A Multigrid Method for Efficiently Training Video Models

Heng Wang, Du Tran, Lorenzo Torresani, Matt Feiszli

Instead of training with fixed temporal and spatial dimensions during the whole training. The paper proposes to scale down these dimensions while scaling up the batch size and learning rate, resulting in both faster training time and even slightly better results. For original scales of T x H x W; the scaling is chosen so that the new scales have a shape t x h x w (iT x jH x kW) and the batch size is scaled up with the combined downscaling rage (b = B x (i x j x k)).

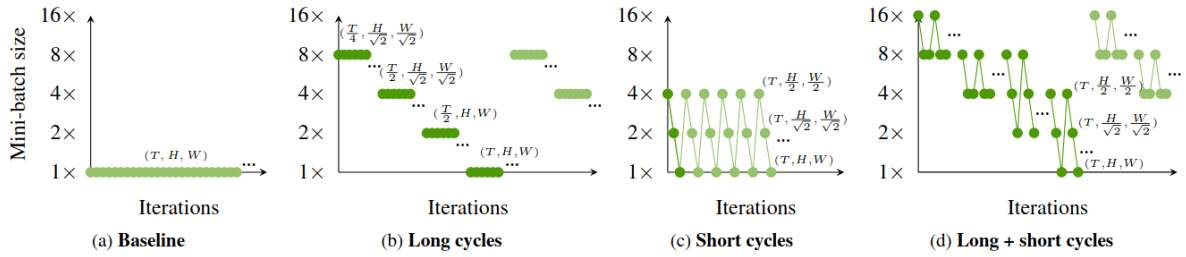

Figure 2. **A general and robust grid schedule (§3.2).** We contrast multigrid training with standard baseline training. (a) **Baseline** training methods typically use a fixed mini-batch shape throughout training. (b) **Multigrid long cycles** loop over inputs from small shapes (with large mini-batch sizes) to large shapes (with small mini-batch sizes), staying on each shape for several epochs. (c) **Multigrid short cycles** rapidly move through a variety of spatial shapes, changing at each iteration. (d) **Multigrid long + short cycles (our default setting)** combines long and short cycles, and moves through shapes at two frequencies simultaneously. **Dark green** points in (b), (c), and (d) correspond to one full period of a long cycle, a full short cycle, and a long+short cycle, respectively.

The paper proposes 3 types of multi-grid scheduling methods; long cycles with 4 possible choices of scales and low frequency, short cycles with 3 possible choices and high frequency, and long+short cycles that combine both.

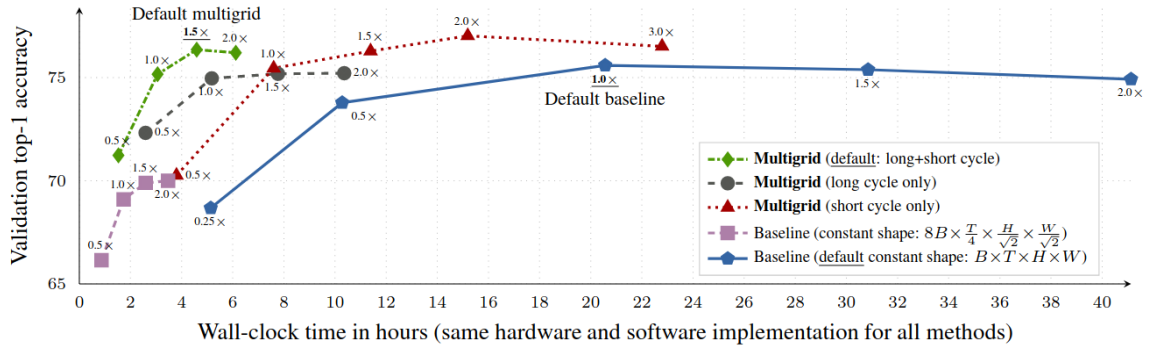

Figure 3. **Multigrid *vs.* baseline training.** Each point corresponds to one model trained with a specific schedule choice. Annotations denote training epochs relative to the baseline 1.0× schedule. For example, '1.5×' denotes training for 1.5× more epochs than the default '1.0×' baseline schedule (112k iterations or ~239 epochs). We see that **all variants of multigrid training achieve a better trade-off than baseline training, which uses a constant mini-batch shape**. Also note that multigrid training can iterate through the same number of epochs more efficiently.