Documents multimédia : description et recherche automatique

M2P GI, examen du 10 avril 2019, 2 heures, documents autorisés.

Les réponses doivent être justifiées de manière concise.

Exercice 1. Descripteurs histogrammes

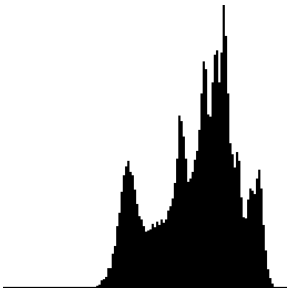

image 1

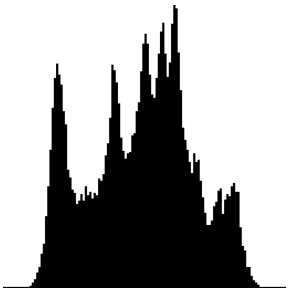

image 2

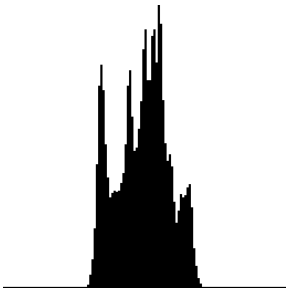

image 3


image 4

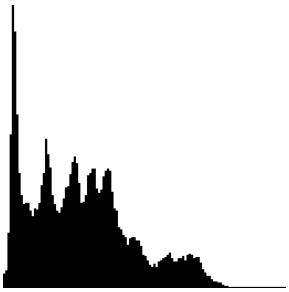

image 5


histogramme a


histogramme b


histogramme c


histogramme d


histogramme e

Question 1.1 : Associez les images ci-dessus avec leurs histogrammes, justifiez vos réponses. Les histogrammes sont monodimensionnels et portent sur l'intensité (luminance) de l'image.

Question 1.2 : Quelle est la dimension d'un descripteur d'image basé sur un histogramme de couleurs tridimensionnel dans l'espace LAB avec respectivement 8, 5 et 5 « bins » pour les composantes L, A et B ?

Question 1.3 : Même question si on calcule en plus cet histogramme par bloc avec un découpage des images suivant une grille bidimensionnelle avec 4 lignes et 4 colonnes ?

Question 1.4 : Parmi ces deux descripteurs (questions 3.2 et 3.3), lequel est sensible à une transformation par symétrie horizontale et lequel est robuste à une telle transformation ?

Question 1.5 : Quelles distances peuvent être utilisées pour évaluer la similarité de ces histogrammes dans le cadre d'une recherche par l'exemple ?

Question 1.6 : Ces distances sont-elles robustes par rapport à un changement de luminosité (brillance ou contraste) ?

Exercice 2. Classification par les k plus proches voisins et évaluation

On dispose d'une collection d'images exemples annotées selon la présence ou non d'un concept cible (une voiture par exemple). On dispose d'une autre collection d'images non annotées et pour lesquelles on voudrait produire des prédictions de présence ou non du même concept cible.

Les images de la collection sont nommées I_n, représentées par un descripteur x_n et annotée par l'étiquette a_n, avec $1 \leq n \leq N$. Les étiquettes a_n prennent la valeur 0 ou 1 selon que le concept cible est visible ou non dans l'image I_n. Les descripteurs x_n sont des vecteurs de nombre réels de dimension fixe d.

La prédiction est faite indépendamment pour chaque image de l'autre collection, on ne considèrera donc que le cas d'une seule image test I également représentée par un descripteur x de même type. On supposera que la distance euclidienne est adaptée pour la comparaison de descripteurs de ce type.

Question 2.1 : Donnez un algorithme pour prédire l'étiquette a correspondant à la visibilité du concept cible dans l'image I selon la méthode des k plus proches voisins (k nearest neighbors) avec $k = 5$ et une règle de décision basée sur une vote majoritaire.

Question 22 : La décision basée sur un vote majoritaire fournit seulement une réponse binaire qui ne permet pas de trier les images non annotées de la plus probablement positive à la plus probablement négative. Proposez une variante de l'algorithme précédent prenant en compte la distance de l'image à annoter à ses plus proches voisins trouvés pour produire un score de confiance réel permettant un tel classement. Plusieurs solutions sont possibles, une seule suffit.

Question 2.3 : Quelles mesures d'évaluation peut-on utiliser pour évaluer un système qui fournit une liste non ordonnée (un ensemble) de résultats ou seulement une valeur booléenne indiquant la pertinence ou non pertinence de chaque résultat (comme dans la question 4.1) ?

Question 2.4 : Quelles mesures d'évaluation peut-on utiliser pour évaluer un système qui fournit une liste ordonnée de résultats ou un score de confiance permettant un tri des résultats du plus probablement pertinent au moins probablement pertinent (comme dans la question 4.2) ?

Exercice 3. Apprentissage profond (deep learning)

La figure 1 donne une vue générale de l'architecture du réseau "VGG-16" :

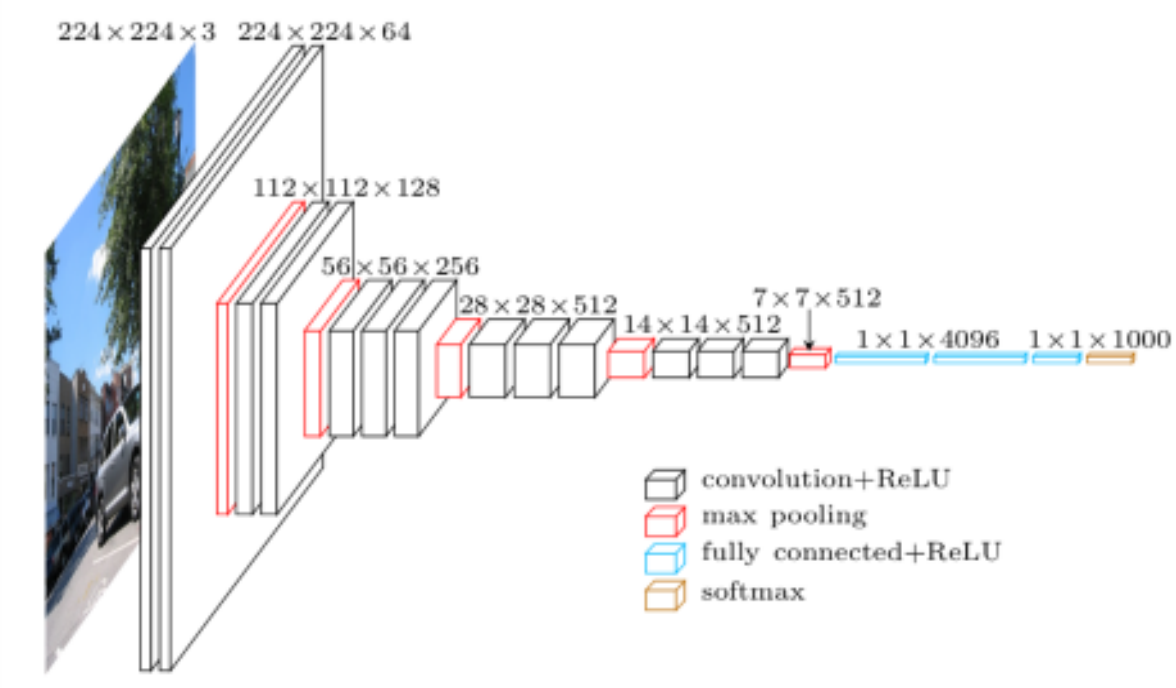


Figure 1 : réseau "VGG-16"

Question 3.1 : combien y a-t-il de couches (ou sous-couches) :

a) de convolution ?
b) de "pooling" ?
c) complètement connectée ?
d) de "softmax" ?

Question 3.2 : couche de "pooling" :

a) quel est le but de cette couche ?
b) comment est-elle implémentée ?
c) quelle autre opérateur associatif peut être utilisé ?
d) quelle alternative pourrait être utilisée à la place de cette couche ?

Question 3.3 : couche de "softmax" :

a) quel est le but de cette couche ?
b) comment est-elle implémentée ?
c) quelle est fonction de coût adaptée à cette couche de sortie ?

Question 3.4 : (sous-)couche "ReLU" :

a) quel est le but de cette couche ?
b) comment est-elle implémentée ?

Question 3.5 : combien y a-t-il de paramètres à apprendre dans :

a) la troisième couche de convolution (112×112×128 → 112×112×128) ?
b) la quatrième couche de "pooling" (28×28×512 → 14×14×512) ?
c) la première couche complètement connectée (7×7×512 → 4096) ?
d) la couche de "softmax" (1000 → 1000) ?

Rappel : toutes les convolutions sont de taille 3×3. Compter aussi les biais.

e) Lesquels de ces nombres dépendent de la taille de l'image d'entrée ?

Question 3.6 : combien y a-t-il de connexions dans :

a) la deuxième couche de convolution (224×224×64 → 224×224×64) ?
b) la deuxième couche complètement connectée (4096 → 4096) ?

Question 3.7 : combien d'opérations flottantes sont-elles effectuées dans :

c) la dernière couche de convolution (14×14×512 → 14×14×512) ?
d) la dernière couche complètement connectée (4096 → 1000) ?

Question 3.8 : comment les effets de bord des convolutions sont-ils gérés dans ce réseau ?