Clasificación de imágenes utilizando la estrategia PHOW en las bases de datos ImageNet y Caltech101

Rubén Darío Bohórquez Cortázar
Universidad de los Andes
Bogota, Colombia
rd.bohorquez10@uniandes.edu.co

Javier Felipe Cifuentes
Universidad de los Andes
Bogota, Colombia
jf.cifuentes10@uniandes.edu.co

Abstract

En este laboratorio se exploró la estrategia para el reconocimiento de objetos PHOW (Pirámide de histogramas de palabras visuales) y su aplicación para la clasificación de imágenes en las bases de datos Caltech 101 e ImageNet. Para realizar esto, se utilizó la biblioteca de funciones VLFeat, desarrollada por Andrea Vedaldi y Brian Fulkerson. Adicionalmente, partiendo de esta estrategia de reconocimiento se desarrolló una función de clasificación basada en la implementación de un modelo óptimo creado a partir del uso de los mejores parámetros para el reconocimiento en PHOW. Para determinar el mejor modelo se realizaron diversos experimentos variando los parámetros del método y evaluando el ACA para cada uno, después de esto, se seleccionaron los parámetros con mayor ACA. Se determinó que los mejores parámetros para ImageNet correspondían a los presentados en la tabla3. Después de esto se realizó un entrenamiento y prueba de la función utilizando los mejores parámetros en PHOW en la base de datos ImageNet, se obtuvó un ACA de 19.72 .Finalmente, se pudo concluir que la estrategia PHOW estudiada esta optimiza para la base de datos Caltech101. Adicionalmente, la selección correcta de los parámetros define el desempeño de la estrategia PHOW, lo que implica que para trabajos futuros se debe realizar un estudio más detallado de los mismos. Por ultimo, se debe resaltar que la estrategia PHOW es muy util para la clasificación de clases en las cuales las imágenes tengan una mayor respuesta al descriptor SIFT.

1. Introducción

La clasificación de imágenes es una de las áreas de la visión por computador más estudiadas en la actualidad, debido a la gran diversidad de problemas de clasificación y reconocimiento de objetos asociados a múltiples áreas de conocimiento tanto en el ámbito académico como en la industria. De forma general, la clasificación de imágenes se basa en el reconocimiento de un objeto específico o de cierto grupo de objetos para la asignación de una etiqueta con la cual se determina a que categoría o categorías pertenece la imagen [1]. Debido a la importancia de la clasificación en el ámbito de la visión, se han desarrollado múltiples algoritmos para la solución de estos problemas y construido diversas bases de datos para la evaluación de los mismos. Entre estos algoritmos se encuentra la estrategia basada en extracción de características PHOW (Pirámide de histogramas de palabras visuales)[2] entre las bases de datos más relevantes, se encuentran Caltech 101 e ImageNet. Debido a la importancia de la estrategia PHOW y de las bases de datos Caltech 101 e ImageNet en la historia del estudio de los problemas de clasificación en el ámbito de la visión por computador, se establece como el objetivo principal de este articulo el estudiar con detalle los distintos parámetros asociados a PHOW para el desarrollo de un modelo adecuado para la clasificación de imágenes en las distintas bases de datos anteriormente nombradas. Para esto, se va a hacer uso de la biblioteca de funciones VLFeat, desarrollada por Andrea Vedaldi y Brian Fulkerson.

2. Materiales y métodos

2.1. Caltech 101

Caltech 101 es una base de datos creada en septiembre de 2003 por Fei-Fei Li, Marco Andreetto, y Marc 'Aurelio Ranzato, para el estudio y desarrollo de algoritmos de clasificación, reconocimiento y categorización. Esta base de datos consta de un total de 9146 imágenes (300 x 200 pixeles) distribuidas en 101 categorías. Es una base de datos desbalanceada en la cual cada categoría tiene entre 40 y 800 imágenes. En la figura1 se pueden observar algunos ejemplos de las imágenes que se pueden encontrar en esta base de datos [3].

Figure 1. Imágenes en la base de datos Caltech101

2.2. PHOW

La estrategia basada en pirámide de histogramas de palabras visuales (PHOW) fue propuesta por Anna Bosch, Andrew Zisserman y Xavier Muñoz en el 2007. Este algoritmo realiza un muestreo denso de puntos con espacio de M pixeles (ventana) a distintas escalas definidas (pirámide) basado en el descriptor SIFT. Esta estrategia está constituida por dos grandes etapas, la primera se basa en la construcción del vocabulario de palabras visuales de una base de datos de imágenes. Para realizar esto, se extraen las características visuales de todas las imágenes con algún método de extracción de características (SIFT). Luego se implementa un algoritmo de agrupamiento (clustering), usualmente se utiliza k-means, el cual crea grupos de características visuales similares entre sí. Se debe resaltar que los centroides de los clusters son las palabras visuales. La segunda etapa es la representación de la imagen mediante un histograma de palabras visuales (HOW). Este histograma se construye mediante una cuantificación de las características extraidas de la imagen con el vocabulario de palabras visuales construido en la primera etapa. Se asigna cada característica a la palabra visual más cercana. Finalmente, con los histogramas de palabras visuales se construye un clasificador. El clasificador más utilizado es la Máquina de Soporte Vectorial o Support Vector Machine (SVM). El entrenamiento de este clasificador consiste en encontrar el hiperplano que maximice el margen de separación entre las dos clases.[1][2]

Se realizó clasificación de imágenes en dos bases de datos con PHOW. Par este fin se utilizó la biblioteca vl_feat de Matlab. En la presente se cargó el código llamado $phow_caltech101.m$, el cual genera la clasificación automática para la base de datos de caltech. Esta se modificó para poder utilizarla en la base de datos de imageNet200. Se realizaron modificaciones en los parámetros de entrada de la función, de manera que se pudiera constatar el efecto de cada parámetro en el ACA. Con este objetivo, se corrió el código varias veces, cambiando solo un parámetro, dejando los otros como los especificados por defecto. Posteriormente, se entrenó un modelo que tuviera los parámetros más efectivos, y se procedió a evaluar test. Finalmente se buscaron las clases con mayor y menor desempeño.

3. Resultados

El resultado para los parámetros de default para la base de datos de imageNet200 es **18.8235%**, mientras que para la base de datos de caltech101 es **68.1046%**. Posteriormente se cambiaron los valores de default para comparar los hiperparametros. Los valores de default escogidos fueron:

Parámetro	**Valor**
C	10
Número de imagenes de test	25
Número de imagenes de train	25
Número de clases	200
Número de palabras	600
División espacial X/Y	[2 4]

Table 1. Hiperparámetros por defecto

A partir de estos valores se alteraron los diferentes hiperparámetros obteniendo diferentes resultados. Se muestran estos en la tabla 2

Parámetro	**Valor**	**ACA**
Default		16.8
C	5	18.38
C	8	18.38
C	15	18.34
Número de imagenes de test	10	21.3
Número de imagenes de test	40	13.725
Número de clases	50	24.96
Número de clases	100	21.92
Número de palabras	300	16.8
Número de palabras	900	21.36
División espacial X/Y	[2 4]	18.38
División espacial X/Y	[3 5]	18.66
División espacial X/Y	[1 2]	17.88

Table 2. Resultados numéricos obtenidos con el cambio de hiperparámetros

Parámetro	**Valor**
C	10
Número de imagenes de test	10
Número de clases	50
Número de palabras	900
División espacial X/Y	[2 4]

Table 3. Mejores valores de hiperparámetros para ImageNet

Del set de test, se hallaron las clases con mayor y menor desempeño.
Las clases con mayor desempeño:

- Web site
- Convertible
- Bookcase

Por otro lado las clases mas dificiles fueron:

- American Staffordshire terrier
- Bedlington terrier
- Chesapeake Bay retriever
- Chihuahua
- Labrador retriever
- Scottish deerhound
- barrow
- bullfrog
- cellular telephone
- green mamba
- hog
- ladle
- malamute
- soft coated wheaten terrier
- swimming trunks
- thunder snake

Finalmente, se obtuvo el resultado del test set. Se obtuvo un ACA final de **19.72%** para la base de datos de imageNet200.

4. Análisis de resultados

Despues de realizar los cambios sobre la función para encontrar como se afecta el resultado de los hiperparametros se encontró que algunos presentan un efecto más notorio que los demás. Por ejemplo, alterar el factor C, necesario para el SVM, no altera en gran medida el ACA obtenido, a diferencia de cambiar la relación entre el número de imágenes de test y de train, el cual modifico el ACA casi en 10%. Dentro de los parámetros más importantes se encuentran, el número de imágenes de test, así como el número de imágenes de train, el número de clases, el número de palabras. Parámetros como el SpatialX/Y o C, no tiene un efecto demasiado notorio en el resultado obtenido. Existen otros parámetros intrínsecos dentro de la función, tales como el tamaño de la ventana deslizante o el número de bins. Sin embargo, estos se encuentran escondidos dentro de las subfunciones de phow.

Como se mostró anteriormente la base de datos de caltech tiene un ACA de 68%, mientras que el ACA de la base de datos imageNet es de 18.8%. Se nota una gran diferencia entre los dos resultados, debido a una serie de factores. El primero de los cuales es que la base de datos difiere en el número de clases presentes. Mientras que clatech101, tiene 101 categorías, la base de datos de imageNet tiene 200 categorías. Para probar con las condiciones por defecto fue necesario despreciar cerca de la mitad de las categorías. Adicionalmente el código está diseñado originalmente para funcionar en la base de datos de caltech, por lo cual presenta factores y parámetros, que maximizan el desempeño de la base de datos, los cuales no van a funcionar opimamente en una base de datos diferente.

De los parámetros alterados, es posible denotar el efecto que se tiene al modificar estos. Primero el número de imágenes de train y de test, presentan un efecto notorio en el resultado. Como se muestra, el total de imágenes siempre es 50, por lo que, si se tiene 10 imágenes de test, ahí 40 imágenes de train. Es de notar que, al disminuir el número de imágenes de test, disminuye el ACA, dado que se tiene menos oportunidad de error. Por otro lado, el parámetro de numero de palabras presenta un efecto de incremento en el ACA, conforme este aumenta. Esto se debe a que el descriptor de las imágenes depende del número de palabras, por lo cual, a un mayor número de palabras se obtendrá un descriptor más preciso de las imágenes, lo cual permite una clasificación más discriminativa. Se debe tener precaución de no exagerar en el descriptor. De la misma manera sucede con el parámetro "SpatialX" y "SpatialY". Ya que estos determinan en qué manera se dará la partición para la pirámide espacial, entre mayor sea este valor, se obtendrá un descriptor conjunto más específico, lo cual permite una clasificación más discriminativa. Adicionalmente se muestra el efecto del número de clases sobre el resultado. Este presenta un comportamiento en el cual, a mayor número de clases, menor resultado. Esto se debe a que entre menor sea el número de clases, mayor es la probabilidad de "atinar" a la clase, es decir, se requiere de clasificadores menos discriminativos para llegar a un resultado relativamente óptimo. Finalmente se puede notar que aumentar el valor de C, disminuye el resultado, sin embargo, es un cambio realmente pequeño. Adicionalmente no hay cambio entre C 5,8 y 10.

El resultado de test mostro un valor mayor al de training. Esto muestra que la seleccion de los parametros optimos tiene un efecto incremental sobre el resultado de ACA. Lo que implica que para trabajos futuros se debe evaluar los parametros mas exhaustivamente, adicionalmente, es necesario probar los paramtros intrinsencos de la funcion.
Las clases Web site, Convertible, Bookcase fueron las mejor clasificadas debido a que para estas clases el descriptor SIFT es muy discriminante. Esto se debe a que las imágenes de estas categorías presentan patrones muy definidos que se describen de manera muy acertada utilizando histogramas de gradientes orientados.

En las clases en las cuales se obtuvieron los más bajos desempeños las imágenes presentes no son descritas de manera eficiente por los histogramas de gradientes orientados, lo que implica que el descriptor SIFT no es discriminante. Esto se debe a que las imágenes presentes en estas clases están fuertemente influenciadas por problemas asociados a la escala y a la oclusión. Lo que dificulta el correcto entrenamiento del clasificador y genera que no se pueda definir un hiperplano óptimo para los support vector machines.

5. Discusión

Para seleccionar los mejores hiperparámetros para ImageNet se realizó un análisis detallado del efecto de cada hiperparémetro sobre la eficiencia del método. Para esto, se realizaron múltiples experimentos en los cuales se variaron los valores de los hiperparámetros para poder de esta forma calcular el rendimiento de la función a través del valor del ACA y determinar el efecto de los mismos. En cada experimento se cambió solo uno de los parámetros, dejando el resto iguales a los valores que tiene la función por defecto, para asegurar que el cambio en el ACA estaba asociado estrictamente a la variación en el parámetro específico analizado. Los valores de ACA obtenidos en cada uno de los experimentos, se pueden observar en la tabla 2.
Una vez obtenido el ACA para cada experimento, se procedió a seleccionar el valor óptimo (dentro del rango de valores analizados) de cada uno de los hiperparámetros. Se debe resaltar que el valor óptimo corresponde al valor del hiperparámetro en el cual la función tiene el mayor rendimiento (ACA). Si al analizar los ACA obtenidos con los nuevos valores para los hiperparámetros se determina que no hubo una mejora significativa, se establece como valor optimo el valor que tiene por defecto la función. Partiendo de esto, los mejores valores obtenidos para ImageNet son los presentados en la tabla 3.
Ahora bien, no todos los mejores parámetros fueron utilizados para realizar el test debido a que no tenía sentido utilizarlos. Por ejemplo utilizar un numero de clases menor aumenta el ACA, no obstante, para evaluar el método de forma objetiva se utilizaron las 200 clases. Esto pasa para otros hiperparámetros, como numero de imágenes de train y test, entre otros. Por este motivo el test se corrió utilizando los parámetros por defecto presentados en la tabla1 cambiando únicamente el número de palabras visuales de 600 a 900. Esto debido a que el número de palabras fue el único parámetro que presento un cambio significativo en el ACA, sin representar una evaluación no objetiva del método.

6. Conclusiones

- El desempeño de la estrategia PHOW está estrechamente ligado con una correcta selección de los hiperparámetros, lo que implica que para obtener mejores resultados se debe realizar un análisis más detallado de los parámetros de la función, que permita definir de forma certera un modelo óptimo para cada una de las bases de datos.
- Se determinó que el parámetro más influyente era el correspondiente al número de palabras visuales debido a que la influencia de parámetros como el número de imágenes de train y test o el número de clases era considerada como evidente y ligada estrechamente a la información presente en base de datos utilizada.
- Al evaluar la función de clasificación con los valores de hiperparámetros óptimos en las imágenes de test de ImageNet no se obtuvieron los mejores resultados, debido a que la estrategia PHOW presentada en la biblioteca VLFeat estaba optimizada para el estudio del problema de clasificación y reconocimiento en la base de datos Caltech101. Esto muestra también porque se obtienen ACAs tan distintos al entrenar los SVM con las clases presentes en cada una de las bases de datos (ImageNet y Caltech101).
- Al utilizar el mejor modelo (hiperparámetros óptimos) para realizar la clasificación de las imágenes de test de ImageNet se obtuvo un ACA mayor que todos los obtenidos en los distintos experimentos realizados y en el modelo especificado por defecto, lo que implica que realizar un estudio detallado para definir los mejores hiperparámetros para PHOW asegura la optimización de los resultados de la clasificación.

References

[1] Castro Piñol, D., Sanabria Macias, F., Marañón Reyes, E., Rodriguez Arias, F. (2016). Reconocimiento de armas en imágenes de rayos X mediante Saco de Palabras Visuales. Revista Cubana de Ciencias Informáticas, 10(1), 152-161.

[2] BOSCH, A., ZISSERMAN, A., and MUNOZ, X. (2007). Image classification using random forests and ferns. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE.

[3] Caltech 101. (n.d.). Retrieved April 06, 2017, from http:www.vision.caltech.eduImageDatasetsCaltech101