Universitat Oberta de Catalunya

uoc.edu

Identificació de sentiments en imatges

Xavi Gimeno Giménez
Màster de ciència de dades
Deep learning i Machine learning

Anna Bosch Rué
Jordi Casas Roma

3 de Gener de 2021

Llicències alternatives (triar alguna de les següents i substituir la de la pàgina anterior)

A) Creative Commons:

Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons

Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial-CompartirIgual 3.0 Espanya de Creative Commons

Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial 3.0 Espanya de Creative Commons

Aquesta obra està subjecta a una llicència de Reconeixement-SenseObraDerivada 3.0 Espanya de Creative Commons

Aquesta obra està subjecta a una llicència de Reconeixement-CompartirIgual 3.0 Espanya de Creative Commons

Aquesta obra està subjecta a una llicència de Reconeixement 3.0 Espanya de Creative Commons

B) GNU Free Documentation License (GNU FDL)

no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU Free Documentation License".

C) Copyright

FITXA DEL TREBALL FINAL

Títol del treball:	*Identificació de sentiments en imatges*
Nom de l'autor:	*Xavi Gimeno Giménez*
Nom del consultor/a:	*Anna Bosch Rué*
Nom del PRA:	*Jordi Casas Roma*
Data de lliurament (mm/aaaa):	*01/2021*
Titulació o programa:	*Màster en ciència de dades*
Àrea del Treball Final:	*Deep learning i Machine learning*
Idioma del treball:	*Català*
Paraules clau	*CNN, Visual Sentiment Analysis, Transfer Learning*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Els elements visual multimèdia com ara imatges, GIFs o vídeos s'han convertit, en l'actualitat, en un element present de forma gairebé inherent dins de les xarxes socials; les qual son, a dia d'avui, l'eina més potent per a comunicar-nos de forma global.

Amb l'increment de la capacitat computacional i la irrupció de les xarxes convolucionals, cada cop podem realitzar tasques més complexes dins del camp de la visió per computador.

L'objectiu d'aquest treball és investigar les arquitectures existents per a la classificació d'imatges i, aprofitant les característiques i capacitat d'aquestes, construir un model capaç d'analitzar el sentiment visual en les imatges.

Aconseguim aquesta fita a través de la utilització de la tècnica de la transferència d'aprenentatge. Aquesta tècnica ens permet re-utilitzar models prèviament entrenats amb grans quantitats de dades, per a l'extracció de característiques d'alt nivell de les imatges a analitzar per a la seva posterior classificació

Els models utilitzats com a base, estan basats en xarxes neuronals convolucionals, les quals representen, a dia d'avui, l'estat de l'art en la classificació d'imatges.

De forma addicional, s'aporten visualitzacions dels elements de la imatge que cada model ha utilitzat per a determinar si el sentiment d'una imatge és negatiu, neutre o bé positiu.

Abstract (in English, 250 words or less):

Visual multimedia elements such as images, GIFs or videos are, currently, inseparable elements within the social networks. Nowadays, these social networks are, one of the most powerful tools to communicate among us in a global way.

With the increase in computational capacity and the emergence of convolutional networks, we can increasingly perform more complex tasks within the field of computer vision.

The aim of this work is to investigate the existing architectures for image classification and, taking advantage of their characteristics and capacity, build a model capable of analyzing the visual feeling in the images. We achieve this goal by using the learning transfer technique.

This technique allows us to re-use previously trained models with large amounts of data, for the extraction of high-level features of the images to be analyzed for later classification.

The models used are based on convolutional neural networks, which today represent state of the art in image classification. Additionally, visualizations of the elements of the image that each model has are provided used to determine whether the feeling in an image is negative, neutral, or positive.

In addition, visualizations of the elements of the image that each model has used to determine if the sensation of an image is negative, neutral, or positive are provided.

Índex

1. Introducció

1.1 Context i justificació del Treball

L'anàlisi automàtic de sentiments és, en l'actualitat, una disciplina de notable rellevància i que presenta diferents reptes per a resoldre. Recentment s'han produït avenços importants en l'anàlisi de sentiments en textos gràcies al processament del llenguatge natural (PLN) com per exemple el model BERT [1] de Google o XLNet [2].

Aquests models a més estan disponibles en diferents *frameworks* i APIs que permeten a qualsevol usuari construir un model d'anàlisi de sentiments en text que obtingui uns resultats equivalents als de l'estat d'art en aquesta disciplina.

D'altra banda, la majoria contingut que es comparteix per Internet i, especialment a través de les xarxes socials, ve acompanyat de contingut visual, ja siguin imatges o vídeos. Aquest fet posa de manifest la importància d'analitzar també els sentiments que generen aquests continguts visuals.

En aquest context on les imatges o altres elements visuals multimèdia com les emoticones, vídeos o GIFs constitueixen de forma natural, el contingut generat a les xarxes, obtenir una eina que permeti analitzar el sentiment que genera una imatge te un gran nombre d'aplicacions diferents àmbits.

En l'actualitat, les xarxes neuronals convolucionals conformen l'estat de l'art pel que fa les tasques de visió per computador, on s'han demostrat com a eines vàlides per tasques de visió per computador com: classificació d'imatges, detecció d'objectes o la transferència d'estils.

L'objectiu d'aquest treball és investigar sobre els models existents per a la classificació d'imatges i crear, a partir d'aquests, un classificador genèric de sentiments en imatges.

1.2 Objectius del Treball

- Obtenció d'un *dataset* amb etiquetes que permeti entrenar el model realitzat en aquest treball.

- Creació d'un model d'aprenentatge automàtic basat en xarxes convolucionals profundes que realitzi una classificació de sentiments.

- Creació d'un model d'aprenentatge automàtic basat en mecanismes d'atenció i xarxes convolucionals profundes que realitzi una classificació de sentiments.

- Comparar i analitzar els models obtinguts per definir amb quin s'obtenen millors resultats per a la tasca proposada.

1.3 Enfocament i mètode seguit

Degut als requisits de *hardware* per a l'entrenament d'un model complet des de zero, s'aplicarà el mètode de *transfer learning*, que permet re-utilitzar un model prèviament entrenat amb el que podem capturar les característiques de més baix nivell de les imatges processades. A més d'aquest model prèviament entrenat, s'afegiran noves capes que permetran realitzar la classificació desitjada. Aquestes darreres capes de xarxes neuronals, s'entrenaran per complet durant l'execució d'aquest treball.

Actualment existeixen diversos models que son capaços de realitzar una classificació d'imatges. Alguns d'aquests models, com el AlexNet [3], han estat entrenats amb el *dataset* ImageNet [4], una col·lecció de més de 15 milions d'imatges genèriques - no pertanyen a cap domini en concret- repartides en més de 22.000 categories .

Donat que l'objectiu és l'obtenció d'un classificador genèric de sentiments per a imatges, es considera una bona estratègia l'ús de la tècnica de *transfer learning* per a re-utilitzar aquests models existents que ja son capaços d'extreure diferents característiques de les imatges processades i centrar-se únicament en el desenvolupament i entrenament de les darreres capes encarregades de realitzar la classificació de sentiments.

1.4 Planificació del Treball

Per a la realització del treball s'utilitzaran les següents plataformes online que posen a disposició dels seus usuaris recursos GPU i TPU, que permeten un millor processament en tasques d'anàlisi d'imatges:

- Google Colaboratory: 12h consecutives de GPU i TPU, subjectes a disponibilitat dels recursos.

- Kaggle: 30h setmanals de GPU i TPU.

Durant les primeres setmanes es realitzarà un estudi per esbrinar quin es l'estat de l'art en aquesta matèria. L'objectiu d'aquesta primera fase és entendre quines son les solucions actuals per a l'anàlisi de sentiments en imatges, conèixer els models existents emprats en aquests solucions i identificar-ne aquells que poden ser útils per a la realització d'aquest projecte.

També durant aquesta primera fase d'anàlisi, es farà una recerca per trobar *Datasets* que puguin ser utilitzats posteriorment durant la fase de desenvolupament.
Un cop es conegui l'estat de l'art i s'hagi obtingut un o més conjunts de dades que puguem utilitzar, s'iniciarà la fase de desenvolupament. Durant aquesta fase l'objectiu serà l'obtenció dels models que realitzaran l'anàlisi visual dels sentiments.

En primer lloc, es desenvoluparà i entrenarà un model que faci un anàlisi utilitzant tota la imatge. Un cop obtingut, es procedirà a la creació i entrenament d'un model basat en mecanismes d'atenció.

La part final de la fase de desenvolupament consistirà en un anàlisi i comparació detallades dels dos models per definir quin és el que millor es comporta per a la tasca que s'està investigant en aquest treball.

El treball es dura a terme tenint em compte la següent planificació:

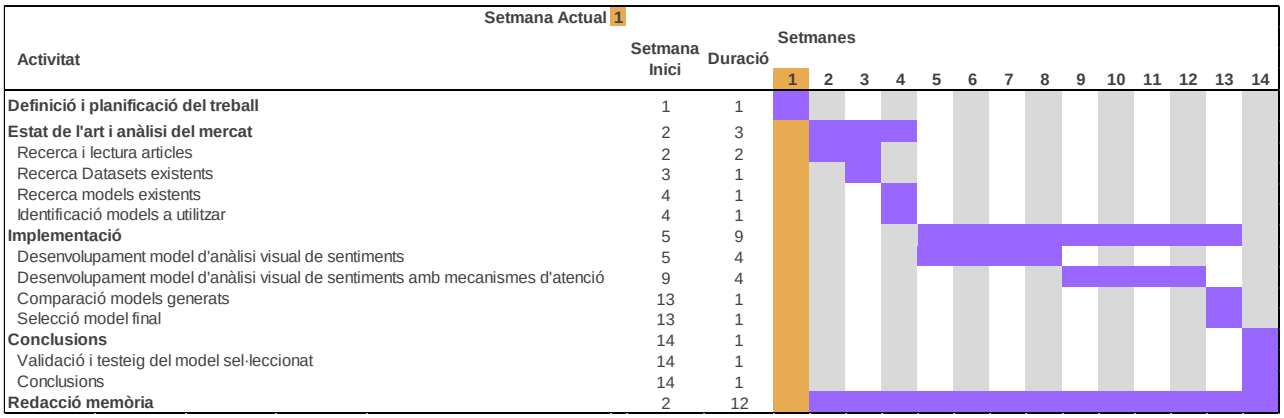


Figura 1: Diagrama de Gantt amb la planificació del projecte

1.5 Breu sumari de productes obtinguts

Durant aquest treball s'han generat dos models basats en xarxes neuronals convolucionals capaços d'identificar la polaritat del sentiment que evoquen, entre tres classes definides: Sentiments negatius, neutres o positius.

Els models han estat generats utilitzant la tècnica de la transferència d'aprenentatge sobre arquitectures existents, en les que s'han afegit una sèrie de capes addicionals per tal de realitzar la classificació desitjada.

Les dos arquitectures que s'han utilitzat com a base dels dos models a generar son ResNet-50 i SE-ResNet-50. Ambdues arquitectures s'expliquen amb detall a la secció 2.2, però cal destacar que el model SE-ResNet-50 representa una

modificació sobre l'anterior (ResNet-50) en el que s'introdueixen mecanismes d'atenció.

L'objectiu del treball, a més de la creació d'un model capaç d'identificar el sentiment subjacent entre les tres categories esmentades, és comparar l'efecte de la introducció d'aquests mecanismes d'atenció en la xarxa neuronal i veure com afecta a la precisió del model i en els elements en el que els models es focalitza alhora de determinar el sentiment de la imatge.

Els diferents elements desenvolupats durant aquest projecte es troben disponibles en el següent enllaç a la plataforma de github.

1.6 Breu descripció dels altres capítols de la memòria

En primer lloc i abans de començar a detallar els models desenvolupats es presenta un capítol on s'analitza l'evolució de l'estat de l'art pel que fa l'anàlisi visual del sentiments. En aquest capítol s'expliquen alguns dels primers models basats en l'anàlisi de les metadades que acompanyen les imatges així com la irrupció dels models basats en xarxes neuronals convolucionals, que conformen l'estat de l'art en l'actualitat.

Posteriorment, es dedica un breu capítol a l'anàlisi dels conjunts de dades existents i disponibles per a la tasca a realitzar. En aquest apartat s'identifiquen els conjunts de dades que seran utilitzats en la resta de capítols i s'explica les tècniques utilitzades per a la seva obtenció.

A continuació, en el capítol 4 s'explica amb detall les diferents activitats realitzades durant el desenvolupament i entrenaments dels models generats durant aquest treball. En aquest capítol veurem l'arquitectura de les xarxes creades, la estratègia emprada per al seu entrenament i els elements obtinguts durant la fase de desenvolupament.

Finalment i abans de les conclusions trobem un capítol amb els resultats obtinguts durant l'entrenament i les resultats de les proves realitzades sobre els models obtinguts.

2. Estat de l'art

L'anàlisi visual dels sentiments és una disciplina d'investigació que ha sorgit recentment. Molts dels treballs en aquesta línia es basen en estudis previs sobre l'obtenció semàntica d'emocions en imatges [5] que intenten relacionar les característiques de baix nivell trobades en les imatges amb emocions, amb l'objectiu d'obtenir i categoritzar automàticament aquestes imatges.

En els darrers anys, l'estat de l'art per a les tasques de visió per computador ha experimentat una transformació gràcies a les xarxes neuronals convolucionals (CNN), les quals han recuperat la popularitat empeses per l'augment de la capacitat computacional, principalment mitjançant l'ús d'unitats GPU.

2.1 Mètodes tradicionals

El primer article sobre l'anàlisi visual dels sentiments és de l'any 2010 [6] i en aquest l'autor estudia la relació entre els sentiments de les imatges representats a través de les seves metadades i el contingut visual de la imatge extreta de la plataforma Flickr. Per a realitzar aquesta tasca, l'autor fa servir el recurs lèxic SentiWordNet [7] amb el que analitza el text que acompanya la imatge (p.ex.: metadades) i assigna un valor numèric a cada imatge que representa com de positiva o negativa és. De l'estudi es desprèn que hi ha una forta correlació entre aquest valor numèric i les característiques visuals de la imatge com els histogrames locals/globals RGB o un *bag-of-visual words* basat en SIFT [8], un descriptor que detecta i descriu regions en les imatges.

Borth et al. [9] construeixen una ontologia de sentiments visual o VSO, (Visual Sentiment Ontology) aplicant la teoria psicològica de la roda de les emocions [10]. El resultat son 1,200 parells nom-adjectiu (ANP) que descriuen conceptes com ara "beautiful flowers", "sad eyes" o "disgusting food".

Amb aquesta representació, aconsegueixen convertir noms neutres com ara '*dog'* en un ANP amb certa connotació sentimental com "*cute dog"*. Aquests parells nom-adjectiu generen conceptes més detectables que no pas adjectius aïllats (com "*beautiful")* que son més abstractes i per tant més complicat d'identificar en les imatges.

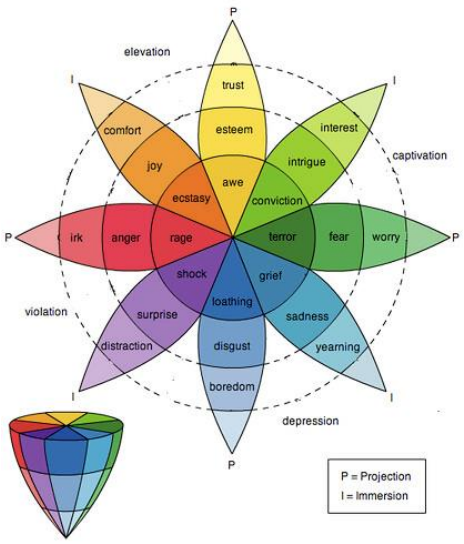


Figura 2: Roda de les emocions

Construït utilitzant aquesta VSO, els autors de l'article presenten SentiBank, una llibreria entrenada per a detectar aquests ANP dins de les imatges aconseguint així una representació de la imatge a mig nivell mitjançant els ANP identificats.

Els ANP detectats pel model son posteriorment avaluats amb els recursos lèxics SentiWordNet [7] i SentiStrength [11] per extreure'n un valor numèric que representi com de positiu o negatiu és el ANP detectat. Els resultats d'aquest treball mostren una millora en l'anàlisi visual del sentiments respecte la que s'obtenia analitzant únicament els textos associats a la imatge.

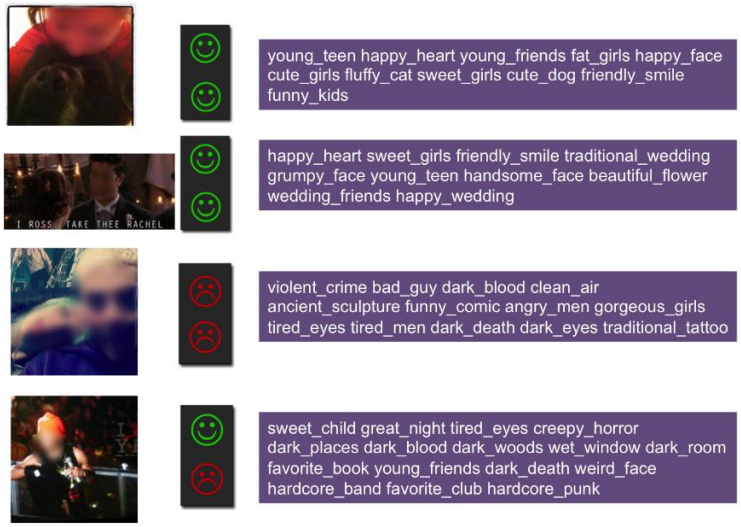


Figura 3: Prediccions SentiBank - Imatge, sentiment (icona superior), predicció (icona inferior) i ANP detectats. (Font: [9])

2.2 Mètodes basats en Deep Learning

Chen et al. [12] introdueixen una aproximació a la tasca de l'anàlisi visual dels sentiments basada en xarxes neuronals convolucionals (CNN) coneguda com a SentiBank 2.0 o DeepSentiBank. En aquest treball realitzen un ajustament fi d'un model previ existent [3] entrenat per a la tasca de classificació d'objectes amb el *dataset* ImageNet[4] amb l'objectiu de classificar la imatge en un dels 2,089 ANP obtinguts de la ontologia VSO creada a [9]. El resultat és una millora de forma significant la detecció de ANP respecte a [9] i en conseqüència un millor anàlisi dels sentiments sobre la imatge.

A l'article *Diving Deep into sentiment* [13] els autors presenten una CNN altre cop pre-entrenada per a la classificació d'objectes en la que es realitza un ajustament fi per a la tasca de predicció visual del sentiment. L'increment del poder computacional a les *GPUs* i la creació de grans col·leccions d'imatges com el *dataset* ImageNet [4] va permetre a les CNN obtenir grans resultats en tasques de computació visual, a més, per a aquest tipus de xarxes s'havia descobert com a eines efectives en experiments de transferència de domini o *transfer learning.* Sota aquesta premissa, l'autor realitza un ajustament fi sobre la xarxa CaffeNet [14], en la que l'última capa és substituïda per una de *fully connected* amb dues neurones, ja que l'objectiu és classificar els sentiments entre positius i negatius. La xarxa (excepte la nova capa afegida), és inicialitzada amb els pesos corresponents a la tasca original: la detecció d'objectes i posteriorment és entrenada utilitzant freqüències d'aprenentatge diferents per a la primera part de la xarxa, encarregada d'aprendre les característiques d'alt nivell de la imatge, i l'última capa, encarregada de la classificació i que utilitzarà una freqüència d'aprenentatge més alta per tal d'assolir una convergència més ràpida.

Sun et al. [15] investiguen en aquest article si descobrir regions afectives i trets locals complementaris poden incrementar el poder de predicció de les CNN utilitzades per a la tasca d'anàlisi visual dels sentiments. En aquest treball, els autors proposen un model en el que es combina la informació extreta de tota la imatge amb una sèrie d'objectes destacats extres de la imatge que probablement contindran més càrrega emocional. Aquests objectes o regions afectives son obtingudes a través d'aquesta eina [16], que genera *N* regions candidates a ser regions afectives.

Des de l'aparició de la xarxa AlexNet [3] l'estat de l'art de les xarxes convolucionals va anar evolucionant a xarxes cada cop més profundes. Mentre que AlexNet comptava amb 5 capes convolucionals, arquitectures posteriors com VGG [17] o Inception [18] tenien 19 i 22 capes convolucionals respectivament.

Malgrat aquest increment en la profunditat de les xarxes, no s'aconsegueixen millors resultats únicament per afegir més capes, degut al conegut problema de la desaparició del gradient [19]. Aquest fenomen s'observa quan es tracta amb múltiples capes ocultes, i dificulta enormement l'aprenentatge de les neurones en les primeres capes de la xarxa.

El problema de la desaparició o l'explosió del gradient és inherent al mètode d'entrenament de les xarxes multicapa, la retropropagació. En aquest mètode l'error comes a la sortida es propaga cap enrere per a calcular en cada capa, com s'han d'ajustar els pesos en funció del gradient. Quan estem entrenant xarxes multicapa, en les primeres capes aquest gradient es calcula a partir d'una funció basada en el producte dels termes obtinguts a totes les capes posteriors. De forma intrínseca, a mesura que entrenem capes més properes a l'entrada ens trobem amb una situació inestable que condueix al problema de l'explosió del gradient o de la desaparició del mateix.

Per a solucionar aquest problema, Kaiming et al. [20] presenten les *Deep Residual Network*, una arquitectura que introdueix la idea d'una connexió directe entre una capa i capes posteriors, com es mostra a la figura següent:

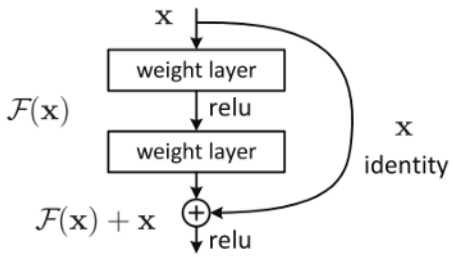


Figura 4: Bloc d'una xarxa residual

En l'article els autors comproven la hipòtesi següent: resulta més fàcil fer que la xarxa s'ajusti a una sortida residual, formada per la sortida desitjada de la xarxa i el valor a l'entrada que no pas ajustar la xarxa a la sortida desitjada únicament. Els autors argumenten que malgrat s'estigui produint un desaparició del gradient durant l'entrenament, es seguirà tenint valor de l'entrada per a la retropropagació a capes anteriors.

El model construït, anomenat del ResNet-50, consisteix en un bloc inicial i quatre fases més formades per un bloc convolucional i un d'identitat com es pot veure en la *Figura 5*.

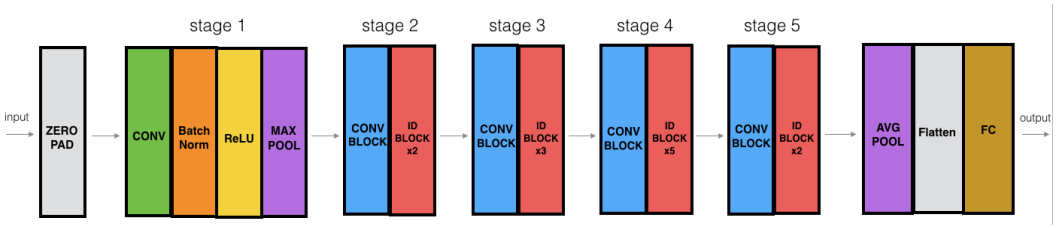


Figura 5: Model ResNet-50 (Font: https://towardsdatascience.com)

Aquest model va ser el guanyador de la competició *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) de l'any 2015 i serà la base per a la creació del nostre model. En concret, per aquest projecte utilitzarem la variant ResNet50, una xarxa amb 25 milions de paràmetres aproximadament i 50 capes CNN a la qual li realitzarem un ajustament fi, partint d'un model prèviament entrenat a través de la tècnica de transferència d'aprenentatge o *transfer learning*, que ha aconseguit bons resultats en treballs anteriors [21].

Des de l'aparició de les *Transformer Networks* [22] els mecanismes d'atenció han experimentat un creixement de popularitat en el camp de l'aprenentatge automàtic i especialment dins del processament del llenguatge natural (PLN). Anàlogament, aquests mecanismes estan començant a gaudir de major atenció en les tasques de visió per computador.

Entenem per mecanismes d'atenció aquelles tècniques que ajuden al model, durant la seva fase d'entrenament a identificar coses o elements més rellevants de forma més efectiva.

Hu et al. [23] presenten una arquitectura que anomenen *Squeeze-and-Excitation (SE)* amb la que pretenen incloure informació global en el procés de decisió de la xarxa. Mentre que les operacions de convolucionals únicament consideren la informació espacial dins d'un determinat radi, el bloc SE agrega la informació de tot el camp receptiu. A la *Figura 6* podem veure una representació d'aquest bloc:

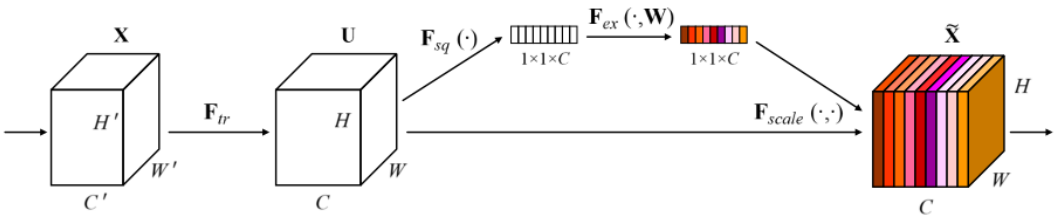


Figura 6: Bloc Squeeze-and-Excitation (Font: [26])

Una de les principals avantatges d'aquest bloc és que resulta molt flexible i pot ser incorporat en xarxes neuronals residuals existents. En l'article, els autors expliquen com pot ser integrada en xarxes conegudes com ResNet, Inception o ResNetXt. Dins de l'article s'explica amb detall el cost computacional que te introduir aquest bloc dins duna xarxa ResNet-50, que resulta ser d'aproximadament 2.5 milions més de paràmetres dels que utilitza la xarxa ResNet-50, que son d'aproximadament uns 25 milions.

En afegir aquest nou bloc dins de l'arquitectura original ResNet, obtenim el mou bloc que anomenen SE-ResNet i que podem veure en la Figura 7:

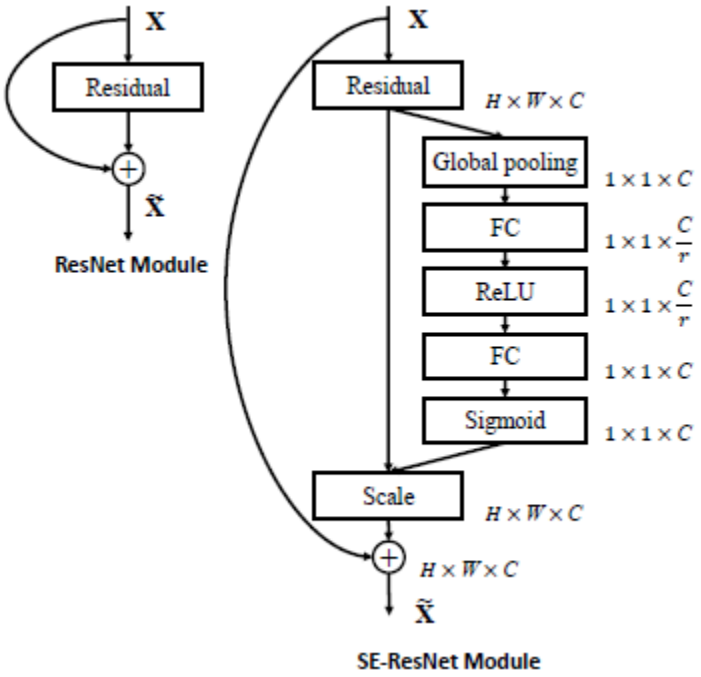


Figura 7: Bloc original ResNet (esquerra) i bloc Squeeze-And-Excitation ResNet (dreta) (Font: [23])

El segon model a construir en aquest projecte estarà basant en el SE-ResNet-50, per tal d'analitzar quin és l'impacte d'afegir mecanismes d'atenció en la tasca d'anàlisi de sentiments en imatges.

Taula 1: Resum de les publicacions sobre anàlisi visual de sentiments revisades en aquest capítol.

Article	**Tipus classificador**	**Entrada**	**Sortida**
Siersdorfer et al. [6]	SVM (Màquines de vector suport)	Característiques visuals extretes manualment	Polaritat del sentiment
Borth et al. [9]	SVM	Conjunts ANP	Polaritat del sentiment i tipus de sentiment
Chen et al. [12]	CNN	Imatge	Conjunts ANP
Campos et al. [13]	CNN	Imatge	Polaritat del sentiment
Sun et al. [15]	CNN + detecció de regions afectives	Imatge	Polaritat del sentiment
Kaiming et al. [20]	CNN	Imatge	Etiqueta Imatge
Hu et al. [23]	CNN + atenció	Imatge	Etiqueta Imatge

3. Datasets existents

Existeixen diverses fonts que poden ser utilitzades per a l'anàlisi visual del sentiment. Aquestes fonts proporcionen imatges etiquetades utilitzant un nombre diferent d'etiquetes en funció de l'objectiu que els creadors del dataset perseguien quan els van crear.

L'enorme creixement de les xarxes socials en els darrers anys han propiciat la possibilitat d'extreure grans col·leccions d'imatges de plataformes com Facebook, Twitter, Flickr o Instragram, que han permet la creació de grans datasets.

El repte però apareix alhora de l'etiquetatge d'aquestes imatges. L'anàlisi dels sentiments en les imatges és més complex que el reconeixement d'objectes, ja que per aquesta darrera tasca les etiquetes estan clarament definides mentre que l'anàlisi dels sentiments comporta un nivell d'abstracció més elevat i està condicionat per la subjectivitat humana davant la imatge, que pot evocar diferents sentiments en diferents persones.

Quanzeng et al. [24] van construir un dataset a partir d'imatges obtingudes de diferents usuaris de Twitter. Per al seu etiquetatge, van utilitzar els serveis de Amazon Mechanical Turk (MTurk), on cinc persones diferents classificaven 1269 imatges en tres sentiments: positiu, negatiu i neutre. Dins d'aquest dataset es poden trobar diferents sub-conjunts d'imatges en funció del nombre d'anotadors que van coincidir alhora d'assignar un sentiment: cinc anotadors, quatre anotadors, etc.

Vadicamo et al. [25] van extreure aproximament 3 milions de *tweets* entre Juliol i Desembre del 2016 de forma aleatòria entre els tots els *tweets* produïts al mon per tal de construir un nou *dataset* per a l'anàlisi visual del sentiments. D'aquests 3 milions de candidats, es van mantenir únicament aquells originals (descartant els re-*tweets*) que contenien almenys 5 paraules en anglès i almenys una imatge. Els *tweets* amb GIFs o vídeos van ser descartats. Utilitzant una arquitectura basada en LSTM-SVM van classificar tots els *tweets* en tres categories: positiu, negatiu i neutre analitzant el text del *tweet*. Posteriorment, es van eliminar imatges corruptes i duplicades i finalment, únicament aquelles imatges en les que la predicció del sentiment basat en el text associat superava un cert llindar, van ser considerades per a formar part del *dataset* final, anomenat T4SA (Twitter for Sentiment Analysis) i que consisteix d'un total de 1,473,394 imatges. De forma addicional, també generen el *dataset* B-T4SA en el que trobem un conjunt balancejat respecte les tres etiquetes utilitzades.


Etiqueta: Neutral


Etiqueta: Negativa


Etiqueta: Positiva

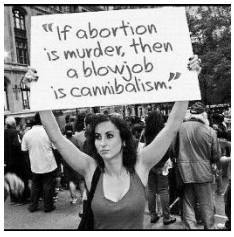
Etiqueta: Neutral

Figura 8:Imatges i etiquetes d'exemple dins del dataset B-T4SA

Utilitzarem el dataset B-T4SA per a l'entrenament i validació dels nostres models.

Els autors de la Visual Sentiment Ontology [9] van definir un banc de proves que contenia 603 tweets etiquetats com a positius o negatius utilitzant el servei de Amazon Mechanic Turk. En aquest servei consistia en tres persones que van analitzar 2000 imatges assignant un sentiment positiu o negatiu a cada una d'elles. Únicament aquelles imatges que van generar consens entre els tres Turkers, un total de 603, van passar a formar part del banc de proves.


Etiqueta: Positiva



Etiqueta: Negativa


Etiqueta: Negativa

Etiqueta: Positiva

Figura 9: Imatges i etiquetes del conjunt Twitter Dataset

Utilitzarem aquest darrer *dataset* per a provar els models generats en aquest projecte, validant així els models generats amb un conjunt de dades diferent.

Taula 2: Datasets que s'utiltizaran durant aquest projecte

Dataset	**Mida**	**Etiquetes**	**Observacions**
B-T4SA	470,586	Positiu, negatiu, neutre	Conjunt balancejat (156,862 de cada tipus)
Twitter Dataset	603	Positiu, negatiu, neutre	470 positives, 133 negatives

4. Metodologia i desenvolupament

El desenvolupament d'aquest treball s'ha realitzat dins de la plataforma de Kaggle, que ofereix el servei Kaggle Kernels, una eina allotjada al núvol que permet als seus usuaris definir models utilitzant diferents entorns de treball com Tensorflow o Pytorch i entrenar-los utilitzant acceleració per *hardware*.

L'entorn de treball seleccionat ha estat TensorFlow, el qual disposa de la API tf.keras que ha estat l'eina utilitzada per a la construcció i entrenament de tots els models generats o testejats durant el desenvolupament. *Tf.Keras* és una API d'alt nivell que permet construir i entrenar models d'aprenentatge profund sobre l'API de Keras [26].

Addicionalment, aquesta API permet instanciar models coneguts com VGG16, Inception o ResNet els quals poden ser creats amb els pesos inicialitzats de forma aleatòria, amb pesos definits per l'usuari, o bé amb pesos pre-entrenats utilitzant el *dataset* ImageNet.

L'entrenament de tots els models s'ha executat utilitzant acceleració per *hardware*, en concret utilitzant les 30h setmanals de GPU (NVIDIA TESLA P100) que posa a disposició Kaggle de forma gratuïta per as seus usuaris.

4.1 Obtenció dels conjunts de dades a utilitzar

Degut a la limitació de recursos computacionals disponibles, es necessari reduir el *dataset* original B-T4SA, que inicialment contenia 156,862 mostres per a cada un de les tres classes.

El conjunt d'imatges B-T4SA està dividit en els següents tres subconjunts:

Taula 3: Dataset B-T4SA

Dataset	**Subconjunt**	**Mida**	**Etiquetes**
B-T4SA	entrenament	368,586	Positiu, negatiu, neutre
	validació	51,000	Positiu, negatiu, neutre
	test	51,000	Positiu, negatiu, neutre

Mitjançant un programa escrit en Python, per a cada subconjunt de la taula anterior s'ha generat un nou subconjunt amb un nombre menor d'imatges. Aquesta selecció ha estat realitzada de forma aleatòria, utilitzant el mètode *seed*

de Python que ens permet inicialitzar un generador de valors aleatoris per tal de poder replicar aquesta generació aleatòria de forma posterior.
Cadascun dels nous subconjunts generats mantenen el seu estat balancejat pel que fa el nombre d'imatges de cada classe incloses en el mateix. La següent taula resumeix el nou *dataset* generat:

Taula 4: Dataset B-T4SA_TFM

Dataset	**Subconjunt**	**Mida**	**Etiquetes**
B-T4SA_TFM	entrenament	60,000	Positiu, negatiu, neutre
	validació	9,000	Positiu, negatiu, neutre
	test	2,000	Positiu, negatiu, neutre

Aquest subconjunt del *dataset* B-T4SA que s'utilitzarà en els següents apartats conté per tant un total de 71,000 imatges, respecte les 470,586 originals.

4.2 Creació dels models a entrenar

En aquest treball s'han emprat dos models base que han servit per extreure les característiques d'alt nivell existents en les imatges i que havien estat prèviament entrenats amb el conegut conjunt d'imatges ImageNet [4]. Aquets dos models base son ResNet-50, un model basat en les xarxes residuals [20] i SE-ResNet-50, una variació del model anterior en el que s'introdueixen les xarxes *Squeeze-and-Excitation* [23].

Ambdós models base han estat modificats aplicant els mateixos canvis sobre les respectives arquitectures originals per tal d'assolir la classificació esperada. Mantenint l'arquitectura del model base, s'ha eliminat la darrera capa densa, utilitzada per a classificar en una de les 1,000 categories definides en el *dataset* ImageNet i s'han afegit noves capes per determinar si una imatge evoca sentiments negatius, neutres o bé positius.

Les figures 10 i 11 mostren els model base i el generat a partir d'aquest per als models base ResNet-50 i SE-ResNet-50 respectivament:

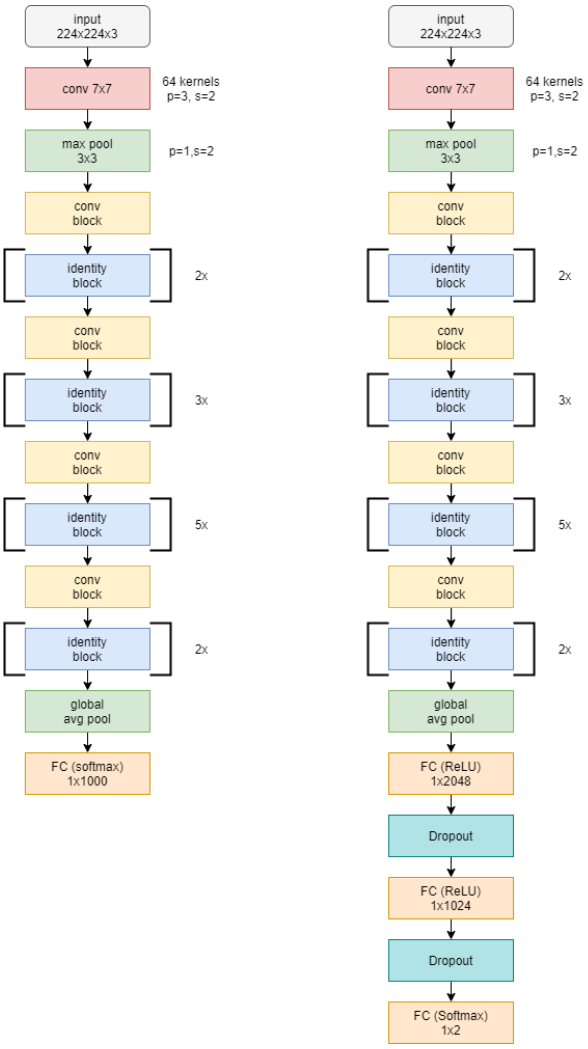


Figura 10: Model ResNet-50 (esquerra) i model generat a partir d'aquest (dreta)
Font: elaboració pròpia

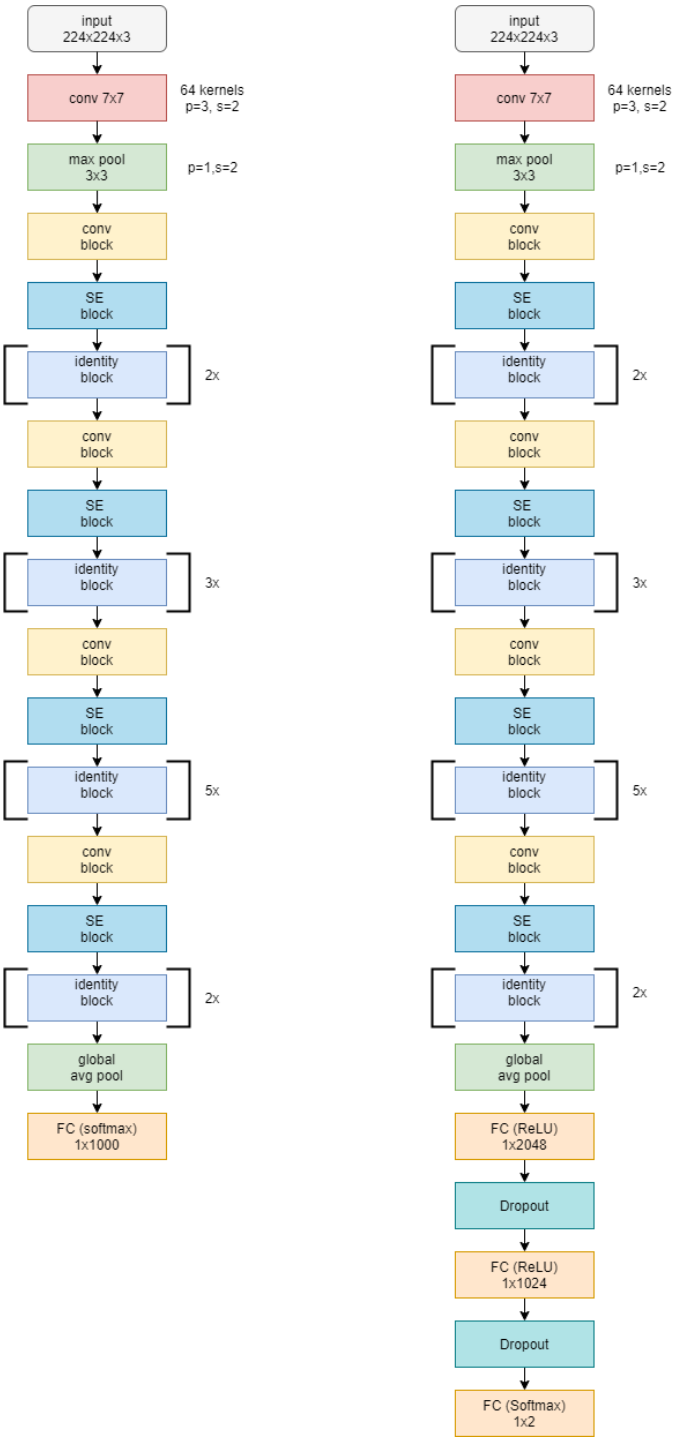


Figura 11: Model SE-ResNet-50 (esquerra) i model generat a partir d'aquest (dreta)
Font: elaboració pròpia

Com es pot observar en les dues figures anteriors, ambdós models han estat creats amb la mateixa estratègia:

S'ha eliminat la darrera capa densa que classificava la imatge en una de les 1,000 categories definides dins del conjunt ImageNet i en el seu lloc s'han afegit dues capes denses , encarregades de combinar la informació

extreta pel model base i identificar els elements que la puguin definir com una imatge negativa, neutre o positiva. Entre mig d'aquestes capes denses s'han afegit capes de *Dropout amb un factor de dropout rate* $p = 0.4$.

La tècnica del *dropout*, que s'utilitza únicament durant l'entrenament del model, consisteix en que durant cada etapa de l'entrenament s'elimina, temporalment, una neurona de l'entrada amb una probabilitat p.

A la pràctica aquest comportament significa que a cada etapa de l'entrenament es modifica l'arquitectura de la xarxa; aquesta modificació, en la que a cada etapa pot tenir informació disponible diferent en funció de si una neurona s'ha desconnectat o no, permet millorar sensiblement la capacitat de generalització de la xarxa neuronal i obtenir un model que no es sobre-especialitzi en el conjunt de dades disponibles durant el seu entrenament.

Pel que fa el model basat en l'arquitectura ResNet50, l'API de Keras ens proporciona tots els elements necessaris per instanciar el model base ResNet-50 entrenat prèviament amb el conjunt d'imatges d'ImageNet.

En canvi, aquesta API no disposa del model SE-ResNet-50, el qual conforma la base del segon model a desenvolupar en aquest treball. Per aquest motiu, durant el desenvolupament s'ha hagut d'instal·lar la llibreria *imatge-classifiers* [27], una llibreria compatible amb l'entorn de treball TensorFlow/Keras que disposa d'un conjunt de models de classificació entrenats prèviament utilitzant el *dataset* d'ImageNet més ampli del que hi ha disponible a través de l'API de Keras de forma estàndard. Dins del conjunt de models disponibles trobem models que utilitzant les xarxes *Squeeze and-Excitation*, com el SE-ResNet-50.

4.3 Ajustament fi per a l'anàlisi visual dels sentiments

Inicialització de la xarxa

A mesura que s'afegeixen capes en una xarxa neuronal i per tant es creen xarxes neuronals més profundes, els models generats requereixen aprendre un nombre molt elevat de paràmetres si aquests son entrenats des de zero. Malgrat les CNN requereixen un menor nombre de paràmetres degut a les connexions disperses entre les neurones de les diferents capes i al fet que, les neurones d'una mateixa capa comparteixen els pesos, les arquitectures proposades contenen un elevat nombre de paràmetres a calcular degut a la seva profunditat i a la presència d'altres tipus de xarxes a més de les convolucionals, com les capes denses o *fully connected.*

Aleshores, per tal de dur a terme entrenaments des de zero es necessiten conjunts de dades molt grans, altrament, el resultat obtingut seria un model que es sobre-ajustaria massa al conjunt de dades utilitzat durant l'entrenament.

Dins del camp de la classificació visual de sentiments però, l'obtenció de conjunts massius de dades etiquetades resulta complex degut a la dificultat de l'etiquetatge. Les grans plataformes socials disposen de diferents APIs amb les que seria possible extreure grans quantitats d'imatges; el problema aleshores el trobem en l'etiquetat de les mateixes, una tasca costosa tant en temps com econòmicament si es decideix contractar algun servei per a un etiquetatge realitzat de forma manual i que a més, obtindria uns resultats que dependrien en gran mesura de la subjectivitat de l'individu o individus que realitzessin l'etiquetatge.

Els models generats en aquest treball tenen un total de ~26 milions de paràmetres per al model basat en ResNet-50 i ~28 milions per al cas en que utilitzem SE-ResNet-50 com a model base. Les 60,000 mostres de les que disposem en el conjunt d'entrenament resulten insuficients per a entrenar des de zero qualsevol dels dos models proposats i per tant, l'estratègia a seguir ha estat la d'utilitzar el mecanisme de transferència d'aprenentatge.

Utilitzant la transferència d'aprenentatge, hem pogut inicialitzar tots els pesos de les capes que pertanyen als models base utilitzant un model prèviament entrenat enlloc d'inicialitzar-los de manera aleatòria. A més, aquestes capes es configuren per a que no siguin entrenades durant el procés d'entrenament del model.

Amb aquesta estratègia, podem utilitzar els conjunts d'entrenament i validació únicament per calcular els pesos de les últimes capes afegides sobre els models base i que seran utilitzades per a realitzar la classificació visual del sentiment, mentre que les capes que pertanyen als models base seleccionats i que s'encarreguen de l'extracció de característiques de les imatges utilitzaran els pesos del model prèviament entrenat a per a dur a terme aquesta tasca i no s'actualitzaran durant l'entrenament.

En els dos models generats, la part de la xarxa que correspon al model base utilitzat (ResNet-50 o SE-ResNet-50), ha estat inicialitzada amb els pesos dels models entrenats amb el *dataset* ImageNet.

Pel que fa als pesos de les capes denses afegides, aquests han estat inicialitzats utilitzant una distribució Glorot Normal [28], la distribució utilitzada per defecte per a les capes denses dins de l'API de Keras.

Entrenament de la xarxa

El rendiment dels models generats i la seva velocitat d'entrenament poden millorar en funció de l'optimitzador utilitzat per seu entrenament. Els optimitzadors

son els algorismes o mètodes utilitzats per a canviar els atributs d'una xarxa neuronal durant el seu procés d'entrenament.

El descens del gradient ha estat el mètode més comú per a l'entrenament de les xarxes neuronals profundes, però el seu mecanisme, que canvia els pesos de les neurones un cop ha calculat el gradient en la totalitat del conjunt de dades d'entrenament, requereix grans quantitats de memòria i a la pràctica el fan invàlid per als entrenaments a realitzar en aquest projecte.

Tanmateix, existeixen variacions del mètode del descens del gradient que permeten esquivar algunes de les seves limitacions així com altres algorismes que realitzen l'actualització dels pesos en funció del moment.

Per a l'entrenament dels models desenvolupats durant aquest treball, s'han utilitzat els següents optimitzadors per tal de comparar el seu efecte en la velocitat de convergència i, sobretot, en el rendiment dels models:

- SGD (*Stochastic Gradient Descent*) [29]: Variació del mètode del descens del gradient en la que es calcula l'error i s'actualitza el model per a cada exemple del conjunt de dades d'entrenament. Aquest optimitzador a més és accelerat amb el moment. El moment acumula el gradient de les etapes anteriors i utilitzen aquesta informació per a determinar el canvi a realitzar en els pesos de les neurones, ja que accelera els gradients del vector en la direcció correcta, facilitant una convergència més ràpida.

- Adam (*Adaptive Moment Estimation*) [30]: Aquest optimitzador calcula un factor d'aprentatge adaptatiu per a cada paràmetre de la xarxa estimant el moment de primer i segon ordre del gradient.

- Nadam (*Nesterov-accelerated Adaptative Moment Estimation*) [31]: Aquest algorisme és la combinació de l'optimitzador *Adam* però utilitza el moment Nesterov.

Els entrenaments s'han executat utilitzant paquets o *batches* de 32 imatges, durant 100 èpoques i utilitzant la tècnica d'*Early stopping* que ens permetia aturar l'entrenament si la pèrdua calculada sobre el conjunt de validació no millorava durant 10 èpoques consecutives.

Degut a les limitacions en els recursos disponibles per a l'entrenament dels models, on com a màxim disposàvem de 9 hores consecutives de d'acceleració per hardware, els entrenaments s'han hagut de realitzar en diferents iteracions o etapes. Per a poder controlar aquesta limitació, s'ha desenvolupat un nou *callback* que aturava els entrenaments quant aquests s'aproximaven a les 9 hores. Posteriorment, en la següent execució, es carregaven els pesos calculats en la sessió anterior i s'iniciava un nou entrenament de 9 hores que eventualment era aturat pel *callback* d'*Early stopping.*

4.4 Explicabilitat

Un dels punts febles de les xarxes neuronals és que sovint actuen com una caixa negra en la que no podem entendre quines característiques o elements son considerats per tal de classificar les mostres en una o altra classe. En el cas de la visió per computador, resulta encara més interessant poder visualitzar quins elements de la imatge han estat més determinants per a la seva classificació.

En l'article *Object detectors emerge in deep scene CNNs* [32] Zhou et al van demostrar que les neurones de diferents capes d'una CNN actuen com a detector d'objectes, malgrat no proporcionar a la xarxa cap informació sobre la localització dels mateixos. L'inconvenient de les arquitectures com les que s'utilitzen en aquest treball és que aquesta informació sobre els objectes es perd quan les capes denses o totalment connectades entren en joc per a realitzar la classificació de la imatge.

En aquest treball s'ha implementat una visualització dels elements que expliquin com s'ha classificat una imatge utilitzant la tècnica *Gradient -weighted Class Activation Mapping (Grad-CAM)* [33], una generalització dels Class Activation Maps (CAM) [34].

El mètode de Grad-CAM, a diferència del CAM que únicament és aplicable a unes arquitectures de CNN específiques que tinguin una capa de *global average pooling* immediatament abans de la capa de predicció (per exemple, capa convolucional → *global average pooling* → softmax layer), és compatible amb qualsevol tipus d'arquitectura i no requereix cap modificació del model existent per tal de poder realitzar les visualitzacions desitjades.

El mètode Grad-CAM utilitza la informació del gradient propagada fins l'última capa convolucional per tal d'entendre la decisió presa pel model. Per obtenir la localització del mapa d'activació per una determinada imatge, es calcula primerament el gradient de la classe predita per a la imatge d'entrada respecte l'activació de la darrera capa convolucional; posteriorment aquest gradient es multiplica per la mitjana de la intensitat del gradient en cada canal, per tal de ponderar la importància d'aquell canal respecte la classe predita pel model. Finalment, i per tal de centrar la visualització únicament en aquells factors rellevants de la imatge es realitza una rectificació lineal (ReLU) per eliminar les contribucions negatives i normalitzar el mapa de calor que crearem entre 0 i 1:

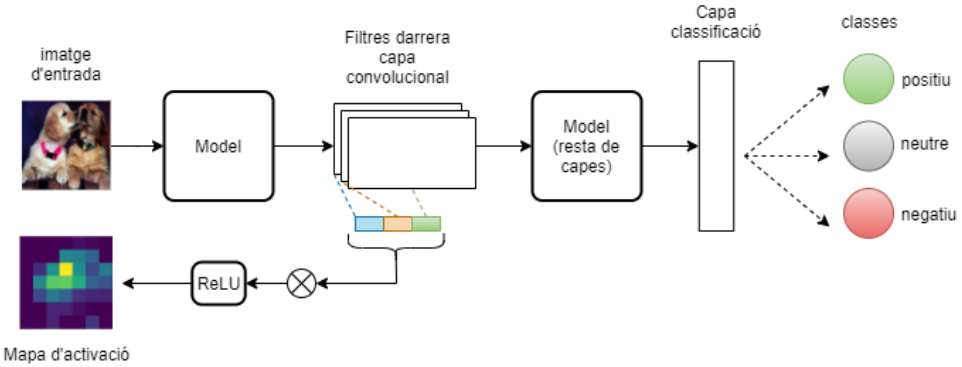


Figura 12: Grad-CAM. (Font: Elaboració pròpia)

5. Resultats

Aquest apartat conté els resultats dels models entrenats que han estat descrits en l'apartat 4. Els dos models han estat entrenats amb els subconjunts d'entrenament i validació descrits en la secció 4.1.

Pel que fa les proves realitzades per tal d'obtenir les mètriques que descriguin els models generats, els models han estat testejats amb el subconjunt de test descrit en el mateix apartat 4.1 i amb el conjunt *Twitter Dataset* esmentat en el capítol 3.

En els apartats 5.1 i 5.2 trobem les mètriques generades pels dos conjunts de tests utilitzats en cadascun dels dos models generats. En aquestes seccions es mostra, en primer lloc, les diferents corbes d'entrenament/validació per als diferents entrenaments realitzats sobre cada model. A continuació es mostra el *classification report*, una eina que permet avaluar la capacitat de predicció dels models i una matriu de confusió que ens ajudarà a visualitzar els errors comesos pels models per al model que ha obtingut millors resultats durant l'entrenament.

A l'apartat 5.3 trobem una comparació dels resultats obtinguts entre els dos models.

Finalment, en el darrer apartat d'aquest capítol s'inclouen les visualitzacions obtingudes amb el mètode Grad-CAM.

5.1 Model basat en ResNet50

En primer lloc, visualitzem les corbes d'entrenament/validació obtingudes durant els entrenaments d'aquest model.

Com s'ha explicat en l'apartat 4.3, cada un dels dos models ha estat entrenat amb tres optimitzadors diferents, utilitzant en tots tres un factor d'aprenentatge igual a 0.001:

- Stochastic Gradient Descent (SGD) amb moment Nesterov

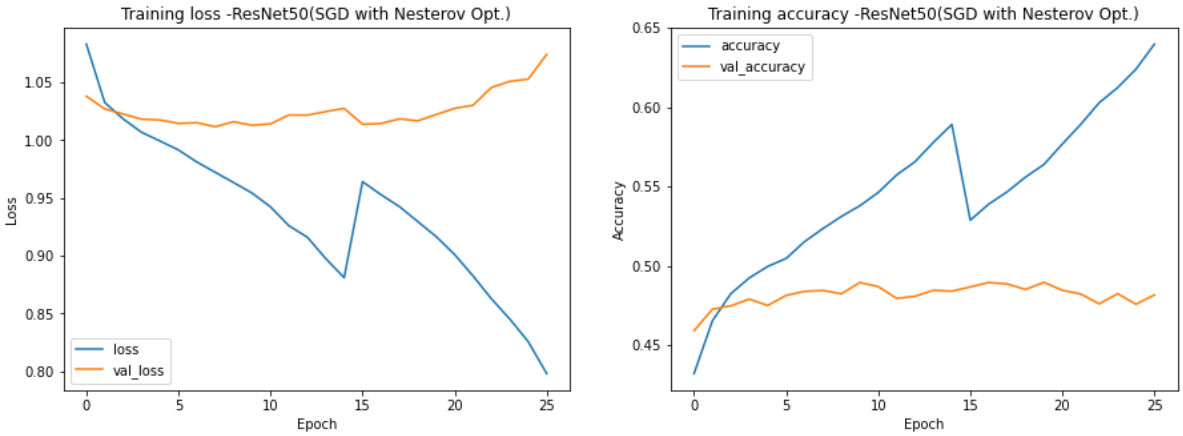


Figura 13: Pèrdua i precisió del model basat en ResNet-50 utilitzant l'optimitzador SGD amb moment Nesterov

- Optimitzador Adam

Figura 14: Pèrdua i precisió del model basat en ResNet-50 utilitzant l'optimitzador Adam

- Optimitzador Nadam

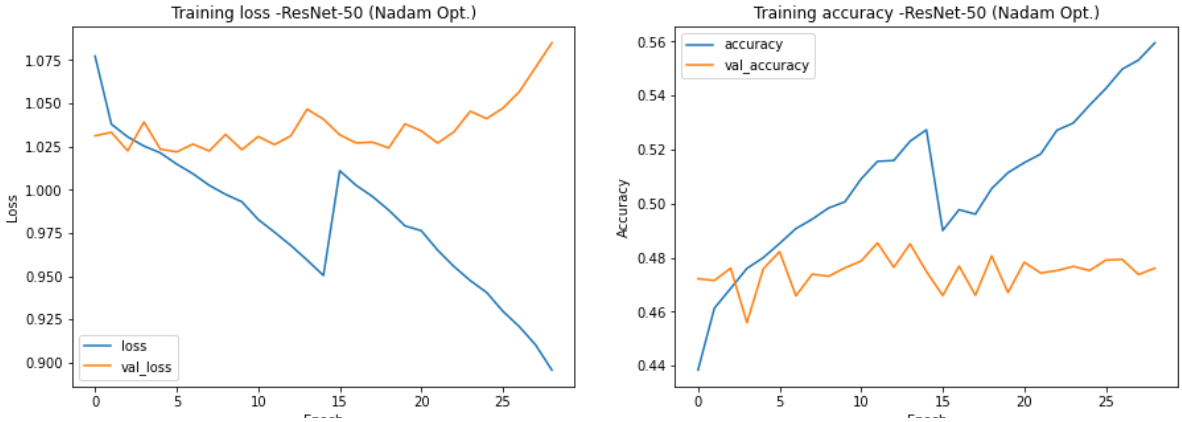


Figura 15: Pèrdua i precisió del model basat en ResNet-50 utilitzant l'optimitzador Nadam

Les gràfiques següents mostren les mètriques obtingudes de forma conjunta per al model basat en l'arquitectura ResNet-50 amb els tres optimitzadors utilitzats:

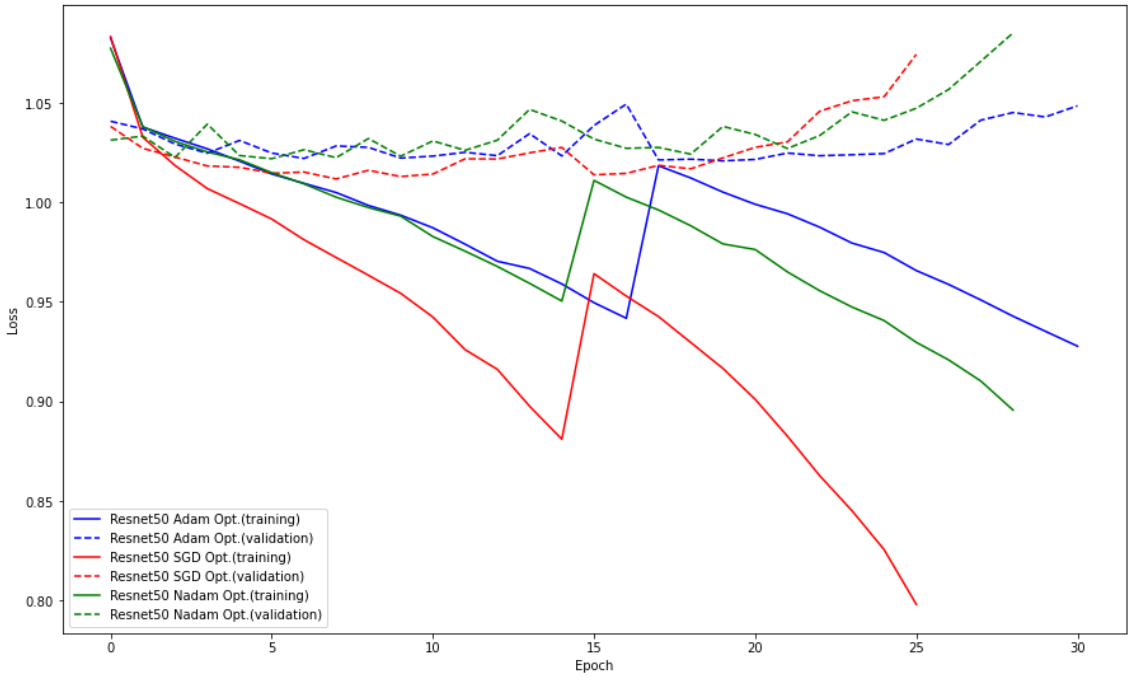


Figura 16: Comparació de la pèrdua entre tots els entrenaments del model basat en ResNet-50

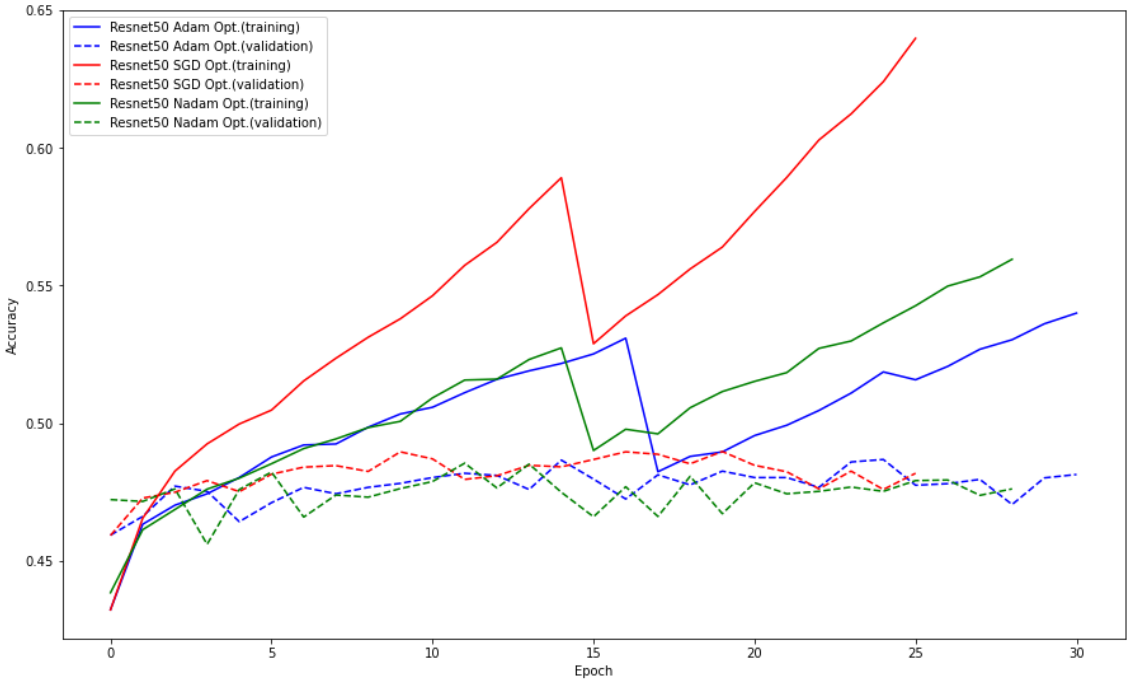


Figura 17: Comparació de la precisió entre tots els entrenaments del model basat en ResNet-50

En la taula 5 podem observar el resultat de l'avaluació dels models utilitzant el conjunt de test:

Taula *5*: Resultats per al model basat en ResNet-50:

Optimitzador	**Pèrdua**	**Precisió**
SGD (moment Nesterov)	0.9972	0.4958
Adam	1.0131	0.4932
Nadam	1.0073	0.4918

Com podem observar en la taula anterior, per aquest model l'optimitzador SGD dona uns resultats lleugerament superiors.

Les figures 18 i 19 mostren el *classification report* i la matriu de confusió respectivament obtingudes durant la validació del millor model obtingut basat en ResNet-50 per al conjunt de test del dataset *B-T4SA_TFM :*

	precision	recall	f1-score	support
0	0.49	0.58	0.53	2000
1	0.51	0.47	0.49	2000
2	0.49	0.44	0.46	2000
accuracy			0.50	6000
macro avg	0.50	0.50	0.49	6000
weighted avg	0.50	0.50	0.49	6000

Figura 18: Classification report

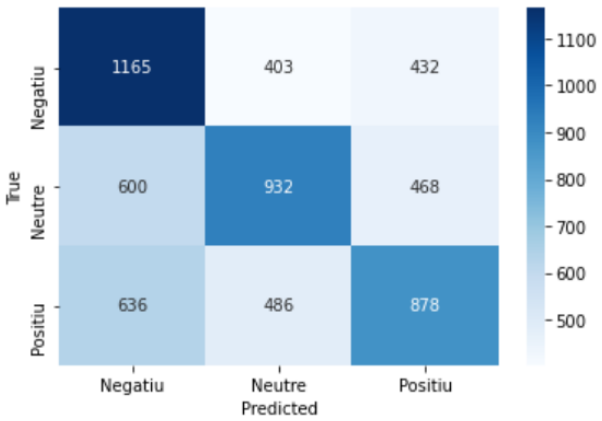


Figura 19: Matriu de confusió

Avaluació amb el conjunt *Twitter Dataset*

A continuació avaluem el millor model entrenat utilitzant el conjunt de dades *Twitter Dataset*, del qual disposàvem d'un total de 603 imatges organitzades en 470 mostres etiquetades amb sentiment positiu i 133 amb etiqueta negativa (en aquest conjunt, no existeix cap mostra amb sentiment neutre).

La figura 20 mostra el *classification report* obtingut pel model compilat amb l'optimitzador SGD:

	precision	recall	f1-score	support
0	0.29	0.52	0.37	133
1	0.00	1.00	0.00	0
2	0.87	0.36	0.51	470
accuracy			0.39	603
macro avg	0.39	0.63	0.29	603
weighted avg	0.74	0.39	0.48	603

Figura 20: Classification report

També mostrem la matriu de confusió a la figura 21, que ens permet observar de forma gràfica el tipus d'error comès pel nostre model:

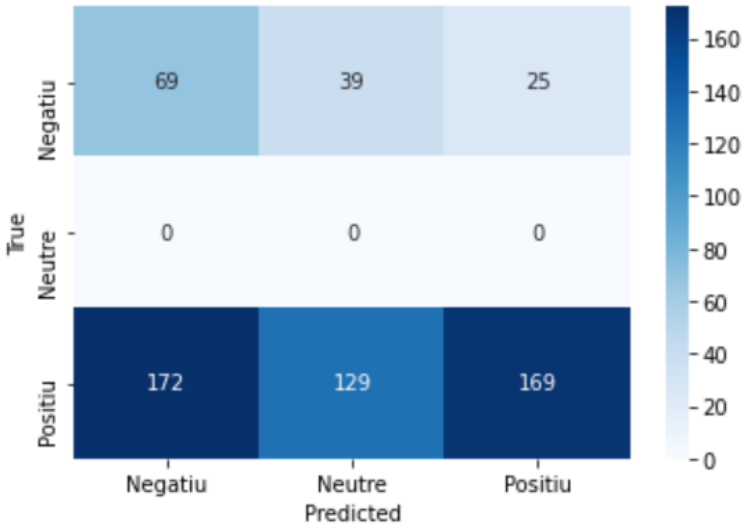


Figura 21: Matriu de confusió

5.2 Model basat en SE-ResNet50

Mostrem les corbes d'entrenament/validació obtingudes durant els entrenaments d'aquest model amb cadascun dels tres optimitzadors:

- Stochastic Gradient Descent (SGD) amb moment Nesterov

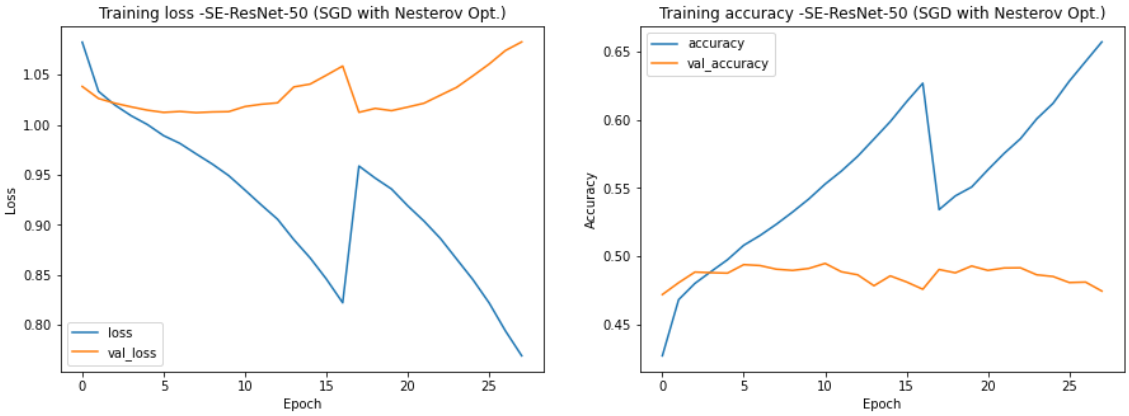


Figura 22: Pèrdua i precisió del model basat en SE-ResNet-50 utilitzant l'optimitzador SGD amb moment Nesterov

- Optimitzador Adam

Figura 23: Pèrdua i precisió del model basat en SE-ResNet-50 utilitzant l'optimitzador Adam

- Optimitzador Nadam

Figura 24: Pèrdua i precisió del model basat en SE-ResNet-50 utilitzant l'optimitzador Nadam

Les gràfiques següents mostren les mètriques obtingudes de forma conjunta per als tres models:

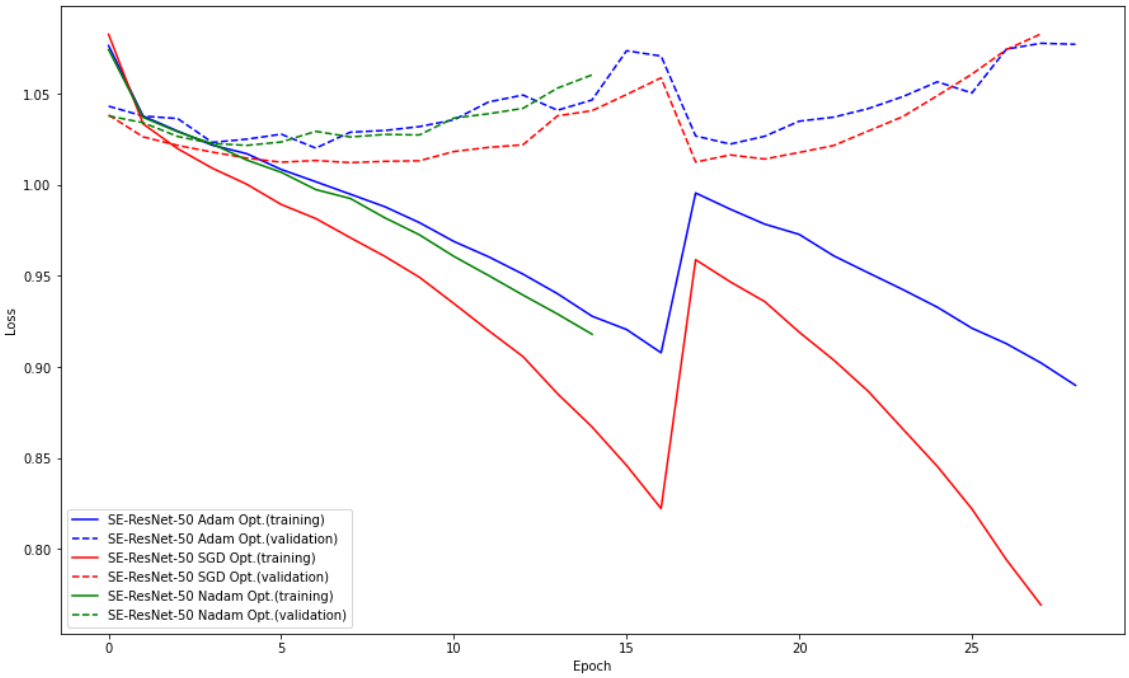


Figura 25: Comparació de la pèrdua entre tots els entrenaments del model basat en SE-ResNet-50

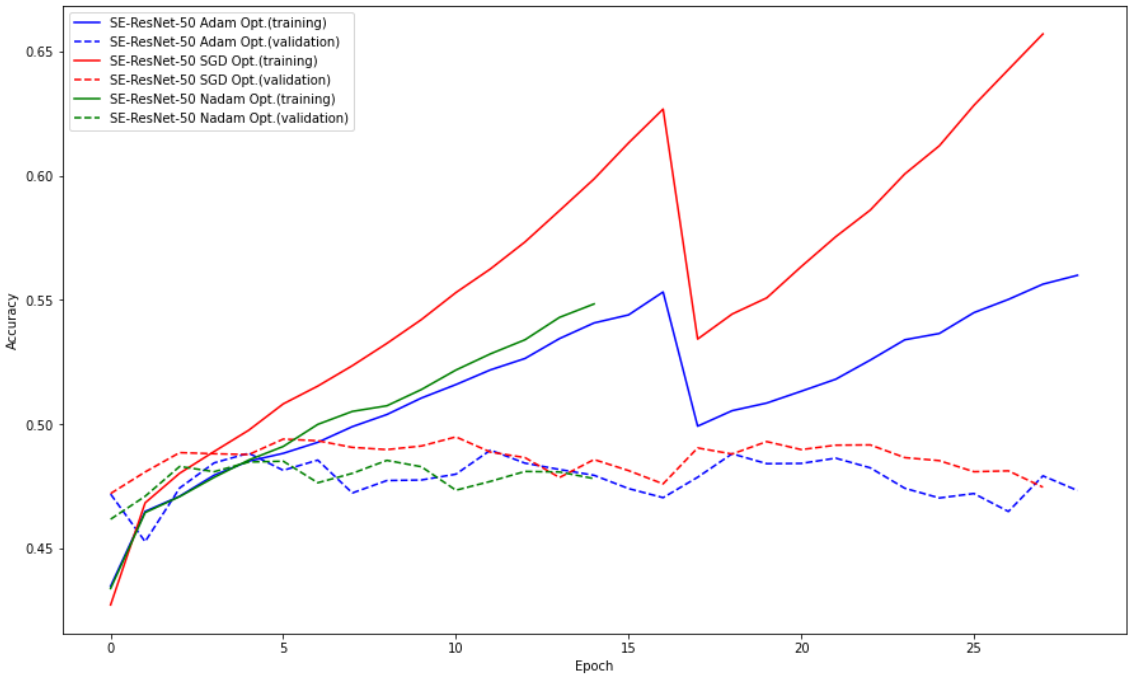


Figura 26: Comparació de la precisió entre tots els entrenaments del model basat en SE-ResNet-50

La taula 6 conté els valors de pèrdua i precisió mesurats sobre el subconjunt de test:

Taula *6*: Resultats per al model basat en SE-ResNet-50:

Optimitzador	**Pèrdua**	**Precisió**
SGD (moment Nesterov)	1.0001	0.4963
Adam	1.0113	0.4945
Nadam	1.0110	0.4962

L'optimitzador SGD torna a donar els millors resultats.

Les figures 27 i 28 mostren el *classification report* i la matriu de confusió respectivament obtingudes durant la validació del millor model:

	precision	recall	f1-score	support
0	0.49	0.54	0.51	2000
1	0.50	0.53	0.51	2000
2	0.51	0.42	0.46	2000
accuracy			0.50	6000
macro avg	0.50	0.50	0.49	6000
weighted avg	0.50	0.50	0.49	6000

Figura 27: Classification report

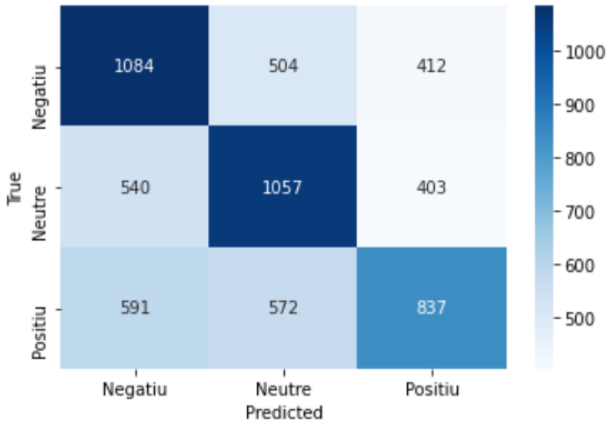


Figura 28: Matriu de confusió

Avaluació amb el conjunt *Twitter Dataset*

Anàlogament a l'apartat 5.1 a continuació avaluem el millor model entrenat utilitzant el conjunt de dades *Twitter Dataset*:

La figura 29 mostra el *classification report* obtingut pel model compilat amb l'optimitzador SGD:

	precision	recall	f1-score	support
0	0.28	0.47	0.35	133
1	0.00	1.00	0.00	0
2	0.87	0.34	0.49	470
accuracy			0.37	603
macro avg	0.38	0.61	0.28	603
weighted avg	0.74	0.37	0.46	603

Figura 29: Classification report

També mostrem la matriu de confusió:

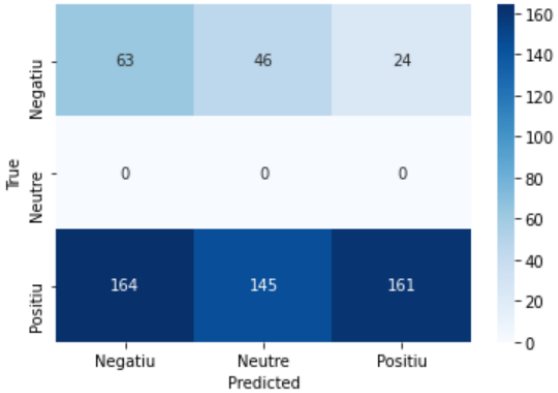


Figura 30: Matriu de confusió

5.3 Comparació dels models

En els dos apartats anteriors hem pogut observar les mètriques obtingudes pels dos models desenvolupats sobre el subconjunt de test del dataset B-T4SA_TFM i el conjunt *Twitter Dataset* per als millors models obtinguts amb cadascuna de les dues arquitectures.

Per tal de comparar les mètriques obtingudes de forma conjunta, es mostren les figures 31 i 32, que superposen els resultats obtinguts en tots els entrenaments realitzats per als dos models:

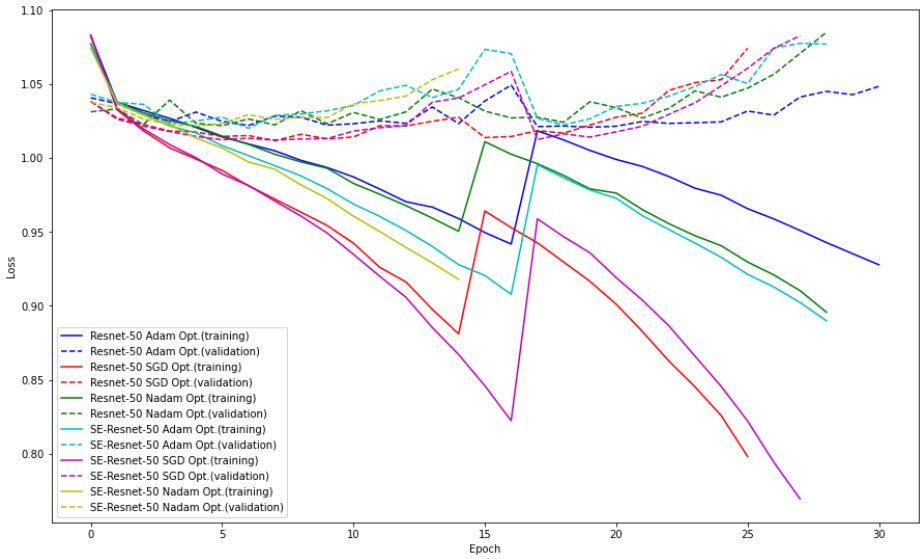


Figura 31: Pèrdua

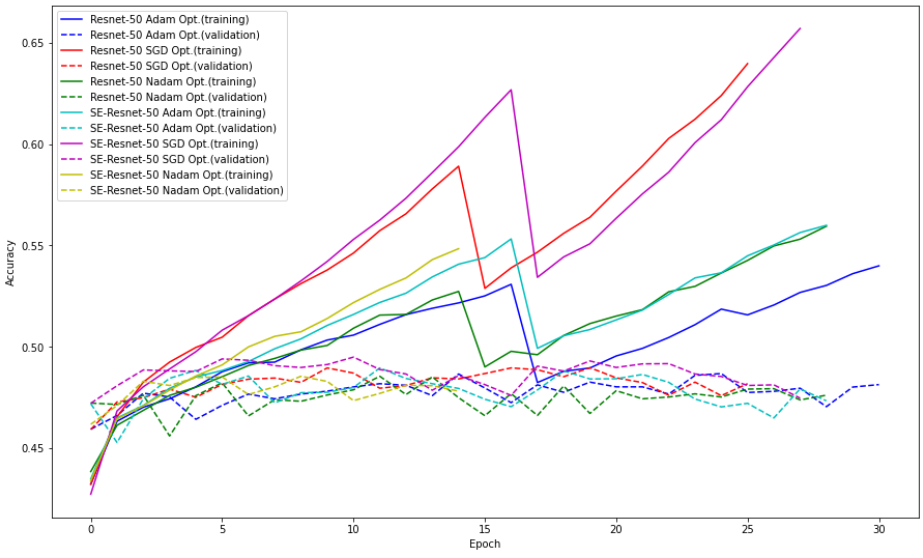


Figura 32: Precisió

En ells dos models obtenen nivells de precisió i pèrdua bastant similars i, en els dos casos, els millors resultats s'han obtingut utilitzant l'optimitzador SGD.

Tots els models tenen una precisió propera al 0.5 sobre el conjunt de dades de test del *dataset* B-T4SA_TFM. Malgrat això, analitzant amb més cura els classification reports obtinguts sobre el subconjunt de test podem veure com en tots els casos, la mètrica F1, que representa una mitjana harmònica entre el *recall* i la precisió, és sempre superior per al cas del sentiment negatiu. Això ens indica que per a tots els models generats, sempre resulta més fàcil classificar de forma correcta una imatge que identifica un sentiment negatiu.

A la taula 7 comparem els millors models obtinguts per a cadascuna de les dues arquitectures amb els resultats obtinguts en l'article [25], on el *dataset* B-T4SA, utilitzat com conjunt de referència per l'entrenament i validació dels models generats en aquest treball, va ser presentat.

Els autors de [25] duen a terme diferents entrenaments utilitzant diferents models, tots ells seguint la mateixa estratègia: Inicialitzen els pesos de la xarxa utilitzant el *dataset* ImageNet i posteriorment hi realitzen un ajustament fi utilitzant les imatges del conjunt d'entrenament i validació del *dataset* B-T4SA. La precisió és calculada posteriorment utilitzant el conjunt de test d'aquest mateix *dataset.*

Taula 7: Resultats dels models generats en aquest treball i dels entrenats en l'article [25]:

Model	**Subconjunt d'entrenament**	**Subconjunt de prova**	**Precisió**
Random Classifier	B-T4SA train	B-T4SA test	0.33
Hybrid-T4SA FT-F			0.499
Hybrid-T4SA FT-A			0.491
VGG-T4SA FT-F			0.506
VGG-T4SA FT-A			0.513
ResNet50_bm	B-T4SA_TFM train	B-T4SA_TFM test	0.4958
SEResNet50_bm			0.4963

Com podem observar en la taula anterior, malgrat treballar amb un subconjunt de dades molt reduït respecte el *dataset* original B-T4SA, els resultats obtinguts s'equiparen als que es van aconseguir utilitzant el *dataset* B-T4SA al complet.

5.4 Visualitzacions

Mostrem les visualitzacions del mètode Grad-CAM, per tal d'observar en quins elements es fixen els models alhora de determinar la polaritat del sentiment en una imatge.

També podem veure l'impacte d'afegir els blocs *Squeeze-and-Excitation (SE)* sobre l'arquitectura del model ResNet-50 i comparar com canvien els mapes d'activació entre aquest model i el de SE-ResNet50.

Visualitzacions arquitectura ResNet-50:

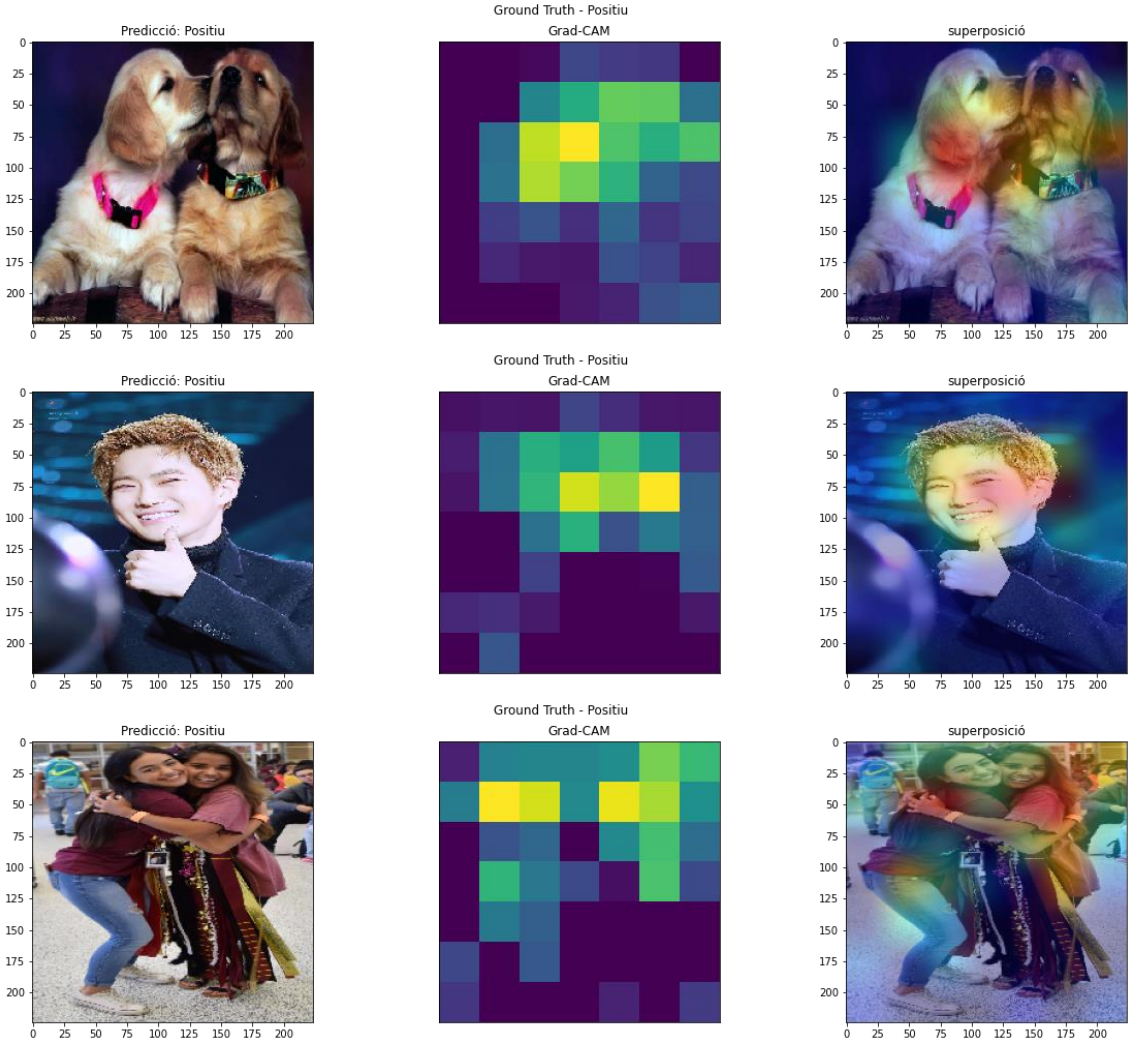


Figura 33: Imatges amb etiqueta positiva

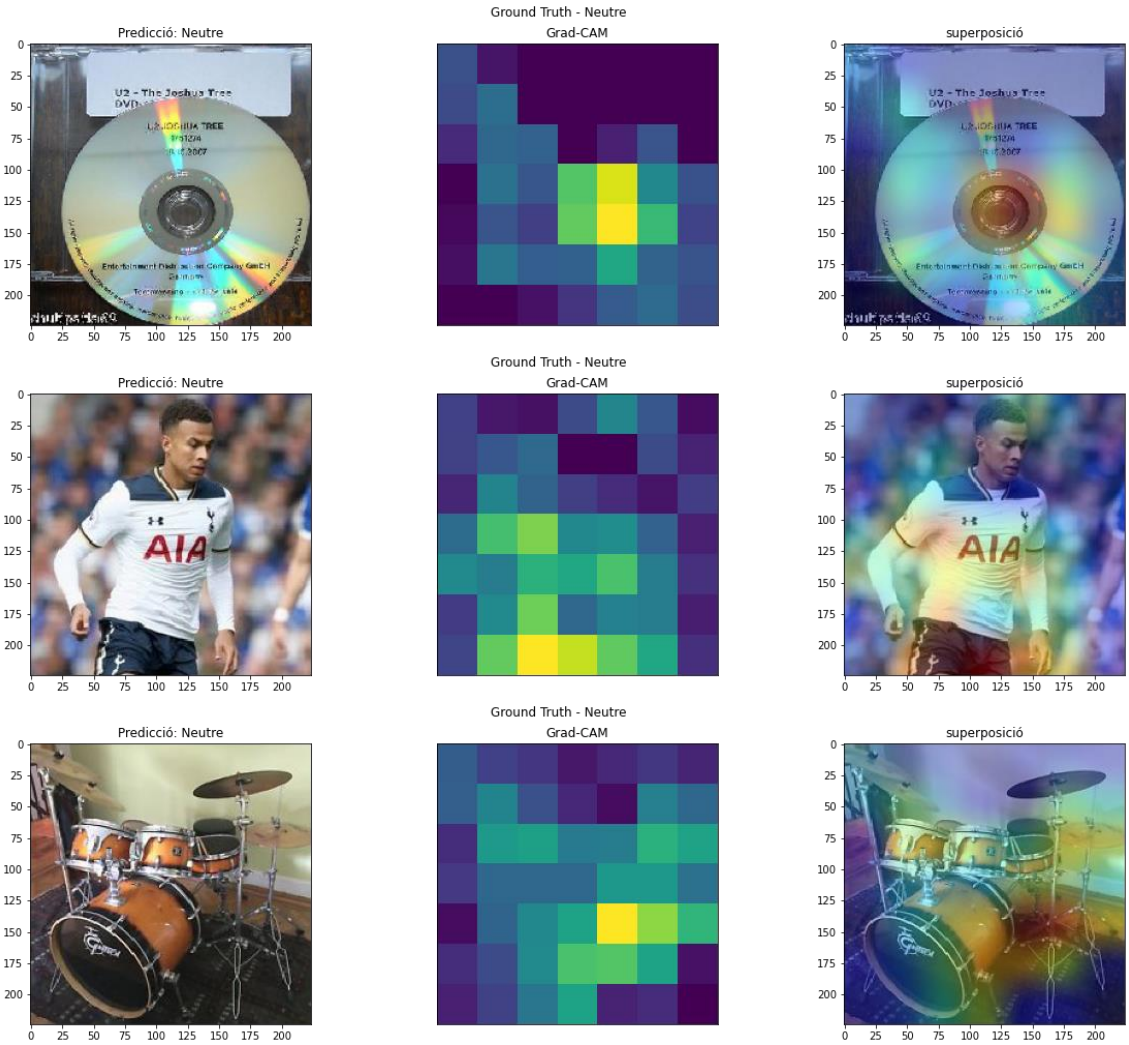


Figura 34: Imatges amb etiqueta neutre

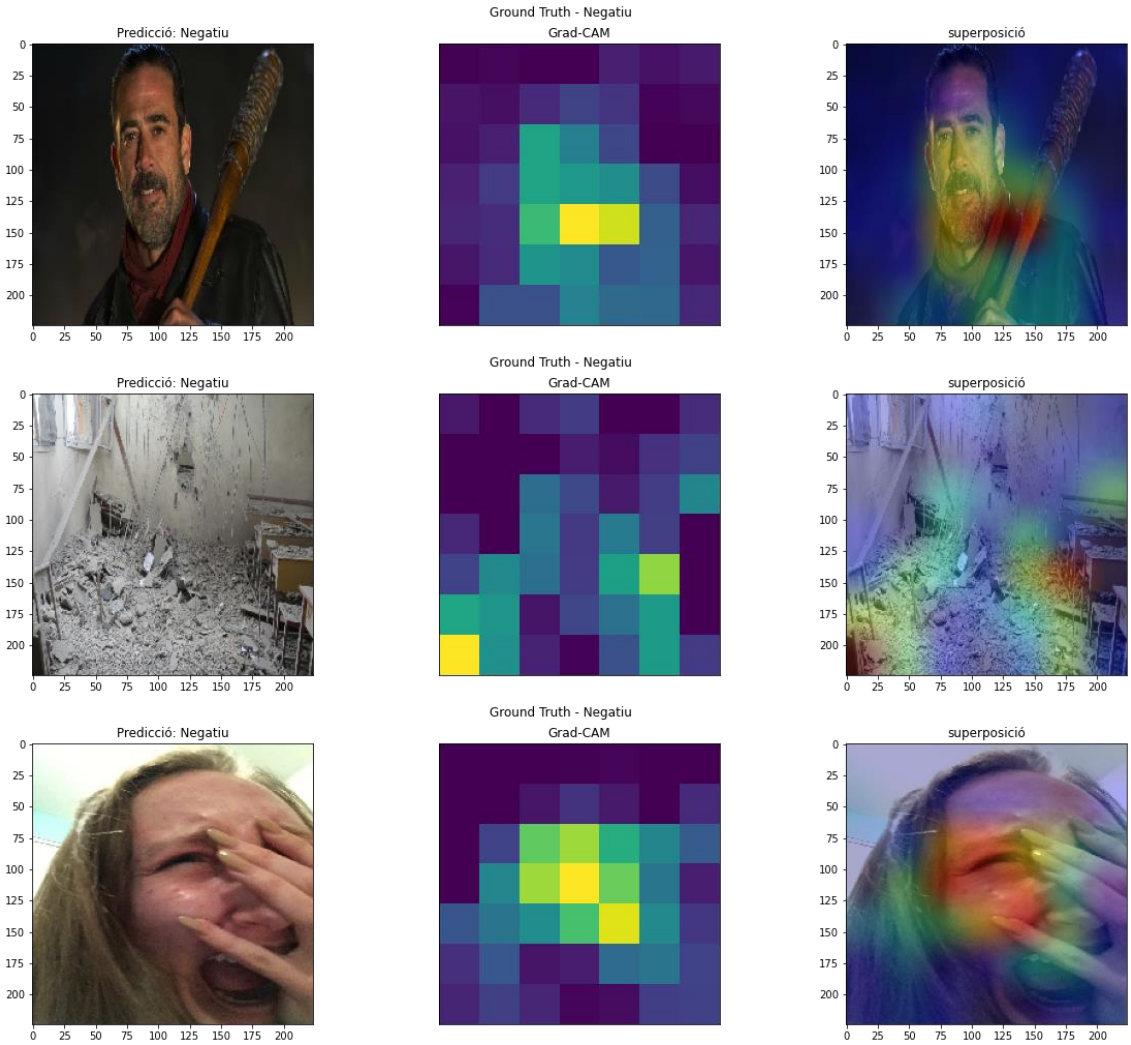


Figura 35: Imatges amb etiqueta negativa

Visualitzacions arquitectura SE-ResNet-50:

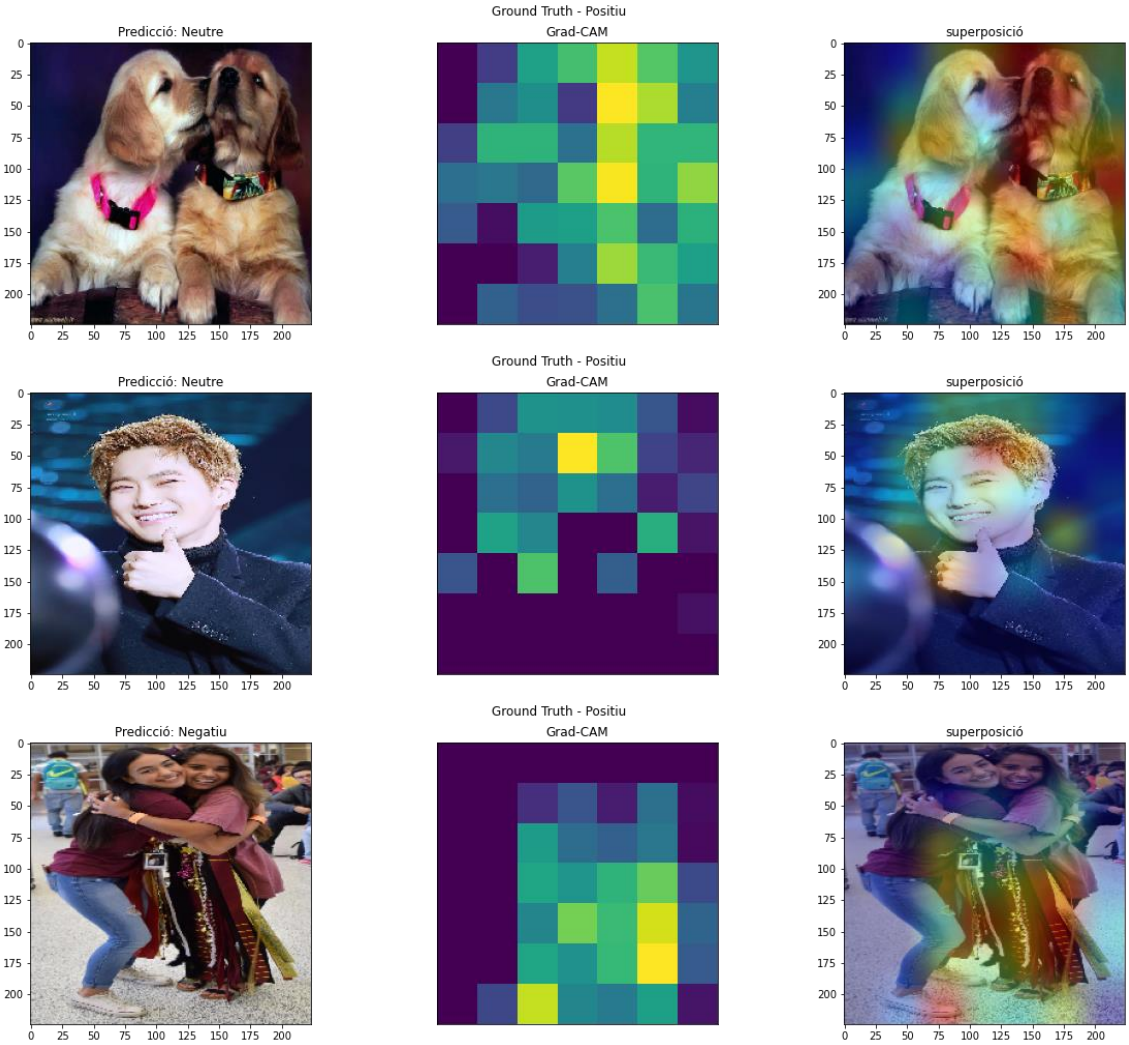


Figura 36: Imatges amb etiqueta positiva

Figura 37: Imatges amb etiqueta neutre

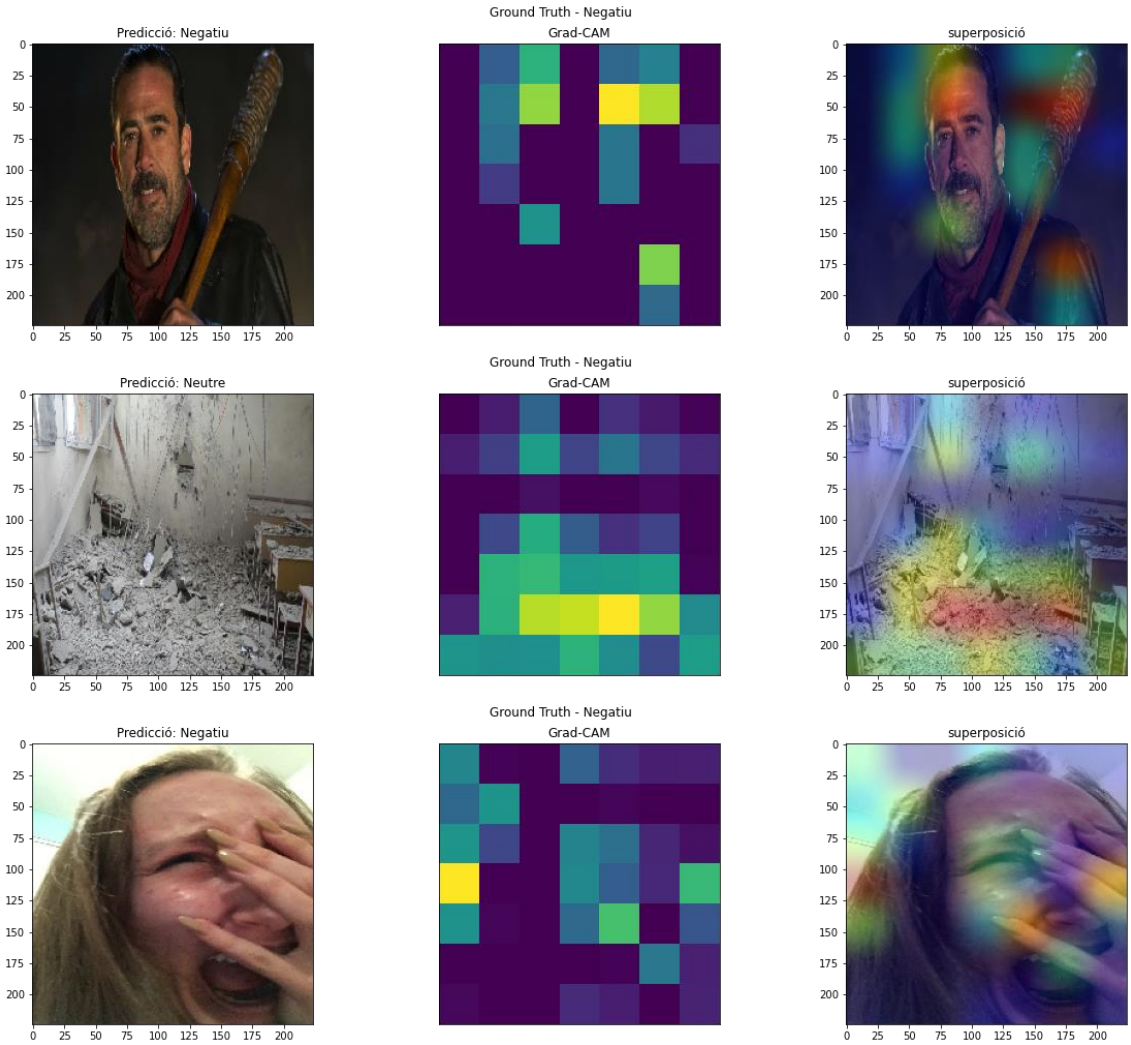


Figura 38: Imatges amb etiqueta negativa

6. Conclusions

La metodologia prevista es va veure modificada, i es va haver de crear un nou conjunt de dades per a l'entrenament del model degut a les limitacions computacionals i a les grans dimensions del *dataset* original. Els models generats compleixen parcialment les expectatives inicials, ja que, com hem vist en el capítol de resultats, les precisions obtingudes disten del que podríem considerar com a un classificador genèric d'imatges, malgrat hem vist que en alguns conjunts, com el cas de *Twitter Dataset*, el comportament era una mica millor.

Els millor model que s'ha obtingut ha estat el basat en l'arquitectura SE-ResNet-50 i compilant aquest utilitzant l'optimitzador SGD. Malgrat tot, la resta de models provats obtenen uns nivells de precisió pràcticament idèntics als obtinguts per aquest model.

S'intueix sobre els resultats de la mètrica F1 sobre les etiquetes negatives que de forma global, resulta més fàcil acordar que el sentiment d'una imatge és negatiu per la presència de certs elements que poden facilitar-ne la seva identificació com a tal. Amb tot, sembla que resulta més complex classificar de forma correcta les imatges de sentiment positiu o neutre, i, de fet, això és quelcom inherent a la identificació dels sentiments en imatges, ja que el sentiment que evoca una imatge dependrà de l'interlocutor que la observa.

L'obtenció de conjunts d'imatges etiquetats per a l'anàlisi visual dels sentiments és una tasca complexa, que sempre estarà condicionada per la subjectivitat dels individus que realitzen l'etiquetatge; ja que la personalitat, la cultura o fins i tot el context poden canviar la percepció que tenim envers la polaritat del sentiment que evoca una imatge.

La disponibilitat d'un major nombre de recursos computacionals ens ajudarà a millorar el model, i poder realitzar un major nombre de proves per a ajustar els models desenvolupats i obtenir millors prediccions. Amb tot, també es necessària l'obtenció d'altres conjunts d'imatges etiquetades, que puguin facilitar un millor aprenentatge del model.

7. Glossari

PLN – Processament del llenguatge natural

És un camp de les ciències de la computació, la intel·ligència artificial i la lingüística que estudia la interacció entre les computadores i el llenguatge humà

CNN – Xarxes neuronals convolucionals

Es tracta de les xarxes neuronals compostes principalment per capes convolucionals

CAM - Class Activation Maps

Mapes d'activació de classe. Son tècniques que permeten visualitzar els elements d'una imatge que han resultat destacats pel model alhora de la seva classificació.

Grad-CAM – Gradient-weighted Class Activation Mapping

Els Grad-CAM es tracten d'una generalització dels CAM que permeten realitzar la visualització esmentada sense restriccions pel que fa l'estructura de la xarxa neuronal estudiada.

SGD – Stochastic Gradient Descent

Optimitzador per a xarxes neuronals utilitzat per actualitzar els pesos durant l'entrenament de la xarxa.

8. Bibliografia

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints, October 2018.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237v2 [cs.CL]. Gener 2020.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS), 2012.

[4] J. Deng , A. Berg, S. Satheesh, H.Su, A. Khosla , L.Fei-Fei . L arge scale viual recognition challenge. www.image-net.org/challenges/LSVRC/2012, 1, 2012.

[5] Carlo Colombo, Alberto Del Bimbo, and Pietro Pala. Semantics in visual information retrieval.IEEEMultimedia, 6(3):38–53, 1999.

[6] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In Proceedings of the 18th ACM international conference on Multimedia, pages 715–718. ACM, 2010.

[7] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinionmining. Proceedings of LREC, volum 6, pàgines 417–422. Citeseer, 2006.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints.IJCV, 60(2):91–110, 2004.

[9] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. ACM Multimedia Conference, Barcelona, Oct 2013.

[10] Robert Plutchik.Emotion: A Psychoevolutionary Synthesis.Harper & Row, Publishers, 1980.

[11] M. Thelwall et al. Sentiment Strength Detection in Short Informal Text.J. of the American Soc. for InformationScience and Tech., 61(12):2544-2558, 2010.

[12] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks.arXiv preprint arXiv:1410.8586, 2014.

[13] Víctor Campos, Amaia Salvador, Xavier Giró-i-Nieto, and Brendan Jou. Diving deep into sentiment:Understanding fine-tuned cnns for visual sentiment prediction. In Proceedings of the 1st International Workshop on Affect; Sentiment
in Multimedia, New York, NY, USA,2015. ACM.
[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev,J. Long, R. Girshick, S. Guadarrama, and T. Darrell.Caffe: Convolutional architecture for fast featureembedding. In ACM MM, 2014.

[15] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In IEEE International Conference on Multimedia and Expo., 2016.

[16] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object?. In Computer Vision and Pattern Recognition(CVPR), 2010 IEEE Conference on. IEEE, 2010.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556,2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ,2015.

[19] Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst.6, 107–116 (1998).

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016. arXiv:1512.03385 http://arxiv.org/abs/1512.03385.

[21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In Computer Visionand Pattern Recognition (CVPR), 2014 IEEEConference, IEEE, 2014.

[22] Vaswani, Ashish, et al. Attention is all you need. *Advances in neural information processing systems*. 2017.

[23] Hu, Jie, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[24] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis
using progressively trained and domain transferred deep networks. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial
Intelligence, ser. http://dl.acm.org/citation.cfm?id=2887007.2887061

[25] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi,and Maurizio Tesconi. Cross-media learning for image sentiment

analysis in the wild. InProceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[26] Chollet, F. Keras. https://keras.io. 2015.

[27] Yakubovskiy, P. Image-classifiers. https://pypi.org/project/image-classifiers/

[28] Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, in PMLR 9:249-256. 2010.

[29] I. Sutskever, J. Martens, G. Dahl, G. Hinton. On the importance of initialization and momentum in deep learning. ICML, 2013.

[30] Kingma, P. Diederik, Ba, J. Adam: A Method for Stochastic Optimization. International Conference for Learning Representations, San Diego, 2015.

[31] Dozat, T. Incorporating Nesterov Momentum into Adam. ICLR, 2016.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.Object detectors emerge in deep scene cnns.International Conference on Learning Representations, 2015.

[33] Selvaraju, R.R., Cogswell, M., Das, A. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis* 128**,** 336–359 (2020).

[34] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. (2015) Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016.