Documents multimédia : description et recherche automatique

M2P GI, examen du 7 avril 2021, 2 heures, calculatrice et documents autorisés.

Les réponses doivent être justifiées de manière concise.

Exercice 1. Descripteurs SIFT

On considère un système de représentation d'images par leur contenu basé sur les histogrammes de descripteurs locaux. Ces descripteurs locaux sont de type "color SIFT", c'est-à-dire des descripteurs SIFT calculés séparément sur les trois canaux R, G et B et concaténés en un seul descripteur. Un descripteur SIFT pour une couleur est calculé avec des histogrammes de gradients selon 8 directions et selon un découpage du voisinage en 4 fois 4 régions.

Question 1.1 : Quelle est la taille d'un tel descripteur local de type "color SIFT" ? Quel espace occupe-t-il en mémoire ou sur un disque (le stockage est en flottant simple précision et en binaire (4 octets par flottant)) ?

Question 1.2 : Le programme qui filtre les points d'intérêt en sélectionne en moyenne 500 par image. Quelle est la taille de la représentation complète d'une image par un ensemble de descripteurs SIFT locaux ? Quelle est la taille de la représentation complète d'une base d'un million d'images ?

Question 1.5 : On décide d'appliquer une analyse en composantes principale (PCA) sur ces descripteurs locaux. Quels avantages et quels inconvénients apporte-t-elle ? On décide de ne retenir que les 96 principales composantes. Quel est le taux de réduction des représentations ? Quelle est la taille finale de la représentation complète de la base ?

Question 1.4 : En quoi consiste l'agrégation de descripteurs locaux, quel en est l'intérêt et quelles sont les différentes méthodes pour le faire ?

Question 2.5 : On souhaite obtenir une représentation globale des images qui ne dépendent pas de la taille et de la complexité de celle-ci. On utilise pour cela une agrégation basée sur le calcul d'un histogramme des points SIFT selon un clustering préalablement effectué. Quelles sont les étapes (succinctement) nécessaires pour cette opération ?

Question 2.6 : On choisit de calculer cet histogramme de SIFT selon 4096 catégories. Quelle est la taille de la représentation complète de la base ?

Question 2.7 : On applique encore une réduction de la taille de la représentation globale par une autre analyse en composantes principale (PCA). On réduit la taille à 512 composantes. On compare cette solution à un calcul d'histogrammes directement sur 512 catégories. Quelle méthode est la plus simple ? Laquelle donnera selon vous les meilleurs résultats et pourquoi ? Qu'en est-il du volume nécessaire pour le stockage des représentations ?

Exercice 2. Machines à vecteurs de support

On donne dans un plan les points suivants comme échantillons de la classe positive : (-2,-4), (0,0), (2,2), (1,-1), (-3,-3), (2,1) et (4,2), et comme échantillons de la classe négative : (-3,4), (1,5), (-3,1), (-5,2) et (-1,5).

Question 2.1 : Dessiner les points correspondant aux classes sur un schéma avec des marques différentes pour chaque classe. Les classes sont-elles linéairement séparables ?

Question 2.2 : Dessiner la droite ayant la marge maximum avec les deux classes ainsi que les deux parallèles s'appuyant sur les points (ou vecteurs) de support.

Question 2.3 : Trouver l'équation de la droite de marge maximum sous la forme $\mathbf{w}^T.\mathbf{x}+b = 0$ avec $\mathbf{x} = (x,y)$ et $\mathbf{w} = (w_x,w_y)$ et telle que $\mathbf{w}^T.\mathbf{x}+b = +1$ et $\mathbf{w}^T.\mathbf{x}+b = -1$ correspondent respectivement aux droites passant par les points de support pour les classes positives et négatives. Notez que si l'on multiplie $\mathbf{w}$ et b par un même facteur, l'équation $\mathbf{w}^T.\mathbf{x}+b = 0$ décrit toujours la même droite.

Question 2.4 : Quelle est la fonction de décision de la machine à points (ou vecteurs) de support (SVM) apprise à partir des données fournies pour les deux classes ? Quelle est la marge du classificateur correspondant ?

Question 2.5 : Utilisez cette fonction pour justifier de la classification des points (4,-2), (-1,1), (1,2), (-2,1) et (-4,4). Classez ces points du plus probablement positif au plus probablement négatif.

Exercice 3. Apprentissage profond (deep learning)

Question 3.1 : En vous inspirant du code du réseau LeNet adapté du tutoriel pytorch pour des images "CIFAR" en couleur de taille 32×32 fourni ci-dessous, proposez le code d'un réseau de type "VGG" comprenant trois blocs de deux couches de convolution chacun, la première couche

comprenant 8 cartes ("feature maps") et les deux premières couches complètement connectées comprenant chacune 256 « neurones ».

```
import torch.nn as nn
import torch.nn.functional as F

class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(3, 6, 5, padding=0)
        self.pool = nn.MaxPool2d(2, 2)
        self.conv2 = nn.Conv2d(6, 16, 5, padding=0)
        self.fc1 = nn.Linear(16 * 5 * 5, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)

    def forward(self, x):
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        x = x.view(-1, 16 * 5 * 5)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x

net = Net()
```

On considère une couche de convolution qui prend en entrée 128 plans de taille 56×56 et qui produit en sortie 256 plans de la même taille avec des filtres de taille 3×3.

Question 3.2 : Combien y a-t-il de couches "neuronales" au total dans ce réseau (celui de la réponse à la question 3.1) et combien de chaque type ?

Question 3.3 : Combien y a-t-il de couches "élémentaires" ou "sous-couches" au total dans ce réseau (celui de la réponse à la question 3.1) et combien de chaque type ?

Pour les questions suivantes, la notation sera adaptée si vous avez des erreurs dans les questions 3.1 à 3.3 mais des réponses cohérentes sont attendues.

Question 3.4 : Combien y a-t-il de paramètres dans première couche de convolution ?

Question 3.5 : Combien d'opérations flottantes sont effectuées pour chaque image d'entrée dans cette première couche de convolution ?

Question 3.6 : Combien y a-t-il de paramètres dans la première et dans la dernière couche complètement connectée ?