UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Classificació Automàtica de Fruites Utilitzant Tècniques d'Aprenentatge Profund

Tesi de Grau

Presentada a la Facultat de

l'Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

per

Adrià Carbó Duch

En compliment parcial

dels requisits per el grau en

ENGINYERIA DE SISTEMES AUDIOVISUALS

Tutors: Josep Ramon Morros i Javier Ruiz Hidalgo

Barcelona, Setembre 2018

Resum

És crucial disposar de sistemes de detecció d'objectes precisos i fiables per a desenvolupar feines d'alt nivell en agricultura com serien fer un mapatge del camp o robotitzar les collites. Aquest document utilitza una Faster-RCNN -que consisteix xarxa de detecció d'objectes de l'estat de l'art- orientada a la detecció de fruites que en aquest treball només seran pomes. La xarxa serà introduïda i explicada. Es fa un anàlisis d'obtenció dels paràmetres d'entrenament i diversos experiments orientats a maximitzar la finesa (accuracy en anglès) del model que es vol obtenir. La xarxa neuronal estarà consistirà en una part preentrenada una part completament per entrenar. Aquest estudi no ha aconseguit equiparar els resultats de treballs anteriors (F1 score > 0.9) però tampoc es pot dir que hagi obtingut mals resultats, com seria un F1-score de 0.85.

Agraïments

L'autor vol agrair el guiatge i motivació, les contribucions, l'ajuda i seguiments constants als seus tutors de Projecte, Ramon Morros i Javier Ruiz. També vol agrair al Fran Roldan i al Jordi Gené-Mola els consells rebuts sobre quin valor assignar a certs paràmetres o quines configuracions utilitzar.

Revision history and approval record

Revision	Date	Purpose
0	07/08/2018	Document creation
1	08/10/2018	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Adrià Carbó Duch	Carboad.14@gmail.com
Ramon Morros	Ramon.morros@upc.edu
Javier Ruiz	j.ruiz@upc.edu

Written by:		Reviewed and approved by:	
Date	07/10/2018	Date	08/10/2018
Name	Adrià Carbó	Name	Ramon Morros
Position	Project Author	Position	Project Supervisor

Taula de continguts

Llistat de figures

Llistat de Taules:

1. Introducció

La detecció de fruites basada en la visió és un component clau per la robotització de l'agricultura. Això ve donat a que permet saber la localització de cada fruita en un camp, saber el rendiment a cada posició del camp i poder-lo mapejar, fet important per als productors ja que facilita fer un ús més eficient dels recursos i incrementar la producció i la qualitat podent tallar el excés de fruita allà on n'hi hagi [1]. No només seria un fet important per als productors, sinó que també s'intenta avançar en el repte de satisfer una demanda d'un increment de la producció i de la qualitat de les fruites en un món en constant creixement de la població [2]. Aquesta robotització també és important degut a que billons de fruites s'han de collir a mà per els treballadors, fet que causa problemes d'esquena, haver de fer llargues pujades, superar dificultats degut a l'altura on la fruita està localitzada. Fer un model de collita mitjanament robotitzat milloraria molt les condicions laborals d'aquests treballadors i minimitzaria els riscos de patir accidents laborals [3]. També està el fet de que els rendiments fins ara s'han calculat en base a un arbre-mostra, i contant manualment el nombre de pomes fet que és molt susceptible a l'error.

Recentment les DNNs han assolit l'estat de l'art en el camp de la detecció d'objectes, aprenent automàticament representacions de característiques que capturen la distribució de les dades, quan anteriorment les característiques s'havien d'extreure fent a mà processos a les dades [4].

Els requisits marcats per aquest treball són detectar pomes en horts i millorar els resultats de l'estat de l'art així com les especificacions són utilitzar DNNs per detecció, utilitzar tècniques de l'estat de l'art i millorar la DNN's afegint un bloc de RPN.

Les dades que utilitzarem seran imatges RGB dels horts de Lleida capturades amb una càmera que amb un vehicle passa pels camps. La meta d'aquest treball és desenvolupar un model d'una Faster-RCNN amb l'objectiu d'igualar l'estat de l'art o bé de millorar-lo, utilitzant tècniques de "transfer learning". El model de la xarxa, partirà de la implementació d'una Faster-RCNN en pytorch de Jianwei Yang, Jiasen Lu, Dhruv Batra i Devi Parikh[1].

Així doncs, aquest treball consisteix en una implementació d'un detector de pomes fent servir una Faster R-CNN utilitzant dades de camps de pomes de la província de Lleida. Forma part d'un projecte que consisteix en robotitzar parcialment els camps de Lleida, però en aquest treball ens centrarem únicament en fer una correcta detecció de pomes. També es farà una comparació dels resultats utilitzant el conjunt de dades de ACFR[2]

El pla de treball ha consistit en fer primer una recerca sobre les tecnologies a utilitzar és a dir recerca en Xarxes Neuronals Convolucionals, trobar una arquitectura de sistema que es pugui aplicar satisfactòriament (s'ha escollit la Faster-RCNN i fer recerca en treballs previs en l'àmbit de la detecció d'objectes [4],[5]. Després vindria implementar (utilitzant tècniques de "*transfer learning*" el model escollit, crear el nostre classificador específic i obtenir resultats intentant posteriorment millorar els resultats de l'estat de l'art.

[1]. Es pot trobar a https://github.com/jwyang/faster-rcnn.pytorch

[2]. Accessible dés de http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/

Degut a incidències en el re-entrenament del model que hem triat, totes les fases del projecte posterior a poder utilitzar perfectament el model s'han vist retardades temporalment. També valdria dir que el fet de preparar el conjunt d'entrenament tampoc havia estat gaire contemplat, així doncs un cop hem implementat el model i l'hem entrenat per conjunts d'entrenament[2] prèviament existents amb dades externes (Conjunt de Dades d'ACFR), s'ha hagut de preparar el conjunt d'entrenament amb les dades pròpies de tal manera que fos el màxim de semblant amb el conjunt d'entrenament, per així poder crear un conjunt de dades que integri les dades del conjunt de dades d'ACFR amb les de Lleida i poder tenir un conjunt de dades més grans que permeti optimitzar encara més el model de Faster-RCNN.

El diagrama de Gantt és el següent:

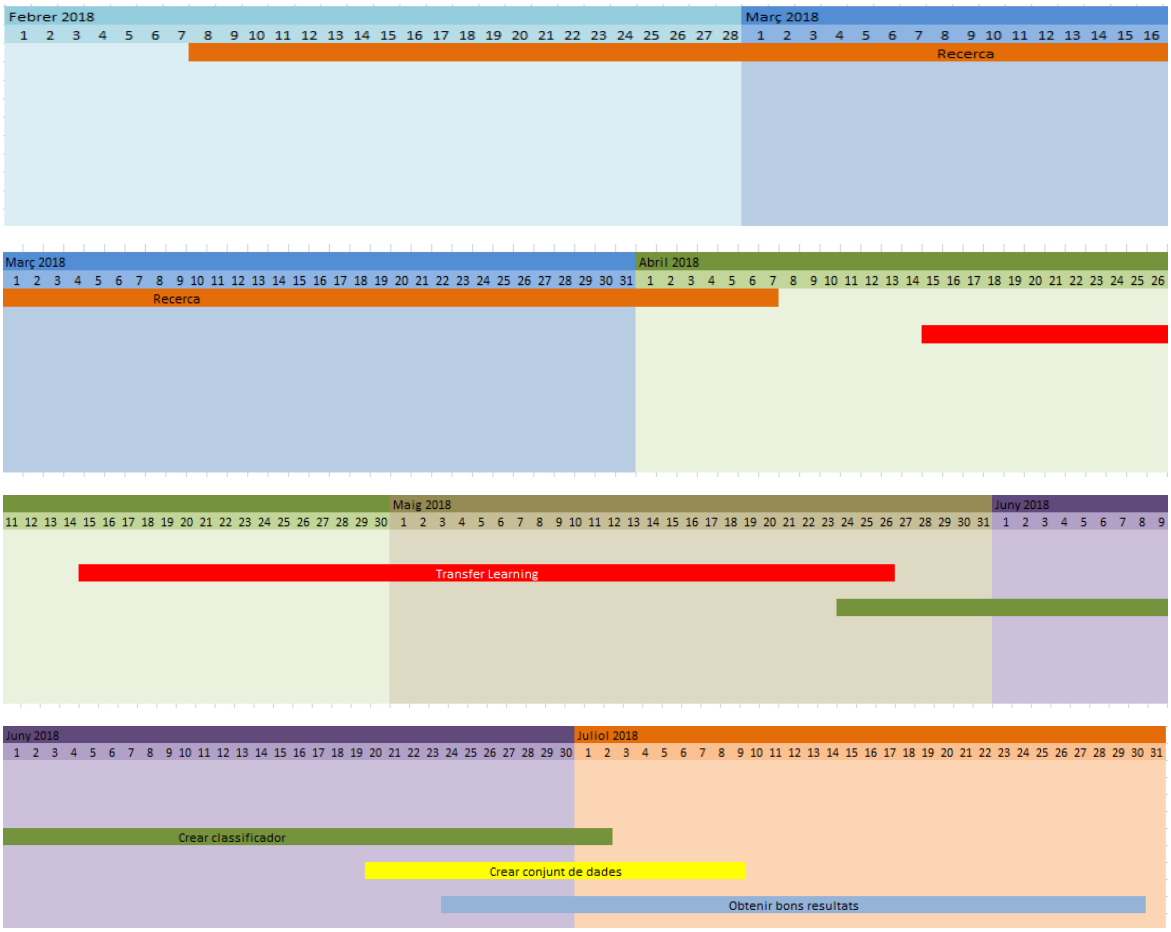

Fig. 1. Diagrama de Gantt del pla de treball escollit

2. Estat de l'art de la tecnologia utilitzada o aplicada en aquesta tesi:

Si detectem a partir d'imatges capturades amb sensors RGB, l'algoritme de detecció és susceptible a condicions variables de llum, oclusions, agrupacions d'elements a detectar, la posició a la que es pren la imatge, entre d'altres.

Els sensors que es poden utilitzar son càmeres de blanc i negre, càmeres RGB, càmeres espectrals o càmeres tèrmiques. Els sensors més utilitzats són les càmeres RGB donat que són els sensors més econòmics en relació a la informació que ens donen, ja que ens donen informació del color, de la forma geomètrica i de la textura, però són bastant susceptibles a les condicions lumíniques. Les de blanc i negre en canvi són independents dels canvis d'il·luminació però no se'n pot extreure tantes característiques. Les càmeres espectrals ens poden donar informació espectral i de color però té el desavantatge de que consumeix molt de temps, i són menys assequibles que les càmeres RGB que poden estar integrades en dispositius d'ús diari (mòbils). Les càmeres tèrmiques són molt útils a l'hora de separar el fons del primer pla (*foreground*), ja que detecten la firma tèrmica de cada element del primer pla. En aquest estudi, només s'han utilitzat imatges RGB, per tant sabem que no hem disposat de moltes característiques. Estudis suggereixen que com més característiques relacionades s'integri, a més finesa podrà arribar el nostre model, per tant una possible millora seria utilitzar imatges capturades amb el màxim nombre de sensors i així el nostre model podrà disposar de més informació [3].

La detecció, típicament està portada a terme fent transformacions de les regions d'una imatge en espais de característiques discriminants, i amb l'ús de classificadors prèviament entrenats associar els espais o bé amb fruites o bé amb fons, com serien branques, fulles o el sòl [4], és a dir fer una segmentació d'una imatge en fruita i fons. Per separar fruites individuals es poden utilitzar tècniques de pre-processat. Aquestes regions d'imatges, dites també regions d'interès (RoIs, acrònim anglès) serien les possibles candidates a ser classificades com a fruita.

Recentment, les R-CNNs [10] van establir resultats de l'estat de l'art en el conjunt de dades de PASCAL-VOC [12]. Les R-CNNs extreuen primer les RoIs utilitzant Cerca Selectiva (Selective Search en anglès) que consisteix en trobar regions d'interès combinant superpíxels. Aleshores les CNNs són utilitzades per classificar les regions i fer una regressió de quadres delimitadors (bounding box en anglès) a la localització de la imatge on hi ha l'objecte detectat i contingut dins del bounding box. El problema de utilitzar cerca selectiva és que és cara en un sentit temporal.

En treballs posteriors, es va proposar el model de la Faster-RCNN [5], que faria una combinació de l'extracció de regions amb la classificació d'objectes i la regressió de bounding boxes en un model unificat de xarxa profunda de detecció d'objectes.

Aquesta xarxa de fi-a-fi (end-to-end) ha comportat millores significants en els resultats de detecció a la vegada que ha reduït significativament els temps d'entrenament i de testejament.

També s'han implementat xarxes orientades a la detecció i classificació d'objectes ([12]) que fan la detecció més ràpida com serien la xarxa YOLO ([13]) i la SSD ([12]) però totes assoleixen un resultat de mAP similar (~78.5%) en el conjunt de dades de PASCALVOC 2007.

Per lo esmentat i perquè ens hem basat en un treball de detecció de fruites que ha utilitzat la Faster-RCNN [4], aquest treball ha estat realitzat amb un model de Faster-RCNN que ha estat implementat inicialment per Jianwei Yang i Jiassen Lu[1].

3. Metodologia / desenvolupament del projecte

A. Adquisició de dades

Les dades utilitzades han estat extretes dels horts de Lleida amb càmeres RGB d'alta resolució muntades en un dispositiu mòbil. Al capturar les imatges a intervals de pocs segons hi ha moltes imatges on hi surten les mateixes pomes (~4 imatges), fet important si per a treballs posteriors es vol fer un mapatge del hort (Figura 2).

Les imatges han estat capturades fent diferents obertures en el diafragma de la càmera i fent diferents temps d'obertura en el obturador, intentant així capturar dades on hi apareguin diferents il·luminacions ja que com hem dit anteriorment les xarxes neuronals de detecció d'objectes que s'entrenen amb imatges capturades amb sensors RGB són susceptibles a canvis d'il·luminació i així s'intenta superar aquest problema (Figura 3).

Ens referirem al conjunt de dades de les imatges preses a Lleida com a conjunt de dades de Lleida.

Fig. 2. Dispositiu mòbil que es mou pel camp amb molts sensors de captura, entre d'altres imatges RGB

Fig. 3. Com es pot veure a la figura, les imatges han estat preses a diferents graus d'iluminació

Les imatges de l'altre conjunt de dades d'imatges de pomes preses en camps d'Austràlia les anomenarem conjunt de dades d'ACFR.

B. Preparació de les dades

Com que en un futur es vol poder ajuntar el conjunt de dades de Lleida amb el conjunt de dades d'ACFR per poder tenir un conjunt de dades de pomes integrat i amb més informació, el que s'ha fet en la preparació de les dades per aquest treball ha estat intentar maximitzar la semblança entre els paràmetres de les imatges[3] de Lleida amb les imatges d'Austràlia.

Prèviament la xarxa va ser utilitzada amb imatges del conjunt de dades PASCAL-VOC on es detectaven i es classificaven diversos objectes i es va modificar perquè pogués funcionar primer amb el conjunt de dades de ACFR, canviant a la part de la xarxa neuronal que gestiona el conjunt de dades les classes de 20 classes a 1 classe (classe poma) i amb les anotacions en format xml. Aleshores en aquesta disposició s'ha preparat també el conjunt de dades de Lleida.

Les imatges originals del conjunt de dades de Lleida era de 1536x2304 i per cada imatge hi havia un nombre elevat d'objectes de detecció. Això repercuteix negativament ja que amb aquesta disposició de poques imatges (120 imatges) amb moltes pomes per imatge (67 pomes de mitjana per imatge), i pel funcionament seqüencial d'optimització dels pesos de la xarxa utilitzada durant l'entrenament, és a dir que s'actualitzen els pesos per cada imatge o grup d'imatges (image batch

[3]. Paràmetres d'imatge es refereix a les dimensions de les imatges, el nombre de píxels que ocupen els objectes de detecció, la informació de les annotacions, el format, etcètera.

en anglès) que transcorre la xarxa, no es poden fer tants passos d'optimització ja que depenen del nombre d'imatges del conjunt d'entrenament. Per solucionar-ho el que s'ha fet ha estat fer 20 retalls amb solapament de 128 píxels a cada imatge i així tenir un conjunt d'entrenament de 1920 imatges de 480x672 píxels (Figura 3). Per realitzar l'entrenament s'ha utilitzat la tècnica d'incrementació de dades d'emmirallar les imatges de tal manera l'entrenament és efectuat amb el doble d'imatges (i les seves anotacions conseqüentment modificades) ja que una imatge emmirallada és vista com una imatge diferent de la imatge original per la xarxa.

Els retalls de les imatges han estat fets amb l'objectiu d'aconseguir que les imatges del conjunt de dades de Lleida tinguin a l'entrada de la xarxa neuronal el mateix nombre de píxels per poma que les imatges del conjunt de dades d'ACFR que són aproximadament de 120x120 píxels/poma abans d'entrar a la xarxa on es redimensionen les dimensions fent que el costat més petit de les imatges sigui de 600 píxels.

Fig. 4. Aquesta figura mostra com han estat fets els retalls i els solapaments per tal de quedar-nos amb 16 imatges retallades per cada imatge original.

Un cop les imatges han estat retallades, les anotacions també s'han separat en funció de quin retall de la imatge ocupa. Aleshores amb els noms de les imatges (que són els mateixos que els de les anotacions però canviant l'extensió del fitxer) s'han creat els conjunts d'entrenament, de validació i de testejament fent particions aleatòries, en ordre, del 60%, del 20% i del 20%.

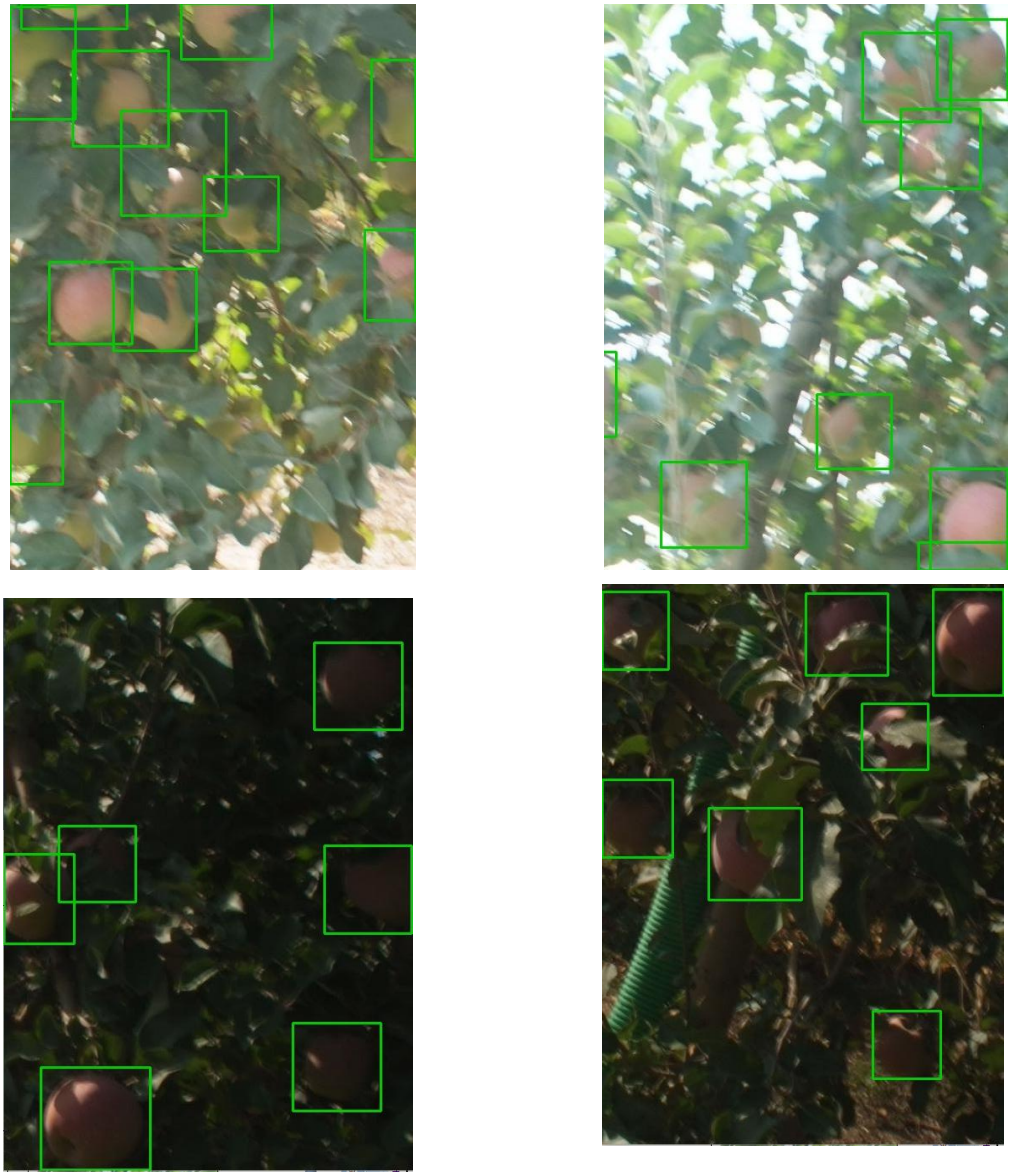

Fig. 5. Imatges retallades amb les seves anotacions superposades.

C. Faster-RCNN

La xarxa de detecció d'objectes Faster-RCNN ha estat utilitzada en aquest treball com a detector d'objectes. Tot i que altres xarxes han assolit resultats de l'estat de l'art i tenen velocitats de detecció més altes la hem escollit ja que en treballs previs ha estat utilitzada [4] i també perquè la velocitat de detecció no ens és rellevant per aquest treball.

Inicialment, la Faster-RCNN va ser implementada com a detector d'objectes en imatges RGB. Està construïda en base a dos mòduls: una RPN que identifica les regions d'interès que probablement contindran l'objecte; un classificador que classifica entre pomes i fons; i un regressor de bounding-boxes,. La diferència entre aquest mètode i el precursor d'aquest (Fast-RCNN) és que en aquest mètode, la part d'extracció de RoIs i la de classificació comparteixen les primeres capes convolucionals. La figura 6 mostra un diagrama de la Faster-RCNN.

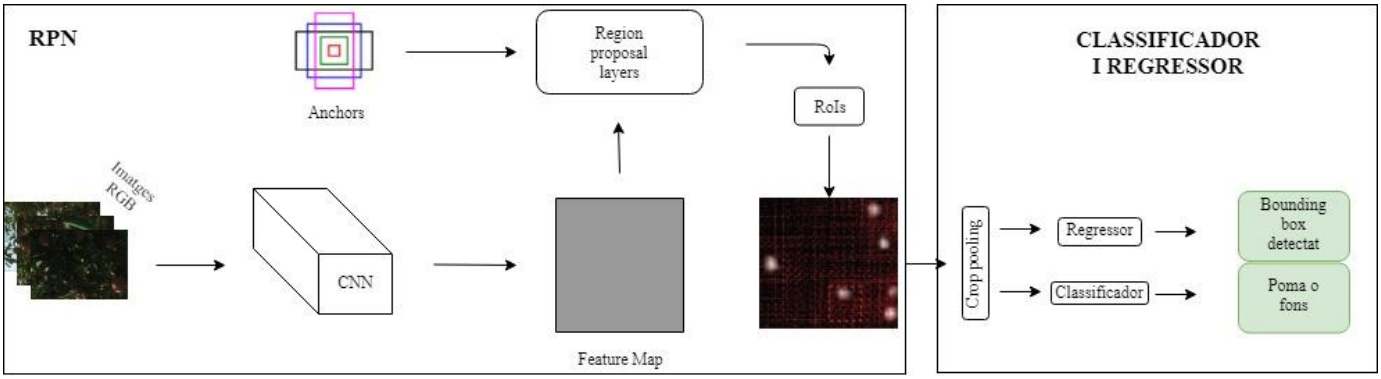

Fig. 6. Diagrama de la Faster-RCNN on es pot observar com a partir del mapa de característiques (feature map en anglès) la capa d'extracció de regions (Region Proposal Layers en anglès) extreu RoIs i aquestes RoIs s'ajunten al feature map, se'ls hi fa Crop Pooling i es classifica si la detecció és poma o fons i es fa la regressió del bounding-box.

D. RPN

Les primeres capes convolucionals de la RPN utilitzen el model de VGG-16 [14] prèviament entrenada amb el conjunt de dades ImageNet que s'afina (fine-tunning en anglès) amb les nostres dades. Abans de la última capa de la part convolucional (max pooling layer) es passa el *feature map* a una capa d'extracció de regions.

Al *feature map* se li aplica una finestra-lliscant de 3x3 píxels i movent-la 1 píxel es recorre tot el *feature map*. A cada posició de la finestra-lliscant es preveuran simultàniament múltiples propostes de regió on el nombre màxim de propostes de regió per cada posició serà determinada per *k*. Així que per cada finestra flotant en sortiran 4*k* coordenades (ja que cada proposta és referenciada per les coordenades d'un punt i l'amplada i l'alçada) i 2*k* qualificacions (que estimarien la probabilitat que la regió sigui una poma o sigui fons). Les *k* propostes estan parametritzades basant-se *k* rectangles de referència anomenats àncores.

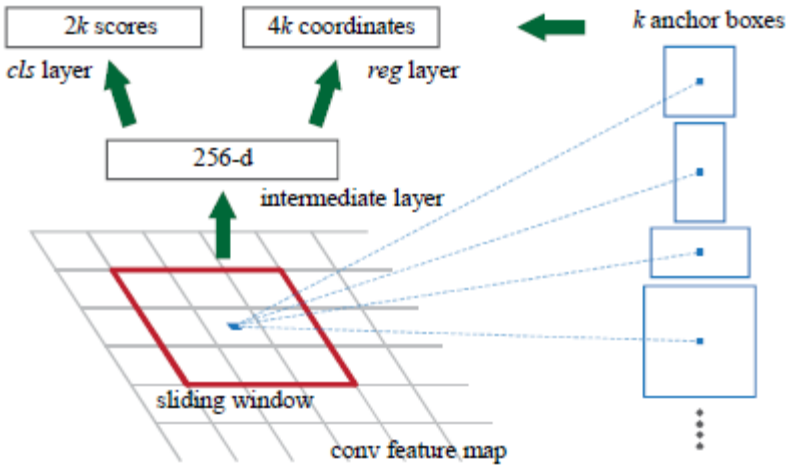

Fig. 7. Esquema on es pot veure el funcionament de la Faster-RCNN [5].

La capa d'extracció de regions extreu *k* propostes per cada finestra-lliscant aplicada sobre el *feature map*. Aleshores aquestes propostes de regions (sobre el *feature map*) és passada a una capa de *Crop Pooling* que extreu les regions retallades i les passa al classificador i regressor.

E. Experiments

Han estat realitzats diversos experiments amb l'objectiu de millorar els resultats obtinguts. Un dels experiments ha estat trobar de manera subòptima els paràmetres d'entrenament òptims com són la taxa d'aprenentatge de la xarxa, la disminució d'aquesta, i l'optimitzador a utilitzar. Aquests experiments s'han avaluat amb les corbes de pèrdues d'entrenament i validació. El procediment subòptim que s'ha seguit ha estat primer amb la taxa d'aprenentatge i la disminució d'aquesta fixades a 0.001 i a 5 i entrenar primer amb l'optimitzador SGD i amb l'optimitzador d'adam. Aleshores s'escull l'optimitzador que hagi fet que el model on s'utilitzi tingui un loss més baix de validació per alternar aleshores la taxa d'aprenentatge comparant els valors de 0.001, de 0.0001 i de 0.00001 fixant la disminució de la taxa a 5. Seguint el mateix procediment de comparar amb les corbes de validació, s'escull la taxa d'aprenentatge que presenti les pèrdues de validació més baixes. Un cop escollits aquests dos paràmetres d'entrenament s'alterna la disminució de la taxa d'aprenentatge comparant una disminució de 3, 4, 5 i 6 i seguint el mateix procediment ja estan el paràmetres d'entrenament escollits. Veure les comparacions a l'apartat de resultats. El mètode escollit és subòptim ja que l'ideal hagués estat comparar tota la combinatòria de paràmetres d'entrenament i escollir la millor combinació però per motius de temps no s'ha pogut realitzar.

Utilitzant els paràmetres d'entrenament (sub)òptims obtinguts a l'experiment anterior s'ha realitzat un altre experiment que ha consistit en entrenar amb aquests paràmetres el model utilitzant les imatge del conjunt de dades d'ACFR i del conjunt de dades de Lleida i la comparació ha estat realitzada amb les corbes d'entrenament i de validació i les corbes de precisió/*recall* i altres mesures d'avaluació com són la mAp i l'F1score.

També s'ha efectuat l'experiment d'afinar (fer *fine*-tunning) la RPN sencera començant a entrenar a partir d'un model ja entrenat de Faster-RCNN.

L'últim experiment realitzat en aquest treball ha estat el de variar la configuració inicial d'àncores, és a dir les *k* propostes de regions que es fan a la capa d'extracció de regions a partir del *feature map*. Aquest experiment podria ser prometedor donat que l'objecte de detecció (pomes) són d'una mida similar i potser es pot estalviar temps proposant les regions justes i necessàries.

Les àncores es configuren a partir d'un escalat i d'una proporció d'amplada i alçada. La base d'una àncora és quadrada amb el costat de 16 píxels. L'escalament inicial és de [8,16,32], és a dir hi haurà àncores quadrades de 16x8=128 píxels de costat, de 16x16 = 256 píxels de costat i de 512 píxels de costat. La proporció entre l'alçada i l'amplada és de [0.5,1,2] el que voldria dir de proporció 1:1 (quadrada), de 1:2 i de 2:1 (rectangulars). Per tant hi hauria 3 àncores quadrades de 128, 256 i 512 píxels de costat; i 6 àncores rectangulars de dimensions diverses, posant de base enlloc de àncores de 16 píxels de costat, àncores de 23 x 12 píxels, 11 x 22 píxels[4] i fer el seu escalat és a dir multiplicar cada dimensió per [8,16,32]. Veure figura 8.

[4]. Si es calcula la mida total d'aquestes àncores prèvies a l'escalat es veu que és de ~256 píxels

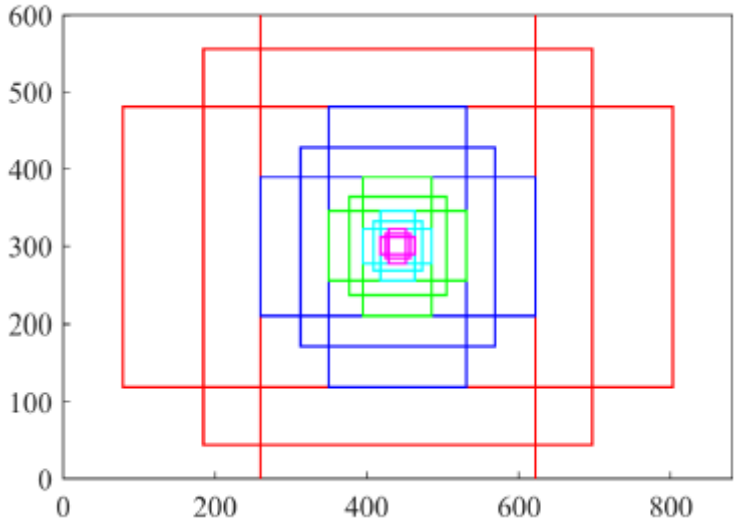

Fig. 8. Mostra de les àncores generades en una posició amb relació a la mida aproximada de la imatge

Aleshores l'experiment amb àncores consistirà en modificar l'escalat i la proporció de dimensions, per tant utilitzarem escalat de [8,16,32], de [4,8,16], de [2,4,8] utilitzant proporció de 1 només (és a dir àncores quadrades) i el mateix escalat utilitzant la proporció original ([0.5,1,2]).

F. Funció de pèrdues

Donat que molts resultats seran avaluats amb la pèrdua només s'ha vist interessant obrir aquest subapartat per explicar com s'ha fet el càlcul de pèrdues. Només es calcula quan s'està fent el procés d'entrenament per el conjunt d'entrenament i també per el de validació.

Al fer l'entrenament, se li assigna una etiqueta de classe binària a cada àncora. Se li assigna una etiqueta positiva a dos tipus d'àncores: (1) les àncores amb major intersecció sobre la unió (Intersection over Union, IoU en anglès), és a dir solapament amb un *bounding-box* de veritat base (*ground-truth*) o bé (2) una àncora que tingui més d'un 0.7 de solapament (IoU). Normalment la segona condició és més que suficient per determinar les mostres positives. Són classificades com a negatives les àncores que tenen un solapament inferior al 0.3 per tots els *ground-truth boxes*. Les mostres que no marquem com a positives ni negatives no contribuiran a l'objectiu de l'entrenament.

Dit això la funció que s'ha minimitzat és la següent:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*). \quad (1)$$

Aquí la *i* és el índex de l'àncora en un petit-grup (*mini-batch* en anglès), i la *p* és la predicció de probabilitat de que l'àncora *i* sigui un objecte. L'etiqueta de *ground-truth* és la *p** que és 1 si l'àncora és positiva i 0 si és negativa. t_i és un vector que representa les coordenades parametritzades dels *bounding-boxes* detectats, i t_i* correspondria a les coordenades parametritzades dels *bounding-boxes* de *ground-truth* associades a una mostra positiva. La pèrdua de classificació L_{cls} és l'error logarítmic entre dues classes. Per les pèrdues de regressió s'ha fet servir $L_{reg} = R(t_i - t_i^*)$ on *R* és la funció robusta de pèrdues (*smooth* L_1) definida a [15].

Els valors N_{cls} i N_{cls} són valors de normalització dels dos termes que compoen la funció. N_{cls} faria referència a la mida del *mini-batch* i el terme *reg* aniria associat al nombre total d'àncores. El paràmetre λ i així els termes *cls* i *reg* estan pràcticament ponderats [5]

4. Resultats

Els resultats del primer experiment que consistia en trobar els paràmetres d'entrenament (optimitzador, taxa d'aprenentatge i disminució d'aquesta) òptims i com ja s'ha explicat s'ha efectuat de manera sub-òptima. La figura 9 il·lustra les tendències de les pèrdues d'entrenament i de validació.

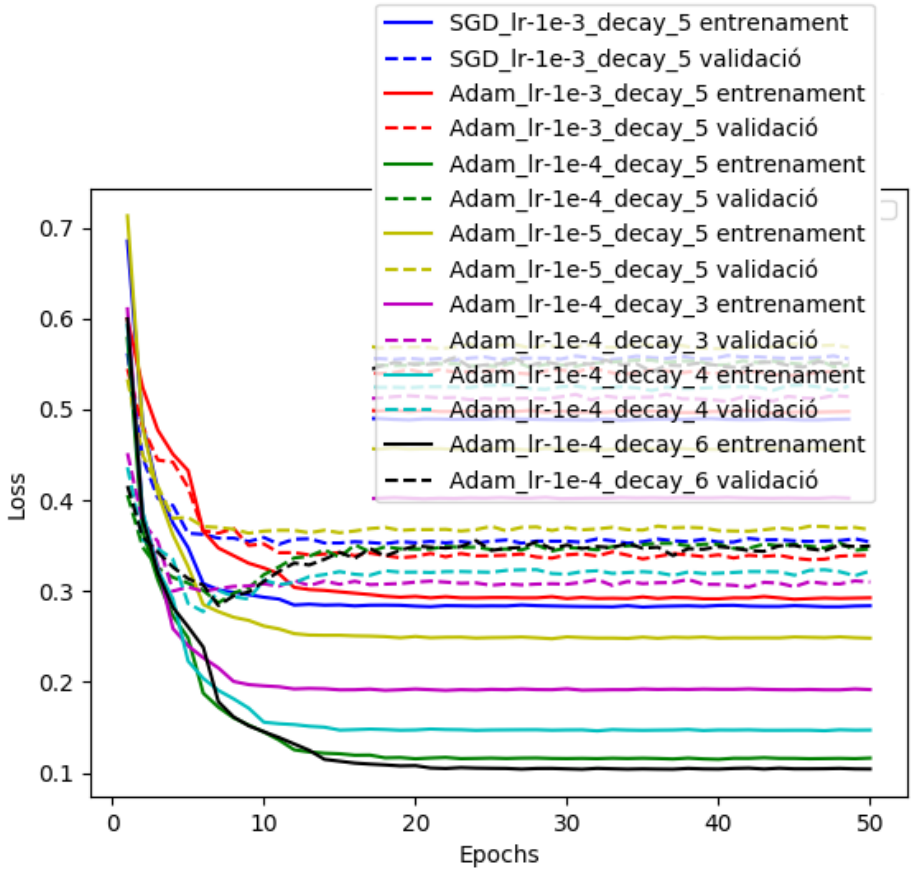

Fig. 9. Gràfica de pèrdues d'entrenament i de validació per múltiples configuracions dels paràmetres d'entrenament (veure la llegenda) sobre el nombre d'èpoques.

Com es pot veure a la figura 9, el model que aconsegueix minimitzar més la pèrdua de validació és l'entrenat amb optimitzador Adam, taxa d'aprenentatge de 0.0001 i disminució de la taxa d'aprenentatge de 4. Aquest model es sobreentrena a partir de l'època 9. Cap model entrenat amb SGD s'ha sobreentrenat. El fet que es sobreentrenin deu estar causat per el fet de tenir una taxa d'aprenentatge molt petita i perquè el nombre de dades pot ser considerat com a baix i el que pot fer la xarxa a l'entrenament és aprendre les "dades d'entrenament de memòria" i per tant generalitza en casos nous com el conjunt de validació. Per poder avaluar el model dibuixat de color turquesa es pren el model a una època prèvia a que la pèrdua pugi (època 9)..El llindar de solapament entre una detecció i el seu *ground-truth box* ha estat de 0.4 degut a que ha estat

[5]. S'ha escollit de 0.65 degut a que s'havien fet proves prèvies d'avaluació amb les corbes de precisió i *recall* i ens ha semblat un valor adient. Al fer corbes de precisió i de recall es variarà aquest llindar.

utilitzat en treballs anteriors sobre reconeixement de fruites de la mateixa manera que també s'hagués pogut prendre un llindar de 0.2 [4], [16].

Els resultats de comparar els conjunts d'entrenament d'ACFR i de Lleida han estat realitzats per diferents configuracions de paràmetres d'entrenament, utilitzant com a mètode optimitzador l'Adam amb una taxa d'entrenament inicial de 0.0001 i l'SGD amb la taxa de 0.001:

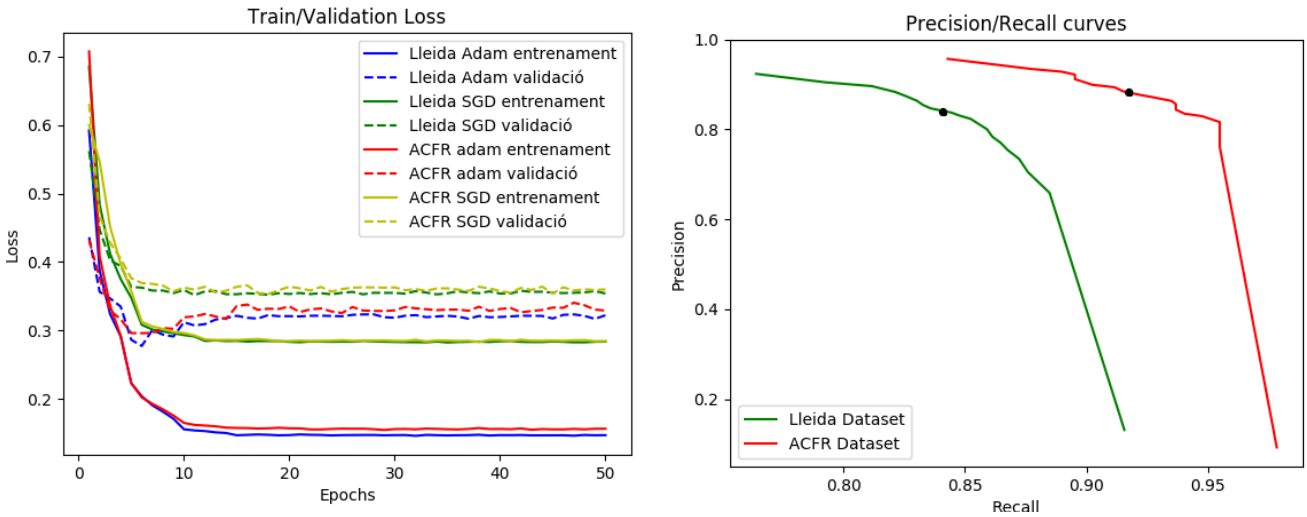

Fig. 10. A l'esquerra, corbes de pèrdues d'entrenament i de validació sobre els conjunts de dades de Lleida i d'ACFR. A la dreta, corbes de precisió i de *recall* en els conjunts de test del conjunt de test de ACFR i de Lleida, i marcat amb un punt negre on fixem el llindar de solapament de 0.6.

Les pèrdues de validació dels models entrenats amb els diferents conjunts tenen comportaments semblants en ambdós conjunts de dades, diferenciant-se només per la configuració d'optimitzador, ja que defineixen el mètode d'aprenentatge. Molts dels models que utilitzen l'Adam es sobreentrenen abans de la desena època com es pot observar a les figures 9 i 10, els que utilitzen el SGD no arriben a sobreentrenar-se però presenten resultats molt similars. En canvi la corba de precisió que presenta el conjunt de test d'ACFR és bastant millor que la de Lleida, mostren que el model entrenat amb ACFR presenta més bons resultats amb les deteccions sobre tots els positius (bon *recall*).

TAULA I

RESULTATS DE DETECCIÓ DE FRUITES UTILTIZANT CONJUNTS DE DADES DE LLEIDA I D'ACFR

Conjunt de dades	mAp(%)	F1 Score(%)
LLEIDA	0.82	0.84
ACFR	0.9	0.9

S'ha realitzat una prova per veure què passaria si enlloc d'entrenar la part d'extracció de dades des de 0 què passaria si li s'entreguessin d'aquesta part preentrenats amb el conjunt de dades d'ACFR a l'entrenament amb conjunts de dades de Lleida. Els resultats no van ser gaire prometedors ja que es va obtenir una mAp de 0.78, per tant s'ha arribat a la conclusió de que no és rellevant entrenar de 0 o afinar els pesos de la part d'extracció de regions. L'entrenament i l'evalucació d'aquest experiment es pot veure a la figura 11.

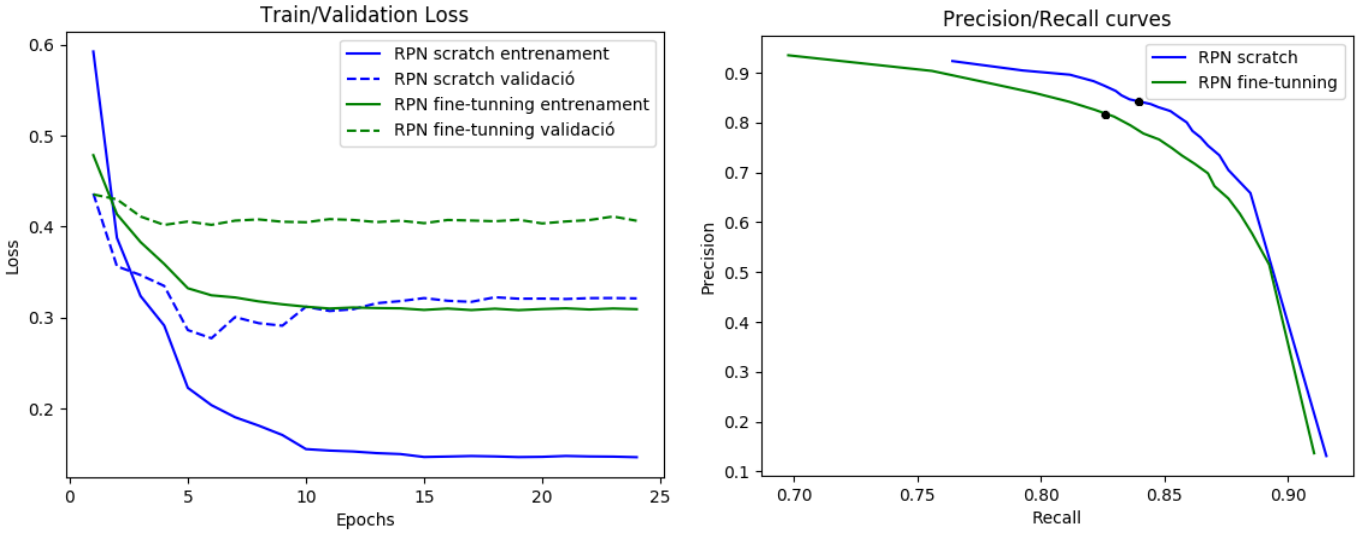

Fig. 11. A l'esquerra, corbes de pèrdues d'entrenament i de validació d'utilitzar els pesos de tota la RPN preentrenats en verd o només entrenats els pesos de la VGG-16 que s'agafa per construïr la RPN (referenciat com RPN des de 0). A la dreta les corbes de precisió i *recall* fent servir el conjunt de testejament a l'època 7 per la corba verda (finetunning) i per l'època 9 per l'altre model, ja que és l'època anterior a que el model e sobreentreni. Els punts en negre de l'esquerra correspondria al llindar de confiança escollit que és de 0.6.

Com es pot veure a la figura 11, entrenar des de zero la part d'extracció de regions és recomanable, donat que la corba de pèrdues de validació dels dos modes (*finetunning* i "des de zero") presenta mínims de pèrdues amb el mode d'entrenar "des de zero". Les corbes de precisió i *recall* reafirmen justament això, presenta més precisió i més *recall* el model entrenat "des de zero".

TAULA II

RESULTATS DE DETECCIÓ DE FRUITES UTILITZANT RPN PREENTRENADA O NO

Mode d'entrenament	Precision(%)	Recall(%)	mAp(%)	F1-score
RPN de 0	0.85	0.84	0.82	0.84
Finetunning de RPN	0.83	0.82	0.81	0.81

L'últim experiment que s'ha realitzat ha estat el explicat anteriorment sobre fer modificacions a l'escalat i a la proporció de les dimensions que se'ls hi aplica a les àncores que proposen regions. La disposició inicial de l'escalat i de la proporció seria, en ordre, [8,16,32] i [0.5,1,2]. Aleshores s'ha explorat fixar primer la proporció entre les dimensions a 1, és a dir àncores quadrades ja que els *bounding-boxes* detectats i de *ground-truth* són quadrats o quasi. Aquest experiment no ha anat gaire bé ja que ens ha donat moltes pèrdues. Per aquest motiu s'han repetit els experiments de modificar les escales però mantenint la proporció original de [0.5,1,2].

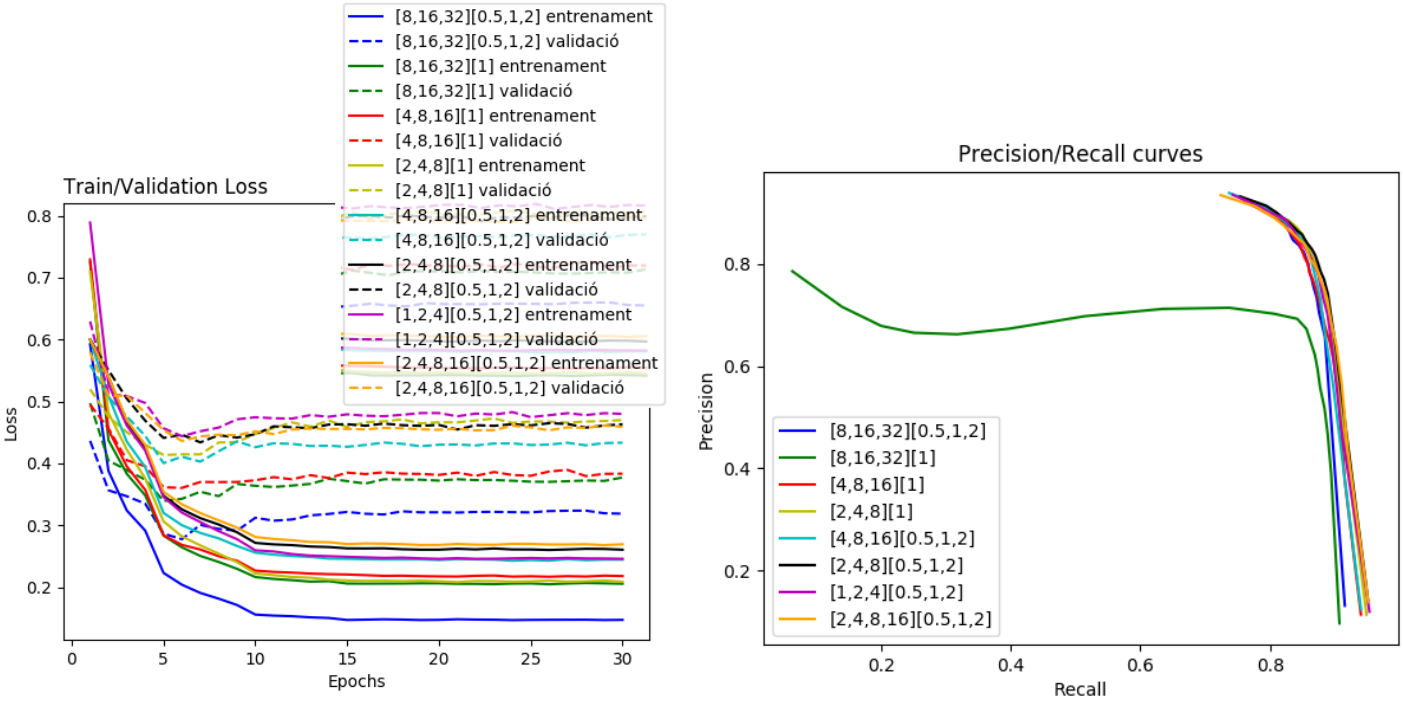

Fig. 12. A l'esquerra les corbes de pèrdua del model en el conjunt d'entrenament i de validació canviant l'escalat i la proporció entre dimensions de les àncores. A la dreta, corbes de precisió i *recall* del mateix experiment. Mostra sobretot els mals resultats que obté la configuració de [8,16,32] [1]

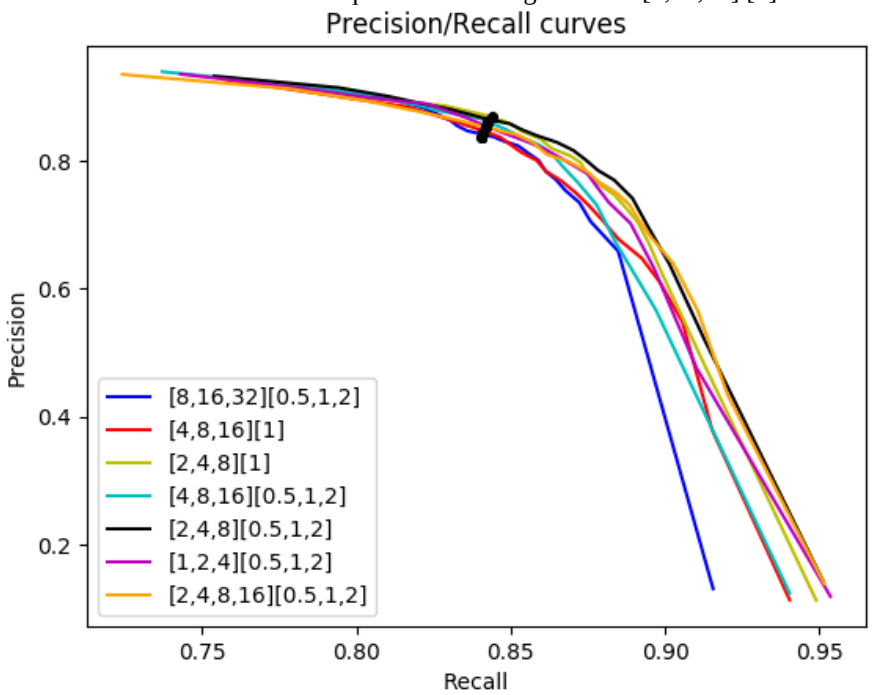

Fig. 13. Gràfica de precision recall dels models que ens interessen (sense la corba verda de la figura 12 a la dreta). El llindar de confiança ha estat establert a 0.6 (veure marques negres).

Tots els models han estat calculats a la època anterior de que les pèrdues comencin a pujar (veure gràfica de la dreta de la figura 12).

El model que utilitza la configuració de [2,4,8][0.5,1,2] ha estat el que ha obtingut més bons resultats. Veure Taula III. Contràriament al que s'havia plantejat al pensar que les anotacions eren totes quadrades potser funcionaria millor si les àncores fossin totes amb la mateixa proporció entre dimensions i l'algoritme d'extracció de regions aniria més ràpid detectant menys àncores a cada

posició de la finestra-lliscant. Això és degut a que l'algoritme NMS que hi ha després de la part *fully-connected* funciona bé per eliminar anotacions redundants i afinar la detecció.

TAULA III

RESULTATS EXPERIMENTS AMB ÀNCORES

Mode (escales i proporcions)	Precision(%)	Recall(%)	mAp(%)	F1-score(%)	Temps entrnmt. /època	Temps test
[8,16,32] i [0.5,1,2]	0.85	0.84	0.82	0.84	11 min 30 seg	55.21 seg
[8,16,32] i [1]	0.71	0.74	0.54	0.73	7 min 40 seg	40.87 seg
[4,8,16] i [1]	0.81	0.85	0.83	0.83	8 min 20 seg	48.54 seg
[2,4,8] i [1]	**0.83**	**0.86**	**0.84**	**0.85**	**7 min 30 seg**	**42.25 seg**
[4,8,16] i [0.5,1,2]	0.87	0.84	0.82	0.85	8 min 45 seg	50.17 seg
[2,4,8] i [0.5,1,2]	**0.86**	**0.85**	**0.83**	**0.85**	**9 min 15 seg**	**66.89 seg**
[1,2,8] i [0.5,1,2]	0.85	0.85	0.82	0.85	8 min 45 seg	67.29 seg
[2,4,8,16] i [0.5,1,2]	**0.84**	**0.85**	**0.83**	**0.85**	**8 min 45 seg**	**57.84 seg**

Com veiem els resultats més positius són els que estan fets amb la configuració d'àncores de [2,4,8,1] i [1] ja que són els que tenen una Precisió del 85%, un *Recall* del 0.86%, i sobretot una mAp de 0.84% i un F1-score del 85%. Com més àncores s'extreuen de cada mode el temps d'entrenament i de test és més gran. Les mesures de temps han estat obtingudes processant en un Titan X GPU.

5. Budget

El temps que s'ha dedicat al desenvolupament d'aquesta tesi ha estat entre 15h i 20h a la setmana. El còmput total d'hores estat de 465 hores. Si això fos remunerat a 8€/hora hauria tingut un cost de 3720€.

6. Conclusions i futur desenvolupament

Aquest treball ha presentat un sistema de detecció de fruites per imatges capturades en horts utilitzant una tècnica d'estat de l'art Faster-RCNN. Les fruites que s'han detectat són pomes. S'ha fet una comparació entre els conjunts de dades d'ACFR i deLleida. El model que ha utilitzat el d'ACFR ha tret més bons resultats (>0.9 dePrecisió, *Recall*, mAp i de F1-score) mentre que el conjunt de dades de Lleida (amb la millor configuració possible) s'ha quedat a, en ordre, 85%, 86%, 0.84% i 85%. No són mals resultats però es poden millorar. S'han aplicat mecanismes per aconseguir millorar els resultats com les tècniques d'incrementar les dades de girar les imatges. També s'ha estudiat la possibilitat de fer una millora agafant els pesos de la capa d'extracció de dades preentrenats però els resultats no han estat gaire bons, per tant és recomanable entrenar-ho des de 0, utilitzant ja els pesos de la part convolucional de la vgg16 preentrenats. S'ha analitzat la diferència que suposa el fet de canviar la configuració d'àncores fet que ha conclòs amb una millora del 2% en mAp i del 1% en F1-Score.

Un futur i prometedor experiment seria provar d'entrenar fixant l'escalament de les àncores a [4,8], [8,16], [2,4], [4], [8], [2], ja que la velocitat d'execució seria major i probablement s'obtinguin resultats semblats als actuals.

Partir d'imatges que no siguin RGB també pot millorar els resultats, donat que els sensors RGB són susceptibles a il·luminacions canviants i que hi ha més mètodes de capturar imatges, es podria sofisticar el model utilitzant imatges que provinguin d'altres sensors. Aquest experiment ja s'ha efectuat a [17] i sembla ser que els resultats milloren significativament

Treballs futurs implementaran un mapatge sencer de l'detectant, identificant, localitzant les fruites i establint un sistema integrat d'horticultura.

Bibliografia:

[1] K. Kapach, E. Barnea, R. Mairon, Y. Edan, i O. Ben-Shahar,"Computer Vision for Fruit Harvesting Robots - state of the Art and Challenges Ahead", *International Journal of Computational Vision and Robotics*, vol. 3, no. 1-2, pp. 4–34, apr 2012.

[2] Karen R. Siegel, Mohammed K. Ali, Adithi Srinivasiah, Rachel A. Nugent, K. M. Venkat Narayan, "Do We Produce Enough Fruits and Vegetables to Meet Global Health Need?", *PLoS ONE*, vol. 9, no. 8, 2014.

[3] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and systems for fruit detection and localization: A review," *Computers and Electronics in Agriculture*, vol. 116, pp. 8–19, 2015.

[4] Suchet Bargoti and James Underwood, "Deep Fruit Detection in Orchards", Submitted to the *IEEE International Conference on Robotics and Automation* 2017.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[6] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *British Machine Vision conference (BMVC)*, 2014.

[7] J. Hosang, R. Benenson, P. Doll´ar, and B. Schiele, "What makes for effective detection proposals?" *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[8] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra, "Object-Proposal Evaluation Protocol is 'Gameable'", *arXiv:1505.05836, 2015.*

[9] S. Bargoti and J. Underwood, "Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards," *Journal of Field Robotics* (accepted, under revision), 2016.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 1, pp. 142–158, 2016.

[11] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2017, 2017.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," pp. 1–14, 2014. [Online]. *Available: http://arxiv.org/abs/1409.1556*

[15] R. Girshick, "Fast R-CNN," *in IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] M. Stein, S. Bargoti, and J. Underwood, "Image Based Mango Fruit Detection, Localisation and Yield Estimation Using Multiple View Geometry," *Sensors*, vol. 16, no. 11, p. 1915, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/11/1915

[17] J. Gené-Mola, "Fruit Detection and Localization from RGB-D Sensors", *Master Thesis Dissertation, UPC, Setembre 2018*

Glossari

DNN – Xarxes neuronals profundes, Deep Neural Networks en anglès

RPN – Xarxa de proposta de regions, Region Proposal Networks en anglès

R-CNN – Xarxa neuronal convolucional basada en regions

Faster-RCNN – R-CNN més ràpida

ACFR – Australian Centre of Field Robotics

RoI – Regió d'interès, Region of Interest en anglès

mAp – mean Average precision, mitjana del promig de precisions

SGD – Disminució estocàstica del gradient, Stochastic Gradient Descent en anglès

IoU – Intersecció sobre la unió, Intersection Over Union en anglès