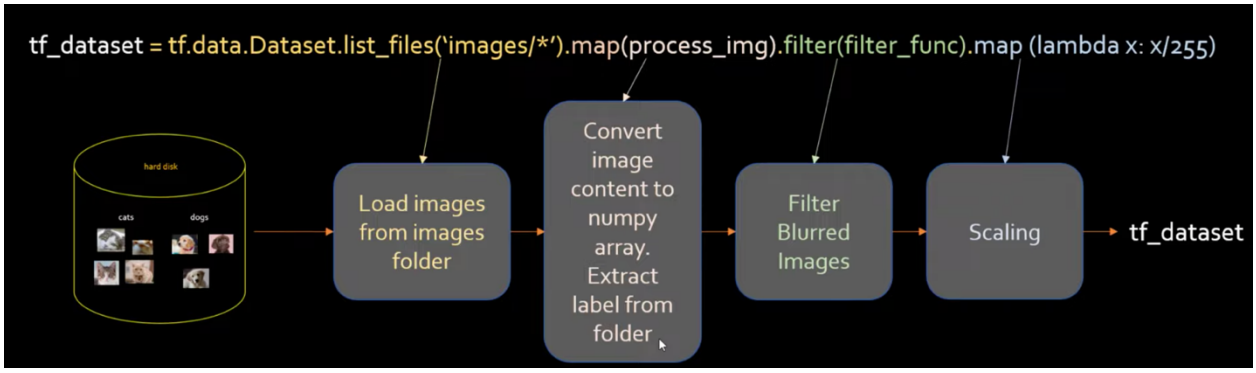
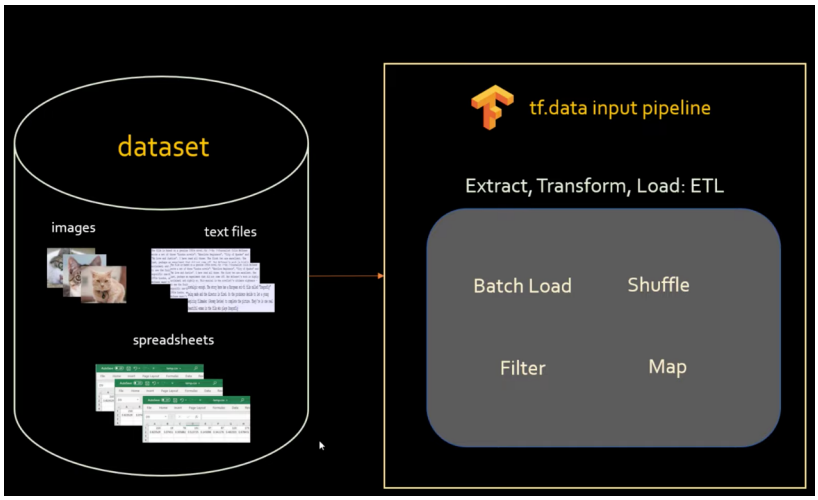
Tensorlfow input pipeline merupakan API untuk membangun input yang kompleks menjadi sederhana dan merupakan fungsi yang dapat digunakan secara berulang. Pada Gambar 1 adalah gambaran penerapan pipeline pada dataset gambar dan cara penggunaakn API nya.



Gambar 1. Pipeline image dataset

Penerapan pipeline pada dataset gambar tujuannya untuk menggabungkan gambar yang dipilih secara acak untuk pelatihan model. Sedangkan pipieline untuk model teks bisa melibatkan ekstraksi simbol dari data teks mentah, mengonversi menjadi penyematan pengidentifikasi dengan table pencarian, dan mengelompokkan urutan dengan panjang yang berbeda. Secara umum penerapan tf.data.input.pipeline untuk dataset gambar dan teks digambarkan pada Gambar 2.



Gambar 2. Blok diagram tf.data.input.pipeline

Tensorflow input pipeline mendukung banyak API untuk melakukan proses transformasi seperti fungsi filter (tf_dataset.filter(filter_func).
Dataset berupa images, *text file*, dan *spreedsheets* dilakukan tranformasi untuk dilakukan *Extract Transform*, dan *Load* atau sering disebut ETL.
Beberapa keuntungan tensorflow input pipeline diantaranya:

1. Dapat menangani dataset yang besar agar mudah digunakan
2. Dapat membaca format data yang berbeda
3. Dapat melakukan transformasi data yang kompleks

Dalam membaca data inputan tf.data.pipilene menggunakan numpy. Semua masukan dimuat pada memori, cara sederhana untuk membuat dataset dari data yaitu dengan mengkonversi terlebih dulu menjadi objek tf.Tensor dan menggunakan Dataset.from_tensor_slices().

Cara membuat pipeline dari data image:

1. Buat dataset dari irisan nama file dan label
2. Acak data dengan ukuran buffer yang sama dengan panjang dataset unutk memastikan shuffle yang baik
3. Parsing gambar dari nama file ke nilai piksel dan store menggunakan fungsi map pada python
4. Gunakan beberapa settingan untuk meningkatkan kecepatan pemrosesan (ini bersifat opsional)
5. Augmentasi data pada gambar
6. Ambil satu batch untuk memastikan bahwa setiap batch dijalankan dengan baik

Cara membuat pipeline dari text:

1. Files format
2. Zip seluruh dataset
3. Membuat vocabulary
4. Membuat padded batches
5. Hitung besaran kalimat
6. Penggunaan lanjutan dengan mengekstraksi karakter

Dari poin pembuatan pipeline pada data *image* dan text. Lantas apa yang dilakukan batch, repeat, dan shuffle dengan tensorflow dataset?

Contoh kasus :
dataset = [1,2,3,4,5,6]

How ds.shuffle() works
dataset.shuffle(buffer_size=3) akan mengalokasikan buffer ukuran 3 untuk memilih entri acak

How ds.repeat() works
ds.repeat() berguna pada saat terdapat kesalahan pada data entri, maka akan inisialisasi ulang dataset ke bentuk awal.

What will ds.batch() produce
ds.batch() akan mengambil entri batch_size pertama dan membuat batch dari data itu sendiri. Karena terdapat ds.repeat() sebelum batch, pembuatan data akan dilanjutkan. Tapi urutan elemen akan berbeda, karena ds.random().