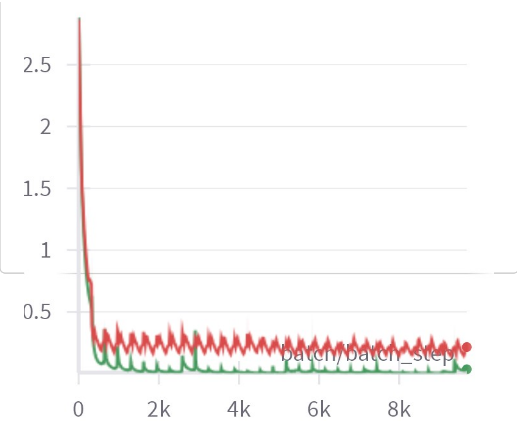


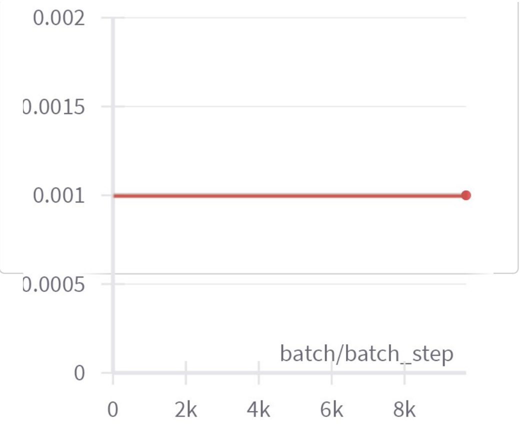
batch/loss



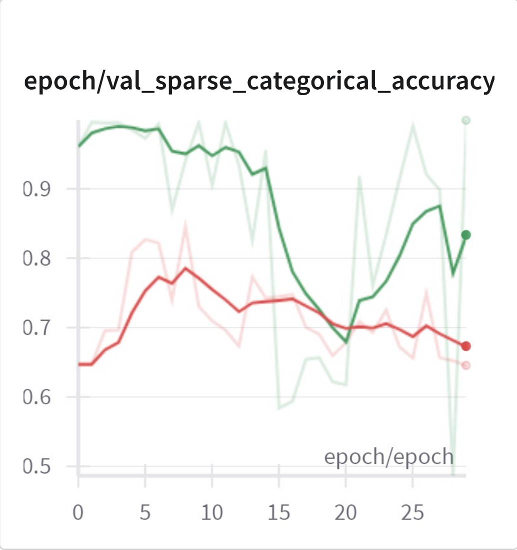
batch/learning_rate

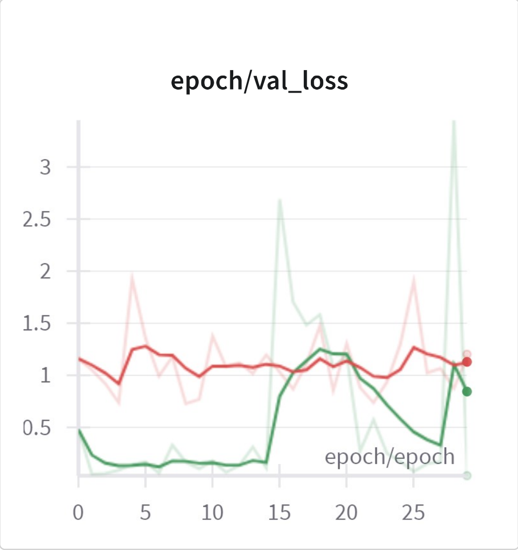


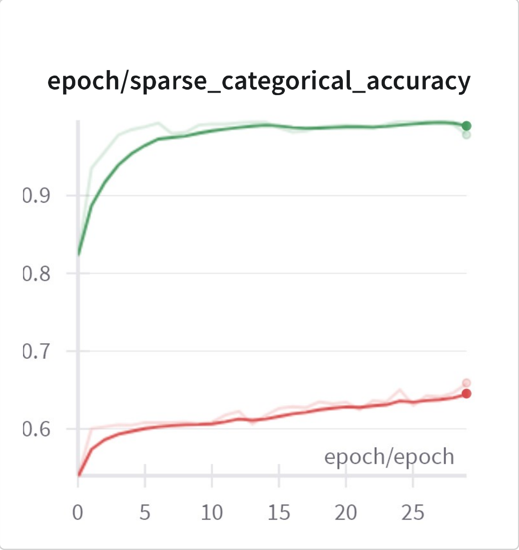




Epoch analysis


epoch/val_sparse_categorical_accuracy


epoch/val_loss


epoch/sparse_categorical_accuracy

Created with ❤️ on Weights & Biases.

https://wandb.ai/mike-fournigault1/astro_iqa/reports/DCN-The-importance-of-the-buffer-size-for-shufling-data--VmlldzoxMjMzOTY5OA