Universitat de Girona
Escola Politècnica Superior

Treball final de grau

Estudi: Grau en Enginyeria en Tecnologies Industrials

Títol: Sistema d'estimació automàtica de carbohidrats en fotografies d'àpats emplatats amb xarxes neuronals

Document: MEMÒRIA I ANNEXOS

Alumne: Martí Gusó

Tutor: Iván Contreras
Departament: Enginyeria elèctrica, electrònica i automàtica
Àrea: Enginyeria de sistemes i automàtica

Convocatòria: Setembre 2022

GRAU D'ENGINYERIA EN TECNOLOGIES
INDUSTRIALS

Sistema d'estimació automàtica de carbohidrats en fotografies d'àpats emplatats amb xarxes neuronals

MARTÍ GUSÓ

Tutor
Iván Contreras

Escola Politècnica Superior
Universitat de Girona
Setembre 2022

SUMARI

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

ACRÒNIMS

T1DM Diabetis mellitus de tipus 1

LSTM memòria a curt i llarg termini

IDF Federació Internacional de Diabetis

SVM Màquina de Suport Vectorial

RESUM

La diabetis és una malaltia greu a escala mundial. Existeixen diverses classificacions d'aquesta malaltia en funció de la severitat. La més severa és la diabetis de tipus 1, la qual comporta una producció d'insulina nul·la per part del pàncrees. La insulina és necessària per controlar els nivells de sucre en sang, i una falta d'aquesta obliga al pacient injectar-se les quantitats adequades. El factor més important a l'hora de calcular la quantitat d'insulina a administrar-se és la quantitat de carbohidrats que s'han ingerit durant les últimes hores. El càlcul de carbohidrats d'un àpat pot resultar complicat per diferents factors, un dels quals pot ser la inexperiència del pacient.

En aquest treball es proposa dissenyar un sistema d'estimació automàtica de carbohidrats a partir d'imatges d'àpats emplatats. L'objectiu és facilitar als pacients de diabetis de tipus 1 el procés de calcular la quantitat de carbohidrats presents en els àpats que ingereixen. D'aquesta manera, es pretén disminuir els errors que es poden cometre portant a terme aquest procés amb mètodes tradicionals i reduir la càrrega que comporta el càlcul als pacients.

Per tal de resoldre aquesta tasca, s'ha utilitzat una de les últimes tecnologies en visió artificial en el camp de l'aprenentatge profund. Concretament, s'utilitza una xarxa de tipus *transformer* aplicada al processament d'imatges, la qual en el darrer any ha demostrat igualar el rendiment de xarxes anteriors i en alguns casos fins i tot millorar-lo. En aquest treball es porta a terme un entrenament supervisat en què es mostra a la xarxa neuronal, també anomenada model, un conjunt d'imatges d'àpats en el qual cadascuna està etiquetada amb la quantitat de carbohidrats que contenen. Un cop entrenat el model, s'avalua la qualitat de les estimacions per tal de valorar si durant l'entrenament ha estat capaç de resoldre correctament la tasca. Les imatges que s'usen durant l'entrenament del model no es fan servir en el procés d'avaluació.

Per entrenar la xarxa neuronal, s'ha buscat el conjunt de dades que millor s'adaptés al problema proposat. Les característiques les quals s'han donat importància han estat la diversitat de mostres, l'abundància d'exemples per mostra, un rang ampli de quantitat de carbohidrats i una alta varietat d'ingredients.

S'ha utilitzat el llenguatge de programació *python* per programar els algorismes que fa possible resoldre la tasca proposada. Els algorismes consisteixen principalment a adquirir, preparar les dades del conjunt de dades, i entrenar el model, així com avaluar-lo. El codi s'ha executat a través de l'aplicació Google Colab, en la seva versió pro, que ofereix un entorn virtual amb un maquinari prou potent per a la tasca que se li exigeix.

CAPÍTOL 1

INTRODUCCIÓ

1.1 Antecedents

La Diabetis mellitus de tipus 1 (T1DM) és un greu problema de salut a escala mundial, suposant una important càrrega per als serveis de salut i és una de les principals malalties que causen la mort. Explicada breument, aquesta malaltia és causada per la destrucció progressiva de les cèl·lules beta del pàncrees. Aquestes cèl·lules són responsables de la producció d'insulina al cos, hormona necessària per a processar els sucres de la sang. Per a compensar la manca de producció d'insulina, els pacients s'han d'administrar dosis per tal de mantenir uns nivells de glucosa en sang adequats. Un dels principals paràmetres que influeixen en el càlcul de la quantitat d'insulina a administrar és el valor de carbohidrats ingerits en les últimes hores. A més, també cal que es controli regularment els nivells de glucosa en sang i l'activitat física a la qual se sotmetin.

Com s'ha esmentat anteriorment, per els pacients de T1DM la estimació dels carbohidrats ingerits en un àpat és crucial per un correcte control de glucosa en sang. Això no obstant, es poden cometre errors produïts per una deficiència d'informació nutricional de l'àpat o en la mida de la porció de l'àpat [1].

En els darrers anys, el desenvolupament de tecnologies en el camp de la intel·ligència artificial ha permès obtenir resultats sense precedents en una gran varietat de problemes. Una d'aquestes tecnologies és l'aprenentatge profund, que a través de l'aplicació de xarxes neuronals artificials permet resoldre tasques de manera més simple i precisa que amb tecnologies disponibles anteriorment. Un dels camps d'aplicació de l'aprenentatge profund és el de la visió artificial. En aquest camp s'hi han desenvolupat diferents arquitectures de xarxes neuronals, entre les quals destaca la xarxa neuronal de tipus convolucional. No obstant això, la recent publicació de les xarxes neuronals de tipus *transformer* mostra que són capaces d'igualar i en alguns casos millorar el rendiment de les xarxes neuronals convolucionals.

Existeixen diferents treballs relacionats amb la tasca d'estimar carbohidrats a partir de fotografies de menjars, els quals proposen diferents metodologies per resoldre la

tasca. En l'apartat de treballs relacionats s'expliquen varis d'aquests treballs, els quals alguns incorporen tècniques d'aprenentatge automàtic.

1.2 Objecte

L'objectiu d'aquest treball és la creació d'un sistema capaç d'estimar la quantitat de carbohidrats presents un àpat donada una fotografia d'aquest. Aquesta eina es portarà a terme amb una de les últimes tecnologies desenvolupades en el camp de l'aprenentatge profund. En concret, s'utilitzarà una arquitectura de xarxa neuronal artificial anomenada *transformer*. Recentment, aquesta arquitectura ha demostrat igualar i en alguns casos donar millors resultats en tasques de processament d'imatge que altres arquitectures anteriors.

El sistema d'estimació de carbohidrats està dissenyat per a què es pugui utilitzar en aplicacions orientades a pacients de T1DM per a ajudar-los a estimar la quantitat de carbohidrats presents en els seus àpats. Aquesta informació és necessària per a calcular la quantitat d'insulina que s'han d'administrar. Per tant, l'objectiu final del treball és crear una eina que millori la qualitat de vida d'aquestes persones.

1.3 Especificacions i abast

Aquest projecte abasta la creació del model d'estimació de carbohidrats en àpats emplatats i la seva avaluació. Es parteix d'un model basat en una arquitectura *transformer*, preentrenat en una tasca de classificació d'imatges. Les dades que s'utilitzen són extretes d'una base de dades d'imatges de plats de menjar etiquetades amb la corresponent informació nutricional necessària.

CAPÍTOL 2

MARC CONCEPTUAL

2.1 Diabetis tipus 1

La T1DM és una malaltia crònica que implica la destrucció de les cèl·lules beta del pàncrees, donant lloc a una deficiència d'insulina [2]. La insulina és una hormona necessària per a poder regular el metabolisme de la captació de glucosa a la sang i mantenir els nivells normals de sucre en sang [3]. Aquesta és l'hormona responsable de transmetre el sucre ingerit als òrgans i teixits musculars. La deficiència d'insulina origina als pacients amb T1DM la necessitat de rebre dosis externes d'insulina mitjançant injeccions d'insulina d'acció ràpida. Les dosis d'insulina es poden administrar a través d'un bolígraf d'insulina o una bomba d'insulina. Aquestes dosis han de ser administrades abans de cada àpat per compensar els sucres ingerits i mantenir els nivells de sucre en sang dins d'uns marges acceptables. A més, els pacients poden necessitar petites dosis d'insulina al llarg del dia per mantenir els nivells mínims. Aquestes dosis es poden administrar en forma de dosis d'acció lenta o a través de bombes d'insulina.

El risc de no controlar correctament aquests nivells de sucre en sang pot provocar un excés o una manca d'aquest. Els termes que s'utilitzen en el camp de la medicina per a referir-se a aquests episodis són hiperglucèmia en cas d'un excés de sucre, i hipoglucèmia en cas d'una deficiència de sucre. Els efectes negatius per a la salut d'aquests episodis són variats, però es poden destacar alguns de més greus. En el cas de la hiperglucèmia, pot derivar en una cetoacidosi diabètica o en un estat hiperosmòtic hiperglucèmic, complicacions les quals són potencialment mortals. La hiperglucèmia de grau 1 es defineix com un nivell de sucre en sang superior a 180 mg/dl, mentre que en la hiperglucèmia de grau 2 és superior a 250 mg/dl [4]. En el cas de la hipoglucèmia, el pacient pot experimentar dèficits neurològics com letargia, pèrdua de memòria o desorientació, i en casos més extrems pot arribar a ser mortal [3]. La hipoglucèmia de grau 1 es defineix com un nivell de sucre en sang inferior a 70 mg/dl, mentre que la hipoglucèmia de grau 2 és inferior a 54 mg/dl.

Segons la Federació Internacional de Diabetis (IDF), hi va haver 537 milions de persones amb diabetis a tot el món l'any 2021 [5], la qual cosa confirma que la diabetis

és una de les emergències sanitàries més grans a escala mundial. Per tant, és imprescindible pal·liar els problemes diaris dels pacients amb T1DM i disminuir la complexitat del seguiment de la malaltia així com la gestió de mantenir els nivells de glucosa en sang dins dels marges segurs. Les millores tecnològiques de les últimes dècades han permès un avenç en les solucions per aquests problemes. Un exemple són els monitors continus de glucosa o l'aplicació de tecnologies d'intel·ligència artificial [6].

Una de les possibles aplicacions d'algorismes d'intel·ligència artificial en aquest camp és el d'estimació de carbohidrats dels àpats ingerits pel pacient, per tal de posteriorment administrar les dosis adequades d'insulina. Els models de visió artificial per entrenament supervisat poden aprendre a relacionar imatges dels àpats amb un valor de carbohidrats. En aquest treball es proposa crear i entrenar un model que relacioni imatges dels àpats amb la quantitat de carbohidrats que contenen aquests. L'objectiu d'aquest model és facilitar als pacients de T1DM el procés de calcular els carbohidrats ingerits.

2.2 Estat de l'art

La visió artificial és un camp científic interdisciplinari que tracta de dotar a sistemes informàtics de la capacitat de comprendre dades en forma d'imatges o vídeos. Aquest camp inclou tasques com l'adquisició, processament, anàlisi i comprensió d'imatges. Es pot entendre la comprensió d'imatges per part d'un model com el procés d'extracció de característiques d'imatges per tal de crear una representació adequada per la tasca que proposada. L'esmentat procés es pot portar a terme amb el suport de la geometria, la física, l'estadística i la intel·ligència artificial.

La visió artificial és útil per resoldre problemes de classificació d'imatges, segmentació d'imatges, detecció d'objectes, modificació d'imatges, transformació de text a imatge i viceversa, entre d'altres. Les imatges amb les quals es treballen poden ser en forma de vídeo, capturades des de diferents càmeres, figures tridimensionals preses amb un escàner 3D o obtingudes a través de dispositius mèdics, entres d'altres.

Una de les xarxes neuronals més utilitzades en el camp de la visió artificial aplicat en imatges és la convolucional. Tradicionalment, les xarxes convolucionals han demostrat ser les millors en tasques de processament d'imatges. En l'última dècada, són les que més s'han usat. No obstant, amb la recent publicació de les xarxes neuronals de tipus *transformer* [7], sembla a ser que les convolucionals podrien esdevenir obsoletes davant d'aquestes. Tot i que inicialment els *transformers* es van dissenyar per a tasques de processament natural del llenguatge, estudis que les apliquen en tasques de visió artificial demostren que poden igualar o fins i tot donar millors resultats [8]. Aquestes noves xarxes són més generalistes que les convolucionals, el que significa que tenen capacitat per comprendre més profundament el problema que se'ls hi pugui proposar. Això no obstant, aquesta característica comporta que necessitin més dades per entrenar-se.

Actualment, en el camp de la medicina la intel·ligència artificial s'utilitza en diferents àmbits, com per exemple en tasques de suport en decisions clíniques i en anàlisi d'imatges. El suport de decisions clíniques ajuda als professionals a prendre decisions en tractaments, medicació o salut mental, entre d'altres. En l'anàlisi d'imatges, la intel·ligència artificial es fa servir per analitzar imatges procedents de procediments com

la tomografia computada, la radiologia o la ressonància magnètica. També es pot aplicar en problemes de predicció i detecció d'esdeveniments clínics com les hipoglucèmies en pacients amb diabetis [9], infarts o atacs neurològics.

2.3 Treballs relacionats

Existeixen diversos estudis que presenten diferents solucions en la tasca d'estimar els carbohidrats presents en un àpat a partir d'una imatge d'aquest. A continuació se'n citen i s'expliquen alguns.

L'aplicació *GoCARB* [10] és un sistema destinat a ser utilitzat per mòbils i per ser usat directament pel pacient. El model de visió artificial que utilitza s'ha entrenat amb 54 plats propis de la gastronomia europea. En l'estudi es compara l'error entre dietistes professionals i el model entrenat en la tasca d'estimació de carbohidrats en les imatges d'entrenament. El model aconsegueix una precisió lleugerament més alta que els dietistes, donant un error absolut mig en les estimacions inferior en les estimacions de les imatges amb què s'ha entrenat. L'usuari ha de capturar dues imatges des de diferents angles del plat amb la càmera del mòbil i han d'estar acompanyades per una targeta convencional per tal de donar una referència de la mida del plat i de l'angle de la fotografia respecte al plat. Per a la tasca d'estimació de carbohidrats, el sistema fa servir un procés consistent dels següents passos: captura de la imatge, detecció del plat, segmentació dels ingredients, classificació dels ingredients, estimació del volum a través del modelatge 3D del plat i finalment càlcul dels carbohidrats amb l'ajuda d'una base de dades de macronutrients dels ingredients (*USDA food composition database*).

En el següent article [11] es presenta un estudi similar que l'esmentat en el paràgraf anterior, però amb imatges d'àpats propis de la gastronomia tailandesa. Es comparen les estimacions del model amb la de dietistes professionals. Els resultats de l'estudi demostren que el model pot competir amb els dietistes, donant en comparació un error generalment més petit en les estimacions. Per a l'estimació de carbohidrats, l'algorisme responsable primer detecta la posició dels diferents ingredients del plat, després segmenta els ingredients per tal de finalment fer una predicció del valor de carbohidrats tenint en compte l'angle en què s'ha capturat la imatge. L'algorisme utilitza tres xarxes neuronals independents: una xarxa convolucional per la detecció d'ingredients, una unitat de segmentació i una xarxa neuronal de regressió per les prediccions finals. El conjunt d'imatges utilitzant per entrenar els models està format per un total de 256.178 ingredients ubicats en 75.232 imatges. Els ingredients es troben classificats entre 175 categories de menjar diferents. Després de l'entrenament, el model té un error quadràtic mig sobre les prediccions de carbohidrats inferior a 10g.

En el següent treball [12] es presenta un conjunt de dades en forma d'imatges de menjar etiquetades amb els macronutrients dels ingredients. Amb la publicació del conjunt de dades, també es presenta un algorisme que estima les calories del menjar de cada imatge. En el conjunt de dades, s'inclouen 2.978 imatges úniques de 19 ingredients diferents més els valors de la massa, el volum, la densitat i les calories. A més, a cada imatge s'hi inclou una moneda com a referència de mida. L'algorisme d'estimació de calories utilitza dues imatges capturades des d'angles diferents per després detectar els menjars presents en el plat i a continuació segmentar aquests ingredients. Un cop fet això en les dues imatges, s'estima el volum del menjar i posteriorment les calories. Per

a la detecció d'ingredients es fa servir una xarxa neuronal convolucional. El volum es calcula a través de l'àrea de l'objecte segmentat, aplicant una proporció amb la mida de la moneda en la imatge. En l'experiment del treball, es demostra que l'error mig relatiu de les estimacions tant de volum com de calories no excedeix el 20% del valor real.

En el següent treball *Nutrition5k* [13] es presenta un nou conjunt de dades per a l'estimació de macronutrients en imatges de plats de menjar així com la proposta d'un algorisme capaç de resoldre aquesta tasca. La publicació d'aquest conjunt de dades és relativament recent, del juny del 2021, i el seu objectiu és proporcionar un conjunt d'imatges d'àpats emplatats més generalista que els existents, amb més ingredients inclosos i més diversitat de plats. S'inclouen els valors de la massa, calories, greixos, carbohidrats i proteïnes per a cada imatge. Està format per 5.000 plats diferents, cadascun etiquetat amb els seus respectius macronutrients i vídeos capturats des de 4 càmeres. Cada càmera pren imatges des d'un angle diferent respecte al pla que reposa al plat. Es captura un vídeo des de cada càmera a la vegada que les càmeres giren al voltant de l'eix vertical centrat en el plat. D'aquesta manera, s'aconsegueix una sèrie d'imatges preses cada una des d'angles diferents. Per cada plat es capturen unes 150 imatges en total, a través de totes les càmeres. Per tant, el nombre total d'imatges que conté el conjunt de dades és d'aproximadament 750.000 . En l'estudi es proposa un algorisme basat en una xarxa neuronal convolucional per a l'estimació dels macronutrients etiquetats en cada imatge. L'error absolut mig relatiu en les estimacions de carbohidrats és del 31,9%.

En el següent treball [14] es presenta el conjunt de dades *FoodSeg-103*, format per imatges d'àpats emplatats amb les segmentacions dels ingredients que hi estan presents. Conèixer l'àrea que ocupa un ingredient en la imatge pot ser útil per a l'estimació del seu volum i, per tant, també és útil per a l'estimació dels macronutrients presents. En el conjunt de dades es troben 7.118 imatges de menjars, amb els corresponents ingredients etiquetats en una de les 103 possibles categories. Es proposen tres xarxes neuronals diferents per a la resolució de la tasca: convolucional, xarxes de memòria a curt i llarg termini (LSTM) i *transformer*. La xarxa que ofereix millor precisió és la LSTM.

En el següent treball [15] es proposa un sistema d'estimació de calories a partir de tres imatges d'un àpat. Tot i que no està destinada a l'estimació de carbohidrats, la metodologia utilitzada es podria aplicar també en aquest problema per la seva semblança. El sistema no fa ús de tècniques d'aprenentatge automàtic, sinó que es basa en operacions matemàtiques per primer classificar el menjar present en les imatges i posteriorment calcular el volum. Per acabar, es calculen les calories presents amb el volum i la informació nutricional del menjar fotografiat i una base de dades. L'avaluació del sistema mostra que és capaç d'estimar el volum dels àpats amb un error absolut màxim del 20%.

FoodCam [16] és una aplicació per a mòbils amb l'objectiu de classificar els ingredients d'àpats a partir d'una imatge d'aquests. El sistema s'inicia a partir de la imatge i els requadres que haurà definit l'usuari al voltant de cada ingredient. A continuació, es segmenta cada requadre per eliminar de la imatge les parts que no continguin menjar per després classificar l'ingredient en una de les possibles categories utilitzant una Màquina de Suport Vectorial (SVM). El sistema assoleix una precisió del 79,2% en classificar els ingredients en una de les 5 classes amb més probabilitat d'entre les 100 possibles.

En el següent treball [17] es proposa un sistema per calcular els volums de menjar en un plat a partir d'imatges amb l'objectiu de ser d'ajuda en la tasca de calcular la

mida de les porcions dels àpats. S'entren al sistema dues imatges de l'àpat preses des d'angles lleugerament diferents acompanyades sempre d'una targeta convencional com a referència de grandària. Llavors, el sistema aplica tres operacions principals. La primera consta de tres parts: calibrar la imatge, el qual suposa detectar els punts comuns en les imatges, calcular les posicions relatives de la càmera i per últim obtenir l'escala a través de les mides de la targeta en les imatges. El segon pas és reconstruir el volum a través de trobar els punts comuns de les imatges rectificades i després utilitzar-los per crear un núvol de punts. L'últim pas consisteix a generar una superfície a partir dels punts trobats anteriorment unint-los de manera que generin àrees triangulars entre ells. Aproximant un pla que representi la base del plat a on reposa el menjar, es calcula finalment el volum estimat de l'àpat. L'error absolut percentual assolit durant l'avaluació del sistema és del 8,2%.

En el següent treball [18] es proposa una metodologia per classificar imatges de diferents menjars utilitzant *Random Forests*, una metodologia d'aprenentatge automàtic que es fan servir en problemes d'aprenentatge supervisat. En aquest cas s'usa per classificar les mostres del conjunt de dades *Food-101*, presentat en el mateix treball i que ofereix 101.000 imatges de menjar etiquetades en una d'entre les 101 classes possibles. Amb la metodologia proposada s'aconsegueix una precisió en la tasca de classificació del 50,76%. Aquesta precisió és molt inferior a l'obtinguda amb models de xarxes neuronal convolucionals introduïts posteriorment, que aconsegueixen precisions per sobre del 95% [19] [20].

L'aplicació LogMeal és una aplicació per a telèfons mòbils que, a través de tecnologies d'intel·ligència artificial, reconeix els diferents ingredients que componen un plat a partir d'una imatge d'aquest. A més, també ofereix informació nutricional dels elements detectats. A part de l'aplicació, l'equip de LogMeal també ofereix altres serveis com una API per tal de poder utilitzar l'aplicació en altres programes, una plataforma destinada als usuaris on es pot monitorar les mostres que s'hagin entrat a l'aplicació, i també un aparell que integra tot el sistema en una mateixa màquina dissenyada per ser utilitzada, per exemple, en restaurants.

CAPÍTOL 3

METODOLOGIA

3.1 Xarxes neuronals artificials

A continuació s'expliquen tres tipus de xarxes neuronals rellevants en l'estructura del model que es proposa per resoldre la tasca d'aquest treball. La primera és el perceptró multicapa, que és una de les xarxes més bàsiques i és àmpliament utilitzada en el camp de la intel·ligència artificial. La segona és la xarxa neuronal convolucional, la qual en els últims anys ha tingut molta rellevància en tasques de visió artificial. En últim lloc, es presenta la xarxa neuronal *transformer*, la qual recentment ha demostrat millorar altres xarxes en alguns àmbits i pot competir amb les xarxes convolucionals en tasques de visió artificial.

3.1.1 Perceptró multicapa

El perceptró multicapa és un tipus de xarxa neuronal on el conjunt de neurones es troben organitzades en capes. Conté una capa d'entrada, una de sortida i una o més capes ocultes. Rep la informació inicial a través de les neurones d'entrada, es processa al llarg de les diferents capes ocultes i finalment les neurones de sortida retornen la informació processada. Cada neurona de la xarxa està connectada amb totes les neurones de les capes adjacents. Aquesta connexió té associada un pes que multiplica el valor de sortida de la neurona anterior. A més de multiplicar-se pel pes, se suma un biaix al valor de sortida de la neurona anterior. La sortida d'una neurona es calcula aplicant una funció d'activació al resultat de multiplicar la sortida de la neurona anterior amb el pes i sumar el biaix. Exemples de funcions d'activació són la sigmoide, la tangent hiperbòlica, l'exponencial normalitzada o SoftMax, la Relu o la Gelu.

L'aprenentatge de la xarxa neuronal es porta a terme amb l'ús d'un algorisme anomenat "backpropagation"o propagació de l'error. Aquest algorisme consisteix a actualitzar els pesos de les connexions i els biaixos de les neurones en funció de l'error que hagin provocat en la sortida. Com més error hagin provocat, més es modificarà el seu valor. L'error es calcula amb una funció que compara la sortida de la xarxa amb el

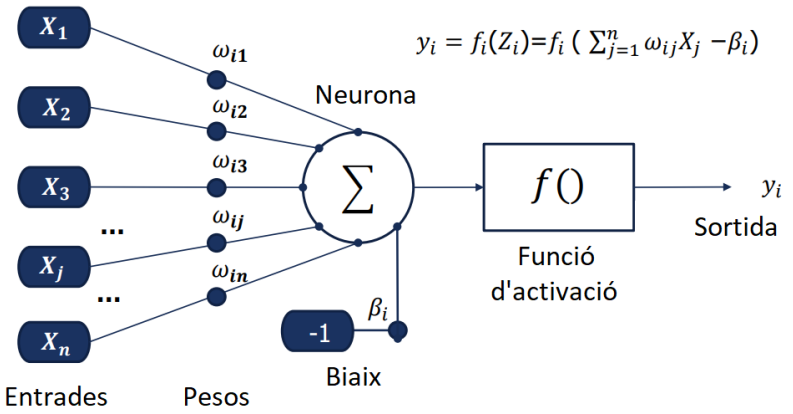

Figura 3.1: Esquema del funcionament d'una neurona en un perceptró multicapa

valor esperat. Aquest algorisme s'aplica en entrenaments supervisats, el qual significa que el model aprèn a resoldre una tasca a partir de comparar el valor de sortida d'aquest amb l'esperat.

El teorema d'aproximació universal indica que una xarxa neuronal de tipus perceptró multicapa amb una sola capa oculta pot aproximar qualsevol funció [21].

3.1.2 Xarxes neuronals convolucionals

Les xarxes neuronals convolucionals han estat dissenyades principalment per tasques de visió artificial [22], tot i que també es poden aplicar en altres camps de la intel·ligència artificial. En tasques de processament d'imatges, aquestes són representades com un tensor a l'entrada de la xarxa. El funcionament d'aquestes xarxes consisteix a aplicar filtres a aquest tensor d'entrada al llarg de tota la seva àrea per cada canal, creant un mapa de característiques per cada filtre que s'aplica. D'aquesta manera, a la següent capa de la xarxa hi ha més mapes de característiques però amb una àrea inferior als mapes anteriors. Finalment, el tensor que s'introdueix a l'entrada acaba passant a ser un vector, el qual recull la informació necessària de la imatge en la tasca que s'hagi entrenat el model.

Les xarxes convolucionals tenen alguns desavantatges quan es comparen amb els *transformers* [23]. Un dels desavantatges que presenten és que en aplicar els filtres no tenen en compte la posició relativa a la imatge. Això provoca que elements d'un objecte desordenats en l'espai puguin provocar una falsa detecció de l'objecte. Per exemple, si es reparteix els elements d'una cara aleatòriament en una imatge, la xarxa podria considerar igualment que en aquesta imatge hi ha una cara 3.4. Aquest problema es pot solucionar, però requereix una elevada potència computacional.

Un altre desavantatge és la dificultat que tenen en relacionar elements llunyans en una imatge. La relació entre dos elements distants pot quedar negligida a mesura que s'apliquen convolucions. Els *transformers* no tenen aquest problema, ja que processen tot el conjunt de la imatge simultàniament, independentment de la distància en què es trobin els elements.

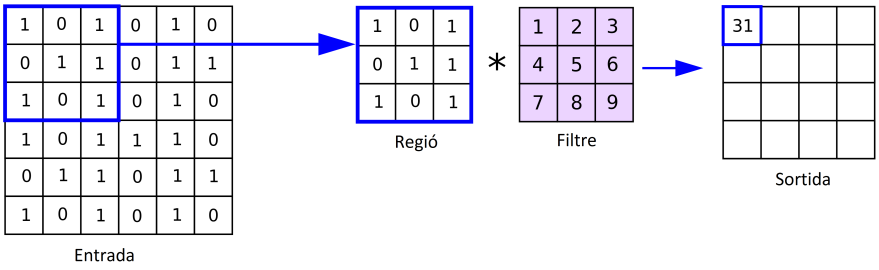

Figura 3.2: Representació del funcionament d'un filtre d'una xarxa neuronal convolucional

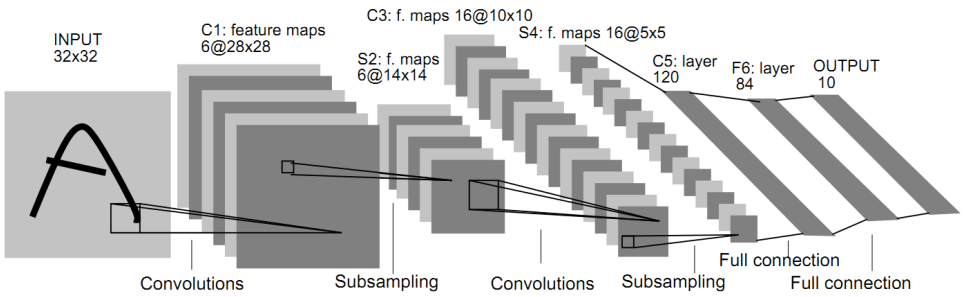

Figura 3.3: Representació del funcionament global d'una xarxa neuronal convolucional amb arquitectura *LeNet-5*

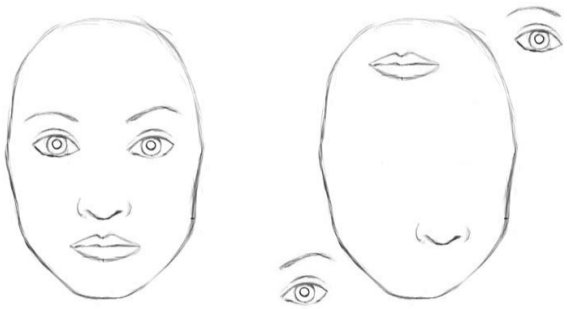

Figura 3.4: Per a una xarxa neuronal convolucional les dues cares són gairabé idèntiques

No obstant això, les xarxes neuronals convolucionals tenen l'avantatge de poder aprendre una tasca amb una quantitat de dades més petita, a canvi de no entendre amb tanta profunditat la tasca. Per contra, els *transformers* necessiten més dades perquè aprenguin però generalitzen millor, prenen així un enteniment més profund del problema.

3.1.3 Transformers

Les xarxes de tipus *transformers* [7] són xarxes neuronals que processen seqüències de mostres. A través d'un mecanisme d'atenció són capaces de quantificar la relació que cada mostra té amb la resta. Un cop calculada l'atenció, es retorna l'entrada codificada en forma de vector. Es va utilitzar per primera vegada en tasques de processat natural del llenguatge o *NLP* (*Natural Language Processing*), més concretament en tasques de traducció.

Aquest tipus de xarxa es va desenvolupar per primera vegada en tasques de traducció de text. La capacitat de comprendre la relació entre dos elements d'una seqüència de mostres independentment de la posició d'aquests suposa una millora respecte xarxes anteriors, com les recurrents, les quals tendeixen a "oblidar·la importància d'un element de la seqüència al llarg de les iteracions. A més, aquesta arquitectura de xarxa neuronal permet més paral·lelització de la seva computació i, per tant, una notable disminució de la potència computacional requerida pel seu entrenament i inferència. Actualment, s'està avaluant el seu ús en altres aplicacions, com per exemple en problemes de processament d'imatges.

L'arquitectura de les xarxes de tipus *transformer* per a la tasca de traducció de texts està composta per dos blocs: el codificador o *encoder* i el descodificador o *decoder*. El primer codifica cada paraula de la frase en forma de vector tenint en compte la seva posició relativa i genera a través del mecanisme d'atenció un vector que recull el significat global de la frase. Durant l'entrenament del model, en el descodificador, es fa el mateix amb la frase equivalent de la llengua a la qual es vol traduir la primera frase. Seguidament, també amb el mecanisme d'atenció, es calcula la relació que té cada paraula de la primera frase amb cada paraula de la segona, donant com a sortida un altre vector. Finalment, s'entra aquest últim vector a una xarxa de tipus *feed-forward* la qual aplicant una funció *SoftMax* retornarà la paraula traduïda a l'idioma desitjat. En la figura 3.5 es mostra un esquema d'aquesta estructura.

El mecanisme d'atenció és el principal component de les xarxes de tipus *transformer* i permet calcular la relació entre les diferents mostres de la seqüència d'entrada. S'utilitzen tres xarxes neuronals durant aquest procés, cada una encarregada de transformar a un vector la mostra amb què s'estigui tractant. El resultat d'aquesta operació són els anomenats vectors *key*, *query* i *value*. El vector *key* conté el significat propi de la paraula, el vector *query* les propietats que tenen relació amb aquesta paraula i el vector *value* representa el significat de la paraula de la manera més adient per a resoldre la tasca proposada. Per trobar la relació entre una paraula amb una altra, es calcula el producte escalar entre el vector *key* de la primera amb el vector *query* de la segona. Si s'aplica aquesta operació amb totes les paraules s'obté la matriu d'atenció, on es troba la relació o "atenció"que el model dóna a les paraules de la frase en el moment de llegir-ne una altra. Es multiplica la matriu d'atenció pels vectors *value* de cada paraula per extraure un vector que recull el significat global de la frase. En la figura 3.6 es mostra un esquema del procés.

Transformers aplicats en tasques de classificació d'imatges

Existeixen diferents variants de models basats en *transformers* per a la tasca de processament d'imatges. En aquest treball es parteix d'un model de classificació d'imatges

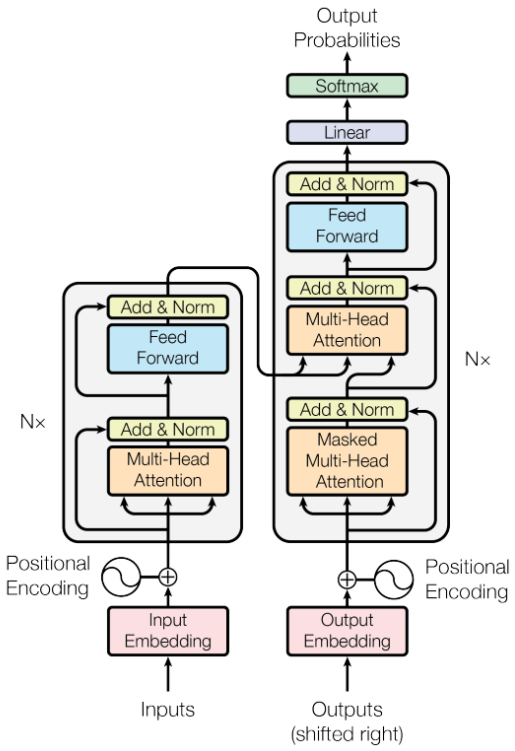

Figura 3.5: Esquema de l'estructura del model d'un *transformer* per a una tasca de traducció de text

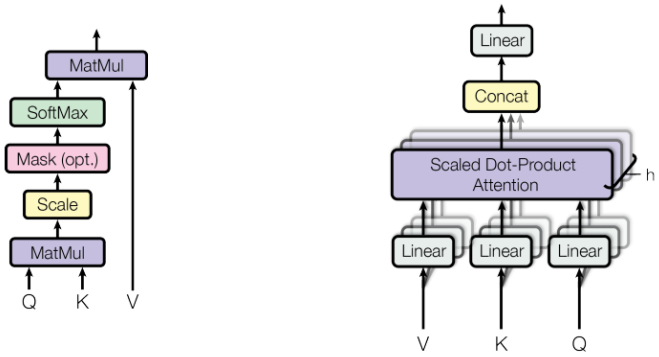

Figura 3.6: A l'esquerre, l'operació de producte escalar. A la dreta, diversos blocs d'atenció en paral·lel

basat en una xarxa transformer [8] per a llavors modificar-lo per la tasca proposada, la qual es tracta d'una regressió. A continuació s'explica el funcionament d'aquest model i la seva arquitectura.

Els *transformers* treballen amb una seqüència de mostres per a trobar la relació entre elles. Per poder processar imatges amb un *transformer*, cal convertir la imatge en una seqüència de mostres. La solució més intuïtiva seria que cada píxel fos una mostra, i el *transformer* computés la relació d'aquest amb la resta de píxels de tota

la imatge. Això no obstant, la gran potència computacional necessària per a portar a terme aquesta solució fa que aquesta sigui inviable. Existeix l'alternativa de dividir la imatge en diferents peces de mides iguals, com si fos un mosaic de quadrats alineats. Per a cada una de les peces, es reordenen els píxels en una sèrie i es multiplica per una matriu posició per donar context de la posició relativa en la imatge. Per tant, cada peça passa a ser un vector multidimensional amb informació de l'emplaçament i aquest serà el format en què les mostres entraran al *encoder*.

Dins del *transformer*, cada mostra és processada per tres xarxes neuronals diferents, cada una generant un vector *key*, *query* i *value* de cada peça. El vector *key* representa les propietats pròpies de la peça, el vector *query* les propietats que tenen relació amb la peça i el vector *value* representa la peça per tal de resoldre la tasca proposada. Amb els dos primers vectors, es pot trobar la compatibilitat que tenen dues peces diferents. Per trobar aquesta compatibilitat es calcula el producte escalar entre el vector *(*key) de la mostra amb els vectors *(*query) de les altres. El resultat d'aquest producte és la compatibilitat entre les dues peces i l'atenció que es dóna a una mostra respecte a una altra. Un cop trobat els valors d'atenció entre les peces, es computa la sortida del *transformer* fent una suma ponderada dels vectors *(*value), on la ponderació és definida per l'atenció. Per tant, la sortida del *transformer* serà un altre vector multidimensional que representarà la imatge. Es pot entendre el conjunt de dimensions del vector sortida com el conjunt de propietats que la xarxa ha considerat més rellevants a l'hora de representar la imatge per a la tasca que se li ha proposat.

En el cas del model ViT per a classificació, el model classifica la imatge en una de les classes possibles. Per aconseguir-ho s'afegeix una capa de neurones després de la sortida del *transformer*. Es connecta cada valor de les n dimensions del vector de sortida del *transformer* amb cada neurona de la capa de classificació. En la capa de classificació es troben les neurones que representen a cada una de les possibles classes a triar. Per tant, cada neurona de la capa de classificació, computada una imatge en el *transformer*, donarà un valor diferent. Aquest valor correspon a la semblança de la imatge amb la classe corresponent. Quant més gran el valor, més semblança té amb aquesta classe. Aplicant la funció *SoftMax* es pot obtenir la neurona que ha retornat el valor més gran. La classe corresponent a aquesta neurona serà la sortida del model de classificació. Les possibles classes que pot retornar el model donat una imatge d'entrada es defineixen durant el disseny del model.

Per a la tasca d'aquest treball s'utilitza una versió modificada d'una xarxa *transformer* inicialment dissenyada per a la classificació d'imatges. En l'apartat de disseny del model s'explica aquesta modificació.

3.2 Preparació de les dades

L'aprenentatge supervisat és un tipus d'aprenentatge automàtic en què es mostra al model la sortida que ha de retornar per a cada entrada. Donades les dades d'entrada i sortida, el model aprendrà durant el procés d'entrenament a trobar la relació entre elles, actualitzant els paràmetres interns de la xarxa neuronal. La tasca de relacionar una entrada qualsevol amb una sortida de valor numèric continu s'anomena de regressió.

En la tasca que es proposa en aquest treball, l'entrada del model és una imatge d'un àpat emplatat i la sortida és un valor que correspon als carbohidrats que conté

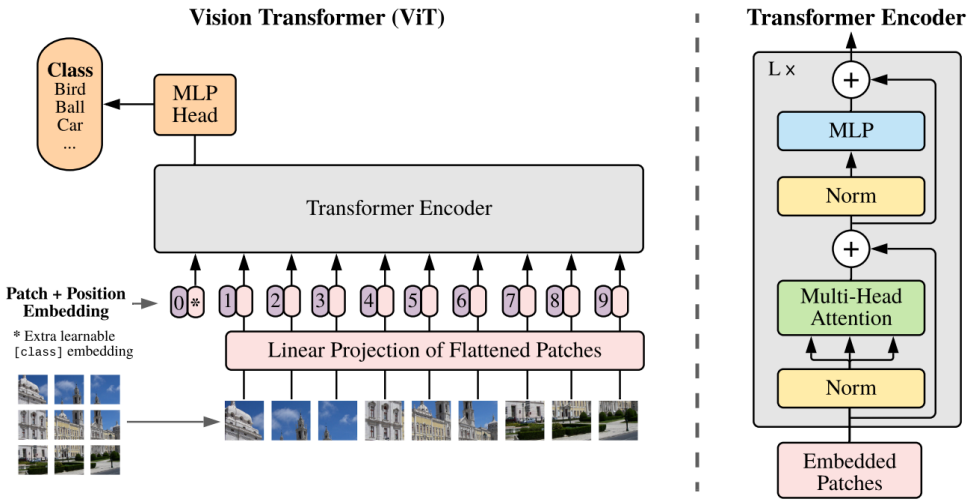

Figura 3.7: Esquema del funcionament de la xarxa de tipus transformer ViT

l'àpat. La base de dades que s'utilitza serà, per tant, un conjunt d'imatges d'àpats emplatats on cada imatge té relacionada una quantitat de carbohidrats. Cal organitzar aquestes dades en la memòria del programa per a després ser usades per l'algorisme d'entrenament durant la seva execució.

La base de dades que s'utilitza en aquest treball és *Nutrition5k* [13]. Ofereix 20.000 vídeos curts de 5.000 plats únics elaborats a partir de 250 ingredients diferents. Per cada plat, ofereix informació en pes de la quantitat de cada ingredient i altra informació nutricional com la quantitat de calories, greixos, proteïnes i carbohidrats.

Macronutrient	**Mitjana**	**Desviació estàndard**	**Desviació mitjana**
Calories	255	220	136
Massa total (g)	215	161	114
Greixos (g)	12,7	13,5	6,93
Carbohidrats (g)	19,4	21,6	10,3
Proteïnes (g)	18,0	20,0	10,7

Taula 3.1: Macronutrients etiquetats en cada mostra de la base de dades Nutrition5k

La base de dades és d'accés públic i es troba disponible a través d'internet. Partint d'un algorisme senzill es poden extreure les imatges a través dels seus enllaços per ser guardades temporalment en la memòria de python. Durant aquest procés, es modifica el format de les imatges de *bgr* a *rgb*. En total, a partir dels vídeos del conjunt de dades es poden extreure aproximadament unes 740.000 imatges úniques. La informació dels macronutrients de cada plat està organitzada en un fitxer Excel. A través de la llibreria *pandas* de python es pot extreure fàcilment aquestes dades. En concret s'extreu la quantitat de carbohidrats per cada plat.

El model treballa amb imatges de 224x224 píxels. Cal modificar aquestes imatges perquè siguin d'aquesta mida, canviant la mida i la relació d'aspecte, en el cas que

sigui necessari. Aquesta operació es fa amb una funció anomenada extractor de característiques. Aplicada aquesta funció a una imatge, aquesta passa a ser un tensor amb els valors normalitzats (entre -1 i 1). El resultat és un tensor de quatre eixos, els quals corresponent a la mida del lot de mostres, a l'altura de la imatge, a l'amplada de la imatge i al nombre de canals (normalment 3 per imatges representades en format *rgb*). A la figura 3.12 es mostren a manera d'exemple vuit imatges presents en el conjunt de dades.

Com que treballar amb totes les imatges resulta en uns temps d'entrenament excessivament llargs, per aquest treball, s'han reduït el nombre total d'imatges utilitzades. Per cada plat s'ha extret només la primera imatge captada en els vídeos de cada càmera. Per tant, per cada plat hi ha 4 imatges. També s'han descartat els plats que contenen valor de carbohidrats per sobre de 100 grams, ja que eren pocs i, per conseqüència, és difícil pel model generalitzar sobre aquests. Aplicades aquestes condicions, el conjunt de mostres està format per unes 17.000 imatges.

A la figura 3.8 es representa a través d'un diagrama de caixa la distribució dels valors de carbohidrats de tots els plats utilitzats durant l'entrenament i avaluació del model. La major part dels plats no superen el valor de 50 grams, però hi ha valors atípics amb quantitats que arriben als 100 grams.

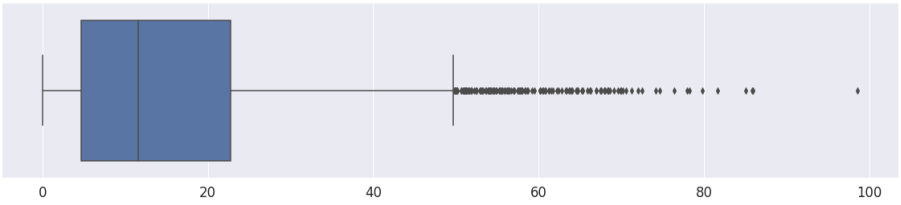

Figura 3.8: Diagrama amb els valors de CHO dels plats utilitzats

Per cada plat, a més d'incloure els valors de diversos macronutrients, també s'inclou la llista d'ingredients del qual es composa. Hi ha un total de 243 ingredients. En el gràfic de barres de la figura 3.9 es mostren els 50 ingredients que més apareixen en els plats, amb el nombre de vegades que apareixen en els diferents plats.

Cada plat està format per entre 1 i 17 ingredients. A la figura 3.10 es pot visualitzar en un gràfic de barres els plats organitzats en funció del nombre d'ingredients que contenen. Una gran part dels plats contenen només un ingredient, i la resta en solen contenir entres 2 i 5. La quantitat de plats que contenen entre 6 i 17 ingredients és reduïda.

A continuació, en la figura 3.11, se sumen les quantitats de carbohidrats que aporta cada ingredient en cada plat. D'aquesta manera es representa en un gràfic de barres els ingredients del conjunt de dades que més carbohidrats aporten en total.

Un cop es té un accés a les dades des del codi, totes les dades necessàries s'organitzen en la memòria en forma d'una variable python de tipus diccionari. Es creen tres apartats dins d'aquest, corresponent als conjunts d'entrenament, validació i test. En cada conjunt s'hi destinen un 80%, 10% i 10% del total de mostres respectivament en l'ordre que s'han esmentat. Aquestes són les fraccions més freqüents que s'utilitzen en el camp de l'aprenentatge automàtic. Dins de cada conjunt, hi ha dos subconjunts corresponents a les imatges i a les quantitats de carbohidrats corresponents al menjar

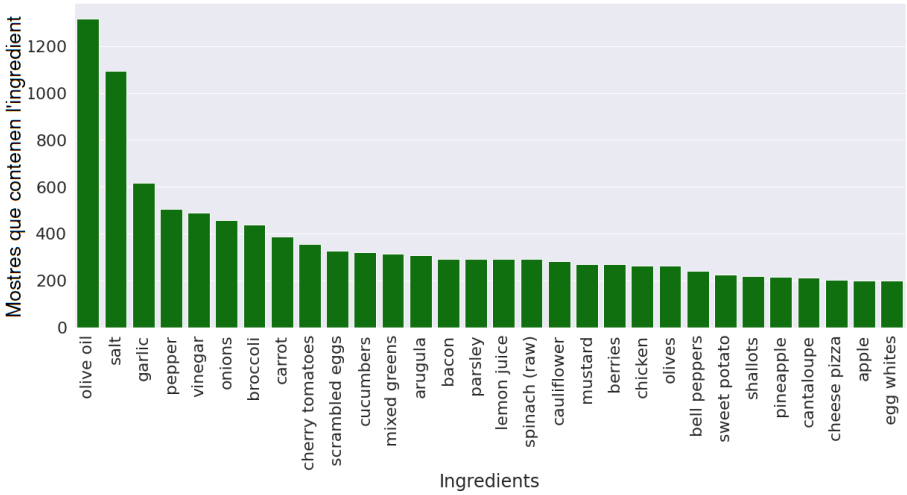

Figura 3.9: Gràfic de barres dels 50 ingredients que més s'utilitzen en els plats

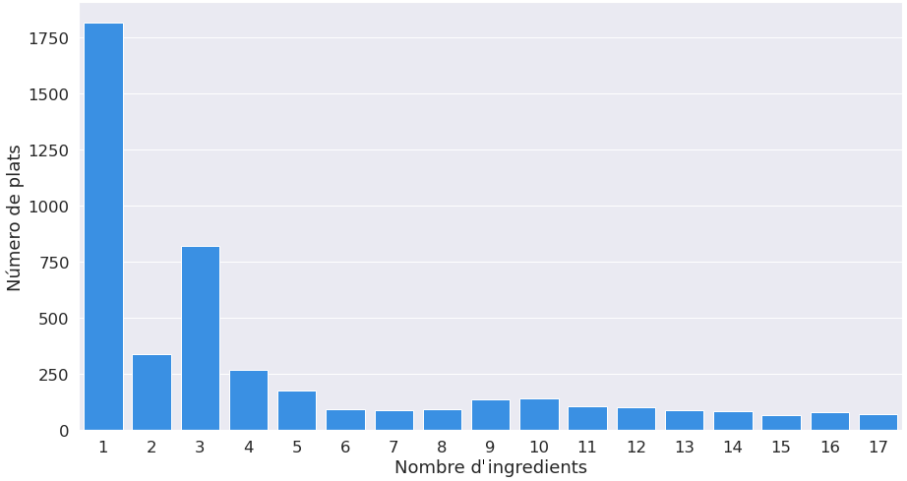

Figura 3.10: Gràfic de barres dels plats organitzats en funció del nombre d'ingredients que els formen

de la mostra. Per a decidir quines mostres van a cada conjunt s'ha seguit el criteri dels creadors del conjunt de dades: classificar aleatòriament en el conjunt d'entrenament el 80% dels plats totals amb les seves respectives imatges i la resta classificar-los en un conjunt de test. D'aquesta manera, s'aconsegueix que cap imatge del mateix plat aparegui en els dos conjunts simultàniament. Per tant, cap imatge d'un plat classificat en el conjunt de test no es mostra en cap moment al model durant l'entrenament. El conjunt de dades està acompanyat per dos fitxers on en un estan apuntades les referències dels plats d'entrenament i en l'altre les de test. Com que en el present treball també es requereix un conjunt de dades de validació, addicionalment a les particions proposades pels creadors, s'ha creat la partició de validació i test prenent les dades intermitentment del conjunt de test inicial. D'aquesta manera, es divideix el conjunt de les dades de test inicial en dos conjunts, el de validació i un nou de test, amb un igual

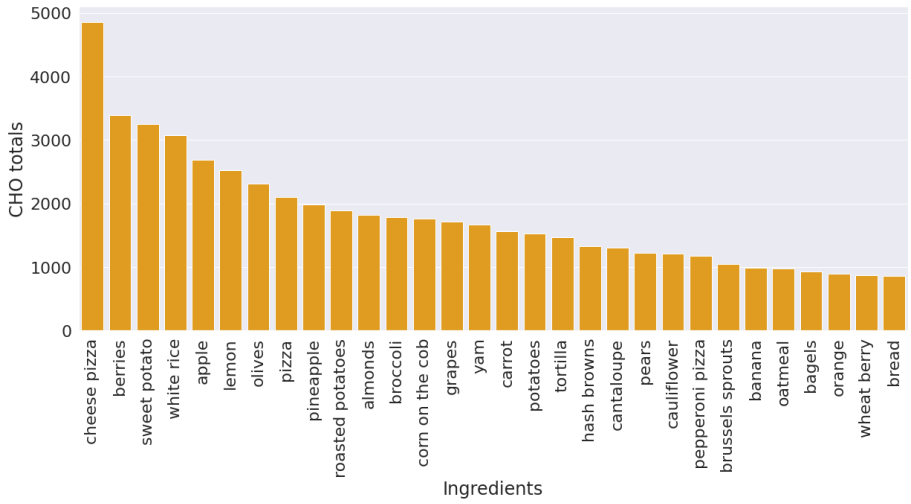

Figura 3.11: Carbohidrats totals per ingredient

nombre de plats.

El conjunt de dades d'entrenament és amb el que pròpiament s'entrena el model. Per altra banda, el conjunt de dades de validació no es mostra al model, però sí que s'utilitza durant el procés d'entrenament. Concretament, es fa servir aquest conjunt per avaluar el rendiment del model en la tasca proposada en certs instants de l'entrenament per tal d'ajustar els hiperparàmetres de l'algorisme d'entrenament. Finalment, el conjunt de dades de test no s'utilitza en cap moment durant l'entrenament sinó que únicament es fa servir per avaluar el model un cop ja està entrenat.

Figura 3.12: Exemples d'imatges presents en la base de dades

A partir d'aquest punt, les dades ja estan preparades per a ser enviades a l'algorisme d'entrenament.

3.3 Disseny del model

3.3.1 Transference learning

El concepte de *transfer learning* és un mètode de l'aprenentatge automàtic que es basa a utilitzar un model entrenat en la tasca resoldre un determinat problema per a entrenar-lo en una nova tasca diferent però relacionada.

Un model prèviament entrenat en una tasca de classificació d'imatges té la capacitat de classificar un conjunt de classes d'imatges diferents, però en el procés d'aprendre a resoldre aquesta tasca, també aprèn a interpretar imatges i a extreure informació d'aquestes. Per tant, es pot utilitzar un model de classificació d'imatges preentrenat per ser novament entrenat en un altre conjunt d'imatges, encara que les imatges no les hagi computat mai. De la mateixa manera, la tasca en què s'ha preentrenat el model inicial no té per què ser exactament la mateixa que la que s'entreni després. D'aquesta manera s'aconsegueix que al model li sigui més fàcil aprendre la nova tasca, i a la vegada requerint menys temps d'entrenament i potència de computació.

En aquest treball s'utilitza un model de classificació d'imatges com a punt de partida per crear un model d'estimació de carbohidrats a partir d'imatges. El model inicial ha estat preentrenat en un gran conjunt d'imatges amb la tasca de classificar-les entre unes determinades possibles classes. Durant el preentranament, el model aprèn a extreure informació de les imatges representant-les com un vector en un espai multidimensional. Cada dimensió dels vectors resultants representa una característica de la imatge, generada amb el criteri del model.

3.3.2 Arquitectura de la xarxa neuronal

El problema que es planteja en aquest treball és el d'obtenir una estimació del valor de carbohidrats presents en un plat a partir d'una imatge d'aquest. Entrenar un model des de zero és una tasca computacionalment molt costosa, per tant, s'utilitza un model preentrenat en una tasca de processament d'imatges i es modifica lleugerament per adaptar-lo a la present tasca.

El model del qual es parteix és el *ViT-B/16* [8]. Ha estat preentrenat en el conjunt d'imatges *ImageNet-21k* [24] en la tasca de classificació. En concret, aquest model s'ha preentrenat en la tasca de classificar 14 milions d'imatges en una d'entre les 21.000 classes possibles. Existeixen altres models basats en *transformers* destinats a aquesta tasca, però s'ha triat aquest perquè dóna uns bons resultats en la tasca esmentada i no té una arquitectura excessivament complicada.

El conjunt de dades *ImageNet-21k* és una col·lecció d'uns 14 milions d'imatges, cadascuna etiquetada en una d'entre 21.841 possibles classes. És un dels conjunts de dades destinats a entrenar models de classificació d'imatges de domini públic més grans que hi ha actualment disponible. La gran varietat d'imatges dota als models que s'entrenen amb el conjunt d'una capacitat generalista de comprensió d'imatges. Aquesta característica permet que aquests models siguin útils per ser usats en altres tasques més específiques.

Cal modificar la sortida del model preentrenat de classificació perquè retorni un valor numèric en comptes d'una de les possibles classes. Per tant, s'elimina l'última capa del model corresponent al classificador i s'afegeix una única neurona en el seu lloc.

Aquesta neurona és la responsable d'estimar els valors de carbohidrats de la imatge d'entrada i està connectada amb cadascun dels diferents valors del vector de sortida del *transformer*. A la figura 3.13 es mostra un esquema de l'última capa del model ViT original a l'esquerra, i a la dreta, l'última capa del model modificat composta per una única neurona.

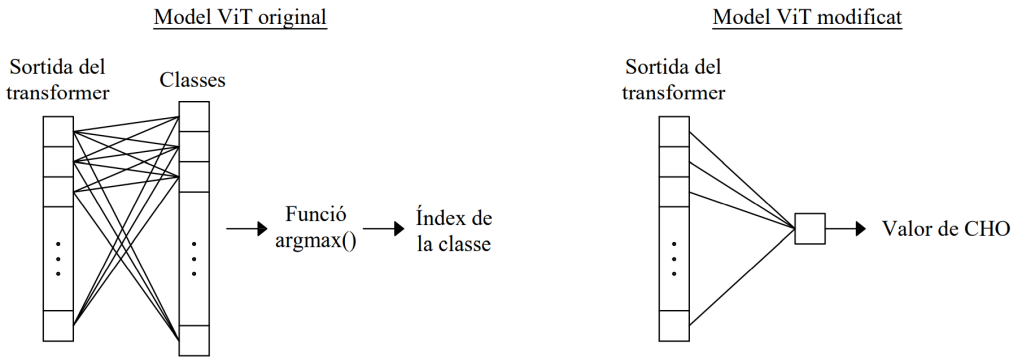

Figura 3.13: Modificació de les capes finals de model ViT per adaptar-lo a la tasca proposada

3.3.3 Entrenament del model

Definides les entrades, les sortides i el model, es pot executar l'algorisme d'entrenament. Es porta a terme un entrenament supervisat en què les sortides del model s'ajusten als valors reals de carbohidrats associats amb cada imatge d'entrada. El procés d'entrenament consisteix a avaluar el model en el conjunt de dades i modificar-lo seguint l'algorisme de la propagació de l'error. Aquesta operació es repeteix al llarg de diverses iteracions, en aquest cas, 1.740 en total.

CAPÍTOL 4

RESULTATS I DISCUSSIÓ

En aquest capítol s'avaluarà el rendiment del model entrenat en la tasca d'estimació de carbohidrats a partir d'imatges d'àpats emplatats. Les dades amb les quals s'avalua el model corresponent a les de la partició de test. Aquestes dades no s'han mostrat durant l'entrenament i, per tant, és la primera vegada que el model les processa. Tots els valors d'errors absoluts d'aquest capítol estan mesurats en grams, unitat amb la qual el model retorna les estimacions.

4.1 Mètriques d'avaluació

La funció d'error que s'ha utilitzat durant l'entrenament del model ha estat l'error quadràtic mig. Aquesta funció penalitza els errors de forma quadràtica, de manera que es penalitzaran considerablement més els errors més grans. Es tria aquesta funció perquè s'ha considerat que per la naturalesa de la tasca, és millor tenir errors petits que alguns de grans.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.1}$$

Per avaluar el model s'ha utilitzat la funció de l'error absolut mig. S'ha considerat que aquesta mètrica és la que mostra d'una manera més intuïtiva el rendiment del model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{4.2}$$

4.1.1 Sobre ajustament

En la gràfica 4.1 es mostra l'evolució de l'error del model en el conjunt de dades d'entrenament al llarg de les iteracions del procés. La mètrica que s'utilitza per calcular l'error que apareix a la gràfica és proporcional a l'error quadràtic mig.

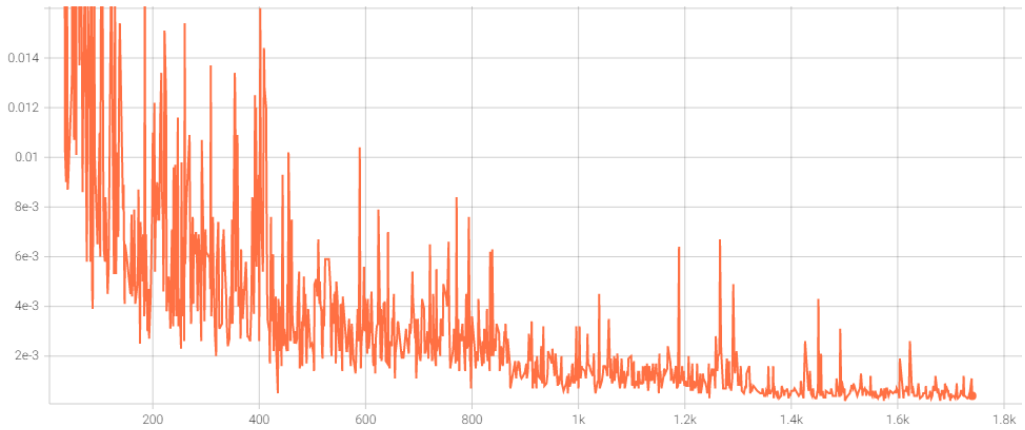

Figura 4.1: Evolució de l'error al llarg de les iteracions del procés d'entrenament

El model s'entrena amb un *learning rate* de valor 1e-4. L'algorisme d'entrenament pot modificar aquest valor per tal d'evitar un sobre ajustament. La mida dels lots de mostres amb el qual el model treballa simultàniament és de 32 mostres. El nombre d'èpoques és de 4. El nombre d'iteracions que portarà a terme l'algorisme d'entrenament és funció del nombre de mostres, la mida del lot de dades i el nombre d'èpoques. El nombre d'iteracions es calcula dividint el nombre total de mostres d'entrenament per la mida de lot i multiplicant aquest resultat pel nombre d'èpoques.

Durant l'entrenament del model es pot presentar el problema del sobre ajustament. Aquest pot aparèixer quan el model s'ajusta massa a les dades d'entrenament de manera que l'error que presenta a les dades de test és molt superior a l'error en les dades d'entrenament. En la figura 4.2 es mostra a manera d'exemple una gràfica de l'evolució de l'error en les dades d'entrenament i test al llarg de les iteracions i on pot ocórrer aquest problema.

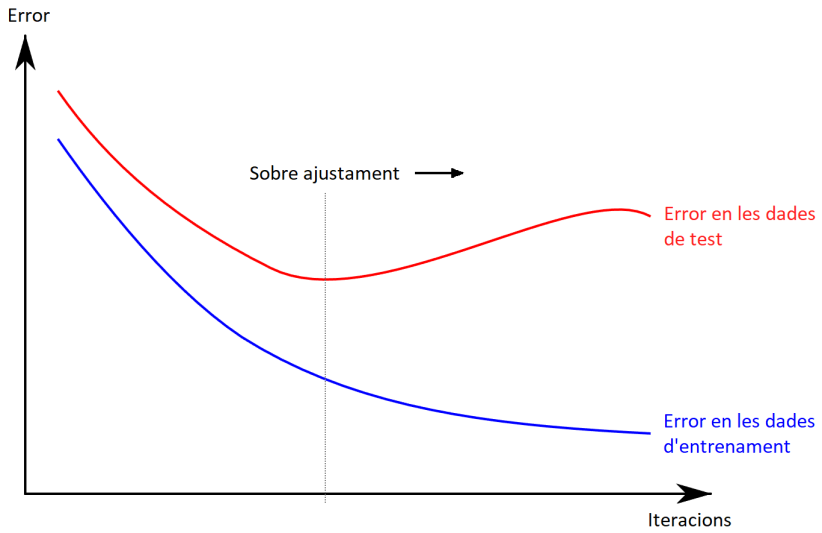

Figura 4.2: Exemple de sobre ajustament del model durant l'entrenament

Per comprovar que el model que hem entrenat en aquest treball no presenti problemes de sobre ajustament, es compara l'error que presenta en avaluar-lo en el conjunt de dades d'entrenament, validació i test. Si l'error no presenta diferències rellevants entre els diferents conjunts, es considera que no ha aparegut sobre ajustament en l'entrenament. La funció d'error que s'utilitza per portar a terme aquesta avaluació és la de l'error absolut mig. A continuació es mostra la taula 4.1 amb els resultats.

Conjunt de dades	**Error absolut mig**
Entrenament	0,85
Validació	4,57
Test	4,65

Taula 4.1: Error absolut mig obtingut en avaluar el model en els diferents conjunts de mostres

Tot i que relativament l'error en les dades d'entrenament és unes cinc vegades inferior respecte a l'error en les dades de test, en valors absoluts, l'error en les dades de test no augmenta significativament respecte a les dades d'entrenament. Per tant, es pot concloure que durant l'entrenament del model no ha aparegut sobre ajustament.

4.2 Predicció de carbohidrats

A continuació es mostra l'error absolut mig (MAE), l'error relatiu de la mitjana dels valors estimats respecte la mitjana dels valors de carbohidrats (ER) i la desviació estàndard (SD) en les prediccions de carbohidrats del model entrenat. També s'inclou l'error quadràtic mig (MSE) i l'arrel de l'error quadràtic mig (RMSE).

Mètriques	**MAE**	**ER**	**SD**	**MSE**	**RMSE**
Valors	4,65	28,07%	6,30	61,37	7,83

Taula 4.2: Resultats de l'avaluació del model

Com es pot observar a la taula, les estimacions de carbohidrats del model s'ajusten als valors reals. Això indica que el model ha après correctament a relacionar cada plat amb la quantitat de carbohidrats que conté. Per tant, es pot afirmar que el model, un cop entrenat, resol correctament la tasca que se li ha proposat.

L'error absolut mig del model basat en una xarxa *transformer* presentat en aquest treball és inferior a l'error del model basat en una xarxa convolucional presentat en el treball fet sobre el conjunt de dades *Nutrition5k* [13] en la seva publicació. Concretament, l'error que presenta la xarxa convolucional és de 6,1. Per tant, la xarxa entrenada en aquest treball redueix l'error absolut mig en un 23,8%.

Cal aclarir que durant l'avaluació del model, els valors de carbohidrats predits inferiors a zero es transformen automàticament a zero. Aquesta és una transformació lògica, ja que no és físicament possible que un plat contingui valors negatius de carbohidrats. Durant l'entrenament del model no s'ha aplicat aquesta transformació.

Per tal de visualitzar la distribució dels valors dels errors absoluts que presenta el model, es representen a la figura 4.3 en una caixa de dispersió. La major part dels

errors tenen un valor absolut d'entre 0 i 10 grams. Això no obstant, hi ha errors que sobrepassen aquest valor i arriben per sobre dels 40 grams.

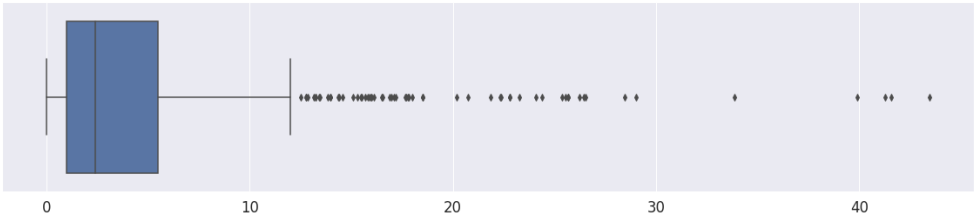

Figura 4.3: Caixa de dispersió de l'error absolut

A la següent figura 4.4 es mostren diversos exemples d'imatges del conjunt de dades de test, cada una etiquetada amb el valor real de carbohidrats i també el valor de carbohidrats predit pel model.

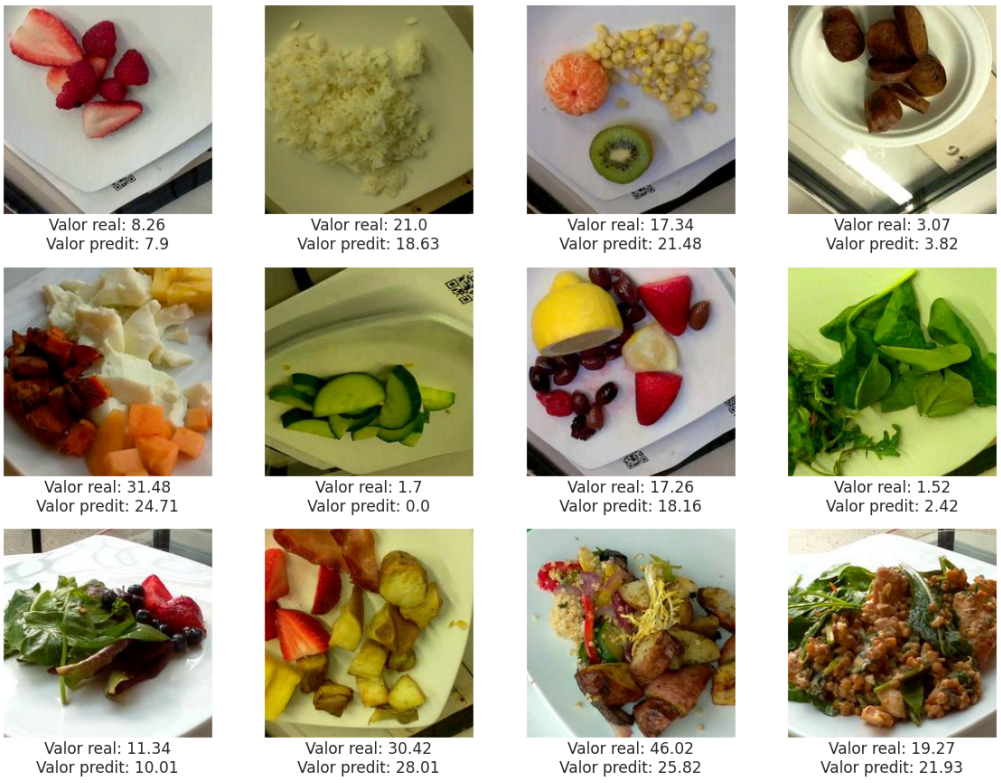

Figura 4.4: Imatges del conjunt de dades etiquetades amb el valor real i el valor predit de carbohidrats

4.2.1 Relació entre l'error relatiu i la quantitat de carbohidrats presents

Es comprova la relació que hi pugui haver entre l'error i la quantitat de carbohidrats presents en el plat. L'objectiu és comprovar si els plats amb més carbohidrats dificulten la resolució de la tasca al model. A continuació es representa en una gràfica els conjunts

d'errors relatius comesos pel model per a cada plat en funció dels carbohidrats presents en el plat.

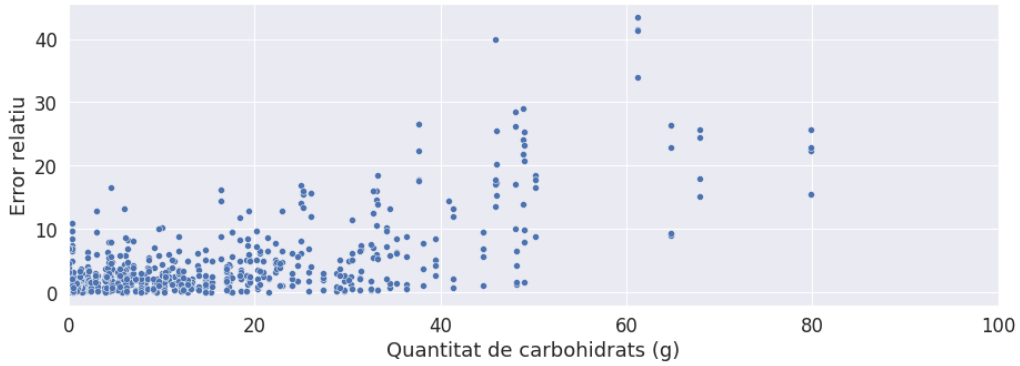

Figura 4.5: Gràfic de dispersió de l'error relatiu en les estimacions en funció de la quantitat de carbohidrats presents en el plat

Com es pot apreciar en el gràfic de dispersió de la figura 4.5 l'error relatiu té tendència a augmentar a mesura que els plats contenen valors de carbohidrats més elevats. Aquest fet indica que els plats amb més quantitat de carbohidrats dificulten al model la tasca d'estimació d'aquests.

4.2.2 Ingredients que presenten errors més significatius

Per la seva naturalesa, alguns ingredients poden resultar al model més difícil d'estimar els carbohidrats que contenen a partir d'una imatge que d'altres. Segons la forma, textura o color, pot variar la dificultat d'identificar-los. La quantitat de carbohidrats per unitat de volum també pot ser un factor rellevant en el rendiment del model d'estimació de carbohidrats. Per aquests motius, resulta interessant comprovar els errors que provoquen en les estimacions del model per cada ingredient. Per fer-ho, es farà una suma per cada ingredient de tots els errors absoluts de totes les estimacions de les imatges en què estiguin presents. A la figura 4.6 es mostren els 50 ingredients que més error provoquen, mostrant per a cada ingredient l'error absolut mig de les estimacions realitzades pel model.

Els ingredients que provoquen errors més grans són el bulgur, els cigrons i l'arròs salvatge.

4.2.3 Error absolut mig en funció del nombre d'ingredients en els plats

El nombre d'ingredients presents en un plat pot afegir complexitat en la tasca d'estimar els seus carbohidrats. Com més ingredients, més elements es troben presents en el plat i cal diferenciar-los per tal de realitzar una estimació precisa. L'augment de complexitat pot fer que el model no sigui capaç de portar a terme estimacions de carbohidrats tan precises com en plats més senzills. En la figura 4.7 es mostra un gràfic de barres on es representa l'error absolut mig en les estimacions dels plats classificats en funció del nombre d'ingredients que continguin.

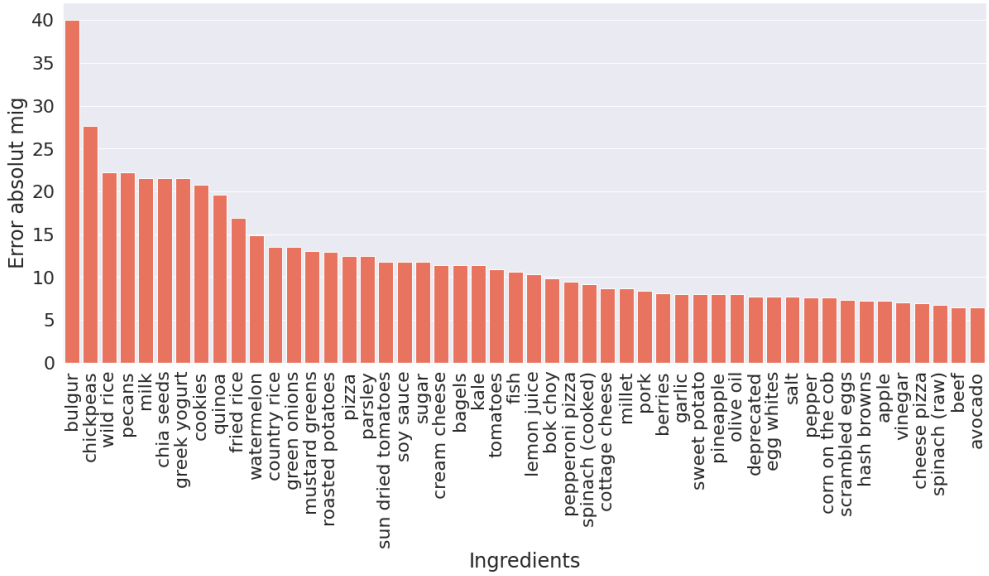

Figura 4.6: Gràfic de barres de l'error absolut mig en les estimacions de les mostres en què un determinat ingredient està present

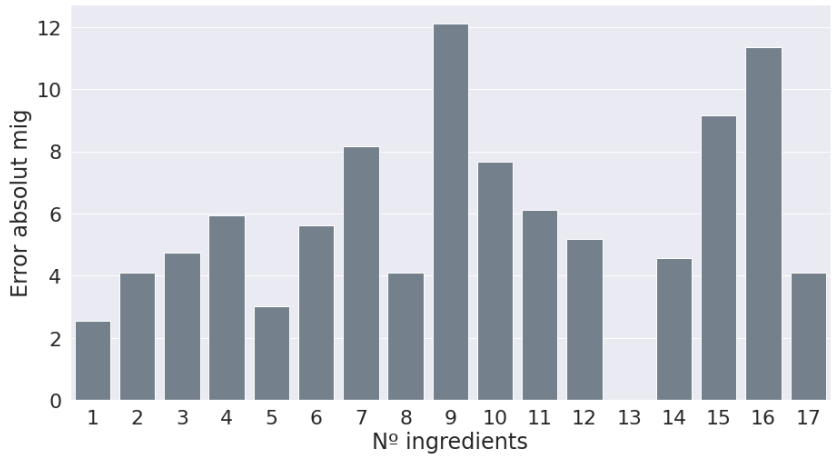

Figura 4.7: Gràfic de barres de l'error absolut mig en les estimacions en funció del nombre d'ingredients presents

Observant la distribució de l'error, no es pot apreciar una relació clara entre el valor de l'error absolut mig de les estimacions de carbohidrats amb el nombre d'ingredients presents en els plats.

4.2.4 Error absolut mig d'un ingredient en funció del nombre de plats en què apareix

En el camp de l'aprenentatge automàtic, una major varietat de mostres pot fer tendir al model a generalitzar millor i, per tant, millorar el seu rendiment en la tasca. En el cas d'aquest treball, s'entén que un major nombre mostres úniques augmenta el

coneixement o habilitat del model per estimar carbohidrats a partir d'imatges dels plats. Per tal de comprovar si aquesta hipòtesi es compleix en aquest problema, comparem l'error absolut mig de les estimacions dels plats per cada ingredient. D'aquesta manera, comprovem si el fet que un ingredient aparegui en més mostres permet al model realitzar estimacions més precises.

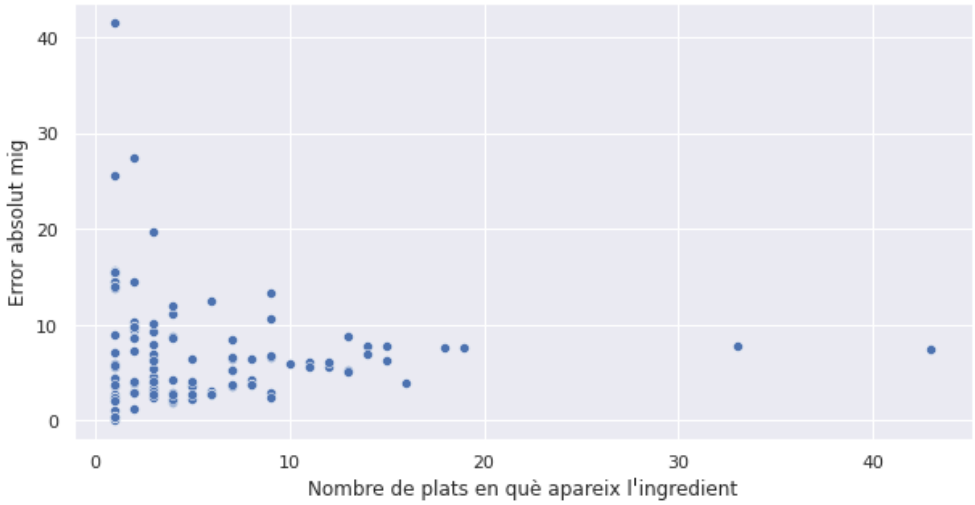

Figura 4.8: Gràfic de dispersió de l'error absolut mig per a cada ingredient en funció del nombre de plats en el qual apareix

En el gràfic no s'observa una disminució de l'error absolut mig en els ingredients que més apareixen. Això no obstant, es pot intuir una estabilització dels valors de l'error absolut mig en els ingredients els quals apareixen en més plats.

4.2.5 Error absolut de les estimacions en funció de la densitat de carbohidrats del plat

Una major quantitat de carbohidrats per unitat de volum en un aliment pot fer que resulti més difícil pel model estimar els carbohidrats que presenta. Per tal de comprovar aquesta hipòtesi, es busca la relació que hi ha entre els errors absoluts de les estimacions amb la densitat de carbohidrats. Aquesta última mesura s'obté dividint la quantitat total de carbohidrats de cada plat, en grams, per la seva respectiva massa, també en grams. Es presenta a la figura 4.9 un gràfic de dispersió amb els errors d'estimació del model, per cada plat, amb la respectiva densitat de carbohidrats.

En la figura s'observa com els plats amb una major densitat de carbohidrats, en alguns casos, presenten errors absoluts més grans. Per contra, els plats amb una quantitat de carbohidrats baixa o nul·la, tendeixen a retornar errors menors.

4.3 Anàlisi de components principals

L'anàlisi de components principals o *PCA*, en anglès, és una tècnica que redueix les dimensions d'una sèrie de dades per tal que aquestes puguin ser representades en

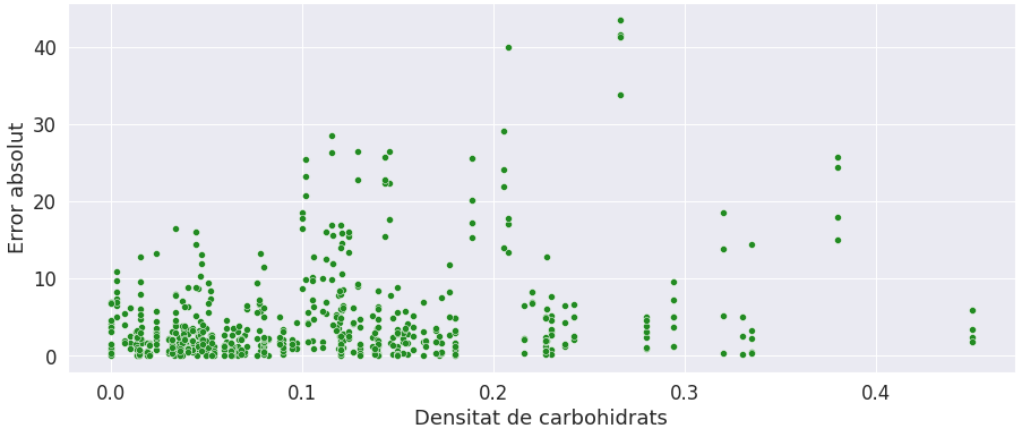

Figura 4.9: Gràfic de dispersió de l'error absolut en funció de la densitat de carbohidrats

gràfiques de menors dimensions. Aquesta tècnica es basa a generar vectors ortogonals principals els quals expliquen la variància del conjunt de dades. La primera component principal correspon a la variància més gran del conjunt de dades, la segona component principal correspon a la segona, i així successivament.

La tècnica de *PCA* pot ser útil en aquest treball per tal de representar gràficament els vectors de sortida de la xarxa neuronal, just abans de ser enviats a l'última neurona encarregada d'estimar la quantitat de carbohidrats del plat. Aquests vectors són la representació de la imatge computada pel model, i, per tant, el producte final del codificador del *transformer*. En aquest cas, aquests vectors són de 768 dimensions. Representant aquests vectors en una gràfica de dues dimensions mitjançant la tècnica de *PCA*, es pot visualitzar aproximadament com el model distingeix les diferents imatges.

A la figura 4.10 es mostren les imatges que es representaran posteriorment en una gràfica en dues dimensions, a partir dels vectors esmentats anteriorment. En total es representen 16 imatges, extretes a partir de quatre plats, dels quals per cada un se selecciona una imatge capturada per cada una de les quatre càmeres. S'han seleccionat plats que presentin característiques notablement diferenciables. El plat 1 i 4 són plats que contenen poc volum de menjar, però els ingredients que els componen tenen una quantitat de carbohidrats relativa molt diferent. El plat 2 s'ha triat perquè presenta varietat d'ingredients i el plat 3 pel gran volum de menjar que conté en comparació a la resta de plats seleccionats.

A la figura 4.11 es pot observar la representació en un espai bidimensional els components principals dels vectors de les imatges. Cada punt és una imatge i el color el relaciona amb un dels quatre plats. Es pot intuir una agrupació dels punts de cada plat en zones diferents, fet el qual demostra que la xarxa neuronal agrupa imatges semblants en zones més properes en l'espai multidimensional dels vectors de sortida.

4.4 Discussió

La naturalesa del problema proposat ofereix algunes dificultats. Estimar el volum de menjar en un plat a partir d'una imatge pot resultar difícil. En molts casos una gran

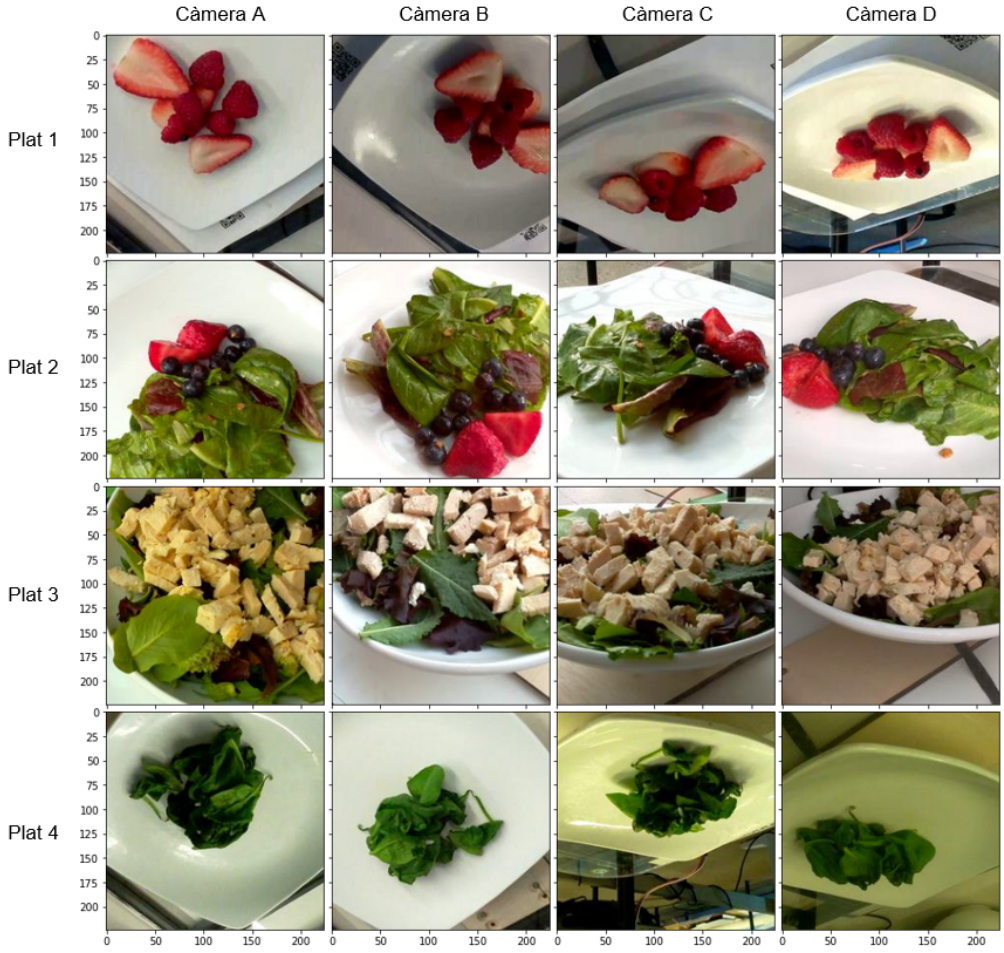

Figura 4.10: Imatges de quatre plats preses des de diferents càmeres

part d'una ració de menjar present en un plat queda tapada pel mateix menjar. Aquest fet pot provocar dificultats importants a l'hora d'estimar la quantitat de les porcions i a la vegada la quantitat de carbohidrats presents. De la mateixa manera, en un plat amb ingredients variats, es pot donar el cas que algun ingredient quedi tapat per un altre, de manera que també dificulti l'estimació de les porcions.

En algunes de les imatges del conjunt de dades una part de la ració de l'àpat fotografiat queda fora dels marges. Això es deu al fet que durant el procés de preparació de les dades, les imatges es retallen per tenir una relació d'aspecte quadrada. Les imatges preses per les càmeres tenen una amplada més gran que l'alçada, en concret tenen una relació d'aspecte de 16:9, ja que tenen una resolució de 1920x1080 píxels. Per altra banda, el model treballa amb imatges de mida 224x224 píxels. En el conjunt de dades també apareixen imatges on per culpa de la forma del recipient i l'angle on es capta la fotografia gran part de la ració de menjar no queda visible. Aquests fets poden provocar errors importants en el model, ja que no és possible fer estimacions precises sobre les parts no visibles del plat. A la figura 4.12 es mostren exemples d'imatges del conjunt de dades que presenten alguns d'aquests problemes.

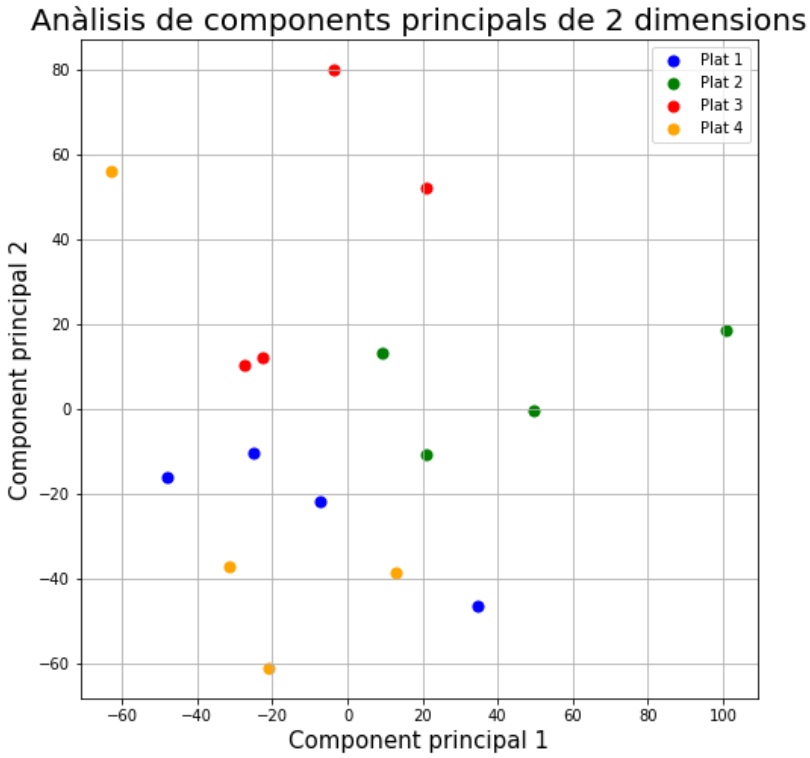

Figura 4.11: Representació dels dos components principals de les imatges

És d'esperar que oferint al model imatges de més qualitat durant l'entrenament, on es puguin veure tots els ingredients correctament, contribueix a una millora del seu rendiment.

Les característiques i ingredients dels plats de menjar, en general, poden ser molt variats. La gastronomia varia arreu del món, existint ingredients diferents en els plats de cada una. El conjunt de dades, tot i haver estat creat amb l'objectiu de ser generalista, no contempla tots els plats que poden existir en el món, ni tampoc tots els ingredients i mètodes de preparació. Per exemple, el conjunt de dades no conté cap imatge de plats de sopes. A la pràctica, és impossible crear un conjunt de dades que preparés al model per tractar amb tots els plats possibles que se li poden presentar. Per tant, cal entendre que el problema al qual s'enfronta aquest treball és molt complex i la solució que es proposa és una aproximació.

El maquinari que s'ha utilitzat per entrenar el model és el que ofereix l'aplicació en línia de Google Colab, en la seva versió pro, el qual es presenta a la següent taula 4.3.

CPU	Intel(R) Xeon(R) CPU @ 2.30GHz
GPU	1xTesla P100-PCIE-16GB , 3584 nuclis CUDA, 16GB HBM2 VRAM
RAM	12.6 GB

Taula 4.3: Maquinari que ofereix l'aplicació google colab pro

Figura 4.12: Exemples d'imatges del conjunt de dades de baixa qualitat

CAPÍTOL 5

RESUM DEL PRESSUPOST

S'inclou aquí un pressupost resumit del cost de la realització d'aquest projecte, on s'hi ha tingut en compte el temps invertit, els costos associats al servei d'emmagatzematge de dades al núvol i els costos de l'entorn virtual on s'executa el codi.

Concepte	Cost unitari	Quantitat	Cost Total
Hores treballades	30 €/h	250 h	7500 €
Emmagatzematge	10 €/mes	5 mesos	50 €
Entorn virtual	42,25 €/mes	5 mesos	211,25 €
		TOTAL	**7761,25 €**

Taula 5.1: Pressupost

El pressupost total del projecte és de SET MIL SET-CENTS SEIXANTA-UN COMA VINT-I-CINC EUROS

CAPÍTOL 6

CONCLUSIONS I FUTUR TREBALL

6.1 Conclusions

En aquest treball s'ha implementat un sistema d'estimació de carbohidrats a partir d'imatges d'àpats emplatats per facilitar el càlcul relacionat amb la dosificació d'insulina a pacients de T1DM. En l'apartat de metodologia s'ha aplicat un *transformer*, una arquitectura de xarxa neuronal artificial d'última tecnologia que promet millorar els resultats en diferents camps de la intel·ligència artificial respecte xarxes anteriors. Els resultats indiquen que aquesta arquitectura és vàlida per a resoldre la tasca proposada en el treball i suposa una millora respecte a solucions creades amb altres xarxes.

S'han aplicat diverses anàlisis al rendiment del model per tal d'avaluar les característiques de les mostres que puguin resultar més difícils pel model estimar els seus carbohidrats. S'ha demostrat gràficament, a través de l'anàlisi de components principals, com el model posiciona en una distància inferior en l'espai multidimensional del vector de sortida de la xarxa les mostres més semblants entre si. També s'ha demostrat l'existència d'imatges en el conjunt de dades de *Nutrition5k* amb una qualitat deficient.

6.2 Futur treball

Actualment, en el camp de la intel·ligència artificial, hi ha un constant desenvolupament de noves arquitectures de xarxes neuronals que milloren el rendiment de les anteriors. Com a conseqüència, resulta interessant aplicar aquestes noves xarxes en aquest treball per tal de millorar el rendiment del model. Existeixen, per exemple, variacions de la xarxa que s'ha utilitzat en aquest treball que afegeixen complexitat a través de diferents mecanismes i que augmenten la precisió en el model.

El conjunt de dades *Nutrition5k* està compost per 5.000 plats diferents. Augmentar el nombre de plats existents, fent aquest conjunt de dades més divers i complet, pot resultar en una millora del rendiment del model. El concepte de fer el model competent per processar tota mena de mostres s'anomena generalitzar.

Es pot afegir complexitat al model dotant-lo amb una nova arquitectura que processi a la vegada altres entrades d'informació útils. A continuació s'expliquen alguns exemples d'entrades que poden millorar el rendiment del model donant-li més coneixement de les característiques de la mostra. El conjunt de dades *Nutrition5k*, acompanyades amb les imatges de cada plat, ofereix una representació tridimensional d'aquest. Aquesta dada, entrada adequadament al model, pot ajudar a quantificar les porcions. Una altra dada que podria resultar ser útil és l'angle en què s'ha pres la fotografia. Per donar informació de les dimensions del plat, també es podria afegir un objecte quotidià de referència al costat del plat, per exemple una targeta de crèdit. Tanmateix, es podria aplicar una segmentació del menjar en el plat per tal d'eliminar el fons innecessari de les imatges i entrar al model de predicció de carbohidrats una imatge més polida del plat.

ANNEXOS

A.1 Codi

Tot el codi que s'ha utilitzat en aquest treball es troba disponible en el directori en línea: https://github.com/MGuso/Treball-Final-de-Grau

BIBLIOGRAFIA

[1] L. Bally, J. Dehais, C. T. Nakas, M. Anthimopoulos, M. Laimer, D. Rhyner, G. Rosenberg, T. Zueger, P. Diem, S. Mougiakakou, and C. Stettler, "Carbohydrate Estimation Supported by the GoCARB System in Individuals With Type 1 Diabetes: A Randomized Prospective Pilot Study," *Diabetes Care*, vol. 40, no. 2, pp. e6–e7, 11 2016. [Online]. Available: https://doi.org/10.2337/dc16-2173 1.1

[2] Melmed, Solmo and S. Polonsky, Kenneth and Larsen, P. Reed and Kronenberg, Henrry M., *Williams. Tratado de endocrinología*, 13th ed. Elsevier, 2017. 2.1

[3] Jameson, J. Larry, *Harrison's Endocrinology*, 3rd ed. Mc Graw Hill Education, 2013. 2.1

[4] A. D. Association, "6. glycemic targets: Standards of medical care in diabetes—2020," *Diabetes Care*, vol. 43, no. Supplement_1, pp. S66–S76, Dec 2019. [Online]. Available: https://doi.org/10.2337/dc20-S006 2.1

[5] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, M. E. Pavkov, A. Ramachandaran, S. H. Wild, S. James, W. H. Herman, P. Zhang, C. Bommer, S. Kuo, E. J. Boyko, and D. J. Magliano, "Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168822721004782 2.1

[6] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," *J Med Internet Res*, vol. 20, no. 5, p. e10775, May 2018. [Online]. Available: http://www.jmir.org/2018/5/e10775/ 2.1

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762 2.2, 3.1.3

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929 2.2, 3.1.3, 3.3.2

[9] S. Oviedo, I. Contreras, C. Quirós, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *International*

Journal of Medical Informatics, vol. 126, pp. 1–8, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505618304970 2.2

[10] M. Vasiloglou, S. Mougiakakou, E. Reber Aubry, A. Bokelmann, R. Fricker, F. Gomes, C. Guntermann, A. Meyer, D. Studerus, and Z. Stanga, "A comparative study on carbohydrate estimation: Gocarb vs. dietitians," *Nutrients*, vol. 10, 06 2018. 2.3

[11] P. Chotwanvirat, N. Hnoohom, N. Rojroongwasinkul, and W. Kriengsinyos, "Feasibility study of an automated carbohydrate estimation system using thai food images in comparison with estimation by dietitians," *Frontiers in Nutrition*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnut.2021.732449 2.3

[12] Y. Liang and J. Li, "Computer vision-based food calorie estimation: dataset, method, and experiment," *CoRR*, vol. abs/1705.07632, 2017. [Online]. Available: http://arxiv.org/abs/1705.07632 2.3

[13] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," *CoRR*, vol. abs/2103.03375, 2021. [Online]. Available: https://arxiv.org/abs/2103.03375 2.3, 3.2, 4.2

[14] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S. C. H. Hoi, and Q. Sun, "A large-scale benchmark for food image segmentation," 2021. [Online]. Available: https://arxiv.org/abs/2105.05409 2.3

[15] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574119211001131 2.3

[16] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287, Jul 2015. [Online]. Available: https://doi.org/10.1007/s11042-014-2000-8 2.3

[17] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3d reconstruction for food volume estimation," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1090–1099, 2017. 2.3

[18] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 446–461. 2.3

[19] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," 2021. 2.3

[20] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021. 2.3

[21] F. Scarselli and A. Chung Tsoi, “Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results,” *Neural Networks*, vol. 11, no. 1, pp. 15–37, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360809700097X 3.1.1

[22] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *ArXiv e-prints*, 11 2015. 3.1.2

[23] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, “Are transformers more robust than cnns?” *CoRR*, vol. abs/2111.05464, 2021. [Online]. Available: https://arxiv.org/abs/2111.05464 3.1.2

[24] T. Ridnik, E. B. Baruch, A. Noy, and L. Zelnik-Manor, “Imagenet-21k pretraining for the masses,” *CoRR*, vol. abs/2104.10972, 2021. [Online]. Available: https://arxiv.org/abs/2104.10972 3.3.2

Universitat de Girona