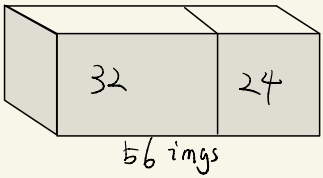
test_batchsize
= max_batchsize = 16

batchsize = 16 × 2

1 job = 1 img

1 GPU



2 epochs for batch A&B

@ Commits / infer_controller.hpp

preprocess the batch A one by one img, then push them into the jobs_ queue one by one. Once 32 jobs are ready in the jobs_ queue, a thread is notified to fetch the ready 32 jobs in the jobs_ queue.

preprocess on GPU

A Step 1

B Step 1

jobs_lock_

jobs_

@ get_jobs_and_wait / infer_controller.hpp

When the jobs_ queue is not empty, cond_.wait will be unblocked. Then 16 (max batchsize) will be picked by a thread and put into the fetched_jobs of the current thread one by one.

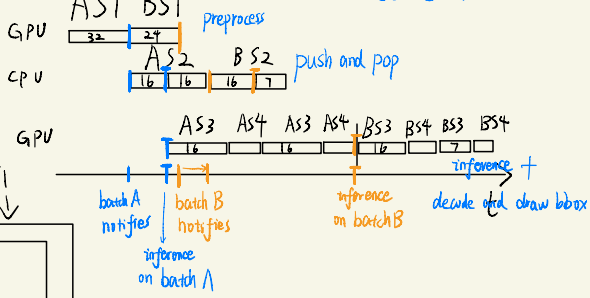
A Step 2

B Step 2

* Timing

| means notify point

T means inference point



also infer batch size

When 16 jobs (max batchsize) are in place, the 16 item can be infered by trt model,

A Step 3

which is followed by the green batch.

fetch_jobs

B Step 3

in a batch (16) one by one

Multiple threads apply to the framework.

A Step 4

Decode on GPU

to CPU

draw bboxes