SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 707

# MODELI DUBOKOG UČENJA ZA RAZLIKOVANJE IZMEĐU UMJETNIČKIH SLIKA I FOTOGRAFIJA

Ivana Stilinović

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 707

# MODELI DUBOKOG UČENJA ZA RAZLIKOVANJE IZMEĐU UMJETNIČKIH SLIKA I FOTOGRAFIJA

Ivana Stilinović

Zagreb, veljača 2025.

**SVEUČILIŠTE U ZAGREBU**
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

Zagreb, 30. rujna 2024.

# DIPLOMSKI ZADATAK br. 707

| | |
|---|---|
| Pristupnica: | **Ivana Stilinović (0036525591)** |
| Studij: | Računarstvo |
| Profil: | Znanost o podacima |
| Mentor: | izv. prof. dr. sc. Alan Jović |

Zadatak: **Modeli dubokog učenja za razlikovanje između umjetničkih slika i fotografija**

Opis zadatka:

Razlikovanje umjetničkih slika od fotografija je neriješeni problem u računalnom vidu. Čak je i čovjeku neki put teško procijeniti radi li se o slici ili o fotografiji. U literaturi su predložene različite tehnike za ovo razlikovanje, od onih zasnovanih na bojama i kompleksnosti elemenata na slikama do metoda zasnovanih na strojnom učenju. U ovom diplomskom radu potrebno je isprobati odgovarajuće modele dubokog učenja koje ne zahtijevaju ekstrakciju ekspertnih značajki sa slika, primjerice one temeljene na konvolucijskim slojevima i one temeljene na mehanizmu pažnje te ponuditi najtočniji model za rješenje ovog problema. Pritom treba pronaći odgovarajuće skupove podataka (primjerice, s platforme Kaggle) i isprobati u srodnoj literaturi predložene modifikacije modela dubokog učenja. Programski kod je potrebno komentirati, a najvažnije isječke koda dokumentirati u tekstu diplomskog rada. Potrebno je vrednovati modele i usporediti rezultate s onima objavljenima u literaturi. Za potrebe provođenja vrednovanja moguće je poslužiti se bilo kojim dostupnim sklopovskim resursima (npr. Google Colab).

Rok za predaju rada: 14. veljače 2025.

*Zahvaljujem svojoj obitelji i prijateljima na podršci i motivaciji tijekom cijelog studija. Također, zahvaljujem svom mentoru izv. prof. dr. sc. Alanu Joviću na korisnim savjetima i smjernicama tijekom izrade ovog rada.*

# Sadržaj

# 1. Uvod

Razlikovanje između umjetničkih slika od fotografija predstavlja izazovan i još uvijek neriješen problem u području računalnog vida. Na prvi pogled može se činiti jednostavnim, no čak ni ljudi ponekad ne mogu lako odrediti radi li se o stvarnoj fotografiji ili o umjetničkoj interpretaciji. Takva klasifikacija ima široku primjenu, od automatskog prepoznavanja i sortiranja slika u digitalnim zbirkama i muzejima do stilskog prijenosa u umjetničkoj obradi, gdje je važno sačuvati vizualni identitet slike i spriječiti neželjene transformacije fotografija u slikarske stilove.

Metode koje su do nedavno razvijane i opisane u literaturi temeljile su se na ručno dizajniranim značajkama kao što su boja, tekstura i rubovi. Iako su ove metode omogućile osnovnu klasifikaciju, bile su ograničene jer nisu mogle prepoznati složenije vizualne obrasce karakteristične za različite umjetničke stilove. Razvojem modela dubokog učenja, osobito konvolucijskih neuronskih mreža (CNN), otvorile su se nove mogućnosti za automatsko prepoznavanje složenih obrazaca u slikama, što omogućuje klasifikaciju bez potrebe za ručnim definiranjem značajki.

U ovom radu analizirani su unaprijed naučeni modeli dubokog učenja, uključujući VGG16, ResNet50 i EfficientNetB0, kako bi se ispitala njihova učinkovitost u razlikovanju umjetničkih slika i fotografija. Osim toga, istražen je mehanizam pažnje CBAM (*Convolutional Block Attention Module*), koji se može dodatno integrirati u modele za poboljšanje prepoznavanja ključnih vizualnih značajki. Dobiveni rezultati uspoređeni su s postojećim istraživanjima u ovom području kako bi se procijenila učinkovitost modela i identificirali daljnji koraci za poboljšanje klasifikacije.

# 2. Teorijske osnove i postojeći modeli

## 2.1. Prethodni pristupi

Raniji pokušaji razlikovanja umjetničkih slika od fotografija temeljili su se na ručno izrađenim značajkama poput boje, teksture, rubova i složenosti slike. Takvi pristupi omogućili su inicijalno razumijevanje ovog problema, no ograničenja su se pokazala u njihovoj prilagodljivosti različitim stilovima slika i njihovoj osjetljivosti na kvalitetu ulaznih podataka.

Prvi značajan rad u ovom području predložili su Cutzu i suradnici [1], gdje su analizirali četiri skupine značajki koje su omogućile razlikovanje umjetničkih slika od fotografija. Među najvažnijim značajkama bile su rubovi boje, prostorna varijacija boje, broj jedinstvenih boja i zasićenost piksela. Fotografije su pokazivale veću prisutnost rubova intenziteta i manje čistih rubova boje, dok su slike sadržavale širu paletu boja i veću zasićenost piksela. Jedan od ključnih pristupa bio je korištenje RGBXY prostora, koji proširuje standardni RGB model boja dodavanjem koordinata $x$ i $y$ koje predstavljaju položaj svakog piksela, omogućujući analizu prostorne distribucije boja unutar slike. Uz to, korištene su značajke teksture dobivene s pomoću Gaborovog filtra, koji mjeri frekvencije i orijentacije tekstura u slici, čime se dodatno razdvajaju umjetnička djela od fotografija. Modeli naučeni na ovim značajkama postigli su točnost između 72% i 81%, ovisno o odabranoj skupini značajki. Kombinacija svih značajki rezultirala je točnošću od 93%, što je pokazalo potencijal ovakvih pristupa za rješavanje problema klasifikacije slika.

Drugi značajan rad predstavili su Carballal i suradnici [2], koji su predložili korištenje složenosti slike kao glavnog kriterija za razlikovanje umjetničkih slika od fotografija. Ovaj pristup uključivao je procjenu pogreške kompresije, primjenu Zipfovog zakona i iz-

računavanje fraktalne dimenzije slike. Zipfov zakon, koji opisuje učestalost elemenata u velikim skupovima podataka, korišten je za analizu distribucije intenziteta piksela i razlika između susjednih piksela [3]. Nadalje, korišteni su detektori rubova u slici, poput Sobelovog i Cannyjevog filtra, za analizu rubova slika. Sobelov filtar identificira osnovne rubove slike mjerenjem promjena u intenzitetu piksela u horizontalnom i vertikalnom smjeru, dok Cannyjev filtar koristi složeniji postupak koji uključuje smanjenje šuma, precizniju detekciju rubova i filtriranje nevažnih detalja [4]. Modeli naučeni na ovim značajkama postigli su točnost od 94,82%, što je nadmašilo većinu prethodnih pristupa.

Unatoč razlikama u metodologiji, oba pristupa oslanjaju se na vizualne značajke koje jasno razlikuju umjetnička djela od fotografija. Dok je pristup Cutzua i suradnika [1] bio fokusiran na specifične vizualne karakteristike poput boje i teksture, Carballal i suradnici [2] proširili su ovu analizu uključivanjem univerzalnih mjera složenosti. Prednosti Carballalovog pristupa uključuju njegovu prilagodljivost različitim stilovima i veću robusnost na promjene u kvaliteti ulaznih podataka, no oba pristupa dijele ograničenje u potrebi za ručno definiranjem značajki, što smanjuje njihovu primjenjivost na širok raspon problema.

Iako su ove metode bile učinkovite, imale su ograničenja u prepoznavanju složenih stilskih razlika. Ručno definirane značajke često su osjetljive na kvalitetu slike i nisu skalabilne za velike i raznolike skupove podataka.

Ovi pristupi pružaju važnu osnovu za daljnji razvoj metoda za klasifikaciju umjetničkih slika i fotografija, no s obzirom na njihova ograničenja, moderne metode sve se više oslanjaju na duboko učenje kako bi automatski prepoznale složene obrasce u slikama i nadvladale ove izazove.

## 2.2. Duboko učenje za analizu slika

Duboko učenje značajno je unaprijedilo analizu slika, omogućujući automatsko prepoznavanje složenih obrazaca i izdvajanje ključnih značajki. CNN-ovi su posebno učinkoviti jer koriste konvolucijske slojeve za prepoznavanje lokalnih značajki, poput rubova i tekstura, dok završni slojevi klasificiraju slike na temelju tih informacija. Ova arhitektura eliminira potrebu za ručnim definiranjem značajki i omogućuje prilagodbu različitim za-

dacima. Detaljan opis CNN-ova istraženih u ovom radu, uključujući VGG16 i ResNet50, bit će predstavljen u poglavlju 3.2.

Primjena CNN-ova u klasifikaciji umjetničkih djela i fotografija već je pokazala izvrsne rezultate. Na primjer, Lopez-Rubio i suradnici [5] koristili su unaprijed naučeni VGG16 model kako bi ostvarili preciznost koja nadilazi prethodne metode temeljene na ručno definiranim značajkama. Zamrzavanjem početnih slojeva mreže i prilagodbom završnih slojeva za zadatak binarne klasifikacije, postigli su izvanredne performanse. Prilagođeni VGG16 model postigao je visoku točnost, veću od 99% na skupovima podataka ImageNet i Kaggle Painter by Numbers, dok je na slikama korištenim u istraživanju Carballala i suradnika [2] ostvario preciznost od 94,2%. Ovakav pristup demonstrira fleksibilnost CNN-ova u prilagodbi različitim zadacima analize slika.

Nove metode integriraju koncepte mehanizama pažnje kako bi se poboljšala učinkovitost modela. CBAM proširuje kanalnu pažnju dodavanjem prostorne komponente, što pomaže modelu da prepozna ključne regije u slici. Kada se kombinira s modernim arhitekturama poput EfficientNetB0, omogućuje poboljšano prepoznavanje vizualnih uzoraka i klasifikaciju slika. Takvi modeli obično koriste potpuno povezane slojeve sa sigmoidnom aktivacijskom funkcijom za završnu klasifikaciju, dok se binarna unakrsna entropija primjenjuje kao funkcija gubitka za optimizaciju modela.

S obzirom na ograničenja ranijih pristupa temeljenim na ručno definiranim značajkama, ovaj rad koristi suvremene modele dubokog učenja, uključujući CNN modele i mehanizme pažnje, kako bi se poboljšala sposobnost razlikovanja umjetničkih slika od fotografija. U nastavku je opisana metodologija implementacije i evaluacije odabranih modela.

# 3. Metodologija

## 3.1. Skup podataka

### 3.1.1. Izvor podataka

Za potrebe ovog istraživanja korištena su dva skupa podataka kako bi se omogućilo učinkovito vrednovanje razlikovanja umjetničkih djela od fotografija. Prvi skup, koji obuhvaća umjetnička djela, preuzet je s platforme Kaggle pod nazivom *Best artworks of all time* [6]. Ovaj skup podataka sadrži slike poznatih umjetnika poput Leonarda da Vincija, Vincenta van Gogha i Pabla Picassa. Slike su razvrstane prema autorima i obuhvaćaju raznolike umjetničke stilove i razdoblja, uključujući renesansu, impresionizam, ekspresionizam i modernu umjetnost. Ovaj skup podataka odabran je zbog raznolikosti i kvalitete, što omogućava modelu da nauči širok spektar karakteristika umjetničkih djela, poput jedinstvenih boja, tekstura i stilova.

Drugi skup podataka koji je korišten u istraživanju dolazi iz skupa podataka COCO (*Common Objects in Context*) [7], koji je poznat po svojoj raznolikosti i kvaliteti fotografija. Skup podataka COCO obuhvaća slike stvarnih scena koje uključuju ljude, objekte i prirodne pejzaže, a često se koristi za zadatke poput detekcije objekata, segmentacije i klasifikacije.

Kombiniranjem ova dva skupa podataka omogućuje se izgradnja modela za binarnu klasifikaciju s ciljem preciznog razlikovanja umjetničkih djela od fotografija. Ovaj pristup osigurava raznoliku i uravnoteženu bazu za učenje modela, čime se poboljšava njegova sposobnost generalizacije i osiguravaju realistični uvjeti za procjenu performansi.

### 3.1.2. Podjela podataka

Za podjelu podataka u ovom istraživanju, podaci su razvrstani u tri disjunktna skupa: učenje, validacija i testiranje. Ukupno je korišteno 8000 slika za učenje, 2000 slika za validaciju i 2000 slika za testiranje. Svaka klasa, odnosno umjetničke slike i fotografije, ravnomjerno su zastupljene u svakom skupu, što znači da svaki od njih sadrži 4000 slika po klasi za učenje, 1000 slika po klasi za validaciju i 1000 slika po klasi za testiranje. Ova podjela osigurava uravnoteženu distribuciju podataka te omogućuje pouzdanu procjenu performansi modela.

Slike umjetničkih djela odabrane su iz skupa podataka *Best artworks of all time* [6], dok su fotografije odabrane iz skupa podataka COCO [7]. Podaci su nasumično podijeljeni unutar svakog skupa kako bi se osigurala raznolikost i ravnoteža.

Kod podjele podataka, cilj je bio osigurati da modeli imaju dovoljno podataka za učenje, dok validacijski skup omogućuje optimizaciju hiperparametara i prilagodbu modela. Testni skup koristi se isključivo za procjenu konačne performanse modela na nepoznatim podacima. Podjela podataka na ovaj način ključna je za sprječavanje pristranosti i osiguranje generalizacijske sposobnosti modela.

Ovaj način podjele podataka odabran je jer pruža uravnotežen i dosljedan okvir za usporedbu rezultata našeg istraživanja s postojećim radovima, dok istovremeno omogućuje nepristrano vrednovanje modela u realnim uvjetima.

### 3.1.3. Priprema podataka za modeliranje

Priprema podataka ključan je korak u procesu učenja modela dubokog učenja jer osigurava da model može učinkovito učiti iz ulaznih slika i postići dobre performanse na nepoznatim podacima. Kako bi se osigurala konzistentnost u obradi, sve slike su transformirane na uniformne dimenzije $224 \times 224$ piksela, što odgovara standardnim ulaznim zahtjevima većine konvolucijskih neuronskih mreža. Osim skaliranja dimenzija, vrijednosti piksela normirane su korištenjem funkcije *Rescaling*, pri čemu su sve vrijednosti transformirane u raspon $[0, 1]$. Ova normalizacija smanjuje varijacije u intenzitetu piksela i osigurava stabilnije učenje modela, osobito pri korištenju optimizatora koji su osjetljivi na različite raspone vrijednosti značajki.

Kako bi se poboljšala sposobnost generalizacije modela i smanjio rizik od prenaučenosti, primijenjene su različite tehnike poboljšavanja (augmentacije) podataka. Slike su podvrgnute nasumičnoj rotaciji do $\pm 10\%$ originalnog kuta, čime se simuliraju različite orijentacije objekata na slici. Osim rotacije, korišteno je nasumično zumiranje u rasponu od -20% do +20% visine i širine slike, što omogućuje modelu prepoznavanje vizualnih uzoraka neovisno o udaljenosti od kamere. Također, slike su horizontalno preslikavane kako bi se povećala varijabilnost podataka reflektiranjem slika. Uz to, primijenjena je nasumična translacija cijele slike do 20% visine i širine slike, čime se simuliraju varijacije u položaju objekata.

Ove transformacije povećavaju raznolikost podataka i pomažu modelu da postane otporniji na promjene u perspektivi, osvjetljenju i orijentaciji slika. Implementacija ovih postupaka slijedi preporuke iz literature [5] kako bi se osiguralo da model ne nauči prepoznati samo specifične karakteristike pojedinih slika, već razvije sposobnost izdvajanja općih vizualnih obrazaca karakterističnih za umjetničke slike i fotografije.

### 3.1.4. Metrike vrednovanja

Za vrednovanje performanci modela korištene su dvije ključne metrike: točnost i F1-mjera. Ove metrike pružaju sveobuhvatan uvid u sposobnost modela da točno klasificira slike kao umjetnička djela ili fotografije.

Točnost mjeri udio ispravno klasificiranih uzoraka u odnosu na ukupan broj uzoraka i definira se kao:

$$\text{Točnost} = \frac{\text{Broj ispravno klasificiranih primjeraka}}{\text{Ukupan broj primjeraka}} \tag{3.1}$$

Budući da je korišten uravnoteženi skup podataka, točnost je prikladna metrika jer omogućuje jednostavnu procjenu ukupne uspješnosti modela. Ova metrika također osigurava usporedivost s drugim istraživanjima koja su koristila isti skup podataka i metodologiju. F1-mjera je harmonijska sredina preciznosti i osjetljivosti te se računa prema formuli:

$$\text{F1-mjera} = 2 \cdot \frac{\text{Preciznost} \cdot \text{Osjetljivost}}{\text{Preciznost} + \text{Osjetljivost}} \tag{3.2}$$

Preciznost i osjetljivost definirane su kao:

$$\text{Preciznost} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Osjetljivost} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

Gdje TP predstavlja broj točno predviđenih pozitivnih uzoraka, FP broj negativnih uzoraka pogrešno klasificiranih kao pozitivne, a FN broj pozitivnih uzoraka pogrešno klasificiranih kao negativne.

F1-mjera pruža ravnotežu između preciznosti i osjetljivosti, čime je posebno korisna u slučajevima kada postoji neravnoteža između klasa ili kada model pokazuje sklonost ka jednoj klasi. U ovom istraživanju omogućava detaljniju analizu performanci modela u razlikovanju umjetničkih djela od fotografija, osobito u kontekstu pogrešnih klasifikacija.

Kombinacija točnosti i F1-mjere omogućuje sveobuhvatnu procjenu performanci modela, pri čemu točnost pruža općeniti uvid u uspješnost klasifikacije, dok F1-mjera osigurava detaljan uvid u ravnotežu između prepoznavanja pravih pozitivnih primjera i izbjegavanja pogrešnih predikcija. Ovakav pristup evaluaciji osigurava pouzdane i usporedive rezultate.

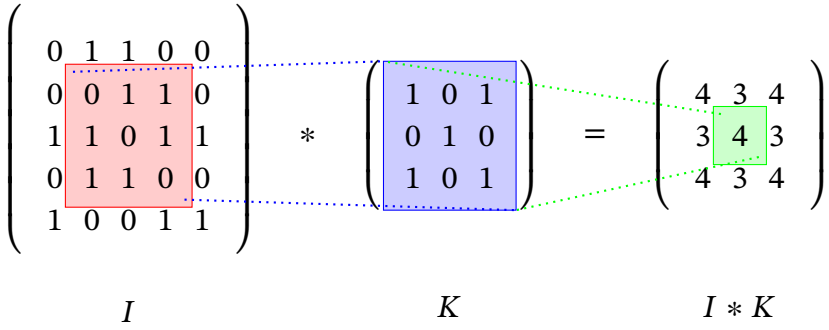
## 3.2. Modeli dubokog učenja

### 3.2.1. Konvolucijske neuronske mreže

Konvolucijske neuronske mreže (CNN) predstavljaju temeljni pristup u analizi vizualnih podataka jer omogućuju automatsko izdvajanje značajki slika bez potrebe za ručnim definiranjem značajki. CNN mreže koriste konvolucijske slojeve za prepoznavanje značajki kao što su rubovi, uzorci i teksture. Osnovni princip rada CNN-a temelji se na konvolucijskoj operaciji, gdje se primjenjuje konvolucijski filtar na ulaznu sliku radi izdvajanja značajki.

$$(I * K)(x, y) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} I(x - m, y - n) K(m, n) \tag{3.4}$$

gdje je $I(x, y)$ ulazna matrica (slika), $K(m, n)$ konvolucijski filtar, a rezultat je nova izlazna matrica $I * K$, koja sadrži prepoznate značajke. Na Slici 3.1. ilustrirana je ova operacija.




**Slika 3.1.** Prikaz množenja matrica filtrom u CNN-u [8].

Konvolucijska operacija prikazana na slici izračunava lokalne značajke ulazne matrice koristeći filtar veličine $3 \times 3$, koji se pomiče preko matrice i računa rezultirajuću matricu $I * K$. Crveno označeno područje u ulaznoj matrici pokazuje trenutačnu lokaciju filtra, dok su plava i zelena područja odgovarajuće veze između filtra i izlazne vrijednosti.

Ovaj proces omogućuje prepoznavanje ključnih značajki slike, poput rubova i tekstura, čime CNN postaje iznimno moćan alat za zadatke poput klasifikacije i detekcije objekata.

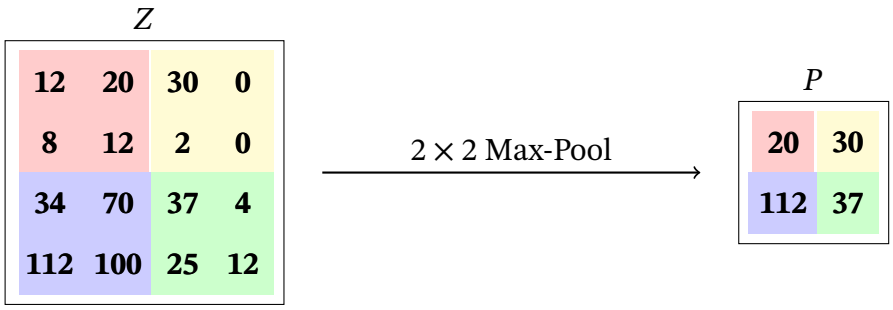
Slojevi sažimanja, poput maksimalnog sažimanja, smanjuju dimenzionalnost značajki dok istovremeno zadržavaju ključne informacije, čime mreža postaje otpornija na pomake, rotacije i varijacije u teksturi unutar slike. Najčešće korišten pristup je maksimalno sažimanje, pri čemu se u zadanom prozoru odabire najveća vrijednost piksela, kao što je prikazano na Slici 3.2.:

$$O(x, y) = \max_{(m,n)\in P} I(x + m, y + n) \tag{3.5}$$

gdje je $O(x, y)$ izlazna vrijednost nakon operacije maksimalnog sažimanja, $I(x+m, y+n)$ predstavlja vrijednosti piksela unutar ulazne slike, dok je $P$ područje prozora sažimanja značajki veličine $k \times k$.

Ovaj proces smanjuje broj parametara mreže, sprječava prenaučenost i omogućuje

bolju generalizaciju na nove podatke. Kombinacija konvolucijskih i slojeva sažimanja omogućuje CNN-ovima da uče hijerarhijske reprezentacije slike od osnovnih značajki poput rubova i tekstura do složenijih struktura poput oblika i objekata.




**Slika 3.2.** Primjer $2 \times 2$ maksimalnog sažimanja [8].

Osim konvolucijskih i slojeva sažimanja, CNN modeli na kraju koriste potpuno povezane slojeve za donošenje konačne klasifikacije. Ovi slojevi služe za kombiniranje značajki iz prethodnih slojeva te ih pretvaraju u distribuciju vjerojatnosti klasa. Na primjer, posljednji sloj CNN-a često koristi aktivacijsku funkciju softmax kako bi predvidio pripadnost slike određenoj klasi, čime se postiže završni korak u procesu analize.

U okviru ovog rada, inicijalni eksperiment proveden je osnovnim CNN modelom, koji je poslužio kao referentna točka (engl. *baseline*) za usporedbu s naprednijim arhitekturama. Cilj je bio analizirati njihovu uspješnost i utvrditi područja za daljnje optimizacije. Iako osnovni modeli CNN-a omogućuju automatsko izdvajanje značajki, njihove performanse mogu se dodatno poboljšati korištenjem unaprijed naučenih arhitektura. Jedan od široko korištenih modela u tu svrhu je VGG16, koji donosi dublju i precizniju analizu vizualnih uzoraka.

### 3.2.2. Mreža VGG16

VGG16 je duboka konvolucijska arhitektura koja se sastoji od 16 slojeva s prilagodljivim težinama, uključujući 13 konvolucijskih slojeva i 3 potpuno povezana sloja. Arhitektura je prvi put predstavljena u radu Simonyana i Zissermana [9] kao dio istraživanja o poboljšanju klasifikacije slika na skupu podataka ImageNet. Model je osmišljen tako da koristi male konvolucijske jezgre veličine $3 \times 3$ sa pomakom od 1, što omogućuje povećanje kapaciteta modela bez značajnog povećanja broja parametara.

Potpuno povezani slojevi na kraju modela odgovorni su za kombiniranje naučenih značajki u konačni izlaz klasifikacije. Prva dva sloja sadrže po 4096 neurona, dok posljednji sloj, koji koristi aktivaciju ImageNet , klasificira slike u određeni broj klasa, ovisno o zadatku. Na primjer, za klasifikaciju na skupu podataka ImageNet, izlazni sloj sadrži 1000 neurona, od kojih svaki odgovara jednoj klasi. Prednost ove arhitekture je jednostavnost dizajna, gdje se slojevi modularno ponavljaju, dok upotreba malih $3 \times 3$ filtara omogućuje modelu da uči složene značajke kroz veću dubinu.

Arhitektura modela prikazana je u Tablici: 3.1.

Tablica 3.1. Arhitektura modela VGG16

| **Sloj** | **Filtri** | **Dimenzije izlaza** |
|---|---|---|
| Ulaz | - | $224 \times 224 \times 3$ |
| Konvolucija | $3 \times 3$ | $224 \times 224 \times 64$ |
| Konvolucija | $3 \times 3$ | $224 \times 224 \times 64$ |
| Maksimalno sažimanje | $2 \times 2$ | $112 \times 112 \times 64$ |
| ... | ... | ... |
| Potpuno povezani | - | $1 \times 1 \times 4096$ |
| Softmax | - | $1 \times 1 \times 1000$ |

Glavno ograničenje VGG16 je veliki broj parametara i visoka računalna složenost, što može otežati njegovu primjenu u scenarijima s ograničenim resursima. Zbog toga se u novijim istraživanjima fokus stavlja na optimirane arhitekture koje nude bolju ravnotežu između preciznosti i učinkovitosti [10].

Prema istraživanju [5], VGG16 je uspješno korišten u zadacima klasifikacije slika, postižući značajna poboljšanja u odnosu na tradicionalne pristupe ručne ekstrakcije značajki. Njegova duboka arhitektura omogućuje učenje složenih vizualnih obrazaca, ali uz visoke računalne zahtjeve, što ga čini manje pogodnim za resursno ograničene sustave.

Kako bi se omogućilo učenje još dubljih mreža i poboljšala sposobnost prepoznavanja složenih vizualnih uzoraka, razvijene su arhitekture koje uvode mehanizme za učinkoviti prijenos informacija kroz slojeve. Jedna od najvažnijih arhitektura koje su značajno unaprijedile klasične modele CNN-ova je ResNet, koja omogućuje dublju analizu poda-

taka bez degradacije performansi.

### 3.2.3. Rezidualna neuronska mreža

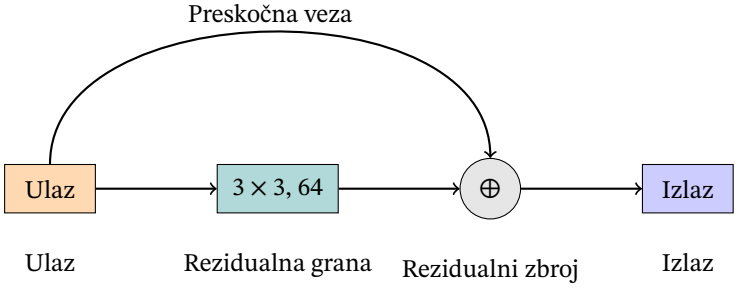
Rezidualna neuronska mreža (ResNet) predstavlja jedan od najvažnijih doprinosa dubokom učenju, koji je značajno unaprijedio performanse modela u klasifikaciji i drugim zadacima računalnog vida. Predstavljen je u radu Hea i suradnika [11], a njegova glavna inovacija leži u uvođenju rezidualnih blokova, koji omogućuju učinkovit prijenos informacija kroz slojeve mreže.

Glavna ideja arhitekture ResNet je dodavanje preskočnih veza (engl. *skip-connections*) između slojeva kako bi se omogućilo da informacije prolaze neometano, čak i kroz vrlo duboke mreže. Matematički, izlaz rezidualnog bloka može se opisati sljedećim izrazom:

$$y = F(x, \{W_i\}) + x \tag{3.6}$$

gdje je $x$ ulazni tenzor, $F(x, \{W_i\})$ niz transformacija (npr. konvolucija i aktivacijske funkcije), a $x$ se direktno dodaje rezultatu transformacija. Ovim pristupom izbjegava se problem eksplozije ili nestanka gradijenata u dubokim mrežama.

Jedna od popularnijih varijanti je ResNet50, koja sadrži 50 slojeva. Arhitektura se temelji na kombinaciji jednostavnih rezidualnih blokova i slojeva koji koriste $1 \times 1$ konvolucije. Ove konvolucije igraju ključnu ulogu u smanjenju dimenzionalnosti prije primjene $3 \times 3$ konvolucija, a zatim povećavaju dimenzionalnost natrag kako bi omogućile učinkovito učenje dubljih mreža. Struktura rezidualnog bloka, uključujući preskočne veze, ilustrirana je na Slici 3.3., koja prikazuje način kombiniranja ulaznih značajki s rezultatima rezidualne grane.




**Slika 3.3.** Rezidualni blok ResNet arhitekture s preskočnom vezom [8].

Kako bi se stabilizirala propagacija gradijenata i poboljšala učinkovitost učenja, ResNet50 koristi normalizaciju po grupi (engl. *batch normalization*), koja osigurava konzistentnu distribuciju podataka unutar svakog sloja tijekom učenja. Time se smanjuje osjetljivost modela na promjene u inicijalizaciji težina i ubrzava konvergencija. Uz normalizaciju po grupi, model koristi aktivacijsku funkciju ReLU kako bi uveo nelinearnost i omogućio bolje učenje složenih značajki.

Usporedno s arhitekturom VGG16, ResNet50 pruža poboljšanu generalizaciju i manji rizik od prenaučenosti, dok u isto vrijeme zadržava visoku razlučivost značajki. Međutim, zbog veće dubine mreže, računalni zahtjevi modela ostaju visoki, što može predstavljati ograničenje u primjeni na uređajima s manjim resursima.

Kako bi se dodatno poboljšala klasifikacija, u nastavku se istražuje EfficientNet te primjena CBAM-a u njegovoj arhitekturi.

### 3.2.4. Mreža EffiecientNet

EfficientNet je moderna konvolucijska arhitektura razvijena s ciljem postizanja visoke točnosti klasifikacije slika uz optimizaciju računalnih zahtjeva. Predstavili su je istraživači Google Braina 2019. godine [12], a glavna inovacija ovog modela leži u kompleksnom skaliranju, kojim se istovremeno optimiraju tri ključne dimenzije neuronske mreže: dubina, širina i rezolucija ulazne slike.

Dok su ranije arhitekture poboljšavale performanse povećavanjem broja slojeva, EfficientNet uvodi princip balansiranog skaliranja, što znači da se svi ključni aspekti mreže povećavaju proporcionalno, umjesto da se povećava samo jedan aspekt. Ovaj proces može se opisati sljedećim izrazima:

$$\text{dubina} = \alpha^d, \quad \text{širina} = \beta^d, \quad \text{rezolucija} = \gamma^d \tag{3.7}$$

gdje je $d$ koeficijent skaliranja, dok su $\alpha$, $\beta$ i $\gamma$ hiperparametri koji definiraju kako se mreža povećava po tim dimenzijama.

Osnovu arhitekture EfficientNet-a čine blokovi MBConv (*Mobile Inverted Bottleneck Convolution Blocks*), koji su optimirani za učinkovitu obradu značajki uz smanjenje broja

parametara. Ovi blokovi kombiniraju ekspanzijske konvolucije, dubinske razdvojene konvolucije i točkaste konvolucije kako bi povećali izražajnu moć modela uz smanjene računalne zahtjeve. Unutar svakog bloka MBConv nalazi se i blok *Squeeze-and-Excitation* (SE), koji omogućuje modelu da dinamički prilagodi značajnost pojedinih kanala slike. Ovaj mehanizam pažnje uključuje dva glavna koraka: sažimanje, koje provodi globalno prosječno sažimanje značajki po kanalima, i pobudu, koja prilagođava značajnost svakog kanala pomoću potpuno povezanih slojeva i sigmoidne aktivacije. Time se ključne značajke slike pojačavaju, dok se manje važne informacije smanjuju. Matematički, izlaz SE bloka definiran je kao:

$$s_c = \sigma(W_2\delta(W_1 z)) \tag{3.8}$$

gdje su $W_1$ i $W_2$ matrice težina, $\delta$ je aktivacija ReLU, a $\sigma$ je sigmoidna funkcija. Konačno, prilagođene značajke množe se s ulaznim tenzorom, čime se modelu omogućuje da poboljša ključne značajke slike:

$$\hat{X}_c = s_c \cdot X_c, \tag{3.9}$$

gdje $\hat{X}_c$ predstavlja prilagođene značajke za svaki kanal $c$. Ovaj korak omogućuje jačanje ključnih značajki slike dok se manje važne informacije prigušuju.

Zahvaljujući ovom ugrađenom mehanizmu pažnje, EfficientNet je sposoban automatski fokusirati pažnju na relevantne značajke slike, poboljšavajući prepoznavanje vizualnih uzoraka. Kombinacija skaliranja mreže i pažnje putem SE blokova omogućuje ovom modelu da postigne vrhunske rezultate u klasifikaciji slika uz značajno smanjenje broja parametara u usporedbi s prethodnim arhitekturama poput VGG16 i ResNet50.

### CBAM – mehanizam dodatne pažnje

Iako EfficientNet već koristi kanalnu pažnju putem SE blokova, moguće je dodatno poboljšati model uvođenjem prostorne pažnje, koja omogućuje bolju lokalizaciju ključnih regija slike. Convolutional Block Attention Module (CBAM) je mehanizam pažnje koji kombinira kanalnu i prostornu pažnju, čime poboljšava izražajnu moć modela uz minimalno povećanje računalnih zahtjeva [13].

CBAM se sastoji od dvije uzastopne komponente: kanalne pažnje, koja modelu omogućuje da identificira najvažnije kanale značajki, i prostorne pažnje, koja model usmjerava na relevantne dijelove slike. Kanalna pažnja koristi kombinaciju globalnog prosječnog i maksimalnog sažimanja, čime se dobiva reprezentacija značajnosti pojedinih kanala unutar slike. Matematički, kanalna pažnja definirana je izrazom:

$$M_c(F) = \sigma\left(W_1\left(\text{ReLU}\left(W_0 \cdot \text{AvgPool}(F)\right)\right) + W_1\left(\text{ReLU}\left(W_0 \cdot \text{MaxPool}(F)\right)\right)\right) \quad (3.10)$$

gdje su $M_c$ maske kanalne pažnje, $F$ ulazni tenzor značajki, AvgPool i MaxPool globalno prosječno i maksimalno sažimanje. Potpuno povezani slojevi sa sigmoidalnom aktivacijom ($W_0$ i $W_1$) dodaju nelinearnost i omogućuju prilagodbu značajnosti svakog kanala.

Nakon kanalne pažnje, CBAM primjenjuje prostornu pažnju, koja dodatno usmjerava model na ključne regije slike. Za razliku od kanalne pažnje, koja obrađuje informacije na razini kanala, prostorna pažnja analizira raspodjelu značajki u različitim dijelovima slike. Prostorna pažnja koristi konvolucijski sloj za identifikaciju značajnih regija slike, a matematički je definirana izrazom:

$$M_s(F) = \sigma\left(\text{Conv2D}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \quad (3.11)$$

gdje su $M_s$ prostorna maska, Conv2D konvolucijski sloj s jezgrom veličine $7 \times 7$, a $[\cdot;\cdot]$ operator spajanja tenzora.

Konačan izlaz modula CBAM kombinira kanalnu i prostornu pažnju serijski, što omogućuje modelu da prvo identificira ključne kanale značajki, a zatim da usmjeri pažnju na relevantne dijelove slike:

$$F' = M_c(F) \odot F, \quad F'' = M_s(F') \odot F' \quad (3.12)$$

gdje su $F'$ izlaz nakon kanalne pažnje, $F''$ izlaz nakon prostorne pažnje, a $\odot$ označava množenje odgovarajućih elemenata dvaju tenzora iste dimenzije.

Modul CBAM može se jednostavno integrirati u postojeće konvolucijske mreže kako bi se dodatno poboljšala njihova sposobnost prepoznavanja bitnih značajki slike. Za raz-

liku od blokova SE, koje EfficientNet već koristi, CBAM dodaje prostornu pažnju, omogućujući mreži da prepozna važne regije unutar slike, a ne samo značajne kanale.

U ovom radu CBAM je dodan na izlaz EfficientNetB0 modela kako bi se istražilo može li dodatna prostorna pažnja poboljšati sposobnost modela da razlikuje umjetničke slike od fotografija. S obzirom na to da EfficientNet već ima ugrađenu kanalnu pažnju kroz SE blokove, dodatak CBAM-a predstavlja proširenje postojećeg mehanizma pažnje s ciljem poboljšanja performansi u prepoznavanju ključnih vizualnih obrazaca.

## 3.3. Detalji implementacije

Implementacija modela dubokog učenja provedena je na platformi Google Colab, koja omogućuje korištenje grafičkih procesora za ubrzanje učenja. U ovom radu korišten je NVIDIA T4 GPU s 16 GB video memorije, što je omogućilo efikasno izvođenje dubokih CNN-ova. Google Colab odabran je zbog svoje dostupnosti i jednostavne integracije s popularnim bibliotekama poput TensorFlow-a i Keras-a.

Za učenje su korišteni modeli VGG16, ResNet50, EfficientNetB0 te CBAM s EfficientNetB0, svi inicijalizirani s unaprijed naučenim težinama na skupu podataka ImageNet. Time je omogućeno prijenosno učenje (engl. *transfer learning*), gdje su prethodno naučene značajke ponovno korištene, čime se značajno ubrzalo učenje i poboljšala generalizacija modela. Posljednji slojevi modela prilagođeni su specifičnom zadatku razlikovanja umjetničkih slika i fotografija, pri čemu su dodani potpuno povezani slojevi s dvije izlazne jedinice te aktivacijskom funkcijom softmax.

Učenje modela provedeno je uz korištenje optimizatora Adam, koji je odabran zbog svoje sposobnosti prilagodbe stope učenja tijekom učenja, čime se omogućuje stabilnija konvergencija. Odabrana početna stopa učenja postavljena je na $10^{-5}$ kako bi se osiguralo postupno poboljšanje modela bez oscilacija. Funkcija gubitka bila je binarna unakrsna entropija, prikladna za binarne klasifikacijske zadatke, dok su kao metričke mjere korištene točnost i F1-mjera, budući da osiguravaju sveobuhvatnu evaluaciju performansi modela.

# 4. Eksperimentalni dizajn i rezultati

## 4.1. Jednostavni CNN model

Prvi model korišten u ovom istraživanju bio je jednostavni CNN model, oblikovan kao referentna točka za usporedbu s složenijim arhitekturama.

Model se sastoji od tri konvolucijska sloja s veličinama filtra $3 \times 3$, uz slojeve maksimalnog sažimanja veličine $2 \times 2$ nakon svakog konvolucijskog sloja. Nakon konvolucijskih slojeva, izlaz se izravnava (engl. *Flatten*) i ulazi u dva potpuno povezana sloja: prvi sadrži 128 neurona s aktivacijskom funkcijom ReLU, dok završni sloj koristi sigmoidalnu aktivaciju za binarnu klasifikaciju. Arhitektura modela prikazana je u Tablici 4.1.

Tablica 4.1. Arhitektura jednostavnog CNN modela.

| **Sloj** | **Vrsta sloja** | **Veličina izlaza** | **Broj parametara** |
|---|---|---|---|
| 1 | Ulaz | $224 \times 224 \times 3$ | 0 |
| 2 | Konvolucijski (ReLU) | $222 \times 222 \times 32$ | 896 |
| 3 | Maksimalno sažimanje | $111 \times 111 \times 32$ | 0 |
| 4 | Konvolucijski (ReLU) | $109 \times 109 \times 64$ | 18.496 |
| 5 | Maksimalno sažimanje | $54 \times 54 \times 64$ | 0 |
| 6 | Konvolucijski (ReLU) | $52 \times 52 \times 128$ | 73.856 |
| 7 | Maksimalno sažimanje | $26 \times 26 \times 128$ | 0 |
| 8 | Izravnavanje (Flatten) | 86, 528 | 0 |
| 9 | Potpuno povezani (ReLU) | 128 | 11.074.560 |
| 10 | Potpuno povezani (Sigmoid) | 1 | 129 |

### Rezultati evaluacije

Rezultati jednostavnog modela CNN-a prikazani su u Tablici 4.2. Model je postigao testnu točnost od 64,85% i F1-mjeru od 0,5053, što ukazuje na njegova ograničenja u generalizaciji.

Tablica 4.2. Rezultati jednostavnog modela CNN-a na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 0,6324 |
| Testna točnost | 64,85% |
| Testna F1-mjera | 0,5053 |

Tijekom učenja, gubitak na skupu za učenje postupno se smanjivao, dok je na validacijskom skupu pokazivao oscilacije, što sugerira da model nije uspio stabilno učiti značajke korisne za klasifikaciju. Slično tome, točnost i F1-mjera na validacijskom skupu nisu pratili poboljšanja na skupu za učenje, što ukazuje na prenaučavanje.

Unatoč ovim ograničenjima, jednostavni CNN model poslužio je kao polazna točka za usporedbu sa složenijim arhitekturama, pružajući osnovu za analizu poboljšanja kod naprednijih modela.

## 4.2. Model VGG16

Model VGG16 korišten je kao unaprijed naučeni model u ovom istraživanju za klasifikaciju slika na umjetnička djela i fotografije. Ovaj model pruža dublju arhitekturu u usporedbi s jednostavnim CNN modelom, s ciljem poboljšanja performansi klasifikacije.

Model je implementiran prema pristupu opisanom u radu [5], pri čemu se prvi slojevi unaprijed naučenog VGG16 modela, koji je prošao učenje na skupu podataka ImageNet, koriste za ekstrakciju značajki. Prve četiri razine modela zamrznute su kako bi se zadržale naučene značajke koje su relevantne za prepoznavanje općih obrazaca, dok su na izlaz VGG16 modela dodani prilagođeni slojevi za binarnu klasifikaciju. Nakon izravnavanja značajki pomoću sloja za izravnavanje, dodan je potpuno povezani sloj s 256 neurona i aktivacijskom funkcijom ReLU. Kako bi se smanjio rizik od prenaučenosti,

uključen je sloj isključivanja neurona (engl. *Dropout*) s omjerom od 50%. Završni sloj je potpuno povezani sloj s jednim neuronom i digmoidalnom aktivacijom za binarnu klasifikaciju. Arhitektura modela prikazana je u Tablici 4.3.

Tablica 4.3. Arhitektura VGG16 modela s prilagođenim slojevima.

| **Sloj** | **Vrsta sloja** | **Veličina izlaza** | **Broj parametara** |
|---|---|---|---|
| 1 | Ulaz | $224 \times 224 \times 3$ | 0 |
| 2-19 | Konvolucijski slojevi (VGG16) | Razne dimenzije | 14.714.688 |
| 20 | Izravnavanje (Flatten) | 25.088 | 0 |
| 21 | Potpuno povezani (ReLU) | 256 | 6.422.784 |
| 22 | Isključivanje neurona (Dropout 50%) | 256 | 0 |
| 23 | Potpuno povezani (Sigmoid) | 1 | 257 |

### Rezultati evaluacije

Dobiveni rezultati za model VGG16 prikazani su u Tablici 4.4. Model je postigao testnu točnost od 96,50% i F1-mjeru od 0,9659, što predstavlja značajan napredak u odnosu na jednostavni model CNN-a.

Tablica 4.4. Rezultati evaluacije modela VGG16 na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 0,0945 |
| Testna točnost | 96,50% |
| Testna F1-mjera | 0,9659 |

Tijekom učenja, gubitak se smanjivao u stabilnom ritmu, dok su točnost i F1-mjera na validacijskom skupu pokazale konzistentan rast, što ukazuje na dobru generalizaciju modela. Ovi rezultati potvrđuju učinkovitost unaprijed naučenih značajki arhitekture VGG16 i njezinu sposobnost prepoznavanja relevantnih uzoraka.

Usporedba s jednostavnim modelom CNN-a pokazuje da dublje arhitekture s prijenosnim učenjem mogu značajno poboljšati performanse bez potrebe za dodatnim složenim prilagodbama modela.

## 4.3. Model ResNet50

Za implementaciju modela ResNet50 iskorišten je unaprijed naučeni model, pri čemu prvih 100 slojeva modela nije bilo uključeno u proces prilagodbe zadatku, kako bi se smanjilo vrijeme učenja i zadržale naučene značajke iz donjih slojeva. Ostatak mreže prilagođen je specifičnom problemu klasifikacije.

Na postojeću mrežu dodani su dodatni slojevi kako bi model mogao izvršiti binarnu klasifikaciju. Ovi dodaci uključuju sloj za globalno prosječno sažimanje značajki, potpuno povezani sloj s 256 neurona i aktivacijskom funkcijom ReLU, sloj isključivanja neurona s omjerom od 50% kako bi se smanjila prenaučenost te završni sloj s jednom izlaznom jedinicom i sigmoidalnom aktivacijom. Tablica 4.5. prikazuje modificiranu arhitekturu modela.

Tablica 4.5. Arhitektura modela ResNet50 s prilagođenim slojevima.

| **Sloj** | **Vrsta sloja** | **Veličina izlaza** | **Broj parametara** |
|---|---|---|---|
| 1 | Ulaz | $224 \times 224 \times 3$ | 0 |
| 2-100 | Zamrznuti slojevi (ResNet50) | Razne dimenzije | 23.587.712 |
| 101 | Globalno prosječno sažimanje | $1 \times 1 \times 1280$ | 0 |
| 102 | Potpuno povezani (ReLU) | 256 | 524.544 |
| 103 | Isključivanje neurona (Dropout 50%) | 256 | 0 |
| 104 | Potpuno povezani (Sigmoid) | 1 | 257 |

### Rezultati evaluacije i optimizacija učenja

Rezultati ResNet50 modela prikazani su u Tablici 4.6. Model je postigao testnu točnost od 52,15% i F1-mjeru od 0,2906, što sugerira ozbiljne probleme s generalizacijom.

Tablica 4.6. Rezultati evaluacije ResNet50 modela na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 0,7058 |
| Testna točnost | 52,15% |
| Testna F1-mjera | 0,2906 |

Iako je gubitak na skupu za učenje stabilno opadao, na validacijskom skupu pokazi-

vao je oscilacije i ostao relativno visok. Sličan obrazac primijećen je i kod točnosti, model je dosegao vrlo visoke vrijednosti na skupu za učenje, dok su rezultati na validacijskom skupu ostali niski, što upućuje na prenaučavanje.

Međutim, tijekom prvog učenja primijećeno je da su validacijska točnost i validacijska F1-mjera dostigli svoje najveće vrijednosti negdje u sredini učenja, nakon čega su postupno opadali. To sugerira da je model počeo gubiti sposobnost generalizacije nakon određenog broja epoha.

Kako bi se optimirao proces učenja i spriječilo nepotrebno produljenje učenja koje vodi do prenaučavanja, primijenjen je mehanizam ranog zaustavljanja. Dodana je funkcionalnost koja prati najveću postignutu točnost na validacijskom skupu, uz postavljeni prag strpljenja od 3 epohe. Na taj način, učenje je automatski prekinuto nakon što se tri uzastopne epohe nisu pokazale kao poboljšanje.

Implementacija ovog pristupa omogućila je optimizaciju vremena učenja i očuvanje najboljih postignutih parametara modela. Iako konačni rezultati prikazani na tablici 4.7. i dalje nisu bili na razini VGG16 modela, primijenjena optimizacija omogućila je izbjegavanje nepotrebnog pogoršanja performansi te osigurala da se model zaustavi u točki kada je najbolje generalizirao na validacijskom skupu podataka.

Tablica 4.7. Rezultati evaluacije optimiziranog modela ResNet50 na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 0,6772 |
| Testna točnost | 58,16% |
| Testna F1-mjera | 0,5378 |

Usporedba s modelom VGG16 pokazuje da ResNet50 nije uspio postići konkurentne rezultate, što može biti posljedica njegove složenije arhitekture i manjeg broja podataka dostupnih za prilagodbu novom zadatku. Međutim, upotreba mehanizma ranog zaustavljanja smanjila je prenaučavanje i omogućila precizniju procjenu pravog kapaciteta ovog modela u kontekstu problema klasifikacije umjetničkih slika i fotografija.

## 4.4. Model EfficientNetB0

EfficientNetB0 odabran je zbog sposobnosti balansiranja točnosti i računalne učinkovitosti. Prvih 100 slojeva modela bilo je zamrznuto kako bi se zadržale unaprijed naučene značajke relevantne za prepoznavanje općih obrazaca, dok su kasniji slojevi ostavljeni za daljnje učenje kako bi se prilagodili specifičnom zadatku razlikovanja umjetničkih slika i fotografija.

Na izlaz modela EfficientNetB0 dodani su prilagođeni slojevi. Prvo je primijenjen sloj za globalno prosječno sažimanje radi smanjenja dimenzionalnosti značajki. Zatim je dodan potpuno povezani sloj s 256 neurona i aktivacijskom funkcijom ReLU, nakon čega je korišten sloj isključivanja neurona s omjerom od 50% kako bi se smanjila mogućnost prenaučavanja. Završni sloj je potpuno povezani sloj s jednom izlaznom jedinicom i sigmoidalnom aktivacijom za binarnu klasifikaciju.

Tablica 4.8. Arhitektura EfficientNetB0 modela s prilagođenim slojevima.

| **Sloj** | **Vrsta sloja** | **Veličina izlaza** | **Broj parametara** |
|---|---|---|---|
| 1 | Ulaz | $224 \times 224 \times 3$ | 0 |
| 2-100 | Osnovni slojevi (EfficientNetB0) | Razne dimenzije | 4.049.972 |
| 101 | Globalno prosječno sažimanje (2D) | $1 \times 1 \times 1280$ | 0 |
| 102 | Potpuno povezani (ReLU) | 256 | 327.936 |
| 103 | Isključivanje neurona (Dropout 50%) | 256 | 0 |
| 104 | Potpuno povezani (Sigmoid) | 1 | 257 |

### Rezultati evaluacije

Rezultati modela EfficientNetB0 prikazani su u Tablici 4.9. Model je postigao testnu točnost od 59,20% i F1-mjeru od 0,653, što ukazuje na bolje performanse od modela ResNet50, ali i dalje značajno slabije u odnosu na VGG16.

Tablica 4.9. Rezultati evaluacije EfficientNetB0 modela na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 0,6554 |
| Testna točnost (Accuracy) | 59,20% |
| Testna F1-mjera | 0,653 |

Analiza metrika tijekom učenja pokazala je da je model stabilno smanjivao gubitak na skupu za učenje, dok su rezultati na validacijskom skupu pokazivali oscilacije. Ovi rezultati sugeriraju da zamrzavanje prevelikog broja slojeva može ograničiti sposobnost modela da se prilagodi specifičnostima podataka u ovom zadatku.

Očekivalo se da će EfficientNetB0 pružiti bolju ravnotežu između točnosti i računalne učinkovitosti, no rezultati sugeriraju potrebu za daljnjom optimizacijom, uključujući prilagodbu broja zamrznutih slojeva i korigiranje stope učenja.

## 4.5. Model EfficientNetB0 s blokom CBAM

Za dodatno poboljšanje performansi osnovnog modela EfficientNetB0, implementiran je blok CBAM, koji kombinira kanalnu i prostornu pažnju. Nakon osnovnih slojeva modela EfficientNetB0, blok CBAM je dodan kako bi se unaprijedile reprezentacijske sposobnosti modela. Prvih 100 slojeva EfficientNetB0 modela bilo je zamrznuto kako bi se sačuvale unaprijed naučene značajke, a ostatak arhitekture prilagođen je specifičnom zadatku klasifikacije.

Na izlaz proširenog modela dodani su standardni slojevi, uključujući sloj za globalno prosječno (2D) sažimanje, potpuno povezani sloj s 256 neurona i aktivacijom ReLU, sloj isključivanja neurona s omjerom 50%, te završni sloj s jednom izlaznom jedinicom i sigmoidalnom aktivacijom za binarnu klasifikaciju. Arhitektura modela prikazana je u Tablici 4.10.

Tablica 4.10. Arhitektura modela EfficientNetB0 s CBAM blokom.

| **Sloj** | **Vrsta sloja** | **Veličina izlaza** | **Broj parametara** |
|---|---|---|---|
| 1 | Ulaz | $224 \times 224 \times 3$ | 0 |
| 2-100 | Osnovni slojevi (EfficientNetB0) | Razne dimenzije | 4.049.972 |
| 101 | Blok pažnje CBAM | $7 \times 7 \times 1280$ | 81.664 |
| 102 | Globalno prosječno sažimanje (2D) | 1280 | 0 |
| 103 | Potpuno povezani (ReLU) | 256 | 327.936 |
| 104 | Isključivanje neurona (Dropout 50%) | 256 | 0 |
| 105 | Potpuno povezani (Sigmoid) | 1 | 257 |

### Rezultati evaluacije

Performance modela vrednovane su na testnom skupu podataka, a rezultati su prikazani u Tablici 4.11. Dodavanje CBAM bloka nije dovelo do poboljšanja performansi u odnosu na osnovni model EfficientNetB0. Naprotiv, testna točnost pala je na 35,35%, dok je F1-mjera iznosila 0,445, što je znatno ispod očekivanih vrijednosti za ovu arhitekturu. Ovi rezultati sugeriraju da blok CBAM nije poboljšao sposobnost modela za prepoznavanje relevantnih značajki u zadatku klasifikacije umjetničkih slika i fotografija.

Tablica 4.11. Rezultati evaluacije EfficientNetB0 modela s CBAM blokom na testnom skupu.

| **Metrička vrijednost** | **Rezultat** |
|---|---|
| Testni gubitak | 1,1983 |
| Testna točnost (Accuracy) | 35,35% |
| Testna F1-mjera | 0,445 |

Analiza metrika tijekom učenja pokazala je da su vrijednosti gubitka na skupu za učenje stabilno opadale, ali je validacijski gubitak oscilirao i zadržavao visoke vrijednosti, što upućuje na problem prenaučavanja. Sličan trend primijećen je i kod točnosti i F1-mjere, model je pokazivao poboljšanja na skupu za učenje, dok su rezultati na validacijskom skupu ostali znatno slabiji.

Osnovni model EfficientNetB0 ostvario je točnost od 59,20% i F1-mjeru od 0,653, što ga čini značajno boljim od verzije s blokom CBAM. Nasuprot tome, prošireni model s blokom CBAM ostvario je pad točnosti za 23,85%, dok je F1-mjera također znatno pala, na 0,445. Dodatno, testni gubitak modela s blokom CBAM iznosio je 1,1983, što sugerira problem konvergencije i lošiju generalizaciju.

Ovi rezultati sugeriraju da povećanje složenosti modela ne jamči automatsko poboljšanje performanci. Naprotiv, može otežati optimizaciju i dovesti do lošije generalizacije, posebno kada je količina podataka ograničena.

Za potencijalna poboljšanja, potrebno je istražiti alternativne metode pažnje, dodatno prilagoditi hiperparametre ili povećati veličinu skupa podataka kako bi se osigurala bolja sposobnost učenja dodatnih parametara modela.

## 4.6. Usporedba rezultata

Tablica 4.12. prikazuje usporedbu performanci različitih modela na testnom skupu. Model VGG16 ostvario je najbolje rezultate, s točnošću od 96,50% i F1-mjerom od 0,9659, što ukazuje na njegovu sposobnost preciznog prepoznavanja značajki koje razlikuju umjetničke slike od fotografija. Njegov uspjeh može se pripisati optimiziranoj arhitekturi s unaprijed naučenim značajkama koje su već dokazano učinkovite u raznim zadacima klasifikacije slika.

S druge strane, ResNet50 je nakon primjene optimizacije i ranog zaustavljanja ostvario poboljšane rezultate u odnosu na inicijalno učenje, postigavši točnost od 58,16% i F1-mjeru od 0,5378. Iako je ovaj rezultat bolji u usporedbi s prethodnim pokušajem, gdje je model ostvario točnost od 52,15%, on i dalje pokazuje probleme s generalizacijom. Poboljšanje nakon primjene ranog zaustavljanja ukazuje na to da je prenaučavanje donekle smanjeno, ali model i dalje nije postigao razinu uspješnosti modela VGG16.

| **Model** | **Točnost (%)** | **F1-mjera** | **Testni gubitak** |
|---|---|---|---|
| $JednostavniCNN$ | $64,85$ | $0,5053$ | $0,6324$ |
| $VGG16$ | $96,50$ | $0,9659$ | $0,0945$ |
| $ResNet50$ | $58,16$ | $0,5378$ | $0,6772$ |
| $EfficientNetB0$ | $59,20$ | $0,6530$ | $0,6554$ |
| $EfficientNetB0 + CBAM$ | $35,35$ | $0,4450$ | $1,1983$ |

Tablica 4.12. Usporedba performansi različitih modela na testnom skupu.

EfficientNetB0 postigao je nešto bolje rezultate od ResNet50-a, s točnošću od 59,20% i F1-mjerom od 0,6530, što sugerira da je njegova optimizirana arhitektura za računalnu učinkovitost imala određenu sposobnost prepoznavanja ključnih značajki. Međutim, iako je EfficientNetB0 poznat po svojoj složenoj skalabilnosti, njegovi rezultati u ovom eksperimentu nisu nadmašili VGG16, što sugerira da dublje i računalno optimirane arhitekture nisu nužno bolje za ovaj specifičan zadatak.

Dodatno, kada je EfficientNetB0 proširen mehanizmom pažnje CBAM, performance

su značajno pale, točnost je iznosila samo 35,35%, uz najviši testni gubitak (1,1983). Ovi rezultati sugeriraju da dodavanje mehanizama pažnje može negativno utjecati na model u zadacima gdje su ključne značajke već inherentno prepoznatljive putem standardnih konvolucijskih slojeva. CBAM se često koristi za zadatke gdje je potrebno pojačati prepoznavanje objekata u složenim scenama, no u ovom slučaju mogao je dodati preveliku kompleksnost i smanjiti sposobnost konvergencije.

Osim kvantitativnih metrika prikazanih u Tablici 4.12., kvaliteta klasifikacije modela može se dodatno analizirati kroz primjere ispravno i pogrešno klasificiranih slika. Za ovu analizu korišten je model VGG16, budući da je ostvario najbolje rezultate na testnom skupu.

Na Slici 4.1. prikazani su primjeri slika koje je model VGG16 ispravno klasificirao kao umjetničke slike i fotografije, dok Slika 4.2. prikazuje primjere gdje model nije ispravno prepoznao kategoriju slike.


**Slika 4.1.** Primjeri ispravno klasificiranih slika.

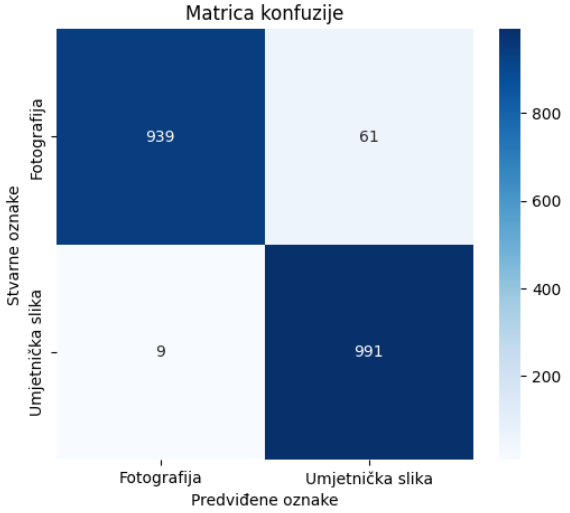

**Slika 4.2.** Primjeri pogrešno klasificiranih slika.

Za usporedbu, analizirani su rezultati jednostavnog modela CNN-a, koji je postigao drugu najbolju točnost. Slika 4.3. prikazuje primjere ispravno klasificiranih slika korištenjem jednostavnog modela, CNN-a, dok Slika 4.4. prikazuje primjere gdje ovaj model nije uspio točno klasificirati slike.

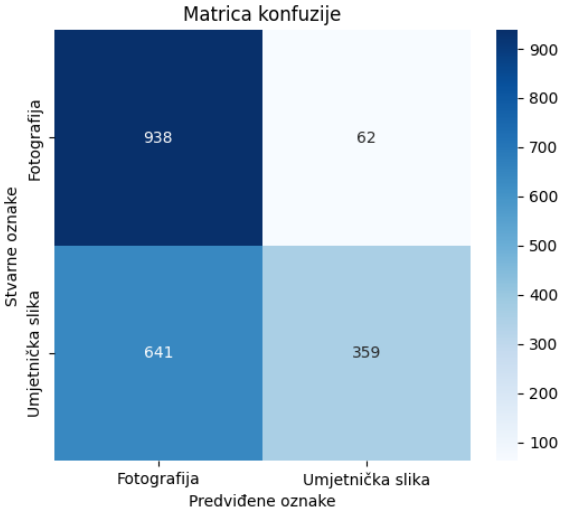

**Slika 4.3.** Primjeri ispravno klasificiranih slika korištenjem jednostavnog CNN modela.


**Slika 4.4.** Primjeri pogrešno klasificiranih slika korištenjem jednostavnog CNN modela.

Dodatni uvid u rezultate modela pružaju matrice zabune prikazane na Slici 4.5. i Slici 4.6.




**Slika 4.5.** Matrica konfuzije za model VGG16.




**Slika 4.6.** Matrica konfuzije za Jednostavni CNN model.

VGG16 pokazuje izrazitu preciznost, s malim brojem pogrešnih klasifikacija. Od ukupnog broja testiranih primjera, samo 61 fotografija je pogrešno klasificirana kao umjetnička slika, dok je kod umjetničkih slika zabilježeno samo 9 pogrešnih predikcija.

Nasuprot tome, jednostavni CNN model pokazuje značajno lošije rezultate pri klasifikaciji umjetničkih slika. Model je pogrešno klasificirao čak 641 umjetničku sliku kao fotografiju, dok je 62 fotografije klasificirao kao umjetničke slike. Ovi rezultati sugeriraju da je jednostavni CNN model imao posebno velike poteškoće u prepoznavanju umjetničkih slika, dok je VGG16, zahvaljujući dubljoj arhitekturi i unaprijed naučenim značajkama, pokazao znatno bolju sposobnost razlikovanja umjetničkih slika od fotografija.

# 5. Rasprava i zaključak

Ovo istraživanje pokazalo je da su modeli dubokog učenja učinkoviti u razlikovanju umjetničkih slika od fotografija, ali s različitim stupnjevima uspješnosti. Model VGG16 postigao je najbolje rezultate, dok su složeniji modeli, poput ResNet50 i EfficientNetB0, pokazali probleme s generalizacijom i prenaučenošću. Implementacija mehanizama pažnje nije donijela očekivana poboljšanja, već je povećala računalne zahtjeve i produljila vrijeme učenja. Ovi rezultati ukazuju na to da dublje i složenije arhitekture nisu nužno bolje za ovaj specifičan zadatak, već je ključna pažljiva optimizacija modela i podataka.

Glavni izazov bila je prenaučenost kod složenijih modela, posebno kod ResNet50, koji je postizao visoku točnost na skupu za učenje, ali je značajno lošije generalizirao na testnim podacima. Ovaj problem mogao bi se smanjiti primjenom naprednijih strategija regulacije (isključivanje neurona, L2 regularizacija) i većeg skupa podataka. Dodatno, povećanje složenosti modela nije rezultiralo boljim performancama, što sugerira da mehanizmi pažnje nisu optimalno prilagođeni ovom tipu klasifikacije, već se bolje primjenjuju na zadatke koji zahtijevaju detekciju ključnih regija unutar slike.

Osim arhitekture modela, ključnu ulogu igrali su skup podataka i računalni resursi. Skup podataka korišten u ovom istraživanju mogao je sadržavati specifične vizualne karakteristike koje CNN modeli nisu optimalno modelirali, što upućuje na potrebu za naprednijim metodama augmentacije i širim skupovima podataka kako bi se poboljšala sposobnost generalizacije modela. Dodatno, dublji modeli zahtijevaju značajne računalne resurse, što može predstavljati ograničenje u praktičnoj primjeni.

Za poboljšanje performansi klasifikacije, buduća istraživanja trebala bi se usmjeriti na primjenu različitih modela Vision Transformera (ViT), koji bolje prepoznaju globalne

odnose u slici, a ne samo lokalne značajke kao modeli CNN-ova. Optimizacija hiperparametara, uključujući prilagodbu stope učenja, zamrzavanje slojeva i primjenu naprednih tehnika regulacije, mogla bi dodatno poboljšati preciznost modela. Također, korištenje ansambla modela može omogućiti kombiniranje različitih arhitektura kako bi se postigla robusnija klasifikacija. Konačno, dublja analiza značajki koje modeli koriste za donošenje odluka mogla bi pomoći u razumijevanju razloga zbog kojih određene arhitekture bolje razlikuju umjetničke slike od fotografija.

Rezultati ovog istraživanja potvrđuju da duboko učenje ima velik potencijal za klasifikaciju umjetničkih slika i fotografija, ali uspješnost modela ovisi o optimizaciji arhitekture, kvaliteti skupa podataka i računalnim resursima. Daljnja istraživanja trebala bi se usmjeriti na primjenu različitih modela Vision Transformera, poboljšanje strategija obrade podataka te eksperimentiranje s naprednim metodama regulacije i optimizacije modela. Ova poboljšanja mogla bi omogućiti preciznije, učinkovitije i robusnije modele za klasifikaciju vizualnih sadržaja.

# Literatura


[1] F. Cutzu, R. Hammoud, i A. Leykin, "Estimating the photorealism of images: distinguishing paintings from photographs", u *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, sv. 2, 2003., str. II–305. https://doi.org/10.1109/CVPR.2003.1211484

[2] A. Carballal, A. Santos, J. Romero, P. Machado, J. Correia, i L. Castro, "Distinguishing paintings from photographs by complexity estimates", *Neural Computing and Applications*, sv. 30, br. 6, str. 1957–1969, Sep 2018. https://doi.org/10.1007/s00521-016-2787-5

[3] W. L. Hosch, "Zipf's law", Encyclopedia Britannica, 2024., pristupljeno: 10 siječnja 2025. [Mrežno]. Adresa: https://www.britannica.com/topic/Zipfs-law

[4] S. Šegvić, "Računalni vid: Elementarne slikovne značajke", Prezentacija, Zagreb, Hrvatska, 2024., Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, pristupljeno: 5. prosinca 2024. [Mrežno]. Adresa: http://adept.zemris.fer.hr

[5] J. M. López-Rubio, M. A. Molina-Cabello, G. Ramos-Jiménez, i E. López-Rubio, "Classification of images as photographs or paintings by using convolutional neural networks", u *Advances in Computational Intelligence*, I. Rojas, G. Joya, i A. Català, Ur. Cham: Springer International Publishing, 2021., str. 432–442.

[6] Ikarus777, "Best artworks of all time dataset", 2025., pristupljeno: 10. studenog 2024. [Mrežno]. Adresa: https://www.kaggle.com/ikarus777/best-artworks-of-all-time

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, i C. L. Zitnick, "Microsoft COCO: Common objects in context", 2014.

[8] TikZ, “Tikz: Neural networks”, Dostupno na: https://tikz.net/neural_networks/, 2025., pristupljeno: 10. veljače 2025.

[9] K. Simonyan i A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, 2015., pristupljeno: 5. prosinca 2024. [Mrežno]. Adresa: https://arxiv.org/abs/1409.1556

[10] S. Šegvić, “Računalni vid: Konvolucijski modeli za klasifikaciju slike”, Prezentacija, Zagreb, Hrvatska, 2024., Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, pristupljeno: 5. prosinca 2024. [Mrežno]. Adresa: http://adept.zemris.fer.hr

[11] K. He, X. Zhang, S. Ren, i J. Sun, “Deep residual learning for image recognition”, 2015. [Mrežno]. Adresa: https://arxiv.org/abs/1512.03385

[12] M. Tan i Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks”, u *Proceedings of the 36th International Conference on Machine Learning (ICML)*, sv. 97. PMLR, 2019., str. 6105–6114. [Mrežno]. Adresa: https://proceedings.mlr.press/v97/tan19a.html

[13] S. Woo, J. Park, J.-Y. Lee, i I. S. Kweon, “CBAM: Convolutional block attention module”, 2018. [Mrežno]. Adresa: https://arxiv.org/abs/1807.06521

# Sažetak

## Modeli dubokog učenja za razlikovanje između umjetničkih slika i fotografija

Ivana Stilinović

Ovaj rad istražuje primjenu modela dubokog učenja za razlikovanje umjetničkih slika i fotografija. Korišteni su unaprijed naučeni modeli konvolucijskih neuronskih mreža, uključujući VGG16, ResNet50 i EfficientNetB0, s ciljem ispitivanja njihove učinkovitosti u binarnoj klasifikaciji. Također je analiziran mehanizam pažnje CBAM kako bi se procijenio njegov utjecaj na performance modela. Eksperimentalni rezultati pokazali su da je VGG16 postigao najbolje rezultate (točnost 96,5%, F1-mjera 96,59%), dok su složeniji modeli, poput ResNet50 i EfficientNetB0, imali poteškoće s generalizacijom. Mehanizam pažnje nije poboljšao klasifikaciju, već je povećao računalne zahtjeve. Zaključeno je da bi dodatna optimizacija hiperparametara, povećanje skupa podataka te istraživanje alternativnih arhitektura, poput Vision Transformera, moglo doprinijeti boljim rezultatima u ovom području.

**Ključne riječi:** duboko učenje, konvolucijske neuronske mreže, klasifikacija slika, umjetničke slike, fotografije

# Abstract


## Deep learning models for distinguishing between paintings and photographs

Ivana Stilinović

This thesis investigates the application of deep learning models to distinguish between art paintings and photographs. Pre-trained CNN models, including VGG16, ResNet50 and EfficientNetB0, were used to test their performance of binary classification. The CBAM attention mechanism was also analyzed to assess its impact on model performance. Experimental results showed that VGG16 achieved the best results (accuracy 96.5%, F1-score 96.59%), while more complex models, such as ResNet50 and EfficientNetB0, had difficulties with generalization. The attention mechanism did not significantly improve the classification, but increased the computational requirements. It was concluded that additional optimization of hyperparameters, increase of dataset, and research of alternative architectures, such as Vision Transformers, can contribute to better results in this field.



**Keywords:** deep learning, convolutional neural networks, image classification, paintings, photographs