Entrenamiento, optimización y validación de una CNN para localización jerárquica mediante imágenes omnidireccionales.

Juan José Cabrera, Sergio Cebollada, Mónica Ballesta, Luis Miguel Jiménez, Luis Payá, Óscar Reinoso
Departamento de Ingeniería de Sistemas y Automática
Universidad Miguel Hernández, Avenida de la Universidad s/n 03202 Elche (Alicante), España
{juan.cabreram, sergio.cebollada, m.ballesta, luis.jimenez, lpaya, o.reinoso}@umh.es

Resumen

El objetivo del presente trabajo es abordar la localización de un robot móvil mediante el entrenamiento de una Red Neuronal Convolucional (CNN) de manera que se obtengan unos resultados óptimos. El problema de localización se aborda de forma jerárquica empleando un sistema catadióptrico omnidireccional y se trabaja directamente con las imágenes capturadas sin pasar a panorámicas, ahorrando así el tiempo de cálculo asociado a este proceso. La localización se lleva a cabo en dos pasos y en ambos se emplea la arquitectura de la CNN con diferentes objetivos. Primero se lleva a cabo una localización gruesa que consiste en identificar la estancia en la que se encuentra el robot por medio de la CNN. Después se realiza una localización fina en dicha estancia, en la cual la CNN es empleada para la obtención de descriptores holísticos a partir de las capas intermedias de la red. Estos descriptores globales permiten encontrar la posición donde se encuentra el robot de manera más precisa por medio de una búsqueda del vecino más cercano, comparando el descriptor correspondiente de la imagen test con los descriptores de las imágenes capturadas en la estancia seleccionada en el primer paso. Con el fin de mejorar el desempeño de la red se recurre a un aumento de datos y a una optimización bayesiana de hiperparámetros. Estas técnicas demuestran ser una solución eficiente y robusta para afrontar el problema de localización tal y como se muestra en la sección de experimentos.

Palabras clave: Localización jerárquica, imágenes omnidireccionales, aprendizaje profundo, optimización bayesiana.

1. INTRODUCCIÓN

En los últimos años, el uso de cámaras omnidireccionales junto con técnicas de visión por computador ha demostrado ser una alternativa sólida para abordar la tarea de localización en robótica móvil. Este tipo de cámaras tienen un campo de visión de 360 grados y un coste relativamente bajo en comparación con otros tipos de sensores. Asimismo, los métodos de descripción holística constituyen una alternativa eficaz para extraer información relevante de las escenas, dado que conducen a algoritmos de localización relativamente sencillos, basados en la comparación directa de pares de descriptores.

En cuanto a creación de mapas, el uso de modelos jerárquicos con descriptores holísticos permite resolver la tarea de localización de forma eficiente. Este método consiste en organizar la información visual de forma jerárquica en diferentes capas de forma que la localización pueda resolverse en varios pasos. En este caso, se proponen dos pasos principales: en primer lugar, una localización aproximada para conocer en qué zona del entorno se encuentra el robot, y en segundo lugar, una localización fina, que se aborda en esa zona preseleccionada. Además, en los últimos años han surgido técnicas basadas en inteligencia artificial (IA) para abordar problemas de visión por ordenador y robótica. Las redes neuronales convolucionales (conocidas comunmente por sus siglas en inglés CNN) permiten reconocer patrones en tipos de datos no estructurados, como imágenes. El proceso de entrenamiento de dichas redes debe ser robusto y variado. De ahí que haya que tener especialmente en cuenta dos cuestiones (a) debe disponerse de un amplio conjunto de datos de entrenamiento y (b) los parámetros de entrenamiento deben seleccionarse con cautela.

La novedad del presente trabajo es un enfoque de localización basado en una CNN que parte de imágenes omnidireccionales. Además, este trabajo presenta un proceso de optimización del entrenamiento de la CNN para llevar a cabo de forma eficiente la tarea de entrenamiento. En general, el objetivo de este trabajo es readaptar y utilizar una CNN con un doble propósito: (1) recuperar en qué habitación se encuentra actualmente el robot (paso de localización gruesa) y (2) refinar esta localización dentro de la habitación seleccionada (paso de localización fina) mediante descriptores de apariencia global obtenidos de capas intermedias de la propia CNN.

El resto del artículo se estructura como se indica a continuación. La sección 2 presenta una revisión de la literatura relacionada. Después, la sección 3 presenta los métodos para entrenar la CNN adaptada. La sección 4 detalla el método de localización, basado en la CNN adaptada. Finalmente, la sección 5 presenta las conclusiones y los trabajos futuros.

https://doi.org/10.17979/spudc.9788497498043.640

2. ESTADO DEL ARTE

En la literatura relacionada se pueden encontrar diversos trabajos que han abordado tareas en robótica y procesamiento de imágenes mediante técnicas de aprendizaje automático [4]. En cuanto al uso de las CNNs en el campo de la robótica móvil, son muchos los autores que han demostrado su éxito utilizando esta herramienta. Por ejemplo, Sinha et al. [16] proponen una re-localización del robot en entornos sin GPS utilizando una CNN para procesar los datos de una cámara monocular. Chaves et al. [7] utilizan una CNN para detectar objetos en imágenes y usan esta técnica para construir un mapa semántico.

En cuanto al uso de la información visual, en línea con trabajos anteriores [6], el presente trabajo se centra en abordar la creación de mapas y localización mediante la obtención de un descriptor único por imagen que contiene información global sobre la misma. Este enfoque holístico ha sido usado en diversos trabajos previos. Originalmente, los descriptores holísticos se basan en métodos analíticos, es decir, parten de una imagen y realizan transformaciones matemáticas para obtener un único vector ($\vec{d} \in \mathbb{R}^{l\times 1}$) con información característica de la imagen.

Sin embargo, en trabajos recientes se ha propuesto el uso de descriptores holísticos que se obtienen a partir de diversas capas de las CNN. En este sentido, las capas intermedias proporcionan descriptores que pueden ser utilizados para caracterizar los datos de entrada. Por citar algunos ejemplos, Arroyo et al. [1] utilizan una CNN que aprende a generar descriptores que son robustos frente a los cambios de estación. Más recientemente, Wozniak et al. [18] proponen el uso de la extracción de características a partir de un clasificador SVM (Máquina de Vector Soporte). Cebollada et al. [5] muestran las ventajas de utilizar descriptores obtenidos de las capas intermedias de una CNN reentrenada para resolver la localización como un problema de detección de la imagen más similar de entre un conjunto de imágenes. Sin embargo, este trabajo propone una CNN basada en imágenes panorámicas. Por lo tanto, para trabajar a partir de imágenes omnidireccionales, se debe realizar una transformación previa a panorámica.

En cuanto al proceso de entrenamiento, las herramientas de aprendizaje profundo requieren de un gran conjunto de datos para obtener comportamientos lo suficientemente robustos. Sin embargo, en algunos casos, el conjunto de datos disponible para el entrenamiento es pequeño y, entonces, el modelo no puede ser entrenado correctamente. Entre las técnicas propuestas para abordar este problema, el presente trabajo se centra en el aumento de datos y en la optimización de los hiperparámetros de entrenamiento. En cuanto a la técnica de aumento de datos, ésta mejora el rendimiento del entrenamiento del modelo aumentando el número de instancias de entrenamiento y evitando el sobreajuste. El aumento de datos consiste básicamente en crear nuevos datos (en este caso, imágenes) aplicando diferentes efectos sobre los originales. Algunos autores han utilizado el aumento de datos para mejorar sus tareas de aprendizaje profundo [8]. Salamon et al. [15] proponen el aumento de datos de audio para superar el problema de la escasez de datos de sonido ambiental. Utilizando esta técnica, son capaces de desarrollar una CNN que es capaz de clasificar este tipo de datos. Sin embargo, estos métodos de aumento de datos no se ajustan a los efectos visuales que pueden producirse cuando el robot se mueve por un entorno de trabajo en condiciones de operación reales.

El correcto entrenamiento de los modelos de aprendizaje automático depende en gran medida de la configuración de los hiperparámetros y, por tanto, del método utilizado para establecerlos. Los métodos de optimización como la búsqueda en rejilla y la búsqueda aleatoria han demostrado superar a otros métodos tradicionales para este problema [2]. Estos métodos han sido capaces de obtener ajustes de hiperparámetros similares o mejores que los establecidos por expertos [3,11]. Como resultado, la optimización de hiperparámetros se ha convertido en un área de investigación importante [2,9]. Durante los últimos años, la optimización bayesiana ha surgido como un enfoque eficiente, logrando resultados satisfactorios [17]. A través de la optimización bayesiana, la optimización de la función de pérdida se considera como una "caja negra", con el objetivo de encontrar $argmin_{x\in X}(f(x))$, donde $x \in X$ son los hiperparámetros y f(x) es la función de pérdida del modelo.

Respecto a la tarea de localización desde un punto de vista jerárquico, trabajos previos han demostrado que el uso de estos modelos con descriptores holísticos e imágenes omnidireccionales lleva a una solución eficiente y robusta para abordar la tarea de localización [6, 13]. Estos trabajos previos consisten básicamente en el cálculo del vecino más cercano en dos capas. En primer lugar, para la capa de alto nivel, se agrupan los descriptores visuales según su similitud, mediante algún método de clustering, y se obtiene un descriptor representativo $R = \{\vec{r}_1, \vec{r}_2, ..., \vec{r}_{n_g}\}$ para cada grupo, donde n_g es el número de grupos. Después, para resolver la tarea de localización, se obtiene una nueva imagen im_{test} y se calcula su descriptor holístico $\vec{d}_{test}$. Este descriptor se compara con todos los representantes R y se retiene el representante más similar $\vec{r}_k$ (paso de localización gruesa o aproximada); después, se realiza una nueva comparación entre $\vec{d}_{test}$ y los descriptores contenidos en el grupo k, $D_k = \{\vec{d}_{k,1}, \vec{d}_{k,2}, ..., \vec{d}_{k,N_k}\}$. Por último, la posición de la imagen im_{test} se estima como la posición en la que se capturó la imagen más similar del grupo k-ésimo (paso de localización fina).

https://doi.org/10.17979/spudc.9788497498043.640

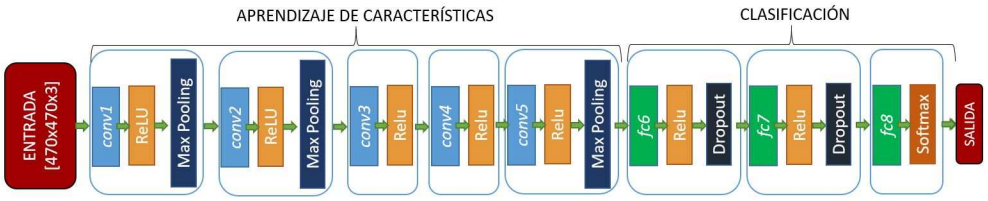

Figura 1: Arquitectura de la CNN. Esta red fue creada a partir de AlexNet, adaptada y reentrenada para clasificar las estancias del dataset de Friburgo.

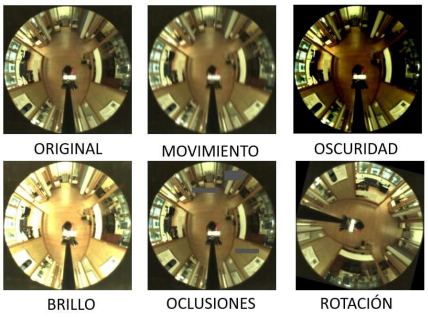

Figura 2: Efectos aplicados a una imagen, ejemplo para aumento de datos.

3. PROCESO DE ENTRENAMIENTO

3.1. ADAPTACIÓN DE LA CNN

La construcción y el entrenamiento de una red desde cero requiere experiencia con arquitecturas de redes, una gran cantidad de datos para el entrenamiento y, por tanto, un tiempo de computación importante. Este trabajo continúa la propuesta realizada en trabajos anteriores [5]: adaptar y entrenar redes preexistentes con un objetivo distinto a aquel para el que inicialmente se diseñaron. En este sentido, se propone partir de AlexNet [12], ya que presenta una arquitectura sencilla y ha sido utilizada con éxito en trabajos anteriores para desarrollar nuevas tareas de clasificación mediante *transfer learning* (como en [5, 10]). A diferencia de estos trabajos anteriores, que hacían uso de imágenes convencionales (no panorámicas y panorámicas), el objetivo del presente trabajo es estudiar la viabilidad de esta arquitectura partiendo de imágenes omnidireccionales. Esta propuesta presenta un doble beneficio: (1) el ahorro de tiempo de computación, ya que no es necesaria una transformación de imágenes omnidireccionales a panorámicas y (2) la obtención de descriptores holísticos basados en imágenes omnidireccionales, que han sido escasamente propuestos en el estado del arte actual. Además, el presente trabajo también desarrolla una optimización robusta de los hiperparámetros con el objetivo de abordar un entrenamiento óptimo del modelo de aprendizaje profundo.

Por tanto, en primer lugar, se modifican algunas capas de la arquitectura AlexNet para adaptar la red a la tarea de clasificación de habitaciones propuesta. En este caso, se redimensiona la capa de entrada, pasando de 227x227x3 a 470x470x3, se sustituyen las capas totalmente conectadas (fc_6), la capa softmax y la capa de clasificación. La capa fc_8 se readapta para dar salida a un vector de nueve componentes. Las capas softmax y de clasificación se sustituyen para calcular las probabilidades entre las posibles estancias del entorno (en este caso 9 estancias) y para calcular la pérdida de entropía cruzada para la clasificación multiclase. En la fig. 1 se muestra la arquitectura utilizada a lo largo de este trabajo. De este modo, tras estas adaptaciones, se reentrena toda la arquitectura, aprovechando los pesos iniciales de AlexNet.

3.2. AUMENTO DE DATOS

Disponer de un gran conjunto de datos de entrenamiento es crucial para el rendimiento del modelo. Sin embargo, a veces, el conjunto de datos de entrenamiento disponible es más pequeño de lo necesario y entonces, el modelo no puede ser entrenado adecuadamente para alcanzar la solución deseada. Para resolver este problema, se ha propuesto la técnica de *Data Augmentation* como método para mejorar el rendimiento del modelo aumentando el número de instancias de entrenamiento y evitando el sobreajuste. El aumento de datos consiste básicamente en la creación de nuevos datos (imágenes) mediante la aplicación de diferentes efectos sobre las imágenes originales, tal y como se muestra en la fig. 2.

Las tranformaciones que se han utilizado son las siguientes. Se han considerado efectos que replican situaciones que pueden ocurrir en entornos reales, cuando el robot debe operar en condiciones desafiantes.

1. **Rotación:** Introducción de rotaciones entre 10 y 350 grados a las imágenes omnidireccionales.

2. **Oscuridad y brillo:** Los valores de baja intensidad son reajustados (incremento) para crear imágenes con más brillo. Por otro lado, para crear un efecto de oscuridad, los valores altos de intensidad son reducidos.

3. **Ruido gaussiano** con varianza de 1e-06 a la imagen en escala de grises.

4. **Oclusiones:** Este efecto simula situaciones reales como por ejemplo que alguna persona u objeto se posicione delante de la cámara, ocluyendo parte de la es-

https://doi.org/10.17979/spudc.9788497498043.640

cena. En este trabajo, simulamos dicho efecto introduciendo imágenes geométricas en escala de grises en lugares aleatorios de la imagen.

5. **Reflexión:** Obtener la fotografía espejo.

6. **Blur effect o efecto movimiento.** Este efecto ocurre cuando la imagen se capturó en movimiento.

3.3. OPTIMIZACIÓN BAYESIANA

La optimización de hiperparámetros que se propone consiste en variar aquellos valores que pueden ser cruciales para abordar el proceso de entrenamiento y que, al mismo tiempo, pueden ser muy diferentes en función del objetivo de la red. Los hiperparámetros que se consideran para evaluar son los siguientes:

Initial Learn Rate. Controla cómo de rápido el modelo se adapta al problema.

Momentum. Permite suavizar el progreso de aprendizaje acumulando el gradiente de los pasos anteriores para determinar la dirección a seguir. De este modo, puede mantener la tendencia global de los puntos, evitando que un mal dato desencadene en un mínimo local.

L2 Regularization o decaimiento de los pesos. Valor escalar positivo que añade un término de regularización para los pesos de la función de pérdida.

4. EXPERIMENTOS

4.1. LOCALIZACIÓN

Este trabajo propone utilizar la CNN como modelo jerárquico con el objetivo de: (a) abordar la localización aproximada como problema de búsqueda de habitaciones (capa de alto nivel) partiendo de la imagen de test y (b) obtener descriptores holísticos de las imágenes de entrada. Los descriptores de las imágenes de entrenamiento formarán la capa de bajo nivel, y permiten resolver una localización fina, abordado como un problema de detección de la imagen más similar, con los descriptores holísticos de las imágenes test (también obtenidos mediante la CNN).

En cuanto a la localización jerárquica, las capas de alto nivel permiten una **localización gruesa** y las capas de bajo nivel una **localización fina**. El paso grueso proporciona una localización más rápida y el paso fino considera información más precisa que se utiliza para realizar la localización fina. La localización jerárquica propuesta se lleva a cabo tal y como muestra el diagrama de la fig. 3. En primer lugar (paso de localización gruesa), se introduce una imagen de test im_{test} en la CNN y se estima la habitación más probable c_i en la que se capturó la imagen a partir de la información de las capas de salida. Al mismo tiempo, la CNN reentrenada también es capaz de proporcionar descriptores holísticos a partir de las capas intermedias. Posteriormente, tras identificar la habitación, se lleva a cabo una localización más precisa (etapa de localización fina). En esta etapa se selecciona uno de los descriptores $\vec{d}_{test}$ y se compara con los descriptores $D_{c_i} = \{\vec{d}_{c_i,1}, \vec{d}_{c_i,2}, ..., \vec{d}_{c_i,N_i}\}$ del conjunto de datos de entrenamiento que pertenecen a la habitación seleccionada c_i mediante la distancia euclídea y se guarda el descriptor más similar $\vec{d}_{c_i,k}$. Por último, la posición en la que se capturó la imagen de prueba se estima como las coordenadas en las que se capturó $im_{c_i,k}$.

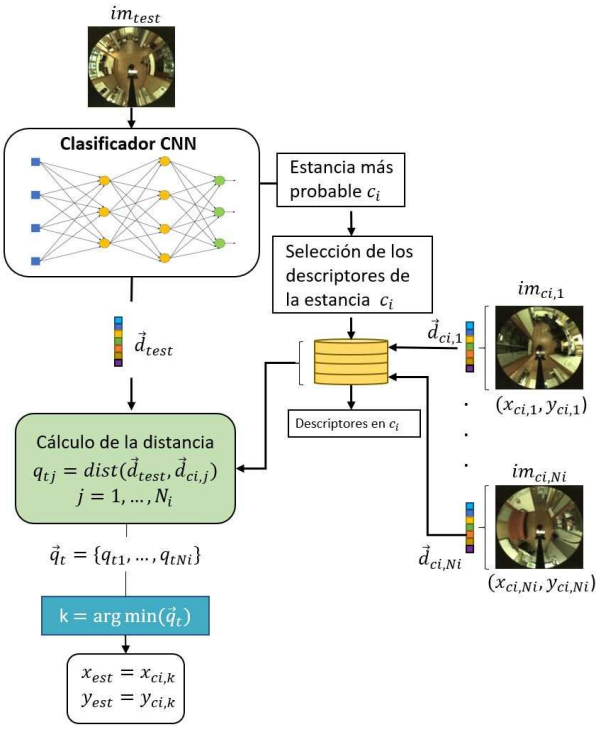

Figura 3: Diagrama de localización jerárquica.

4.2. BASE DE DATOS DE FRIBURGO

Las imágenes utilizadas en el presente trabajo se obtuvieron del conjunto de datos de Friburgo, incluido en la base de datos COLD (COsy Localization Database) [14]. Este conjunto de datos contiene imágenes omnidireccionales capturadas mientras el robot recorría varias trayectorias a través de las 9 estancias del entorno.

El robot aborda la tarea de captura de imágenes en condiciones reales de funcionamiento, es decir, personas que aparecen y desaparecen de la escena, cambios en el mobiliario, efectos de desenfoque, cambios dinámicos, etc. Las imágenes utilizadas fueron capturadas bajo tres condiciones de iluminación (nublado, soleado y noche).

El dataset de nublado se muestrea con el objetivo de obtener un conjunto de datos resultante con una distancia media de 20 cm entre imágenes consecutivas, lo

https://doi.org/10.17979/spudc.9788497498043.640

cual dará como resultado el conjunto de entrenamiento 1, con 519 imágenes. Adicionalmente, a este dataset se le realizará el aumento de datos descrito en la sección 3.2, dando lugar al conjunto de entrenamiento 2, que contiene 213.504 imágenes. Estos conjuntos de datos se usarán, individualmente, para entrenar las CNNs. De este modo, será posible conocer el efecto del aumento de datos en el desempeño de la red. Por su parte, se consideran diferentes conjuntos de test: el conjunto test 1, que contiene imágenes capturadas en condiciones de nublado (ruta del robot diferente a la de entrenamiento 1), con un total de 2.595 imágenes; el conjunto test 2, que contiene todas las imágenes capturadas en condiciones soleado (2.807 imágenes) y el conjunto test 3, con todas las imágenes capturadas por la noche (2.876 imágenes). Por tanto, el entrenamiento de la red se realiza, en todo caso, con imágenes capturadas en condiciones de iluminación nublado, y el test se realizará en tres condiciones distintas: nublado, soleado o noche, con lo que será posible testar la robustez de la red ante este tipo de cambios de iluminación en el entorno.

4.3. ENTRENAMIENTO DE LA CNN

El proceso de identificación de la estancia en que se encuentra el robot se ha abordado como un problema de clasificación. Para ello, se reentrena la arquitectura descrita en el apartado 3.1 con el objetivo de identificar la habitación donde se capturó la imagen de entrada. Las funciones de pérdida y de optimización empleadas para los entrenamientos son la *Cross Entropy* y el *Stochastic Gradient Descent* (SGD) respectivamente. En cuanto a la optimización de los hiperparámetros mediante la optimización bayesiana, se han realizado cinco experimentos, en los cuales se ha variado el conjunto de datos de entrenamiento, el número de puntos explorados y los valores de los hiperparámetros. Los cinco experimentos abordados para entrenar la CNN son los siguientes:

Experimento 1: Entrenamiento con el conjunto de entrenamiento 1. Hiperparámetros a optimizar: Initial Learn Rate, Momentum y L2 Regularization. En este experimento se realizaron 30 combinaciones diferentes de valores de hiperparámetros.

Experimento 2: Entrenamiento con el conjunto de entrenamiento 2 (conjunto aumentado) y los hiperparámetros que se encontraron como óptimos para el experimento 1.

Experimento 3: Entrenamiento con el conjunto de entrenamiento 2. Se consideran diferentes valores de Momentum para estudiar la influencia de este parámetro. En total, se testan 8 valores.

Experimento 4: Entrenamiento con el conjunto de entrenamiento 2. Se consideran diferentes valores de Initial Learn Rate y Momentum para estudiar la influencia de estos parámetros. En total, se testan 8 valores.

Experimento 5: Entrenamiento con el conjunto de entrenamiento 2. Se consideran diferentes valores de Momentum y L2 Regularization para estudiar la influencia de estos parámetros. En total, se testan 30 valores.

A modo de ejemplo, la fig. 4 muestra el valor de la función objetivo tras realizar el entrenamiento del experimento 5. Tras realizar los experimentos para la optimización del entrenamiento, la tabla 2 muestra el rango de hiperparámetros y los valores óptimos obtenidos para cada experimento. Una vez entrenadas las redes con la configuración óptima de cada experimento, se han introducido los tres conjuntos de test en las redes resultantes. La precisión en la clasificación (porcentaje de veces que la red devuelve la habitación correcta en que se capturó la imagen de entrada) se muestra en la Tabla 1, en la que se detalla las redes resultantes de cada experimento de entrenamiento y la precisión alcanzada con cada uno de los tres conjuntos test.

Como podemos ver, el mejor resultado para las condiciones de soleado se obtiene en el experimento 1, que es el único que no considera aumento de datos en el entrenamiento. Un análisis más detallado del funcionamiento muestra que la CNN tiende a cometer errores en una habitación específica en la que los rayos de sol atraviesan las ventanas e invaden la habitación de una forma especialmente apreciable en las imágenes. Este hecho, en combinación con los efectos del aumento de datos, provoca confusión en el algoritmo reduciendo la precisión global pero no podemos concluir que el aumento de datos provoque que la red no funcione bien en condiciones soleadas en general. En el experimento 2 se obtienen los mejores resultados para la noche y los días nublados aunque los experimentos 3 y 4 también proporcionan un desempeño comparativamente bueno. Además, considerando conjuntamente las tres condiciones de iluminación, la mejor solución se presenta en el experimento 3.

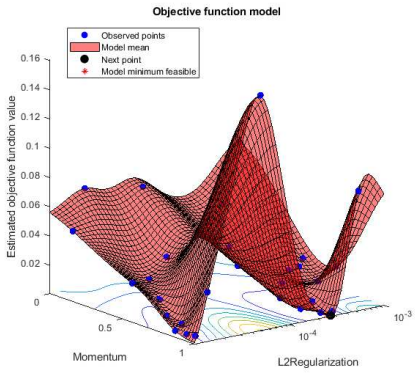

Figura 4: Optimización del entrenamiento del experimento 5.

https://doi.org/10.17979/spudc.9788497498043.640

Cuadro 1: Resultados en la identificación de estancias.

Exp	Exactitud test de clasificación (%)		
	Nublado	Noche	Soleado
1	97.42	96.91	92.27
2	98.92	98.78	89.6
3	98.77	98.4	91.73
4	98.46	98.37	89.31
5	94.72	98.02	86

Cuadro 2: Valores de los hiperparámetros obtenidos a través de la optimización bayesiana.

Exp	Hiperparámetros	Rango	Óptimo
1	Initial Learn Rate	[1e-4, 1]	0.006
	Momentum	[0.5, 1]	0.539
	L2 Regularization	[1e-10, 1e-2]	3.87e-9
2	Initial Learn Rate	0.006021	-
	Momentum	0.53961	-
	L2 Regularization	3.873e-9	-
3	Initial Learn Rate	1e-3	-
	Momentum	[0, 1]	0.911
	L2 Regularization	1e-4	-
4	Initial Learn Rate	[1e-5, 1e-2]	0.007
	Momentum	[0, 1]	0.384
	L2 Regularization	1e-4	-
5	Initial Learn Rate	1e-3	-
	Momentum	[0, 1]	0.979
	L2 Regularization	[1e-5, 1e-3]	3.06e-4

4.4. LOCALIZACIÓN JERÁRQUICA EMPLEANDO DESCRIPTORES HOLÍSTICOS

Como se ha mencionado anteriormente, el método de localización propuesto para abordar esta tarea se basa en un enfoque jerárquico. Éste consiste en utilizar descriptores holísticos obtenidos de una capa intermedia de la CNN entrenada. Esta localización se aborda en dos pasos. El primer paso es la localización aproximada, que consiste en realizar la tarea de recuperación de la habitación mediante la CNN reentrenada, cuyo resultado se ha mostrado en la subsección 4.3. El segundo paso es la localización fina y consiste en estimar la posición de captura mediante un método de búsqueda del vecino más cercano utilizando descriptores holísticos. Entre las diferentes capas intermedias, hemos decidido estudiar las capas totalmente conectadas 6 y 7, ya que los experimentos preliminares mostraron que presentaban mayor robustez ante cambios de iluminación, en especial la capa fc_6, cuyo desempeño se analizará más profundamente.

Teniendo estos hechos en cuenta, los métodos de localización jerárquica considerados son los siguientes:

Método 1: $CNN_0 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento no fue optimizado (se seleccionaron hiperparámetros por defecto). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

Método 2: $CNN_0 + fc_7$. Localización aproximada mediante una CNN cuyo entrenamiento no fue optimizado (se seleccionaron hiperparámetros por defecto). Localización fina mediante el descriptor holístico extraído de la capa fc_7.

Método 3: $CNN_1 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento fue optimizado (se seleccionaron hiperparámetros obtenidos en el experimento 1). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

Método 4: $CNN_2 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento fue optimizado (se seleccionaron hiperparámetros obtenidos en el experimento 2). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

Método 5: $CNN_3 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento fue optimizado (se seleccionaron hiperparámetros obtenidos en el experimento 3). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

Método 6: $CNN_4 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento fue optimizado (se seleccionaron hiperparámetros obtenidos en el experimento 4). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

Método 7: $CNN_5 + fc_6$. Localización aproximada mediante una CNN cuyo entrenamiento fue optimizado (se seleccionaron hiperparámetros obtenidos en el experimento 5). Localización fina mediante el descriptor holístico extraído de la capa fc_6.

El error de localización se mide como la distancia euclídea entre la posición estimada y la posición real en la que se capturó la imagen test (dada por el ground truth del conjunto de datos). Además, se evaluará el error medio de localización separadamente para cada condición de iluminación, ya que el objetivo es estudiar la robustez del método propuesto frente a los cambios de iluminación en el entorno de trabajo. Los resultados obtenidos se muestran en la fig. 5.

Un análisis detallado de los datos muestra que, como era de esperar, el error de localización está relacionado con la precisión de la clasificación. Es decir, cuando la CNN es más capaz de identificar correctamente la estancia en el paso de localización aproximada, el error de localización es menor. En términos generales, las condiciones de iluminación soleadas son las que resultan más desafiantes para la red, y el error medio más alto se encuentra en la CNN asociada al entrenamiento 5, que presentó el menor ratio de éxito entre todos los procesos de entrenamiento abordados. En cuanto a los mejores valores para nublado y noche, se dan en

https://doi.org/10.17979/spudc.9788497498043.640

CNN_2,CNN_3 y CNN_4, cuyo entrenamiento se basó en la optimización bayesiana.

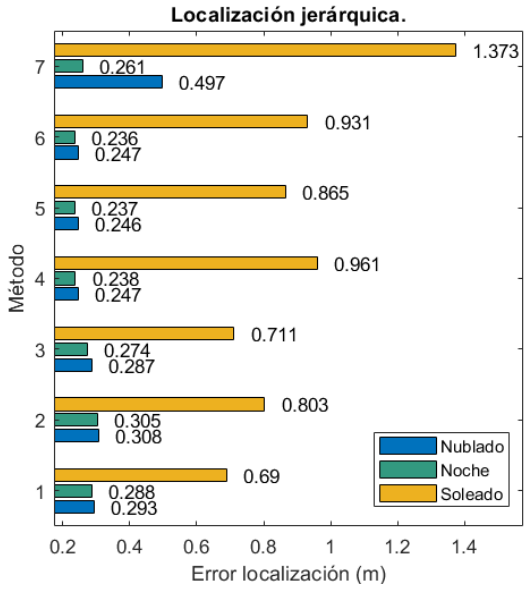

Figura 5: Resultados de la localización jerárquica.

5. CONCLUSIÓN

En este trabajo, hemos evaluado el uso de una técnica de aprendizaje profundo para construir modelos topológicos jerárquicos para la localización.

A lo largo del presente trabajo se ha evaluado el uso de dos técnicas para mejorar el proceso de entrenamiento de la CNN y el uso de la red para resolver la localización mediante un método basado en un enfoque de localización jerárquica.

En cuanto al proceso de optimización del entrenamiento, la optimización bayesiana es capaz de mejorar el proceso de entrenamiento de la CNN en general, ya que la tasa media de éxito aumenta cuando se considera este enfoque. Además, el aumento de los datos también permite obtener CNNs con mejores resultados. En cuanto a los resultados obtenidos en condiciones de iluminación nocturna, se han mejorado considerablemente, llegando a alcanzar un ratio de éxito similar al de las imágenes capturadas en las mismas condiciones que el conjunto de entrenamiento (es decir, en condiciones de nublado). Sin embargo, todas las redes tienden a presentar un peor desempeño con imágenes capturadas en condiciones soleadas. Tras un profundo análisis, llegamos a la conclusión de que este aumento de los errores se debe a que las imágenes capturadas en una de las salas se ven especialmente afectadas por los rayos. Este hecho, junto con los efectos del aumento de datos, provoca confusión durante el proceso de entrenamiento y, por tanto, reduce la precisión global.

En cuanto al reentrenamiento de una CNN preentrenada, ésta presenta buenos resultados para realizar una tarea de recuperación de habitaciones partiendo de imágenes omnidireccionales. Este resultado presenta una novedad en el campo, ya que hasta el momento, son escasos los trabajos que habían propuesto un modelo de aprendizaje profundo basado en arquitecturas convencionales e imágenes omnidireccionales para fines de localización. Además, las capas intermedias también son capaces de proporcionar vectores de información que pueden ser utilizados para obtener descriptores de apariencia global.

Este método ha sido evaluado y ha demostrado ser adecuado para abordar la tarea de localización. Los resultados obtenidos muestran que el error de localización es considerablemente bajo en condiciones de nublado y noche ya que el error mínimo obtenido es de 25 cm (teniendo en cuenta que la distancia media entre imágenes de entrenamiento es de 20 cm). En cuanto a las condiciones de iluminación soleada, el error de localización es mayor, debido sobre todo a la menor tasa de éxito en la clasificación, dada por la especial influencia de los rayos de luz en una de las salas. Por el contrario, los efectos producidos por la oscuridad se han reducido completamente, ya que el error medio de localización asociado a dicha condición es prácticamente igual al obtenido sin cambios de iluminación.

En futuros trabajos, nos centraremos en reducir el error de localización en condiciones de iluminación soleadas. Además, ampliaremos el uso de las técnicas de aprendizaje profundo para la localización mediante el uso de diferentes herramientas como los autoencoders, redes LSTM o convolucionales no rectangulares. Por último, también nos gustaría abordar la tarea de localización mediante CNNs en exteriores.

English summary

Training, optimization and validation of a CNN for hierarchical localization using omnidirectional images.

Abstract

The aim of this work is to address the localization of a mobile robot by training a Convolutional Neural Network (CNN) in order to obtain optimal results. The localization problem is approached in a hierarchical way by using an omnidirectional catadioptric system and working directly with the captured images without panoramic conversion, saving the computational time associated with this process. Localization is carried out in two steps, both using the CNN architecture for different purposes. First, a rough localization is carried out, which consists of

https://doi.org/10.17979/spudc.9788497498043.640

identifying the room in which the robot is located by means of the CNN. Then a fine localization is performed in the room, in which the CNN is used to obtain holistic descriptors from the intermediate layers of the network. These global-appearance descriptors allow finding the position where the robot is located more precisely by means of a nearest neighbour search, comparing the corresponding descriptor of the test image with the descriptors of the images captured in the room selected in the first step. In order to improve the accuracy of the network, data augmentation and Bayesian hyperparameter optimisation are used. These techniques prove to be an efficient and robust solution to tackle the localization problem as shown in the experiments section.

Keywords: Hierarchical Localization, Omnidirectional imaging, Deep Learning, Bayesian Optimization.

Referencias

[1] ARROYO, R., ALCANTARILLA, P. F., BERGASA, L. M., AND ROMERA, E. Fusion and binarization of cnn features for robust topological localization across seasons. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Oct 2016), pp. 4656–4663.

[2] BERGSTRA, J., AND BENGIO, Y. Random search for hyper-parameter optimization. *Journal of machine learning research 13*, Feb (2012), 281–305.

[3] BERGSTRA, J., YAMINS, D., AND COX, D. D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures.

[4] CEBOLLADA, S., PAYÁ, L., FLORES, M., PEIDRÓ, A., AND REINOSO, O. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. *Expert Systems with Applications 167* (2021), 114195.

[5] CEBOLLADA, S., PAYÁ, L., FLORES, M., ROMÁN, V., PEIDRÓ, A., AND REINOSO, O. A deep learning tool to solve localization in mobile autonomous robotics. In *ICINCO 2020, 17th International Conference on Informatics in Control, Automation and Robotics (Lieusaint-Paris, France, 7-9 July, 2020)* (2020), Ed. INSTICC.

[6] CEBOLLADA, S., PAYÁ, L., ROMÁN, V., AND REINOSO, O. Hierarchical localization in topological models under varying illumination using holistic visual descriptors. *IEEE Access 7* (2019), 49580–49595.

[7] CHAVES, D., RUIZ-SARMIENTO, J., PETKOV, N., AND GONZALEZ-JIMENEZ, J. Integration of cnn into a robotic architecture to build semantic maps of indoor environments. In *International Work-Conference on Artificial Neural Networks* (2019), Springer, pp. 313–324.

[8] DING, J., CHEN, B., LIU, H., AND HUANG, M. Convolutional neural network with data augmentation for sar target recognition. *IEEE Geoscience and remote sensing letters 13*, 3 (2016), 364–368.

[9] FALKNER, S., KLEIN, A., AND HUTTER, F. Bohb: Robust and efficient hyperparameter optimization at scale. *arXiv preprint arXiv:1807.01774* (2018).

[10] HAN, D., LIU, Q., AND FAN, W. A new image classification method using cnn transfer learning and web data augmentation. *Expert Systems with Applications 95* (2018), 43–56.

[11] KOTTHOFF, L., THORNTON, C., HOOS, H. H., HUTTER, F., AND LEYTON-BROWN, K. Auto-weka: Automatic model selection and hyperparameter optimization in. *Automated Machine Learning: Methods, Systems, Challenges* (2019), 81.

[12] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[13] PAYÁ, L., PEIDRÓ, A., AMORÓS, F., VALIENTE, D., AND REINOSO, O. Modeling environments hierarchically with omnidirectional imaging and global-appearance descriptors. *Remote Sensing 10*, 4 (2018), 522.

[14] PRONOBIS, A., AND CAPUTO, B. COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR) 28*, 5 (May 2009), 588–594.

[15] SALAMON, J., AND BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters 24*, 3 (March 2017), 279–283.

[16] SINHA, H., PATRIKAR, J., DHEKANE, E. G., PANDEY, G., AND KOTHARI, M. Convolutional neural network based sensors for mobile robot relocalization. In *2018 23rd International Conference on Methods Models in Automation Robotics (MMAR)* (Aug 2018), pp. 774–779.

[17] SNOEK, J., RIPPEL, O., SWERSKY, K., KIROS, R., SATISH, N., SUNDARAM, N., PATWARY, M., PRABHAT, M., AND ADAMS, R. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning* (2015), pp. 2171–2180.

[18] WOZNIAK, P., AFRISAL, H., ESPARZA, R. G., AND KWOLEK, B. Scene recognition for indoor localization of mobile robots using deep cnn. In *International Conference on Computer Vision and Graphics* (2018), Springer, pp. 137–147.

https://doi.org/10.17979/spudc.9788497498043.640