Universitat de Girona
Escola Politècnica Superior

Treball Final de Màster

Estudi: Màster en Ciència de Dades

Títol: Classificació de troballes en radiografies de tòrax

Document: Resum

Alumne: Pau Olivés Tarrés

Tutor: Robert Marti Marly
Departament: Arquitectura i tecnologia de computadors
Àrea: Arquitectura i tecnologia de computadors

Convocatòria (mes/any): Setembre 2022

Resum

El COVID-19 és un virus que ha tingut un impacte a nivell mundial durant aquests últims dos anys amb un índex de mortalitat elevat. És per això que tant la comunitat mèdica com la de intel·ligència artificial ha dedicat molts esforços a fer front a aquesta pandèmia.

S'ha demostrat que les xarxes neuronals convolucionals poden ajudar als metges amb l'anàlisi d'imatge mèdica per assistir en el diagnòstic o la detecció de malalties. En el cas del COVID-19, els esforços de l'aprenentatge automàtic s'estan centrant en la detecció de lesions en els pulmons causades pel virus.

El grup de recerca ViCOROB de la Universitat de Girona va proposar un projecte sobre aquesta temàtica. El projecte tracta de treballar sobre imatges de radiografies de tòrax i està dividit en tres parts que són: classificació de lesions, detecció de lesions i una tècnica de preprocess innovadora simulant els passos que realitzen els radiòlegs a l'hora d'analitzar imatges.

L'objectiu d'aquest treball de final de màster és contribuir en aquest projecte enfocant-nos en la part de classificació. Per tal de fer això s'ha proposat crear models de xarxes convolucionals utilitzant Transfer Learning d'arquitectures ja existents. Apart d'això, s'han proposat diversos experiments de classificació de radiografies per tal de veure quins donen millors resultats i a la vegada comparar les diferents arquitectures a cada experiment.

Les dades que hem utilitzat provenen de la competició de Kaggle pública "SIIM-FISABIORSNA COVID-19 Detection" que va acabar l'Agost de 2021. Aquesta competició en ofereix 6334 imatges de Train i 1263 imatges de Test. També ens proporciona dos fitxers CSV, un a nivell d'estudi i un a nivell d'imatge. El fitxer a nivell d'estudi ens diu la classe de lesió que s'haurà de classificar i el fitxer a nivell d'imatge ens dóna unes coordenades que són la regió amb la lesió que s'ha de detectar. Com que nosaltres ens centrem amb la classificació, farem servir principalment el fitxer a nivell d'estudi. En aquest cas, les classes són Negatiu, Aparença Típica, Aparença Atípica i Aparença Indeterminada.

Tot i agafar les dades de la competició, l'objectiu del treball no és en cap moment fer enviaments, ja que aquest treball es centra en la classificació i la competició requereix tant classificació com detecció per tal de fer un enviament. Per tant, en el treball no hem utilitzat les dades de Test de la competició i en el seu lloc hem dividit les dades de Train i hem agafat 80% per entrenament i 20% per validació.

El primer pas del treball ha sigut fer un anàlisi d'aquestes dades. El primer que hem observat ha sigut que estan molt desbalancejades, ja que quasi la meitat de les instàncies pertanyen a la classe d'Aparença Típica. Una altre cosa que hem vist és que hi ha algunes imatges que no són de classe Negatiu i tot i així no tenen

cap regió marcada com a lesió. Aquests casos no sabem si es tracten d'outliers o no, ja que a la competició no en fan cap referència, per tant els hem conservat. Un del anàlisis realitzats ha estat un mapa de calor de les regions amb lesions de cada classe. Això ens indicaria si les lesions estan situades en llocs concrets, però al mirar els mapes de calor veiem que estan repartides pels pulmons i no ens aporta molta informació rellevant.

El següent pas ha estat realitzar un preprocés de les dades. En aquest cas lo principal que hem fet és escalar les imatges perquè totes tinguin mida 224x224. Això ho fem perquè els models requereixen que totes les imatges d'entrada tinguin la mateixa mida. Pels últims experiments també hem escalat les imatges a 512x512 i hem aplicat una tècnica de preprocés innovadora creada pel grup ViCOROB que consisteix en ampliar la regió del pulmons, eliminant així tot el voltant que es podria considerar soroll.

Les arquitectures de Transfer Learning utilitzades en els experiments de classificació són la EfficientNetB0 que és un model que escala molt bé i té un cost de computació més baix que altres models. La ResNet50 que és una de les més utilitzades a l'hora de realitzar Transfer Learning. La VGG16 que és relativament senzilla, però que en el seu moment va resultar una millora important per les xarxes convolucionals. La Inception V3 que es basa en fer la xarxa ample en lloc de profunda. La Xception que és una modificació de la Inception portada al extrem. La DenseNet121 que és una xarxa molt profunda i evita l'overfitting fent connexions entre les primeres capes i les capes més avançades. Finalment, un cop tenim tots els models, hem creat un model d'ensamblatge ajuntant els models que han donat un resultat similar.

Al llarg dels experiments ens hem trobat amb una inconsistencia entre la funció model.evaluate() i model.predict(). Si fem servir la funció d'avaluació del model que ens ofereix Keras obtenim un resultat, però si fem servir la funció de predir etiquetes i traiem resultats a partir de la matriu de confusió, obtenim un resultat diferent. La diferència ha estat sempre del voltant de 20%.

El primer experiment realitzat ha sigut per tenir una base de la que partir. Hem utilitzat les imatges de 224x224, una mida de batch de 64 i hem entrenat els models durant 5 epochs. Els resultats d'aquest primer experiment ens mostra que els models EfficientNetB0, ResNet50 i VGG16 són els que ens donen millors resultats. En canvi, els models Inception V3 i DenseNet121 han estat els pitjors, ja que han predit totes les etiquetes como a una sola classe. Al crear el model d'ensamblatge, hem excluit aquests dos models, ja que afectarien negativament al resultat. Tot i això, el model d'ensamblatge no ha aconseguit donar millors resultats que els altres models per separat.

El següent experiment realitzat consisteix en fer una classificació binaria entre Positiu i Negatiu en lloc de les quatre classes proposades per la competició. Això ens podría permetre que, en cas de que la classificació fos bona, aplicar

després un segon model que classifiqui entre les tres classes de positiu. Malauradament, els resultats no han estat prou bons com per justificar crear aquest segon model. Els resultats han sigut similars a l'anterior experiment, amb la diferència que aquesta vegada la DenseNet121 ha estat la que millors resultats ha donat, seguida per les tres millors del primer experiment. La Inception V3 i la Xception ho han etiquetat tot com una sola classe, per tant s'han exclòs del model d'ensemblatge, el qual no ha aconseguit superar els resultats de la DenseNet121.

Després vam seguir amb un experiment que consisteix en realitzar Fine Tuning. Això consisteix en realitzar un entrenament amb el model base congelat i seguir-lo d'un altre entrenament amb el model descongelat. Les epochs que hem dedicat a cada entrenament són 3 i 15 respectivament. Aquest experiment té un alt cost de computació, el qual ens ha obligat a reduir la mida de batch a 16 i no ens ha permès crear el model d'ensamblatge. El que hem observat en aquest experiment és que, a excepció de la EfficientNetB0, tots els models han tingut overfitting. Tot i això, observem que a l'hora de classificar, cap model s'ha centrat únicament en una sola classe. Per tant, si aconseguissim reduir l'overfitting, aquest experiment té potencial de millorar els resultats dels altres. El problema és que l'experiment té un alt cost i temps de computació, i com que havíem de fer altres experiments, vam decidir seguir amb aquells.

El proper experiment és igual que les dos primers, amb la diferència que utilitzem imatges perprocessades amb l'algorisme que amplia la zona dels pulmons. Els resultats de la classificació entre les quatre classes ha estat molt similar als del primer experiment, amb la diferència de que els models han repartit més les etiquetes entre les quatre classes. En canvi, els resultats de la classificació binaria sí que mostren una lleugera millora respecte el segon experiment realitzat. Això mostra que aquesta tècnica de preprocés pot ajudar a millorar els resultats, encara que sigui lleugerament. L'últim experiment realitzat consisteix en el mateix que el primer experiment, però amb les imatges escalades a 512x512. Aquest experiment ha estat el que ha tingut major cost de computació i un altre cop hem necessitat reduir la mida de batch a 16. Tot i això, l'entrenament de cada model durava varies hores i com que teníem un temps limitat, vam haver d'aturar l'experiment després d'entrenar el segon model. Tot i així, podem observar com no només hem tingut molt d'overfitting amb els models entrenats, sinó que també han obtingut uns resultats molt dolents.

D'aquest treball extraiem les conclusions de que els models EfficientNetB0, ResNet50 i VGG16 són els que millor s'adapten a tots els experiments. El model DenseNet121 ha donat bons resultats en alguns experiments i mals resultats en altres. Els models Inception V3 i Xception són els que pitjors resultats han donat en tots els experiments. També hem observat que el preprocés d'ampliar la regió dels pulmons ha millorat lleugerament els resultats dels primers experiments.

L'experiment amb Fine Tuning ens ha mostrat que té potencial de millora i en un futur seria interessant dedicar millors recursos per intentar superar l'overfitting que hem trobat.

Com a treball futur, s'hauria de continuar amb el projecte realitzant experiments de detecció de lesions i implementar la tècnica de preprocés innovadora.