

DATA SCIENCE

TUGAS 7 : Tensorflow/images

Nama : Sagita Pra Kosa (22120044)

Ferry Juliardi (22120026)

Erlangga Satria Mustofa (22120050)

Alysia Dewi Nur Masyithoh (22120052)

Muh. Haikhal Zahri Widya Putra (22120084)

MK : Data Science

Dosen : Nia Saurina, S.ST., M.Kom.

Soal :

<https://www.tensorflow.org/tutorials/images/cnn>

Tampilkan data yang ada

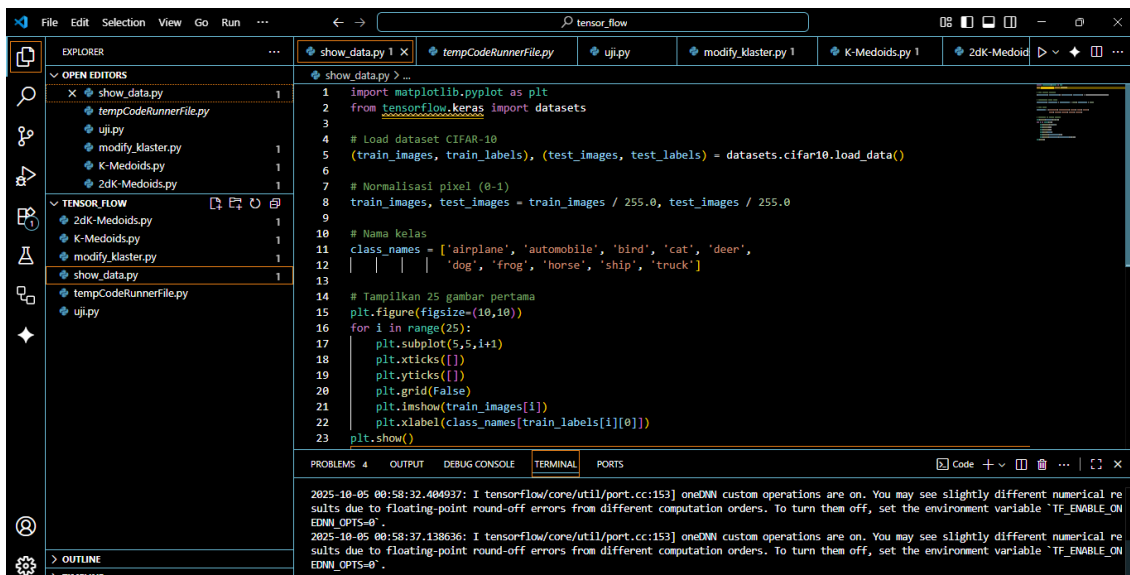
Modifikasi klaster

Bandingkan dan jelaskan hasil klastering K Medoids

Code Tugas 7 : https://github.com/gitapra0111/DataScience_TensorFlow_7

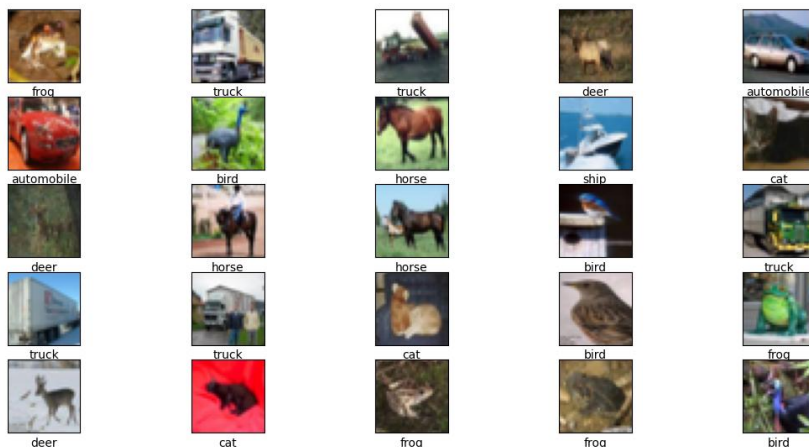
TUGAS 7 TENSORFLOW (CLUSTER)

1. Show_Data/Menampilkan Data



```
1 import matplotlib.pyplot as plt
2 from tensorflow.keras import datasets
3
4 # Load dataset CIFAR-10
5 (train_images, train_labels), (test_images, test_labels) = datasets.cifar10.load_data()
6
7 # Normalisasi pixel (0-1)
8 train_images, test_images = train_images / 255.0, test_images / 255.0
9
10 # Nama kelas
11 class_names = ['airplane', 'automobile', 'bird', 'cat', 'deer',
12 | | | 'dog', 'frog', 'horse', 'ship', 'truck']
13
14 # Tampilkan 25 gambar pertama
15 plt.figure(figsize=(10,10))
16 for i in range(25):
17     plt.subplot(5,5,i+1)
18     plt.xticks([])
19     plt.yticks([])
20     plt.grid(False)
21     plt.imshow(train_images[i])
22     plt.xlabel(class_names[train_labels[i][0]])
23     plt.show()
```

Hasil Data Yang di tampilkan



2. Modifikasi Cluster. Bandingkan dan jelaskan hasil klastering K Medoids

Bayangkan Anda memiliki sebuah kotak berisi 5.000 foto mainan yang berantakan (ini adalah data CIFAR-10 Anda). Anda ingin menyortir foto-foto ini ke dalam 10 tumpukan berbeda berdasarkan kemiripannya, tetapi Anda tidak tahu label setiap foto. Anda meminta komputer untuk melakukannya secara otomatis.

Tujuan Modifikasi :

Bayangkan Kita punya dua manajer, "Manajer K-Means" (kode asli) dan "Manajer K-Medoids" (kode modifikasi), yang ditugaskan untuk menyortir 5.000 foto ke dalam 10 tumpukan. Keduanya punya strategi yang berbeda:

1. Strategi Lama (K-Means):

- Untuk setiap tumpukan, Manajer K-Means akan membuat satu "gambar rata-rata" imajiner yang menjadi pusat tumpukan itu. Gambar ini adalah hasil pencampuran semua gambar di tumpukan itu, sehingga mungkin terlihat buram dan tidak menyerupai gambar asli manapun.

2. Strategi Baru (K-Medoids) - INILAH MODIFIKASINYA:

- Manajer K-Medoids tidak membuat gambar imajiner. Sebaliknya, ia akan memilih satu gambar asli dari dalam tumpukan untuk dijadikan "kapten" atau "perwakilan terbaik" (disebut Medoid). Gambar ini adalah gambar yang paling mirip dengan semua gambar lain di tumpukan yang sama.

Tujuan dari modifikasi ini adalah untuk bereksperimen dan membandingkan:

"Manakah strategi yang menghasilkan pengelompokan yang lebih baik atau lebih masuk akal untuk data gambar saya? Strategi yang pusatnya adalah 'gambar rata-rata' (K-Means), atau strategi yang pusatnya adalah 'gambar perwakilan asli' (K-Medoids)?"

Dengan K-Medoids, hasil clusternya seringkali lebih mudah diinterpretasikan karena pusatnya adalah contoh nyata, dan metodenya lebih tahan terhadap gambar-gambar aneh (outlier) yang bisa merusak perhitungan "rata-rata" pada K-Means.

Jadi, modifikasi klaster di sini adalah mengubah otak atau metode inti yang digunakan komputer untuk mengelompokkan data, dari K-Means menjadi K-Medoids, untuk melihat apakah kita bisa mendapatkan hasil yang lebih baik atau lebih berguna.

Bagian	Fungsi
K_RANGE = range(5, 16)	Mencoba K dari 5 sampai 15 (modifikasi dari versi awal yang hanya 10).
Loop for k in K_RANGE:	Melatih model K-Medoids untuk setiap nilai K.
silhouette_score, adjusted_rand_score, normalized_mutual_info_score	Menghitung metrik internal dan eksternal tiap model.
best_k dan best_clusters	Menyimpan konfigurasi K dengan hasil terbaik.
Plot	Menampilkan perbandingan semua metrik antar nilai K.

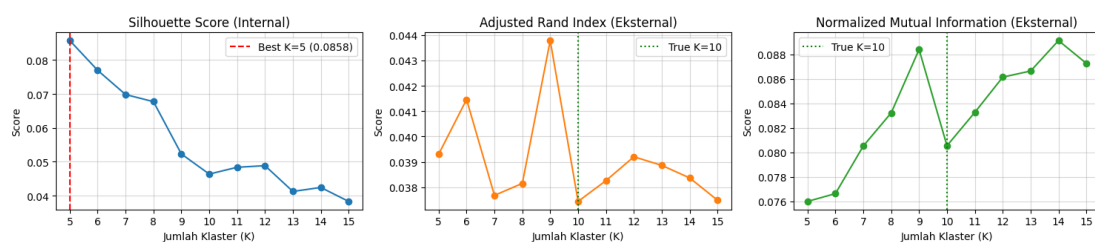
```
[INFO] Dipakai 5000 gambar untuk K-Medoids.
[INFO] Varian menjelaskan PCA(50): 0.846

[INFO] Mulai pengujian K-Medoids untuk K dari 5 hingga 15...
- K=5: Silhouette=0.0858, ARI=0.0393, NMI=0.0760
- K=6: Silhouette=0.0770, ARI=0.0415, NMI=0.0766
- K=7: Silhouette=0.0698, ARI=0.0377, NMI=0.0805
- K=8: Silhouette=0.0677, ARI=0.0382, NMI=0.0832
- K=9: Silhouette=0.0523, ARI=0.0438, NMI=0.0884
- K=10: Silhouette=0.0464, ARI=0.0374, NMI=0.0805
- K=11: Silhouette=0.0484, ARI=0.0383, NMI=0.0833
- K=12: Silhouette=0.0489, ARI=0.0392, NMI=0.0861
- K=13: Silhouette=0.0412, ARI=0.0389, NMI=0.0866
- K=14: Silhouette=0.0424, ARI=0.0384, NMI=0.0891
- K=15: Silhouette=0.0383, ARI=0.0375, NMI=0.0873
```

- Dari dataset CIFAR-10 yang totalnya 50.000 gambar, hanya pakai subset 5.000 gambar agar proses lebih ringan.
- Setelah direduksi dengan PCA ke 50 dimensi, sebanyak 84.6% informasi (varian) dari data asli masih dipertahankan.

Artinya hasil reduksi masih cukup representatif.

Perbandingan Metrik Evaluasi K-Medoids (K=5 hingga 15)



K=5 → kluster paling “rapi” secara struktur internal (Silhouette tertinggi).

K=9–14 → sedikit lebih cocok dengan label asli (NMI/ARI lebih tinggi).

Jadi, kalau fokus pada *pemisahan alami data*, **K=5** adalah pilihan terbaik.

```
[DEEP DIVE] K=5
=====
- Silhouette : 0.0858
- ARI        : 0.0393
- NMI        : 0.0760
- Jumlah item/cluster: {0: 1477, 1: 801, 2: 868, 3: 560, 4: 1294}
```

Hasil di atas menunjukkan perbandingan tiga metrik evaluasi klustering untuk setiap jumlah kluster () yang diuji.

1. Silhouette Score (Kualitas Internal Klaster)

Apa itu? Metrik ini mengukur seberapa **padat** klaster dan seberapa **terpisah** satu klaster dengan klaster lainnya. Nilai ideal mendekati 1.0.

Hasil:

- **Nilai Tertinggi:** dengan skor **0.0858**.
- **Tren:** Skor Silhouette cenderung **menurun drastis** seiring bertambahnya (dari 5 hingga 10).
- **Interpretasi:** Klustering dengan adalah yang **paling optimal** dari sudut pandang internal (kepadatan/pemisahan) di antara semua yang diuji. Namun, secara keseluruhan, skor sangat **rendah** (jauh dari 1.0), menunjukkan bahwa klaster-klaster yang terbentuk **tidak terlalu padat** dan **saling tumpang tindih** satu sama lain di ruang fitur (ruang 50 dimensi PCA).

2. Adjusted Rand Index (ARI) (Kesesuaian Eksternal)

Apa itu? Metrik ini mengukur seberapa **cocok** pembagian klaster dengan label kelas **asli** (seperti yang dilakukan oleh manusia: kucing dikelompokkan dengan kucing). Nilai ideal mendekati 1.0.

Hasil:

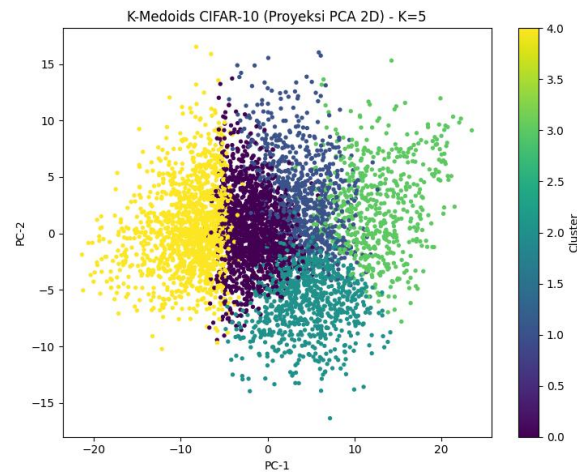
- **Nilai Tertinggi:** dengan skor **0.0438**.
- **Interpretasi:** Semua skor **sangat dekat dengan nol** (antara 0.0374 hingga 0.0438). Skor yang mendekati nol (atau negatif) menunjukkan bahwa hasil klustering **hampir acak** atau **tidak memiliki korelasi yang signifikan** dengan label kelas asli CIFAR-10. Klaster yang terbentuk tidak berhasil memisahkan gambar berdasarkan jenis objeknya (misalnya, Klaster 1 berisi campuran anjing, kucing, dan burung).

3. Normalized Mutual Information (NMI) (Kesesuaian Eksternal)

- **Apa itu?** Mirip dengan ARI, metrik ini mengukur seberapa banyak **informasi bersama** antara klaster dan label asli. Nilai ideal mendekati 1.0.

Hasil:

- **Nilai Tertinggi:** dengan skor **0.0891**.
- **Interpretasi:** Seperti ARI, skor NMI **sangat rendah** (jauh dari 1.0). Ini mengonfirmasi bahwa klastering *K-Medoids* pada fitur PCA 50D **gagal** menangkap struktur kelas asli yang ada pada data CIFAR-10.
- Menariknya, NMI tertinggi terjadi pada , bukan (jumlah kelas asli) atau (Silhouette terbaik). Ini menunjukkan kerumitan dan ketidakstabilan hasil klastering ini.



[Komposisi Label Asli per Cluster]										
	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
Cluster										
0	74	156	177	148	193	173	168	176	92	120
1	88	45	101	101	71	102	79	101	53	60
2	100	104	31	27	38	24	8	89	222	225
3	183	36	62	45	22	21	18	21	110	42
4	60	119	148	165	195	168	246	99	43	51

[Label dominan per cluster]

- Cluster 0: dominan 'deer' ($193/1477 = 13.1\%$)
- Cluster 1: dominan 'dog' ($102/801 = 12.7\%$)
- Cluster 2: dominan 'truck' ($225/868 = 25.9\%$)
- Cluster 3: dominan 'airplane' ($183/560 = 32.7\%$)
- Cluster 4: dominan 'frog' ($246/1294 = 19.0\%$)

[illegible]