Speed up OPs with System Optimizations

In the previous notebook, we are already familiar with how to develop your own OP.

For most text-only OPs, the running speed of them is usually very fast. But for multimodal datasets and OPs, we often need to integrate some excellent existing models to help clean/synthesize data. Therefore, the efficiency of these OPs might become a bottleneck in the whole data processing pipeline due to some compute-intensive subprocedures and large model inference.

Luckily, Data-Juicer already conducts several system optimization on OPs to speed up their processing, such as CUDA support, OP fusion, and so on. In this notebook, we will check how to enable these optimization strategies in OPs and how they work.

Test Dataset Preparation

Here we prepare a test dataset to check system optimizations on OPs, which contains 10k image-text pairs.

We just need to download to the current directory.

In [6]:

```
!wget http://dail-wlcb.oss-cn-wulanchabu.aliyuncs.com/data_juicer/tutoria
```

```
--2024-08-09 06:40:56--  http://dail-wlcb.oss-cn-wulanchabu.aliyuncs.com/d
ata_juicer/tutorial_test_data.tar.gz
Resolving dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-wulan
chabu.aliyuncs.com)... 39.101.35.6
Connecting to dail-wlcb.oss-cn-wulanchabu.aliyuncs.com (dail-wlcb.oss-cn-w
ulanchabu.aliyuncs.com)|39.101.35.6|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 466663769 (445M) [application/gzip]
Saving to: ‘tutorial_test_data.tar.gz’

tutorial_test_data. 100%[===================>] 445.04M  24.5MB/s    in 13s

2024-08-09 06:41:09 (34.5 MB/s) - ‘tutorial_test_data.tar.gz’ saved [46666
3769/466663769]
```

Then we can check the basic format of this dataset.

In [1]:

```
import os
from jsonargparse import dict_to_namespace
from data_juicer.format.load import load_formatter
# replace the dataset path to your correct version
dataset_path = './tutorial_test_data/tutorial_test.jsonl'

formatter = load_formatter(dataset_path, text_keys='text')
dataset = formatter.load_dataset(global_cfg=dict_to_namespace({
    'dataset_dir': os.path.dirname(dataset_path),
    'video_key': 'videos',
    'image_key': 'images',
    'audio_key': 'audios',
}))
print(dataset)
```

```
/usr/local/python310/lib/python3.10/site-packages/tqdm/auto.py:21: TqdmWar
ning: IProgress not found. Please update jupyter and ipywidgets. See http
s://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
Generating jsonl split: 10000 examples [00:00, 943112.45 examples/s]
2024-08-09 14:57:27.519 | INFO     | data_juicer.format.formatter:unify_fo
rmat:185 - Unifying the input dataset formats...
2024-08-09 14:57:27.520 | INFO     | data_juicer.format.formatter:unify_fo
rmat:200 - There are 10000 sample(s) in the original dataset.
Filter: 100%|██████████| 10000/10000 [00:00<00:00, 90626.52 examples/s]
2024-08-09 14:57:27.633 | INFO     | data_juicer.format.formatter:unify_fo
rmat:214 - 10000 samples left after filtering empty text.
2024-08-09 14:57:27.633 | INFO     | data_juicer.format.formatter:unify_fo
rmat:237 - Converting relative paths in the dataset to their absolute vers
ion. (Based on the directory of input dataset file)
Map: 100%|██████████| 10000/10000 [00:00<00:00, 18639.65 examples/s]
2024-08-09 14:57:28.172 | INFO     | data_juicer.format.mixture_formatter:
load_dataset:137 - sampled 10000 from 10000
2024-08-09 14:57:28.174 | INFO     | data_juicer.format.mixture_formatter:
load_dataset:143 - There are 10000 in final dataset
Dataset({
    features: ['id', 'text', 'images'],
    num_rows: 10000
})
```

CUDA Support

For those OPs who apply some other models to help to clean, filter, synthesize data, it usually takes lots of time for model inference on CPUs. Therefore, Data-Juicer supports CUDA for these OPs to enable parallel computation on GPUs.

Here we take a image-related OP `image_aesthetics_filter` as the example to show you how CUDA is supported in Data-Juicer. This OP scores images in the dataset from the aesthetics perspective with an existing model `shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE`. The implementation of this OP can be found here.

We prepare a test data recipe to apply this OP on the test dataset. For the first version, we run this OP in CPU mode.

In [4]:

```
cpu_recipe = '''
project_name: 'cpu_mode_test'
dataset_path: './tutorial_test_data/tutorial_test.jsonl'
export_path: './outputs/cpu_mode/res.jsonl'
np: 24

# we need to experiment for several times, so we need to close cache mana
# to guarantee all these exps run the OP indeed.
use_cache: false

# you can replace the model path with a local path on your machine
process:
  - image_aesthetics_filter:
      hf_scorer_model: shunk031/aesthetics-predictor-v2-sac-logos-ava1-l1
      min_score: 0.3
      max_score: 1.0
      accelerator: 'cpu'  # run the model inference with CPUs
'''

with open('cpu_mode.yaml', 'w') as fout:
    fout.write(cpu_recipe)
```

Then we run this recipe with `dj-process` command.

In [1]:

```
!dj-process --config cpu_mode.yaml
```

```
2024-08-09 16:01:05 | WARNING  | data_juicer.config.config:405 - Cache man
agement of datasets is disabled.
2024-08-09 16:01:05 | WARNING  | data_juicer.config.config:416 - Set temp
directory to store temp files to [None].
2024-08-09 16:01:06 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/cpu_mode.yaml]
into the work_dir [/root/projects/kdd_tutorial_notebooks/outputs/cpu_mode]
2024-08-09 16:01:06 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒══════════════════════════╤═════════════════════════════════════════════
═════════════════════════════════════════════════════════════════════╕
│ key                      │ values
│
╞══════════════════════════╪═════════════════════════════════════════════
═════════════════════════════════════════════════════════════════════╡
│ config                   │ [Path_fr(cpu_mode.yaml, cwd=/root/projects/kdd
_tutorial_notebooks)]                                                │
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ hpo_config               │ None
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ data_probe_algo          │ 'uniform'
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ data_probe_ratio         │ 1.0
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ project_name             │ 'cpu_mode_test'
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ executor_type            │ 'default'
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ dataset_path             │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data/tutorial_test.jsonl'                                     │
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ export_path              │ '/root/projects/kdd_tutorial_notebooks/output
s/cpu_mode/res.jsonl'                                                 │
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ export_shard_size        │ 0
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ export_in_parallel       │ False
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ keep_stats_in_res_ds     │ False
│
├──────────────────────────┼─────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ keep_hashes_in_res_ds    │ False
```

```
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ np                       │ 24
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ text_keys                │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ image_key                │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ image_special_token      │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ audio_key                │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ audio_special_token      │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ video_key                │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ video_special_token      │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ eoc_special_token        │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ suffixes                 │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ use_cache                │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ ds_cache_dir             │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ cache_compress           │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ use_checkpoint           │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ temp_dir                 │ None
```

```
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ process                 │ [{'image_aesthetics_filter': {'accelerator':
'cpu',                                                               │
│                         │                              'any_or_all': 'a
ny',                                                                │
│                         │                              'audio_key': 'au
dios',                                                              │
│                         │                              'cpu_required':
1,                                                                  │
│                         │                              'hf_scorer_mode
l': 'shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE', │
│                         │                              'image_key': 'im
ages',                                                              │
│                         │                              'max_score': 1.
0,                                                                   │
│                         │                              'mem_required':
0,                                                                  │
│                         │                              'min_score': 0.
3,                                                                   │
│                         │                              'num_proc': 24,
│
│                         │                              'stats_export_pa
th': None,                                                          │
│                         │                              'text_key': 'tex
t',                                                                 │
│                         │                              'video_key': 'vi
deos'}}]                                                            │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ ray_address             │ 'auto'
```

```
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/cpu_mode'                                                           │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ timestamp               │ '20240809160105'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ dataset_dir             │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data'                                                        │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╛
2024-08-09 16:01:06 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
2024-08-09 16:01:06 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-09 16:01:06 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
2024-08-09 16:01:07 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-09 16:01:07 | INFO     | data_juicer.format.formatter:200 - There
are 10000 sample(s) in the original dataset.
Filter (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 60270.87
examples/s]
2024-08-09 16:01:07 | INFO     | data_juicer.format.formatter:214 - 10000
samples left after filtering empty text.
2024-08-09 16:01:07 | INFO     | data_juicer.format.formatter:237 - Conver
ting relative paths in the dataset to their absolute version. (Based on th
e directory of input dataset file)
Map (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 59091.10 exa
mples/s]
2024-08-09 16:01:08 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 10000 from 10000
2024-08-09 16:01:08 | INFO     | data_juicer.format.mixture_formatter:143
- There are 10000 in final dataset
2024-08-09 16:01:08 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
2024-08-09 16:01:10 | INFO     | data_juicer.core.executor:164 - Processin
```

```
g data...
Adding new column for stats (num_proc=24): 100%|##########| 10000/10000 [0
0:00<00:00, 61090.70 examples/s]
image_aesthetics_filter_compute_stats (num_proc=24): 100%|##########| 1000
0/10000 [08:21<00:00, 19.93 examples/s]
image_aesthetics_filter_process (num_proc=24): 100%|##########| 10000/1000
0 [00:00<00:00, 62824.34 examples/s]
2024-08-09 16:09:33 | INFO     | data_juicer.core.data:193 - OP [image_aes
thetics_filter] Done in 503.301s. Left 9987 samples.
2024-08-09 16:09:33 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 503.302s.
2024-08-09 16:09:33 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-09 16:09:33 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 191.
92ba/s]
2024-08-09 16:09:33 | INFO     | data_juicer.core.exporter:140 - Export da
taset into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 206.
68ba/s]
```

As we can see, the data processing for the whole dataset costs **503.302s (8min23s)** in total in CPU mode using 24 subprocesses, which is relatively not very fast. Because if we want to process some larger-scale datasets (e.g. CC3M with nearly 3 million image-text pairs), it might take several days, let alone those larger and slower models.

Now let's replace the accelerator of this OP with CUDA and check its speed. We run this experiment on a machine with 8 GPUs. As we benchmarked this model before, one aesthetics predictor model requires about 1500MB GPU memory, so we need to set the "mem_required" in addition.

In [2]:

```
cuda_recipe = '''
project_name: 'cpu_mode_test'
dataset_path: './tutorial_test_data/tutorial_test.jsonl'
export_path: './outputs/cpu_mode/res.jsonl'
np: 24

# we need to experiment for several times, so we need to close cache mana
# to guarantee all these exps run the OP indeed.
use_cache: false

# you can replace the model path with a local path on your machine
process:
  - image_aesthetics_filter:
      hf_scorer_model: shunk031/aesthetics-predictor-v2-sac-logos-ava1-l1
      min_score: 0.3
      max_score: 1.0
      accelerator: 'cuda'  # run the model inference with CUDA
      mem_required: '1500MB'  # set the GPU memory requirements for this
'''

with open('cuda_mode.yaml', 'w') as fout:
    fout.write(cuda_recipe)
```

Then we run this recipe with `dj-process` command.

In [3]:

```
!dj-process --config cuda_mode.yaml
```

```
2024-08-09 16:11:22 | WARNING  | data_juicer.config.config:405 - Cache man
agement of datasets is disabled.
2024-08-09 16:11:22 | WARNING  | data_juicer.config.config:416 - Set temp
directory to store temp files to [None].
2024-08-09 16:11:22 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/cuda_mode.yam
l] into the work_dir [/root/projects/kdd_tutorial_notebooks/outputs/cpu_mo
de]
2024-08-09 16:11:22 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒══════════════════════════╤═══════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╕
│ key                      │ values                                         
│
╞══════════════════════════╪═══════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╡
│ config                   │ [Path_fr(cuda_mode.yaml, cwd=/root/projects/kd
d_tutorial_notebooks)]                                              │
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ hpo_config               │ None                                           
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_algo          │ 'uniform'                                      
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_ratio         │ 1.0                                            
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ project_name             │ 'cpu_mode_test'                                
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ executor_type            │ 'default'                                      
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ dataset_path             │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data/tutorial_test.jsonl'                                    │
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_path              │ '/root/projects/kdd_tutorial_notebooks/output
s/cpu_mode/res.jsonl'                                                │
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_shard_size        │ 0                                              
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_in_parallel       │ False                                          
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ keep_stats_in_res_ds     │ False                                          
│
├──────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ np                      │ 24
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ use_cache               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ temp_dir                │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ process                 │ [{'image_aesthetics_filter': {'accelerator':
'cuda',                                                              │
│                         │                              'any_or_all': 'a
ny',                                                                │
│                         │                              'audio_key': 'au
dios',                                                              │
│                         │                              'cpu_required':
1,                                                                  │
│                         │                              'hf_scorer_mode
l': 'shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE', │
│                         │                              'image_key': 'im
ages',                                                              │
│                         │                              'max_score': 1.
0,                                                                   │
│                         │                              'mem_required':
'1500MB',                                                           │
│                         │                              'min_score': 0.
3,                                                                   │
│                         │                              'num_proc': 24,
│
│                         │                              'stats_export_pa
th': None,                                                          │
│                         │                              'text_key': 'tex
t',                                                                 │
│                         │                              'video_key': 'vi
deos'}}]                                                            │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ ray_address              │ 'auto'
│
├──────────────────────────┼──────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ debug                    │ False
│
├──────────────────────────┼──────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ work_dir                 │ '/root/projects/kdd_tutorial_notebooks/output
s/cpu_mode'                                                           │
├──────────────────────────┼──────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ timestamp                │ '20240809161122'
│
├──────────────────────────┼──────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ dataset_dir              │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data'                                                          │
├──────────────────────────┼──────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ add_suffix               │ False
│
╘══════════════════════════╧══════════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╛
2024-08-09 16:11:22 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
2024-08-09 16:11:22 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-09 16:11:22 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
2024-08-09 16:11:23 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-09 16:11:23 | INFO     | data_juicer.format.formatter:200 - There
are 10000 sample(s) in the original dataset.
Filter (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 71484.76
examples/s]
2024-08-09 16:11:24 | INFO     | data_juicer.format.formatter:214 - 10000
samples left after filtering empty text.
2024-08-09 16:11:24 | INFO     | data_juicer.format.formatter:237 - Conver
ting relative paths in the dataset to their absolute version. (Based on th
e directory of input dataset file)
Map (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 66390.57 exa
mples/s]
2024-08-09 16:11:24 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 10000 from 10000
2024-08-09 16:11:24 | INFO     | data_juicer.format.mixture_formatter:143
- There are 10000 in final dataset
2024-08-09 16:11:24 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
```

```
2024-08-09 16:11:26 | INFO     | data_juicer.core.executor:164 - Processin
g data...
Adding new column for stats (num_proc=24): 100%|##########| 10000/10000 [0
0:07<00:00, 1373.10 examples/s]
image_aesthetics_filter_compute_stats (num_proc=24):   0%|          | 0/10
000 [00:00<?, ? examples/s]/usr/local/python310/lib/python3.10/site-packag
es/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download`
is deprecated and will be removed in version 1.0.0. Downloads always resum
e when possible. If you want to force a new download, use `force_download=
True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
```

```
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
```

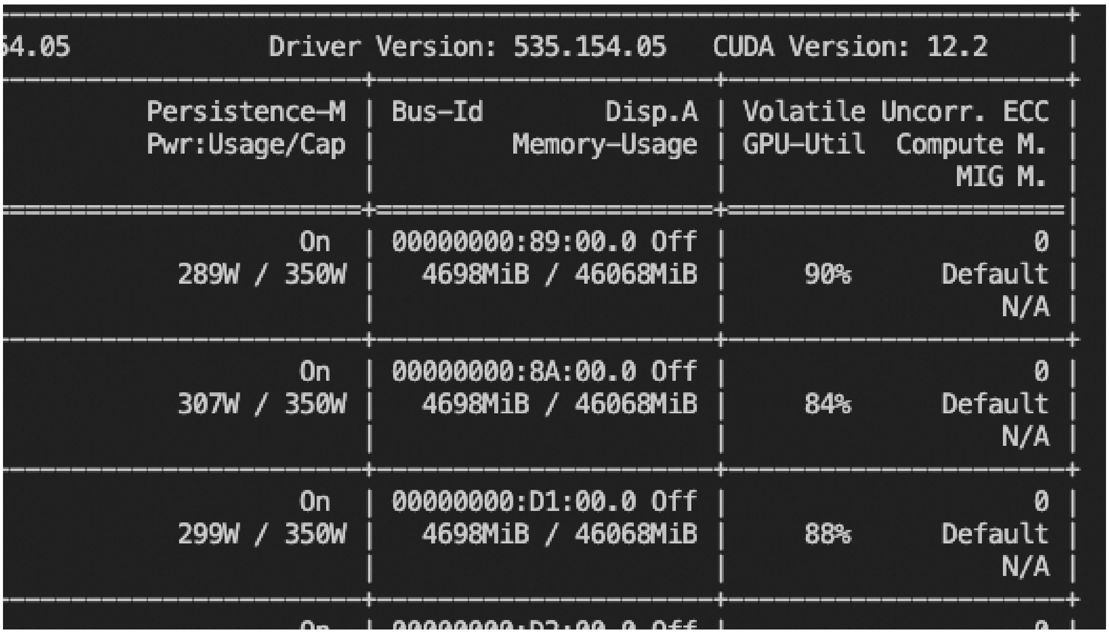
```
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
```

```
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
```

```
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
image_aesthetics_filter_compute_stats (num_proc=24): 100%|##########| 1000
0/10000 [00:26<00:00, 375.50 examples/s]
image_aesthetics_filter_process (num_proc=24): 100%|##########| 10000/1000
0 [00:07<00:00, 1345.44 examples/s]
2024-08-09 16:12:08 | INFO     | data_juicer.core.data:193 - OP [image_aes
thetics_filter] Done in 42.509s. Left 9987 samples.
2024-08-09 16:12:08 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 42.510s.
2024-08-09 16:12:08 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-09 16:12:08 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 193.
95ba/s]
2024-08-09 16:12:09 | INFO     | data_juicer.core.exporter:140 - Export da
taset into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 196.
94ba/s]
```

As we can see, the data processing for the whole dataset only costs **42.510s** in total in CUDA mode using 24 subprocesses, which is much faster than CPU mode and the speed-up ratio is nearly **12x**.

In CUDA mode, we use the same number of subprocesses as in CPU mode. For 8 GPUs, Data-Juicer allocate 3 models on each GPU (GPU mem occupation 4600MiB is nearly 3 times of the memory requirement of a single model).



```
--------------------------------------------------------------------------+
54.05            Driver Version: 535.154.05   CUDA Version: 12.2     |
-------------------------+----------------------+----------------------+
          Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
                        |                      |               MIG M. |
========================+======================+======================|
                     On | 00000000:89:00.0 Off |                    0 |
            289W / 350W |   4698MiB / 46068MiB |     90%      Default |
                        |                      |                  N/A |
------------------------+----------------------+----------------------+
                     On | 00000000:8A:00.0 Off |                    0 |
            307W / 350W |   4698MiB / 46068MiB |     84%      Default |
                        |                      |                  N/A |
------------------------+----------------------+----------------------+
                     On | 00000000:D1:00.0 Off |                    0 |
            299W / 350W |   4698MiB / 46068MiB |     88%      Default |
                        |                      |                  N/A |
------------------------+----------------------+----------------------+
                     On | 00000000:D2:00.0 Off |                    0 |
```

Theoretically, we can continue to increase the number of subprocesses to fully make use of GPUs and obtain more speed-up.

Data-Juicer set the `accelerator` of those model-based OPs to `cuda` in default, so users don't need to reset it during data prcessing in general. But sometimes for some new models in new OPs, users need to benchmark the GPU utilization for their models first and set it as the default value of `mem_required` parameter.

OP Fusion

Some OPs in Data-Juicer might share the same computation subprocedures. For example, both `video_aesthetics_filter` and `video_frames_text_similarity_filter` OPs need to sample several frames from the videos. For the same video, extrating the same frames for multiple times in different OPs is actually redundant computation procedure and a waste of time.

When OP fusion is enabled, Data-Juicer will find these redundant computation procedure automatically from different OPs. These procedures will only be computed for once in the first OP that contains this procedure and the results of them will be stored in the context fields. For later OPs that contain the same procedure, they only need to read the results from the context fields instead of computing them again, which saves lots of computation costs.

Here we prepare a simple example for OP fusion to see how to enable OP fusion in your recipes. First, we prepare a data recipe including several OPs that contain the same subprocedures with OP fusion disabled.

In [1]:
```
op_fusion_disabled_recipe = '''
project_name: 'op_fusion_disabled_test'
dataset_path: './tutorial_test_data/tutorial_test.jsonl'
export_path: './outputs/op_fusion_disabled_mode/res.jsonl'
np: 24
```

```
# we need to experiment for several times, so we need to close cache mana
# to guarantee all these exps run the OP indeed.
use_cache: false

# you can replace the model path with a local path on your machine
process:
  - image_aesthetics_filter:
      hf_scorer_model: shunk031/aesthetics-predictor-v2-sac-logos-ava1-l1
      min_score: 0.3
      max_score: 1.0
      mem_required: '1500MB'  # use the default cuda accelerator and set
  - image_aspect_ratio_filter:
  - image_nsfw_filter:
      hf_nsfw_model: 'Falconsai/nsfw_image_detection'
      score_threshold: 0.5
'''

with open('op_fusion_disabled_mode.yaml', 'w') as fout:
    fout.write(op_fusion_disabled_recipe)
```

Then we run this recipe:

In [2]:

```
!dj-process --config op_fusion_disabled_mode.yaml
```

```
2024-08-12 09:39:35 | WARNING  | data_juicer.config.config:405 - Cache man
agement of datasets is disabled.
2024-08-12 09:39:35 | WARNING  | data_juicer.config.config:416 - Set temp
directory to store temp files to [None].
2024-08-12 09:39:35 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/op_fusion_disa
bled_mode.yaml] into the work_dir [/root/projects/kdd_tutorial_notebooks/o
utputs/op_fusion_disabled_mode]
2024-08-12 09:39:35 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═════════════════════════╤════════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╕
│ key                     │ values
│
╞═════════════════════════╪════════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╡
│ config                  │ [Path_fr(op_fusion_disabled_mode.yaml, cwd=/ro
ot/projects/kdd_tutorial_notebooks)]                                │
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ hpo_config              │ None
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ project_name            │ 'op_fusion_disabled_test'
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ dataset_path            │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data/tutorial_test.jsonl'                                    │
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/op_fusion_disabled_mode/res.jsonl'                                 │
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ keep_hashes_in_res_ds    │ False
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ np                       │ 24
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ text_keys                │ 'text'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ image_key                │ 'images'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ image_special_token      │ '<__dj__image>'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ audio_key                │ 'audios'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ audio_special_token      │ '<__dj__audio>'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ video_key                │ 'videos'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ video_special_token      │ '<__dj__video>'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ eoc_special_token        │ '<|__dj__eoc|>'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ suffixes                 │ []
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ use_cache                │ False
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ ds_cache_dir             │ '/root/.cache/huggingface/datasets'
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ cache_compress           │ None
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ use_checkpoint           │ False
│
├──────────────────────────┼──────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
```

```
│ temp_dir                  │ None
│
├──────────────────────────┼────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ open_tracer               │ False
│
├──────────────────────────┼────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ op_list_to_trace          │ []
│
├──────────────────────────┼────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ trace_num                 │ 10
│
├──────────────────────────┼────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ op_fusion                 │ False
│
├──────────────────────────┼────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ process                   │ [{'image_aesthetics_filter': {'accelerator': N
one,                                                                  │
│                           │                              'any_or_all': 'a
ny',                                                                  │
│                           │                              'audio_key': 'au
dios',                                                                │
│                           │                              'cpu_required':
1,                                                                    │
│                           │                              'hf_scorer_mode
l': 'shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE', │
│                           │                              'image_key': 'im
ages',                                                                │
│                           │                              'max_score': 1.
0,                                                                     │
│                           │                              'mem_required':
'1500MB',                                                             │
│                           │                              'min_score': 0.
3,                                                                     │
│                           │                              'num_proc': 24,
│
│                           │                              'stats_export_pa
th': None,                                                            │
│                           │                              'text_key': 'tex
t',                                                                   │
│                           │                              'video_key': 'vi
deos'}},                                                              │
│                           │  {'image_aspect_ratio_filter': {'accelerator':
None,                                                                │
│                           │                                'any_or_all':
'any',                                                                │
│                           │                                'audio_key':
'audios',                                                              │
│                           │                                'cpu_require
d': 1,                                                                   │
│                           │                                'image_key':
'images',                                                             │
│                           │                                'max_ratio':
3.0,                                                                   │
│                           │                                'mem_require
d': 0,                                                                   │
│                         │                                'min_ratio':
0.333,                                                              │
│                         │                                'num_proc': 2
4,                                                                  │
│                         │                                'stats_export_
path': None,                                                       │
│                         │                                'text_key': 't
ext',                                                              │
│                         │                                'video_key':
'videos'}},                                                         │
│                         │ {'image_nsfw_filter': {'accelerator': None,
│
│                         │                        'any_or_all': 'any',
│
│                         │                        'audio_key': 'audios',
│
│                         │                        'cpu_required': 1,
│
│                         │                        'hf_nsfw_model': 'Falc
onsai/nsfw_image_detection',                                       │
│                         │                        'image_key': 'images',
│
│                         │                        'mem_required': 0,
│
│                         │                        'num_proc': 24,
│
│                         │                        'score_threshold': 0.
5,                                                                  │
│                         │                        'stats_export_path': N
one,                                                               │
│                         │                        'text_key': 'text',
│
│                         │                        'video_key': 'video
s'}}]                                                                 │
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/op_fusion_disabled_mode'                                          │
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────────────────────────────────┤
```

```
│ timestamp                 │ '20240812093935'
│
├───────────────────────────┼──────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ dataset_dir               │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data'                                                         │
├───────────────────────────┼──────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────┤
│ add_suffix                │ False
│
╘═══════════════════════════╧══════════════════════════════════════════════════
═════════════════════════════════════════════════════════════════════╛
2024-08-12 09:39:35 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
2024-08-12 09:39:35 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-12 09:39:35 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
2024-08-12 09:39:36 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-12 09:39:36 | INFO     | data_juicer.format.formatter:200 - There
are 10000 sample(s) in the original dataset.
Filter (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 70468.23
examples/s]
2024-08-12 09:39:36 | INFO     | data_juicer.format.formatter:214 - 10000
samples left after filtering empty text.
2024-08-12 09:39:36 | INFO     | data_juicer.format.formatter:237 - Conver
ting relative paths in the dataset to their absolute version. (Based on th
e directory of input dataset file)
Map (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 65612.99 exa
mples/s]
2024-08-12 09:39:37 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 10000 from 10000
2024-08-12 09:39:37 | INFO     | data_juicer.format.mixture_formatter:143
- There are 10000 in final dataset
2024-08-12 09:39:37 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
preprocessor_config.json: 100%|##########| 325/325 [00:00<00:00, 1.88MB/s]
config.json: 100%|##########| 724/724 [00:00<00:00, 5.90MB/s]
model.safetensors: 100%|##########| 343M/343M [00:26<00:00, 12.9MB/s]
2024-08-12 09:40:09 | INFO     | data_juicer.core.executor:164 - Processin
g data...
Adding new column for stats (num_proc=24): 100%|##########| 10000/10000 [0
0:07<00:00, 1388.96 examples/s]
image_aesthetics_filter_compute_stats (num_proc=24):   0%|          | 0/10
000 [00:00<?, ? examples/s]/usr/local/python310/lib/python3.10/site-packag
es/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download`
is deprecated and will be removed in version 1.0.0. Downloads always resum
e when possible. If you want to force a new download, use `force_download=
True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
```

```
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
```

```
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
```

```
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
```

```
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
image_aesthetics_filter_compute_stats (num_proc=24): 100%|##########| 1000
0/10000 [00:26<00:00, 374.28 examples/s]
image_aesthetics_filter_process (num_proc=24): 100%|##########| 10000/1000
0 [00:07<00:00, 1343.78 examples/s]
2024-08-12 09:40:51 | INFO     | data_juicer.core.data:193 - OP [image_aes
thetics_filter] Done in 42.616s. Left 9987 samples.
image_aspect_ratio_filter_compute_stats (num_proc=24): 100%|##########| 99
87/9987 [00:00<00:00, 10886.18 examples/s]
image_aspect_ratio_filter_process (num_proc=24): 100%|##########| 9987/998
7 [00:00<00:00, 54821.22 examples/s]
2024-08-12 09:40:54 | INFO     | data_juicer.core.data:193 - OP [image_asp
ect_ratio_filter] Done in 3.052s. Left 9955 samples.
2024-08-12 09:40:54 | WARNING  | data_juicer.utils.process_utils:70 - The
required cuda memory of Op[image_nsfw_filter] has not been specified. Plea
se specify the mem_required field in the config file, or you might encount
er CUDA out of memory error. You can reference the mem_required field in t
he config_all.yaml file.
image_nsfw_filter_compute_stats (num_proc=24):   0%|          | 0/9955 [0
0:00<?, ? examples/s]/usr/local/python310/lib/python3.10/site-packages/hug
gingface_hub/file_download.py:1132: FutureWarning: `resume_download` is de
precated and will be removed in version 1.0.0. Downloads always resume whe
n possible. If you want to force a new download, use `force_download=True
`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
image_nsfw_filter_compute_stats (num_proc=24): 100%|##########| 9955/9955
[00:17<00:00, 565.86 examples/s]
2024-08-12 09:41:13 | WARNING  | data_juicer.utils.process_utils:70 - The
required cuda memory of Op[image_nsfw_filter] has not been specified. Plea
se specify the mem_required field in the config file, or you might encount
er CUDA out of memory error. You can reference the mem_required field in t
he config_all.yaml file.
image_nsfw_filter_process (num_proc=24): 100%|##########| 9955/9955 [00:07
<00:00, 1391.43 examples/s]
2024-08-12 09:41:21 | INFO     | data_juicer.core.data:193 - OP [image_nsf
w_filter] Done in 26.626s. Left 9917 samples.
2024-08-12 09:41:21 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 72.298s.
2024-08-12 09:41:21 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-12 09:41:21 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 108.
90ba/s]
2024-08-12 09:41:21 | INFO     | data_juicer.core.exporter:140 - Export da
```

```
taset into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 164.
71ba/s]
```

As we can see, the data processing for the whole dataset without OP fusion costs **72.298s** in total. Three OPs cost 42s, 3s, 26s respectively, and they are run independently. All of them share the same subprocedure of loading images from the disk.

Then, let's enable the OP fusion:

In [3]:

```
op_fusion_enabled_recipe = '''
project_name: 'op_fusion_enabled_test'
dataset_path: './tutorial_test_data/tutorial_test.jsonl'
export_path: './outputs/op_fusion_enabled_mode/res.jsonl'
np: 24

# we need to experiment for several times, so we need to close cache mana
# to guarantee all these exps run the OP indeed.
use_cache: false

# enable the OP fusion
op_fusion: true

# you can replace the model path with a local path on your machine
process:
  - image_aesthetics_filter:
      hf_scorer_model: shunk031/aesthetics-predictor-v2-sac-logos-ava1-l1
      min_score: 0.3
      max_score: 1.0
      mem_required: '1500MB'  # use the default cuda accelerator and set
  - image_aspect_ratio_filter:
  - image_nsfw_filter:
      hf_nsfw_model: 'Falconsai/nsfw_image_detection'
      score_threshold: 0.5
'''

with open('op_fusion_enabled_mode.yaml', 'w') as fout:
    fout.write(op_fusion_enabled_recipe)
```

And run this recipe:

In [3]:

```
!dj-process --config op_fusion_enabled_mode.yaml
```

```
2024-08-12 10:12:57 | WARNING  | data_juicer.config.config:405 - Cache man
agement of datasets is disabled.
2024-08-12 10:12:57 | WARNING  | data_juicer.config.config:416 - Set temp
directory to store temp files to [None].
2024-08-12 10:12:57 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/root/projects/kdd_tutorial_notebooks/op_fusion_enab
led_mode.yaml] into the work_dir [/root/projects/kdd_tutorial_notebooks/ou
tputs/op_fusion_enabled_mode]
2024-08-12 10:12:57 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═════════════════════════╤═══════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╕
│ key                     │ values
│
╞═════════════════════════╪═══════════════════════════════════════════════
════════════════════════════════════════════════════════════════════╡
│ config                  │ [Path_fr(op_fusion_enabled_mode.yaml, cwd=/roo
t/projects/kdd_tutorial_notebooks)]                                 │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ hpo_config              │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ project_name            │ 'op_fusion_enabled_test'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ dataset_path            │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data/tutorial_test.jsonl'                                    │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/op_fusion_enabled_mode/res.jsonl'                                  │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ keep_hashes_in_res_ds     │ False
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ np                        │ 24
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ text_keys                 │ 'text'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ image_key                 │ 'images'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ image_special_token       │ '<__dj__image>'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ audio_key                 │ 'audios'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ audio_special_token       │ '<__dj__audio>'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ video_key                 │ 'videos'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ video_special_token       │ '<__dj__video>'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ eoc_special_token         │ '<|__dj__eoc|>'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ suffixes                  │ []
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ use_cache                 │ False
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ ds_cache_dir              │ '/root/.cache/huggingface/datasets'
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ cache_compress            │ None
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ use_checkpoint            │ False
│
├───────────────────────────┼────────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
```

```
│ temp_dir                  │ None
│
├───────────────────────────┼──────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ open_tracer               │ False
│
├───────────────────────────┼──────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_list_to_trace          │ []
│
├───────────────────────────┼──────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ trace_num                 │ 10
│
├───────────────────────────┼──────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ op_fusion                 │ True
│
├───────────────────────────┼──────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ process                   │ [{'image_aesthetics_filter': {'accelerator': N
one,                                                                │
│                           │                               'any_or_all': 'a
ny',                                                                │
│                           │                               'audio_key': 'au
dios',                                                              │
│                           │                               'cpu_required':
1,                                                                  │
│                           │                               'hf_scorer_mode
l': 'shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE', │
│                           │                               'image_key': 'im
ages',                                                              │
│                           │                               'max_score': 1.
0,                                                                    │
│                           │                               'mem_required':
'1500MB',                                                           │
│                           │                               'min_score': 0.
3,                                                                    │
│                           │                               'num_proc': 24,
│
│                           │                               'stats_export_pa
th': None,                                                          │
│                           │                               'text_key': 'tex
t',                                                                 │
│                           │                               'video_key': 'vi
deos'}},                                                            │
│                           │  {'image_aspect_ratio_filter': {'accelerator':
None,                                                             │
│                           │                                 'any_or_all':
'any',                                                              │
│                           │                                 'audio_key':
'audios',                                                             │
│                           │                                 'cpu_require
d': 1,                                                                  │
│                           │                                 'image_key':
'images',                                                             │
│                           │                                 'max_ratio':
3.0,                                                                  │
│                           │                                 'mem_require
d': 0,                                                                  │
│                         │                                  'min_ratio':
0.333,                                                               │
│                         │                                  'num_proc': 2
4,                                                                   │
│                         │                                  'stats_export_
path': None,                                                        │
│                         │                                  'text_key': 't
ext',                                                               │
│                         │                                  'video_key':
'videos'}},                                                          │
│                         │  {'image_nsfw_filter': {'accelerator': None,
│
│                         │                          'any_or_all': 'any',
│
│                         │                          'audio_key': 'audios',
│
│                         │                          'cpu_required': 1,
│
│                         │                          'hf_nsfw_model': 'Falc
onsai/nsfw_image_detection',                                        │
│                         │                          'image_key': 'images',
│
│                         │                          'mem_required': 0,
│
│                         │                          'num_proc': 24,
│
│                         │                          'score_threshold': 0.
5,                                                                   │
│                         │                          'stats_export_path': N
one,                                                                │
│                         │                          'text_key': 'text',
│
│                         │                          'video_key': 'video
s'}}]                                                                  │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/op_fusion_enabled_mode'                                            │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────────────────────────────────────────────────────┤
```

```
│ timestamp                 │ '20240812101256'
│
├──────────────────────────┼───────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ dataset_dir               │ '/root/projects/kdd_tutorial_notebooks/tutoria
l_test_data'                                                              │
├──────────────────────────┼───────────────────────────────────────────────────
──────────────────────────────────────────────────────────────────────┤
│ add_suffix                │ False
│
╘══════════════════════════╧═══════════════════════════════════════════════════
══════════════════════════════════════════════════════════════════════╛
2024-08-12 10:12:57 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
2024-08-12 10:12:57 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-12 10:12:57 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
2024-08-12 10:12:58 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-12 10:12:58 | INFO     | data_juicer.format.formatter:200 - There
are 10000 sample(s) in the original dataset.
Filter (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 69510.02
examples/s]
2024-08-12 10:12:58 | INFO     | data_juicer.format.formatter:214 - 10000
samples left after filtering empty text.
2024-08-12 10:12:58 | INFO     | data_juicer.format.formatter:237 - Conver
ting relative paths in the dataset to their absolute version. (Based on th
e directory of input dataset file)
Map (num_proc=24): 100%|##########| 10000/10000 [00:00<00:00, 64799.31 exa
mples/s]
2024-08-12 10:12:59 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 10000 from 10000
2024-08-12 10:12:59 | INFO     | data_juicer.format.mixture_formatter:143
- There are 10000 in final dataset
2024-08-12 10:12:59 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
2024-08-12 10:13:03 | INFO     | data_juicer.ops.op_fusion:113 - Ops are f
used into one op OpFusion:(image_aesthetics_filter,image_aspect_ratio_filt
er,image_nsfw_filter).
2024-08-12 10:13:03 | INFO     | data_juicer.core.executor:164 - Processin
g data...
Adding new column for stats (num_proc=24): 100%|##########| 10000/10000 [0
0:07<00:00, 1345.09 examples/s]
fused_op_compute_stats (num_proc=24):   0%|          | 0/10000 [00:00<?, ?
examples/s]/usr/local/python310/lib/python3.10/site-packages/huggingface_h
ub/file_download.py:1132: FutureWarning: `resume_download` is deprecated a
nd will be removed in version 1.0.0. Downloads always resume when possibl
e. If you want to force a new download, use `force_download=True`.
```

```
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
```

```
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/huggingface_hub/file_dow
nload.py:1132: FutureWarning: `resume_download` is deprecated and will be
removed in version 1.0.0. Downloads always resume when possible. If you wa
nt to force a new download, use `force_download=True`.
  warnings.warn(
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
```

```
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
```

```
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
```

```
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
/usr/local/python310/lib/python3.10/site-packages/torch/_utils.py:831: Use
rWarning: TypedStorage is deprecated. It will be removed in the future and
UntypedStorage will be the only storage class. This should only matter to
you if you are using storages directly.  To access UntypedStorage directl
y, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
fused_op_compute_stats (num_proc=24): 100%|##########| 10000/10000 [00:34<
00:00, 291.42 examples/s]
fused_op_process (num_proc=24): 100%|##########| 10000/10000 [00:07<00:00,
1318.92 examples/s]
2024-08-12 10:13:54 | INFO     | data_juicer.core.data:193 - OP [fused_op]
Done in 50.515s. Left 9917 samples.
2024-08-12 10:13:54 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 50.516s.
2024-08-12 10:13:54 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-12 10:13:54 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 100.
71ba/s]
2024-08-12 10:13:54 | INFO     | data_juicer.core.exporter:140 - Export da
taset into a single file...
Creating json from Arrow format: 100%|##########| 10/10 [00:00<00:00, 190.
08ba/s]
```

After OP fusion, the data processing for the whole dataset only costs **50.515s, less than 70% of the original time cost.** And these three OPs are fused into only ONE OP named "fused_op" here. It's worth noticing that the common subprocedure "loading images from disk" of these 3 OPs is acutally not a computation-intensive procedure, but over 30% time can be saved. According to our experiments, for some more computation-intensive subprocedure (e.g. sampling and extracting frames from videos), much more time (up to 50%) can be saved.

OP fusion can only be applied on Filters for now because only Filters are commutative so we can reorder Filters to fuse them more conveniently. Besides, OP fusion is disabled in default. Users should access the data recipe before enable the OP fusion to check how many OPs can be fused and decide whether to enable the OP fusion.

And when developing new OPs, if it has some common subprocedure with other OPs and we want make it fusible, we need to wrap this subprocedure in the implementation of OPs and try to load the results of them from context first. If there are new subprocedures, we also need to add a new type of intermediate context variable to the `InterVars` in `utils/constant.py` and register a new registry group for it in `ops/op_fusion.py`. For more details about this part, please refer to the DeveloperGuide document in Data-Juicer.

Conclusion

In this notebook, we learn how to apply several system optimizations on OPs, including CUDA support, OP fusion, and so on. And we compare the processing efficiency before and after these optimizations to understand the speed-up from them intuitively.