Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary

P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth

Dr. Enrique Sucar[1] Victor Hugo Arroyo Dominguez[1]

[1] Instituto Nacional de Astrofísica Óptica y Electrónica

15 de junio de 2009

Tabla de Contenido

1. Resumen
2. Introducción
3. Marco teórico
4. Utilizando un algoritmo EM para aprender un diccionario
5. Aplicando y refinando el diccionario
6. Resultados experimentales
7. Discusión
8. Referencias

Resumen I

Máquina de traducción

Se describe un modelo para reconocimiento de objetos como una máquina de traducción. En este modelo, el reconocimiento es un proceso de describir regiones de imágenes con palabras.

Resumen II

Proceso

1. Las imágenes son segmentadas en regiones.
2. Se clasifican en tipos de regiones utilizando una variedad de características.
3. Se aprende el mapeo entre los tipos de regiones con las palabras clave suministradas en la imagen, utilizando un metodo basado en EM.

Agrupamiento

Se muestra como agrupar palabras que individualmente son difíciles de predecir en grupos que se pueden predecir correctamente.

Introducción I

Existen tres tipos importantes de teoría acerca del reconocimiento de objetos. Los tipos de métodos existentes en reconocimiento de objetos no estan dirigidos a resolver algunos temas en particular.

- ¿Qué cuenta como un objeto?
- ¿Cuáles objetos son fáciles de reconocer y cuales son difíciles?
- ¿Cuáles objetos son indistinguibles utilizando nuestras características?

Introducción II

Se tiene una imagen, consistente de regiones y un conjunto de texto.

- Se sabe que el texto va con la imagen.
- No se sabe cual palabra va con con cual región.

En este trabajo se muestra como se puede aprender esta correspondencia utilizando una variante del algoritmo EM [1], Este enfoque ataca algunos problemas importantes en el reconocimiento de objetos.

Algoritmo EM

Definición

El algoritmo expectation-maximization (EM) se usa en estadística para encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables. El algoritmo EM alterna pasos de esperanza y maximización.

- (paso E), donde se calcula la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables.
- (paso M), donde se calculan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E.

Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite.

Utilizando un algoritmo EM para aprender un diccionario

- Se segmentaran las imágenes en regiones y se aprenderá a predecir palabras utilizando regiones. Cada región será descrita por un conjunto de características.
- Sin embargo, las caracteristicas asociadas con las regiones de una imagen no ocupan un espacio discreto.
- Un *"blob"* se refiere a la etiqueta asociada con una región.

- El problema es utilizar el conjunto de datos de entrenamiento para construir una tabla de probabilidad, la cual ligue los blob con las palabras.
- Esta tabla es la probabilidad condicional de una palabra dado un blob.

Utilizando un algoritmo EM para aprender un diccionario II

Esencia del algoritmo EM

- La dificultad de aprender la tabla de probabilidad es que el conjunto de datos no proporciona correspondencia explícita. No se sabe que palabra corresponde a que región.
- Esto sugiere la siguiente estrategia iterativa:
  - Utilizar un estimado de la tabla de probabilidad para predecir las correspondencias.
  - Utilizar las correspondencias para refinar el estimado de la tabla de probabilidad.

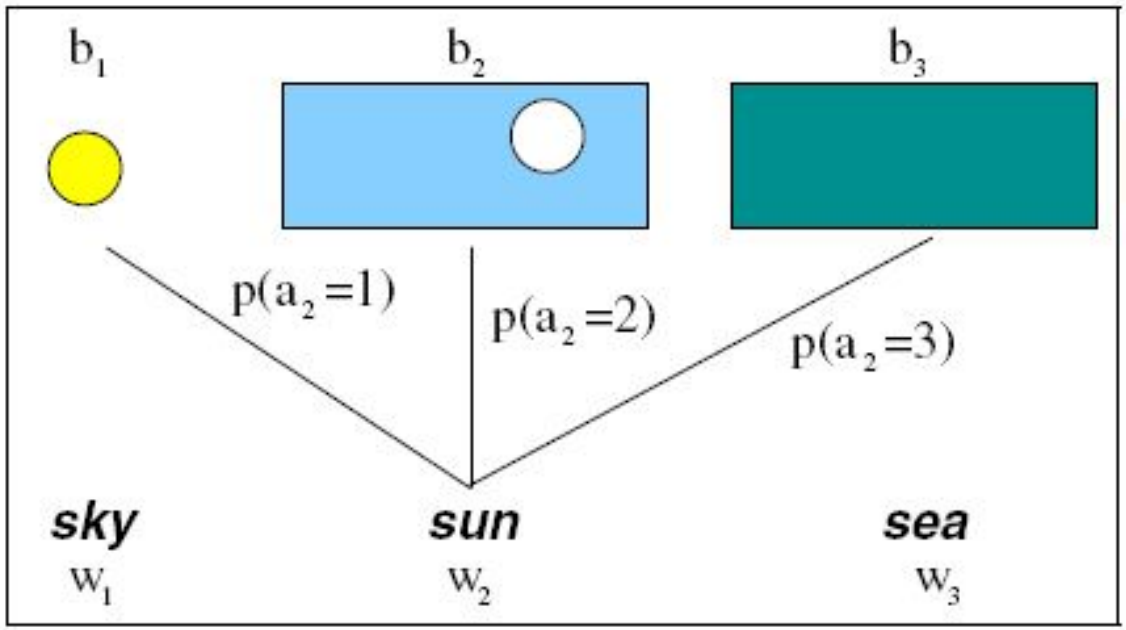
Utilizando un algoritmo EM para aprender un diccionario III

Se pueden describir las imágenes haciendo lo siguiente:

1. Clasificando los segmentos para encontrar los correspondientes *blobs*.
2. Encontrando la palabra correspondiente para cada *blob* seleccionando la palabra con mayor probabilidad.

Utilizando un algoritmo EM para aprender un diccionario IV

Para trasladar los *blobs* a palabras, se necesita estimar la probabilidad de que en la imagen *n*, un *blob* en particular se asocie con una palabra en particular. Esto se hace con cada una de las imagenes.



Utilizando un algoritmo EM para aprender un diccionario V

Estimación de la maxima probabilidad con EM

Se quiere encontrar los parametros de máxima probabilidad, se puede lograr esta estimación utilizando un algoritmo EM, el cual itera entre los siguientes dos pasos:

- Paso E: Calcular el valor esperado de la funcion completa de log-probabilidad con respecto a la distribución de las variables de asignación.
- Paso M: Encontrar el nuevo argumento máximo.

Aplicando y refinando el diccionario

Existen algunas variantes importantes disponibles.

Controlando el vocabulario negandose a predecir

- En particular, algunos *blobs* puede que no pronostiquen ninguna palabra con alta probabilidad, tal vez por que son muy pequeños para tener una identidad propia.
- Es natural establecer un umbral y requerir que se cumpla la siguiente condicion antes de predecir la palabra

$$p(palabra \mid blob) \rangle umbral$$

- Esto es equivalente a asignar una palabra nula a un *blob* que su mejor palabra predecida este debajo del umbral.

Aplicando y refinando el diccionario II

Agrupando palabras indistinguibles

- Algunas palabras puede que sean visualmente indistinguibles, como gato y tigre o tren y locomotora.
- Esto sugiere agrupar las palabras que son similares.
- Cada palabra en el conjunto es reemplazada con la etiqueta del grupo.
- Esto implica que dos palabras serán similares si generan blobs de imágenes similares en frecuencias similares.

Resultados experimentales

Caracteristicas del conjunto de datos

- Se entreno utilizando 4500 imágenes del conjunto de datos *“Corel”*.
- Hay 371 palabras en total en el vocabulario, cada imagen tiene 4-5 palabras clave.
- Las imágenes son segmentadas utilizando *“Normalized cuts”*.
- Solo regiones mas grandes que un umbral son utilizadas.
- Las regiones son agrupadas en 500 blobs utilizando *k-means*.
- Se utilizan 33 características para cada región (color de la región, desviación estandar, ..., tamaño de la región).

Evaluando la descripción (anotación) I

Método de evaluación

Cada imagen en la prueba es descrita automáticamente, tomando cada región mayor a un umbral, cuantificando la región en un blob, y utilizando el diccionario para determinar la palabra más probable para el blob; si la probabilidad de la palabra dada el blob es mayor al umbral, entonces la imagen es descrita por la palabra.

Resultados base

Solo 80 palabras de las 371 pudieron ser predecidas. Se asignó el umbral mínimo de probabilidad a cero, así cada blob predijo una palabra.

Evaluando la descripción (anotación) II

El efecto de re-entrenar

Debido a que solo se predijeron 80 palabras, se puede reducir el vocabulario a solo esas palabras, y ejecutar el algoritmo EM nuevamente. Los resultados obtenidos son muy similares a los resultados base. Sin embargo se pueden predecir palabras con mayor recuerdo y mayor precisión.

El efecto de la probabilidad nula

Se compararon los valores de recuerdo y precisión para los datos de entrenamiento y pruebas en algunas palabras seleccionadas. Los resultados son muy similares para ambos tipos de datos. Cuando se incrementa el umbral nulo lo suficiente, algunas palabras no pueden ser predecidas, pero las que se tienen son mas fiables.

Evaluando la descripción (anotación) III

El efecto de agrupar las palabras

Se compararon los valores de recuerdo y precisión despues de agrupar las palabras, los valores de recuerdo de las palabras agrupadas es mayor que los valores de recuerdo para las palabras solas. Se muestra que los resultados son mejores cuando se tienen relaciones semánticas o visuales fuertes, como hoja-flores-plantas-vegetales.

Evaluando la correspondencia I

Método de evaluación

Debido a que el conjunto de datos no contiene información de correspondencia, es dificil revisar la correspondencia para grandes volumenes de datos. Cada imagen de prueba debe ser vista para decidir si una anotación o una región es correcta. Inevitablemente esta prueba es subjetiva.

Evaluando la correspondencia II

Resultados base

- Se trabajó con un conjunto de 100 imágenes para revisar los resultados de correspondencia.
- El rango de predicción se calcula contando el promedio de veces que el blob predice la palabra correctamente.
- Para algunas palabras buenas se obtuvo hasta el 70 % de predicción correcta.
- Es mas dificil evaluar el rango en el cual las regiones se pierden.
- Se pueden predecir algunas palabras como cielo, arbol, pasto siempre correctamente en la mayoria de las imágenes.

Evaluando la correspondencia III

El efecto de la predicción nula

Algunas palabras con baja probabilidad se les asigna probabilidad nula, pero las palabras con alta probabilidad permanecen igual. Esto incrementa la predicción correcta para las palabras buenas.

El efecto del agrupamiento de palabras

Generalmente las palabras similares se agrupan en una sola, esto incrementa la predicción debido a que palabras indistinguibles estan ahora agrupadas.

Discusión

- Este método es atractivo debido a que permite atacar problemas inaccesibles en el reconocimiento de objetos.
- Es invariante con respecto a las características.
- Existe la dificultad con los algoritmos de aprendizaje de diccionarios que el prejuicio en la correspondencia puede ocasionar problemas.

Referencias I


DUYGULU, P., BARNARD, K., DE FREITAS, J., AND FORSYTH, D.
Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary.
2002, pp. 349–354.

Gracias.