Projecte: Report

Tècniques Avançades de la Intel·ligència Artificial

2022/23

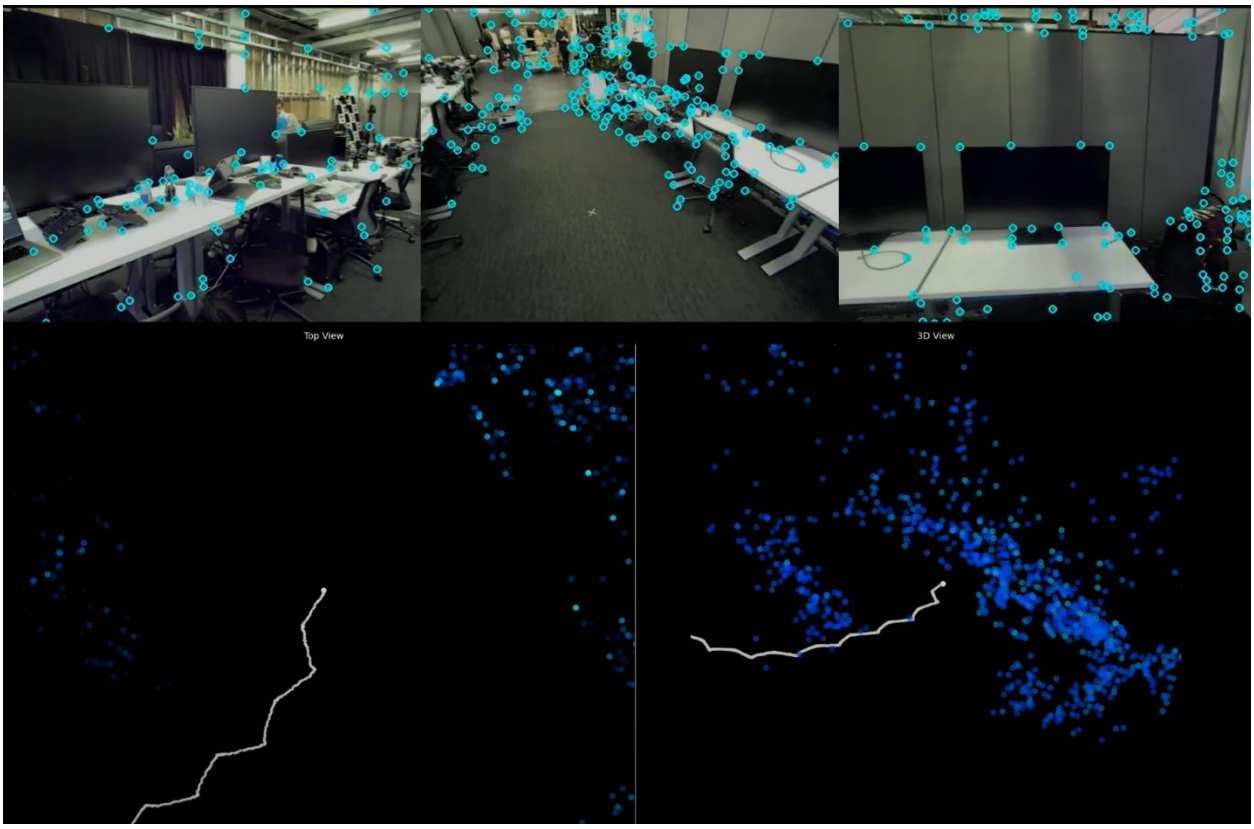
Martí Mas Fullana

Índex

1. Descripció del Problema

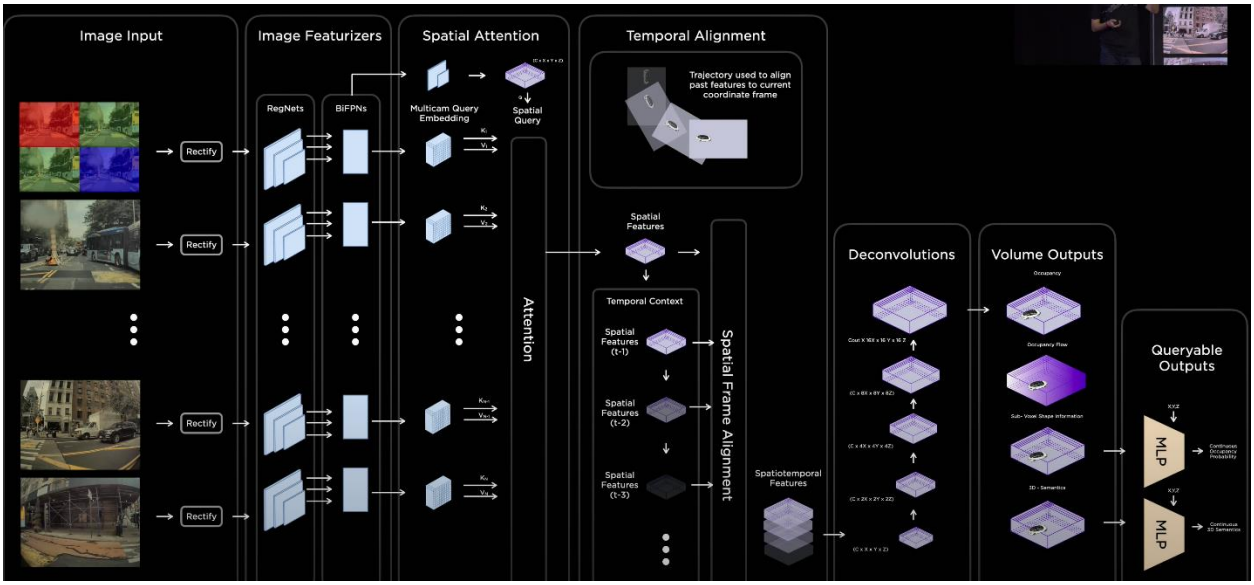
1a Part

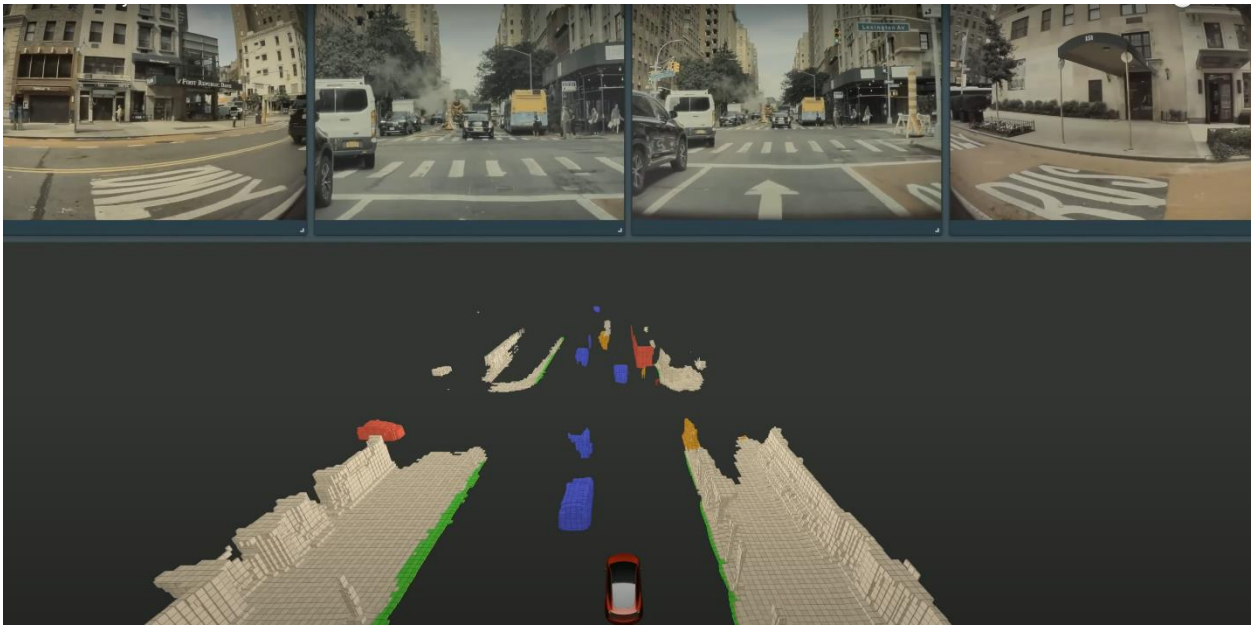
La meva proposta és crear un sistema de deep learning que estimi la posició i orientació de la càmera donat un vídeo, com es mostra a l'exemple:



2a Part (com a possibilitat)

Ja que aquest camp sembla que hi ha molta recerca feta, també proposo la possibilitat d'estendre la pràctica si la estimació de posició resultés ser relativament senzilla. Aquesta extensió consistiria en intentar replicar el sistema que va presentar Tesla a https://youtu.be/ODSJsviD_SU?t=4516





Aquest sistema prediu, a partir d'una seqüència d'imatges, la probabilitat que un voxel estigui ocupat dins d'un volum particionat a cubs (en la presencació a més ensenyen que classifiquen aquests voxels en unes 4 categories, pero jo aquest pas no el faria).

Tot això ho faré amb imatges generades amb Blender ja que això em permetria tenir moltes dades amb els valors correctes que la IA hauria de predir (blender sap la posició exacte de la càmera, per exemple) per poder fer entrenament totalment supervisat.

Per començar em centraré en la primera part i ignoraré la segona part.

2. Introducció i Motivació

Recentment la recerca en xarxes neuronals que incorporen equacions diferencials per emular el comportament de les neurones biològiques ha demostrat resultats prometedors en modelatge seqüencial (Neural ODE [1], Liquid Networks [2], CfC [3]). Amb la publicació de [3] s'aconsegueix una aproximació per la solució d'aquestes EDOs que promet incorporar escalabilitat a aquests models.

A pesar que el problema de odometria visual (1a Part) ja està molt investigat, en aquest projecte m'enfocaré a aplicar les últimes tecnologies de IA a aquest problema. Concretament l'objectiu serà incorporar les aportacions de [3] al sistema desenvolupat a [4] per aconseguir un sistema end-to-end que sigui, idealment, robust, general i explicable, al mateix temps que es respecta el principi de causalitat. Això ho faré reemplaçant el mòdul RNN de [4] per [3].

3. Dades / Coneixement del que es disposa

Jo personalment ja disposo d'experiència prèvia entrenant sistemes de deep learning per identificació d'objectes (localització i classificació dins la imatge) en temps real.

Per entrenar i avaluar el sistema utilitzaré el dataset de odometria visual de KITTI [5]. Concretament utilitzaré el subconjunt de clips en escala de grisos degut a limitacions de memòria i còmput. Aquest subconjunt consisteix en 22 seqüències de vídeo estereoscòpic (tot i que jo utilitzaré només 1 càmera com a input). 11 de les 22 seqüències disposen de ground truth per poder entrenar, mentre que les 11 restants no tenen ground truth i s'utilitzen per avaluar el model.

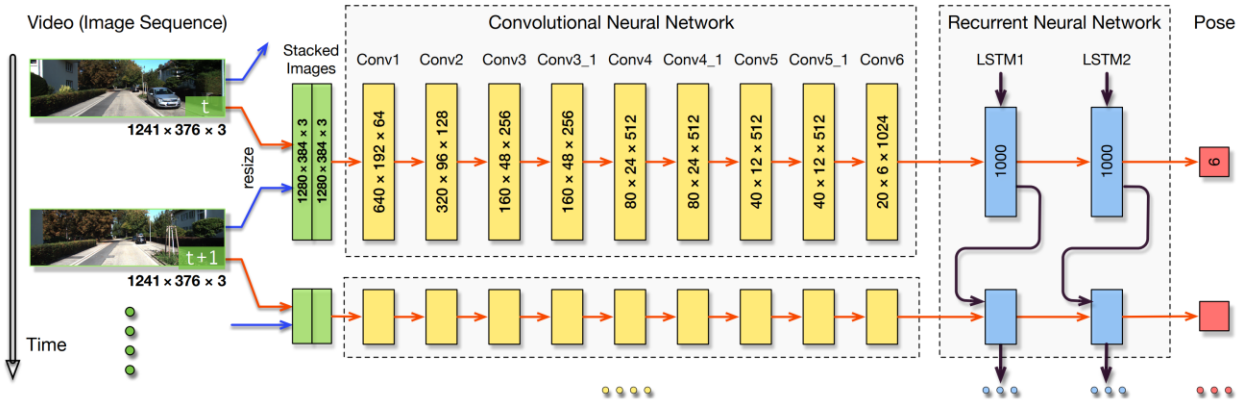
Aquest subconjunt pesa 22 GB, per tant hi ha 11 GB de vídeos estereoscòpics etiquetats per entrenar. En teoria això consistiria en 11 vídeos, però com que només entrenaré en mode monocular, caldrà experimentar si val la pena introduir els vídeos de l'ull dret i l'ull esquerra com si fossin vídeos diferents, duplicant la quantitat de clips a efectes pràctics; o si els vídeos son massa similars entre ells i causen overfitting.

Utilitzaré la mínima quantitat de pretractament: normalitzar per tal de tenir els valors dels píxels entre 0 i 1 enlloc de 0-255. Depenen de les limitacions de hardware pot ser necessari reduir la resolució de les imatges, o fins i tot dividir un mateix clip en varies passades per mantenir l'ús de VRAM sota control.

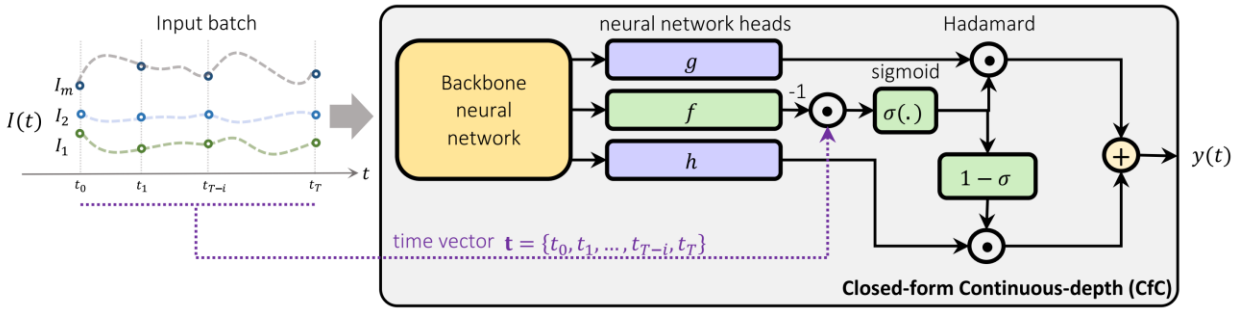
4. Proposta

L'objectiu és incorporar les aportacions de [3] al sistema desenvolupat a [4] per aconseguir un sistema end-to-end que és, idealment, robust, general i explicable, al mateix temps que es respecta el principi de causalitat. Això ho faig reemplaçant el mòdul RNN de [4] per [3].

Partint del model implementat a [4]:



L'objectiu serà reemplaçar el mòdul RNN per el mòdul de [3]:



A la pràctica això significarà que el "backbone" del segon diagrama podrà ser el CNN del primer diagrama (ja que aquest "backbone" no és part de les aportacions de [3], i per tant cal experimentar quina és la millor solució donades les restriccions de memòria, còmput i temps).

Caldrà retocar la implementació d'ambdós models per tal de donar suport a imatges en blanc i negre i poder connectar correctament les parts de les diferents xarxes (à la Frankenstein).

5. Experiments

6. Conclusions i línies futures

7. Referències

[1] R. T. Q. Chen, Y. Rubanova, J. Bettencourt i D. Duvenaud, 19 June 2018. [En línia]. Available: https://arxiv.org/abs/1806.07366.

[2] R. Hasani, M. Lechner, A. Amini, D. Rus i R. Grosu, «arXiv,» 8 June 2020. [En línia]. Available: https://arxiv.org/abs/2006.04439.

[3] R. Hasani, M. Lechner, A. Amini, L. Liebenwein, A. Ray, M. Tschaikowski, G. Teschl i D. Rus, «Nature,» 15 November 2022. [En línia]. Available: https://doi.org/10.1038/s42256-022-00556-7.

[4] S. Wang, R. Clark, H. Wen i N. Trigoni, «arXiv,» 25 September 2017. [En línia]. Available: https://arxiv.org/abs/1709.08429.

[5] A. Geiger, P. Lenz i R. Urtasun, «Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,» de *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.