Compte-Rendu

Deep Learning TP 1-2
Description d'images par SIFT et Bag of Words

Yining Bao **Hanshuo WANG** **Yongtao WEI**

1. Comparaison des méthodes

1.1. Méthodes utilisées (définition, visualisation)

On utilise principalement SIFT et Bag of Words dans ce TP.

SIFT signifie Scale Invariant Feature Transformation, qui est une description utilisée dans le domaine du traitement d'images. Cette description est invariante à l'échelle, peut détecter des points clés dans l'image et est un descripteur de caractéristique local.

Le modèle BOW, également appelé "sac de mots", dans la recherche d'informations, le modèle BOW suppose que pour un texte, l'ordre des mots, la grammaire et la syntaxe sont ignorés, et il n'est considéré que comme une collection de mots, ou une combinaison de mots.L'apparence de chaque mot dans le texte est indépendante et ne dépend pas de l'apparence d'autres mots. Il peut également être utilisé dans le domaine de l'image pour réaliser une classification d'image.

1.2. Leur pertinence

L'ensemble de l'exercice est divisé en trois parties. La première consiste à calculer le SIFT d'une image, qui est son information de point clé. La deuxième partie effectue une classification en kmeans des points clés calculés pour obtenir le dictionnaire. Chaque mot de ce dictionnaire exprime une sorte d'information dans une image. La troisième partie consiste à utiliser ce dictionnaire généré pour la prédiction d'images.

1.3. Leurs limites

L'algorithme SIFT présente également quelques lacunes. La méthode construit un vecteur à 128 dimensions à partir des points caractéristiques, puis fait correspondre le vecteur, de sorte que l'image doit rencontrer suffisamment de textures, sinon le vecteur à 128 dimensions construit ne sera pas trop distinctif, ce qui peut facilement conduire à des discordances, telles que Correspondance d'images, reconnaissance de carte des étoiles, etc., il n'y a pas de texture autour de ces points caractéristiques de l'image et l'algorithme SIFT est complètement invalide pour le moment.

2. Réponses aux questions

2.1. SIFT

2.1.1. Calcul du gradient d'une image

1. Montrer que les masques Mx et My sont séparables, c'est-à-dire qu'ils peuvent s'écrire $Mx = hy \times hx^T$ et $My = hx \times hy^T$ avec hx et hy deux vecteurs colonne de taille 3 a déterminer.

Par observation, on est capable de trouver

$$hx = \frac{1}{2}[-1 \quad 0 \quad 1], \quad hy = \frac{1}{2}[1 \quad 2 \quad 1]$$

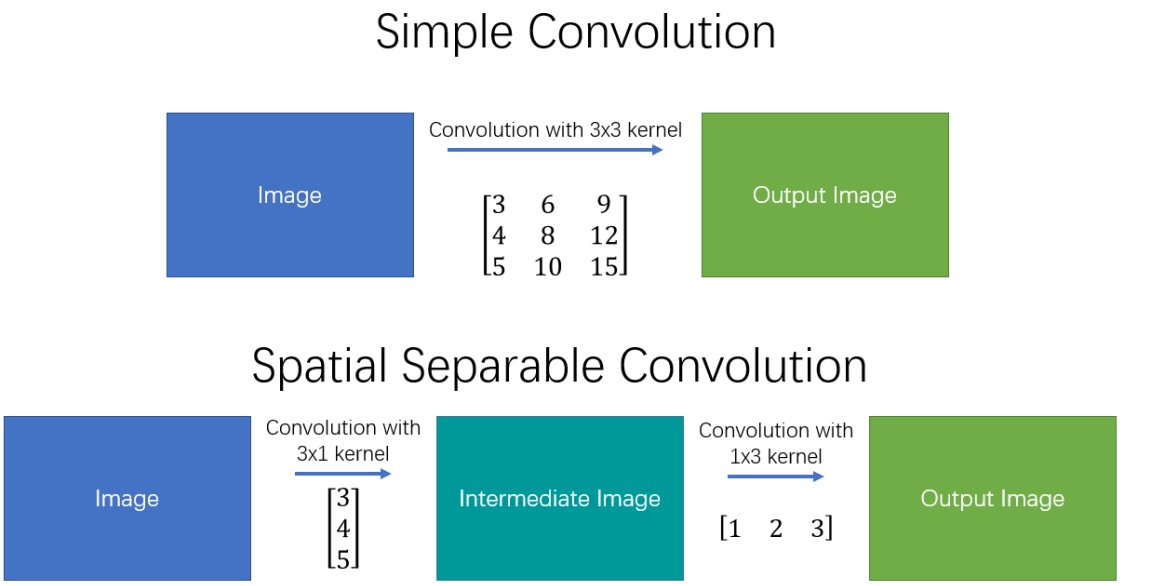
pour que

$$Mx = hy \times hx^T = \frac{1}{4}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

$$My = hx \times hy^T = \frac{1}{4}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

2. Quel est l'intérêt de séparer le filtre de convolution ?

Au lieu de faire une convolution avec 9 multiplications, nous faisons deux convolutions avec 3 multiplications chacune (6 au total) pour obtenir le même effet. Avec moins de multiplications, la complexité de calcul diminue, et le réseau peut fonctionner plus rapidement.



https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728

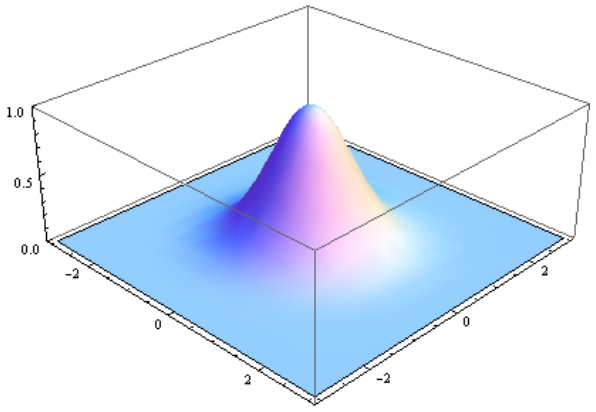
Conclusion

Dans cette section, nous utilisons la convolution séparable pour calculer le gradient de l'image. Il existe deux principaux types de convolutions séparables : les convolutions séparables spatiales et les convolutions séparables en profondeur. La convolution séparable spatiale est ainsi nommée parce qu'elle traite principalement les dimensions spatiales d'une image et d'un noyau : la largeur et la hauteur. (L'autre dimension, la dimension "profondeur", est le nombre de canaux de chaque image). Notre cas ici, le noyau de Sobel, est l'une des convolutions les plus célèbres qui peuvent être séparées dans l'espace, utilisé pour détecter les bords.

2.1.2. Calcul de la représentation SIFT d'un patch

3. Quel est le rôle de la pondération par masque gaussien ?

Le filtrage gaussien consiste à appliquer une moyenne pondérée à l'ensemble de l'image, où la valeur de chaque point de pixel est obtenue par une moyenne pondérée de sa propre valeur et de celle des autres pixels de son voisinage. Le filtrage gaussien présente une distribution gaussienne horizontalement et verticalement, il est donc en mesure de donner plus de poids au point central afin de maintenir sa propre identité, tout en rendant l'opération de moyennage moins influente sur les pixels environnants. Ce processus de flou préserve l'effet de bord plus haut que les autres masques de flou, tels que les filtres de flou d'égalisation.



Dans notre cas, nous utilisons un masque gaussien pour Gn, qui réduit le bruit dans les données de gradient et augmente leur lisse tout en conservant le maximum d'informations sur le gradient. Par conséquent, avec la fonction de suppression du bruit et des images floues, SIFT maintient également un certain degré de stabilité contre les changements d'angle de vue, les transformations affines et le bruit.

4. Expliquez le rôle de la discrétisation des orientations.

Les directions du gradient d'origine sont des variables continues dans la gamme des nombres réels de 0° à 360°, et si nous ne les discrétions pas, les coordonnées horizontales de l'histogramme R_{enck} auraient un nombre infini de valeurs, ce qui serait impossible à réaliser. Nous devons donc discrétiser les directions. Et le diviser en 8 directions est juste parfait. Si on le divise en trop de directions, le calcul prendra beaucoup de temps, et si on le divise en seulement 4 directions, le résultat ne sera pas assez précis.

5. Justifiez l'intérêt des différents post-processings appliqués au SIFT.

Après le post-processing, nous avons supprimé les vecteurs de gradient trop petits, qui sont les bruits pour trouver les points clés. La réduction de ce bruit réduit également la quantité de données à traiter en conséquence.

6. Expliquez en quoi le principe du SIFT est une façon raisonnable de décrire numériquement un patch d'image pour faire de l'analyse d'image.

1. Après SIFT, l'image est traité selon le gradient de chaque pixel y compris la direction et l'amplitude, il présente le caractère de image donc il peut maintenir le même caractère après le changement de rotation, l'échelle, luminosité et aussi il y a peu d'influence sur le changement de perspective, le transformation affine et le bruit grâce à l' utilisation de masque gaussien.

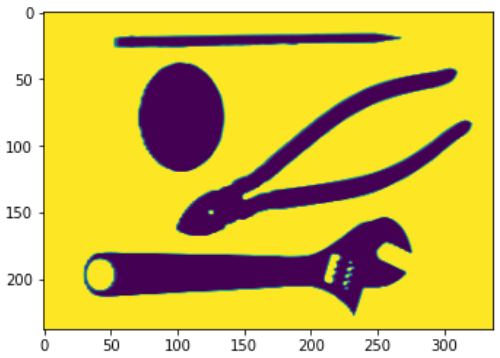
2. L'image de sift est unique. Deux images différentes ne peuvent pas avoir la même image de sift, et des images similaires auront des images de sift similaires.

3. que même quelques objets peuvent générer un grand nombre de vecteurs de caractéristiques SIFT.

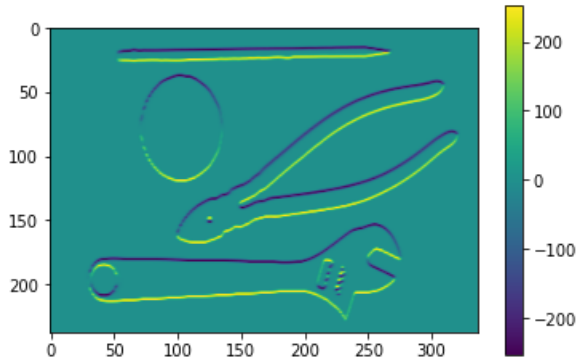
4. Il possède un bon caractère distinctif (caractère distinctif), est riche en informations et convient à une mise en correspondance rapide et précise dans une grande base de données de caractéristiques.

7. Interpréter les résultats que vous avez obtenus dans cette partie.

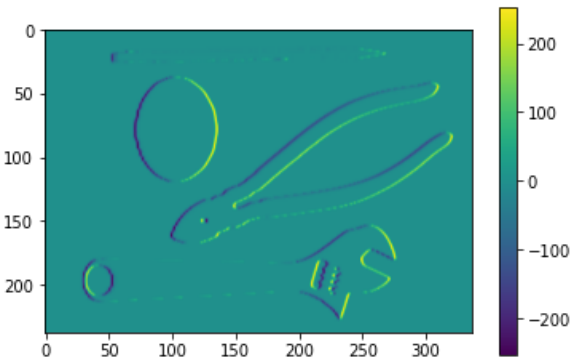
Nous affichons d'abord l'image originale :



Nous calculons le gradient Ix et Iy de chaque pixel par la fonction **def compute_grad()**, et obtenons les figures suivante :



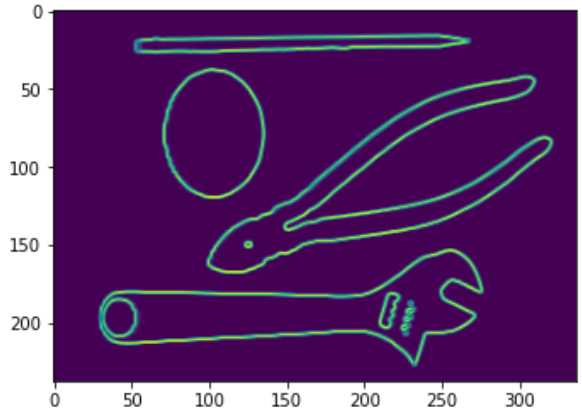
Ix



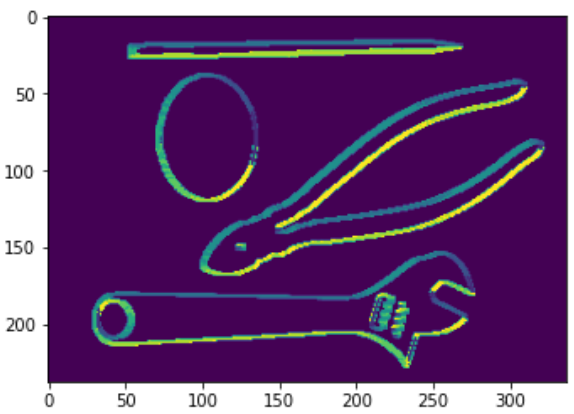
Iy

Selon les résultats, on a vu que dans Ix, les bords dans direction x est affiché plus marqué parce que le traitement de gradient est dans la direction x; et dans Iy, les bords dans la direction y est affiché plus marqué.

Ensuite, nous utilisons la fonction **compute_grad_mod_ori()** pour calculer Gn et Go, et obtenons les figures suivante :



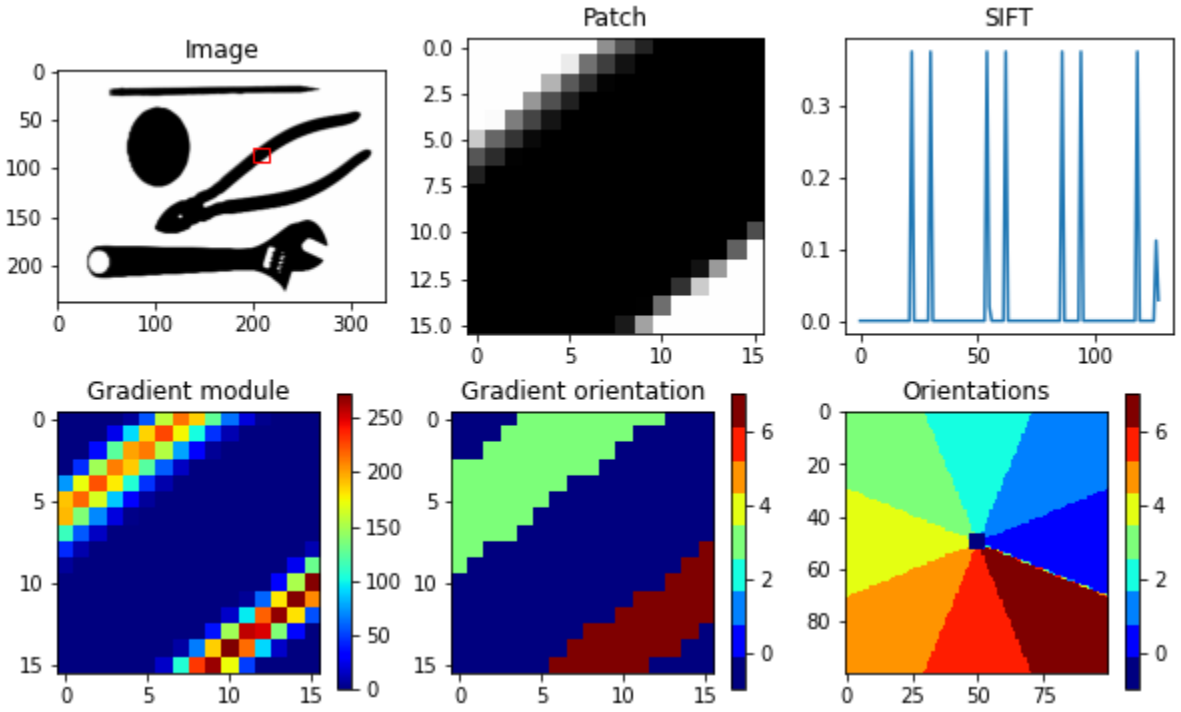
Gn



Go

Gn à gauche est la magnitude du gradient, et à partir du diagramme, nous pouvons voir que seuls les bords de l'objet sont soulignés, c'est parce que seuls les bords de l'objet ont un changement de gradient, les restes sont des blocs de couleur uniforme sans changement de gradient.
Et, puisque la bordure est essentiellement de la même couleur (vert), nous avons pu vérifier que les gradients à différentes positions de la bordure variaient de façon similaire, ce qui est le cas dans l'image originale.

Go à droite indique la direction du gradient, et parce que le gradient change dans différentes directions, nous pouvons voir que les bords de l'objet sont colorés différemment. Regardons le stylo en haut de l'image. Il a un bord inférieur vert et un bord supérieur bleu, car la direction du gradient du bord de la ligne ou du bord supérieur est la même et le gradient entre eux diffère de 180°.



Enfin, avec la fonction **compute_histogram(g_n, g_o)** et la fonction **compute_sift_region(Gn, Go, mask)**, nous obtenons l'image sift, et les points qui montent soudainement dans l'image sift correspondent aux points clés de l'image originale.

2.2. Dictionnaire visuel

8. Justifiez la nécessité du dictionnaire dans le processus général de reconnaissance d'image que nous sommes en train de mettre en place.

Le dictionnaire concerne plusieurs mot clés, et chaque mot clé est une barycentre qui est une représentation d'un type de descripteur SIFT pour ce cluster. Pendant le processus de reconnaissance d'image, on calcule plusieurs descripteurs SIFT et alors on classe a des SIFT type et pendant la recherche la plus proche de SIFT, on peut trouver aussi les features d'image. Donc en général, les SIFT sont des mot clés et plusieurs mot clés qui forment un image.

9. Considérant les points {xi}i=1..n assignés à un cluster c, montrer que le centre du cluster qui minimise la dispersion est bien le barycentre (moyenne) des points xi :

$$\min_c \sum_i ||x_i - c||_2^2$$

Le barycentre c cluster est le moyenne de l'ensemble des points xi de cluster. On peut considérer que xi-c est l'erreur entre le point et le cluster, donc c'est la moyenne de tous les points xi qui minimisent la somme des erreurs.

10. En pratique, comment choisir le nombre de clusters "idéal" ?

Le nombre de clusters est un hyperparamètre à choisir avec plusieurs méthodes ,soit Elbow méthode qui concentre sur le pourcentage des variances, soit la cross-validation en évaluant les

classification et etc. Alors, il s'agit d'un apprentissage non supervisé, les labels pour le training set sont inconnus et c'est difficile d'évaluer les performances du modèle.

11. Pourquoi l'analyse des éléments du dictionnaire doit-elle se faire à travers des exemples de patchs et pas directement ?

Le barycentre du modèle est un moyen pour l'ensemble des SIFT de clusters, il est difficile d'observer directement les éléments du dictionnaire, l'affichage de SIFT est difficile à comprendre, car on ne peut pas reconstruire le patch avec les descripteurs. Donc en associant le barycentre et des régions d'image, on peut observer le cluster représenter quel région d'une image.

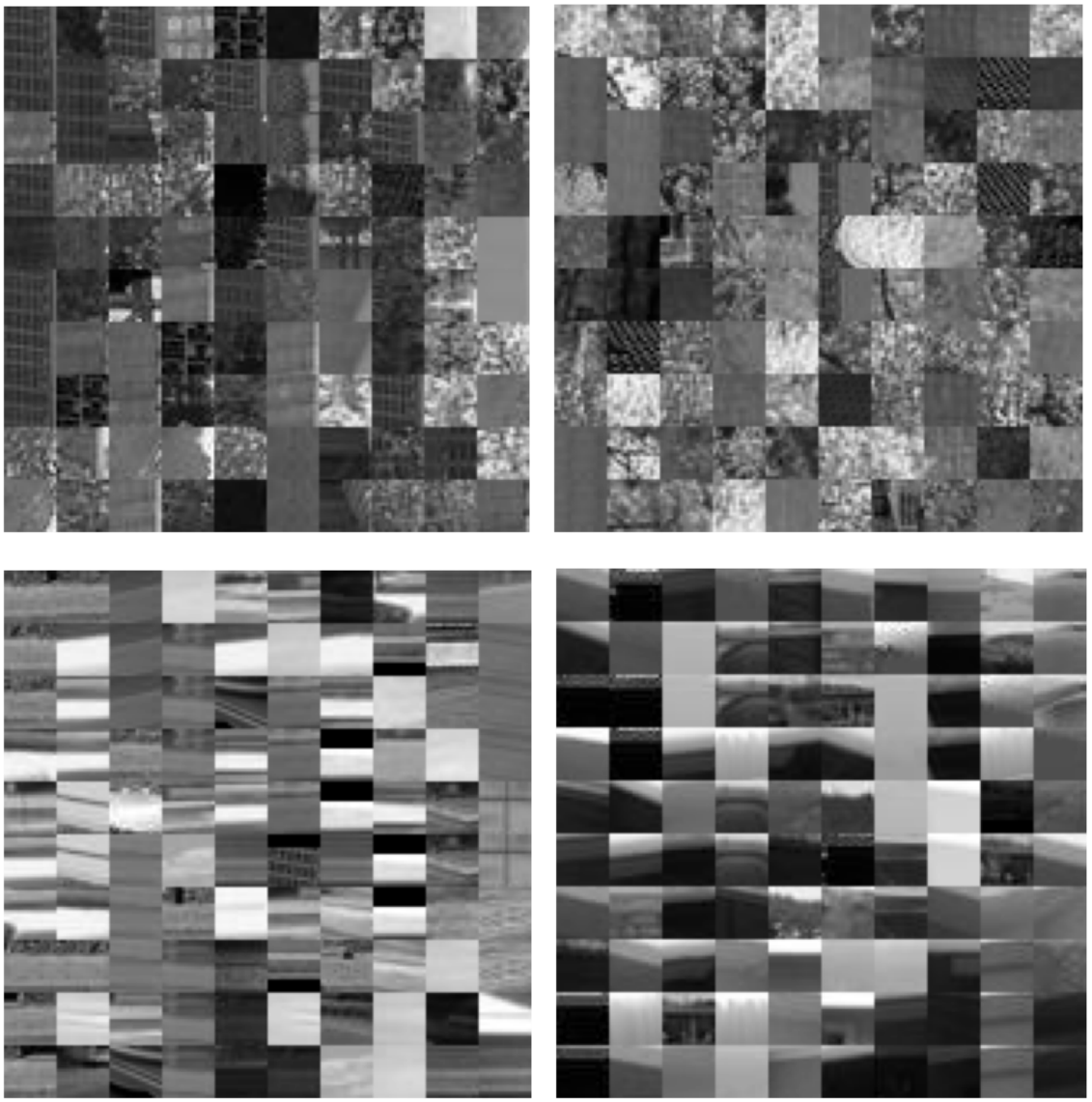
12. Reportez dans votre rapport les visualisations des patches proches des centres, et commentez les résultats que vous aurez obtenus.

Tout d'abord, on calcule kmeans avec 1000 clusters auxquels on ajoute le vecteur nul. Alors, on prend aléatoirement 30 images et on récupère les régions et SIFT, puis on calcule le barycentre et on cherche le cluster le plus proche.

Voici le résultat de 100 régions d'image aléatoire:

On peut observer le panneau du magasin, la roue etc.

Alors avec le barycentre, on trouver alors les top 100 des régions d'image associés et aléatoires, Voici quelques résultats.



2.3. Bag of Words (BoW)

13. Finalement, que représente concrètement notre vecteur z pour une image ?

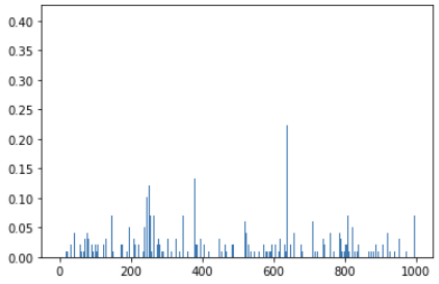
z représente un histogramme. Nous avons un total de 1000 clusters, notre BOW contient donc 1000 mots. Pour une image, on calcule la distance correspondant aux points clés et aux mots du dictionnaire, et on obtient le nombre de fois où chaque mot apparaît dans l'image, afin de réaliser le classement de l'image.

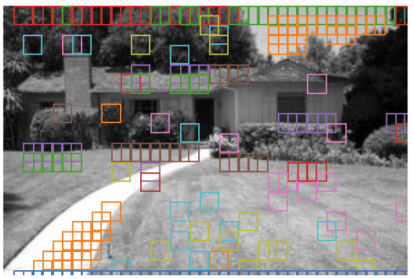
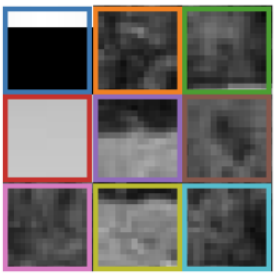
14. Montrez et discutez les résultats visuels obtenus.

Le chiffre de gauche représente hi. Puisque nous avons sélectionné 1001 groupes lors de l'utilisation de l'algorithme k-means, nous avons 1001 éléments dans l'histogramme, et l'histogramme indique le nombre de fois où chaque élément apparaît. Et ce chiffre a été normalisé.

L'image du milieu est l'image et le patch de notre choix. Les différentes couleurs du patch indiquent un mot du dictionnaire auquel il correspond.

L'image de droite montre les différents mots du dictionnaire distingués par leur couleur.



15. Quel est l'intérêt du codage au plus proche voisin ? Quel(s) autre codage pourrait-on utiliser ?

Cette forme de codage peut calculer à quel mot du dictionnaire appartient chaque point clé. On calcule la distance de chaque point clé au mot le plus proche dans le dictionnaire, et dessine un histogramme sous la forme de 0 et 1. 0 signifie que ce point clé n'appartient pas au jième mot, et 1 signifie que ce point clé appartient à ce mot.

16. Quel est l'intérêt du pooling somme ? Quel(s) autre pooling pourrait-on utiliser ?

La mise en commun des sommes consiste à superposer les histogrammes générés par tous les points clés ensemble. Cela permet de distinguer complètement le type d'image et les informations qu'elle contient.

Les autres méthodes comprennent Max Pooling, Average Pooling etc. Ces méthodes peuvent réduire la quantité de données et augmenter l'efficacité de calcul.

17. Quel est l'intérêt de la normalisation L2 ? Pourrait-on utiliser une autre normalisation ?

L'opération de normalisation L2 place tous les points entre 0 et 1, ce qui facilite leur calcul et la comparaison de la relation entre eux. Les autres méthodes de normalisation incluent :

L1 Normalisation

- Normalisation linéaire
- normalisation min-max
- Normalisation du score Z