Interpretabilnost Modela –

Integrisani Gradijent

# Uvod

Razvoj dubokih neuronskih mreža doneo je izuzetne rezultate u mnogim poljima, ali zbog svoje izrazite kompleksnosti je doveo u pitanje **interpretabilnost**. Iako modeli postižu visok stepen tačnosti, često se ne zna iz kog razloga je doneta određena odluka od strane modela. Priroda ove *crne kutije* otežava poverenje korisnika, donošenje odluka u osetljivim oblastima i mogućnost validacije od strane stručnjaka.

Koncept **Explainable AI (XAI)** uvodi načine pomoću kojih se ponašanje modela može bolje razumeti kroz objašnjenja predikcija. Cilj **objašnjivosti** nije samo transparentnost već i povećanje poverenja u model, olakšavanje dijagnosifikovanja grešaka i podrška u regulativnim okvirima gde je potrebno jasno opravdanje modelskih odluka. Objašnjivost je naročito važna u realnim scenarijima, gde samo visoka tačnost modela nije dovoljna – potrebno je i objašnjenje zašto je određena odluka doneta, da bi korisnici modela imali poverenje u isti i na kraju ga i koristili.

Prema *Munn i Pitman (2022)*, objašnjivost se može posmatrati iz više perspektiva:

* **Praktičari** (*practitioners*) su istraživači i inženjeri koji razvijaju modele i žele da razumeju njihovo ponašanje,
* **Posmatrači** (*observers*) su donosioci odluka i regulatorna tela kojima je potreban uvid radi poverenja i bezbednosti,
* **Krajnji korisnici** (*end users*) su oni koji često nemaju tehničko znanje, ali žele da razumeju i veruju modelskim predikcijama.

Ove grupe imaju **različite zahteve** u pogledu objašnjivosti, što dodatno naglašava značaj razvoja metoda koje nude jasna i razumljiva objašnjenja odluka modela.

Razvoj metoda objašnjivosti doneo je veliki broj pristupa ovom problemu – od jednostavnih vizuelnih tehnika do formalnih matematičkih izvođenja. Ipak, tu se posebno izdvajaju **metode atribucije osobina** (*feature attribution methods*), koje se fokusiraju na to da odrede značaj pojedinačnih ulaznih vrednosti koje su ključne za odluku modela. Ove metode su posebno važne, u našem slučaju, gde intuicija o tome kako mreža donosi odluke je često izrazito teško dostižna.

U okviru atribucionih metoda, **integrisani gradijent** (*Integrated Gradients, IG)* predstavlja jedan od najuticajnijih i najviše korišćenih pristupa. Prednost IG je u tome što kombinuje jednostavnost implementacije sa teorijom, jer ispunjava ključne aksiome kao što su **senzitivnost** i **implementaciona invarijantnost**. IG metod se zasniva na ideji da se **uticaj pojedinačne osobine procenjuje kroz akumulaciju gradijenata duž puta od neke referentne tačke** (*baseline*) **do stvarnog ulaza**. Time se obezbeđuje *stabilnije* i *interpretabilnije* objašnjenje od metoda koje posmatraju samo trenutne gradijente.

U praksi IG je naročito primenjiv u računarskoj viziji, gde omogućava generisanje toplotnih mapa koje jasno ukazuju na delove ključne za klasifikaciju. Na taj način domenski eksperti mogu da potvrde da li model *posmatra* regije koje treba ili da uoče potencijalne *pristrasnosti* i *greške* u ponašanju modela. Baš iz razloga što IG ima sposobnost da spoji teoriju sa praktičnom vizuelizacijom, je postao jedna od glavnih tehnika u istraživanju i primeni u industriji kada se radi od interpretabilnosti modela.

# Metode i materijali

## Algoritam integrisanih gradijenata

Jedan od najvećih izazova u oblasti interpretabilnosti dubokih neuronskih mreža jeste da se razumeju doprinosi pojedinačnih ulaznih karakteristika u konačnoj predikciji modela. Klasične metode oslanjanja na gradijente često imaju problem sa nestabilnošću i nepoštovanjem osnovnih aksioma koje metoda treba da zadovolji, kao što su osetljivost (*sensitivity*) i invarijantnost implementacije (*implementation invariance*). IG predstavljaju metodologiju koja rešava ove nedostatke pomoću formalnog okvira koji dodeljuje značaj ulaznim karakteristikama kroz proces integracije gradijenata duž putanje od baznog ulaza do stvarnog ulaza.

Neka je dat diferencijabilni model koji predstavlja **funkciju neuronske mreže**, a **vektor ulaznih podataka**. Takođe se definiše i **bazni ulaz** , koji najčešće odgovara slici bez informacija (npr. crna slika) ili vektoru nula. **Integrisana atribucija komponente** se definiše kao:

Ova formula prikazuje **doprinos -te ulazne karakteristike** tako što meri **koliko promena vrednosti u odnosu na bazu utiče na predikciju modela**, akumulirano duž linije interpolacije između i . U vektorskoj formi, integrisani gradijenti se mogu zapisati kao:

gde operator označava **element-po-element množenje** (*element-wise/Hadamard množenje*), a **gradijent funkcije u odnosu na ulaz**.

Problem nastaje u tome što se integral *ne može izračunati analitički*, pa se koristi aproksimacija putem diskretizacije. Najčešće se primenjuje Rimanova suma, gde se interval deli na jednakih delova. Tada se integrisani gradijenti aproksimiraju kao:

Na ovaj način se integral menja sa *konačnom sumom gradijenata* izračunatih za tačke na interpolacionoj putanji između baznog i stvarnog ulaza. Broj koraka je ključni parametar metode. Mali broj koraka vodi ka gruboj aproksimaciji i potencijalnim greškama u atribuciji, dok preveliki broj koraka znatno povećava računsku složenost, sa čim smo se i susreli u kodu pri promeni ovog parametra. U praksi se najčešće koristi opseg od do koraka, u zavisnosti od potrebne tačnosti i raspoloživosti resursa.

A graph of a number of data

AI-generated content may be incorrect.

*Aproksimacija integrala u zavisnosti od parametra* ***N*** *(u našem slučaju* ***m****)*

Integrisani gradijenti su jedna od najčešćih tehnika koje se koriste kada govorimo o interpretabilnosti, jer pored lakog prikaza koji se može sprovesti znajući samo arhitekturu modela, poseduje i neke jako bitne osobine. To su:

* **Potpunost** (*completeness*) prema kome je zbir svih atribucija jednak razlici između predikcije na stvarnom ulazu i predikcije na baznom ulazu ,
* **Osetljivost** (*sensitivity*) koja nalaže da ulazna karakteristika koja menja predikciju mora imati nenultu atribuciju,
* **Invarijantnost implementacije** (*implementation invariance*) po kojoj dva modela koja implementiraju istu matematičku funkciju, nezaivsno od razlika u arhitekturi, moraju proizvesti identične atribucije.

Zbog ovih svojstava, metoda IG se smatra jednim od **najrobusnijih** i **najčešće primenjivanih** alata u interpretabilnosti neuronskih mreža.

## Arhitektura korišćenog modela

Za ilustraciju metode IG korišćena **je MobileNetV3-Large** mreža, koja pripada konvolutivnim neuronskim mrežama dizajniranim za efikasnu obradu slika sa ograničenim resursima. MobileNetV3 je razvijena kombinovanjem automatizovanog pretraživanja neuralnih arhitektura (NAS) i ručno projektovanih optimizacija, sa ciljem postizanja optimalnog balansa između tačnosti i brzine izvođenja.

Ona kombinuje principe dubinskih separabilnih konvolucija sa unapređenjima u vidu **SE** (*Squeeze-and-Excitation*) modula, kao i prilagođenim nelinearnostima poput **h-swish** funkcije. Zbog ovoga je postignuta značajna redukcija u broju parametara, samim tim i računskih operacija, dok se zadržava visok nivo tačnosti.

A diagram of a block diagram

AI-generated content may be incorrect.

*Ilustracija osnovnog MobileNetV3 bloka (bottleneck struktura)*

U osnovi arhitekture nalaze se bottleneck blokovi koji sadrže tri ključne faze:

1. **Ekspanzija dimenzionalnosti** ulaznog tenzora pomoću konvolucije.
2. **Dubinska konvolucija** ( ili ), koja primenjuje filtere odvojeno po kanalima, značajno smanjujući broj parametara u odnosu na standardne konvolucije.
3. **Projekcija unazad** na nižu dimenzionalnost pomoću konvolucije.

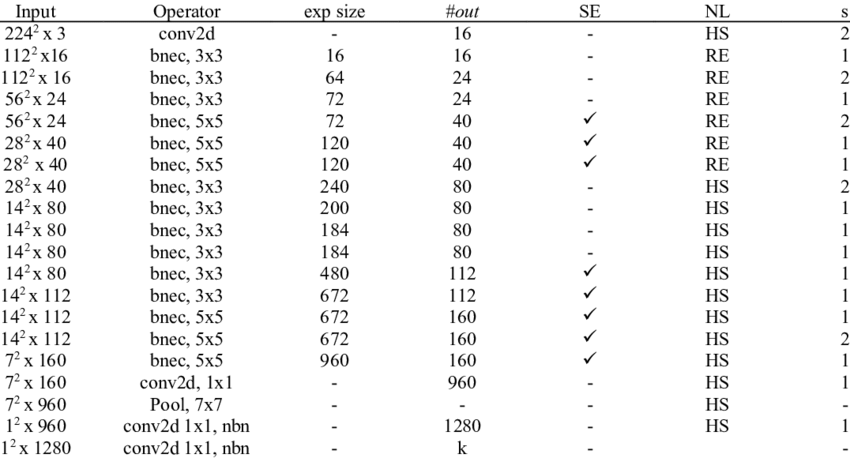
Pored toga, u određenim blokovima uključeni su SE moduli, koji implementiraju mehanizam pažnje na nivou kanala. Oni omogućavaju modelu da dinamički rekalibriše značaj pojedinačnih kanala pomoću pooling-a, male potpuno povezane mreže i sigmoidne aktivacione funkcije. Na taj način, mreža fokusira svoju snagu na najrelevantnije karakteristike.

A diagram of a bottleneck

AI-generated content may be incorrect.

*Pregled kompletne arhitekture MobileNetV3 sa prikazanim SE modulima i tokom podataka*

Tabela prikazana na slici ispod daje **detaljan pregled** arhitekture MobileNetV3-Large, uključujući ulazne i izlazne dimenzije svakog sloja, tip operacije, veličinu ekspanzije, broj filtera i prisustvo SE modula.



*Tabelarni prikaz arhitekture MobileNetV3-Large sa opisom slojeva*

Ključna inovacija MobileNetV3 leži u kombinaciji:

* **Dubinskih separabilnih konvolucija** koje minimizuju broj parametara,
* **SE modula** koji uvode mehanizam pažnje na nivou kanala,
* **Optimizovanih nelinearnosti** koje poboljšavaju performanse pri niskoj preciznosti računanja,
* Kao i pažljivo dizajniranoj arhitekturi koja **balansira između tačnosti i efikasnosti**.

Zahvaljujući ovim karakteristikama, MobileNetV3-Large postiže performanse uporedive sa znatno većim modelima, uz znatno manju složenost, zbog čega je i izabran za evaluaciju metode interpretabilnosti IG.

## Podaci korišćeni za testiranje

Za eksperimentalnu ilustraciju metode IG korišćen je manji skup slika koji je preuzet sa **Caltech-101** kolekcije Caltech Dataset, koja sadrži preko 9,000 slika raspoređenih u 101 kategoriju objekata. U ovom radu su selektovane 4 slike iz različitih kategorija: morska zvezda (*starfish*), riba (*bass*), leptir (*butterfly*) i avion (*airplanes*). Konkretni fajlovi korišćeni za evaluaciju su:

* starfish/image\_0007.jpg
* bass/image\_0001.jpg
* butterfly/image\_0001.jpg
* airplanes/image\_0001.jpg

Treba napomenuti da klase iz Caltech-101 skupa **ne odgovaraju** u potpunosti sa klasama iz ImageNet-a na kojima je MobileNetV3-Large treniran. U Caltech skupu bi odgovarajuća klasa za *goldfish* bila *bass*, za *monarch* bila *butterfly* i za *warplane* bila *airplanes*. Možemo videti da je ImageNet mnogo detaljniji dataset, i da ovde dolazi do proširivanja sa 101 klase na 1000, tako da klase koje dobijamo za prediktovanje nisu odgovarajuće folderima u kojima se slike nalaze, ali to je i očekivano jer se radi o drugom skupu podataka. Uprkos ovome model je **uspešno predvideo klase slika**, dajući nam još detaljnije informacije, kao npr. da je vrsta leptira sa slike *monarch*, a ne samo da je leptir. Ovo svakako nije bila poenta rada, i nema neke razlike u odnosu na to da smo sami pronalazili slike po internetu, ili koristili neke iz svojih privatnih kolekcija, već je poenta u analizi rezultata koja je u nastavku rada.

# Rezultati i diskusija

Iako se klase u ovom skupu ne poklapaju u potpunosti sa klasama na kojima je treniran MobileNetV3 (ImageNet), rezultati pokazuju da model uspešno prepoznaje slike kroz **semantički bliske kategorije**. U nastavku su prikazani pojedinačni slučajevi.

## Primer 1: Riba (*bass*) → Predikcija: *goldfish* (36.06%)

A close up of a fish

AI-generated content may be incorrect.

Na slici ribe model daje predikciju klase *goldfish* sa uverenjem od 36.06%. Ovaj rezultat je očekivan jer postoje takođe ostale klase riba koje su ‘*preuzele’* deo ukupne verovatnoća, ali ono što je najbitnije da je ‘*pobedila’* prava klasa.

A screenshot of a fish

AI-generated content may be incorrect.

Metoda IG pokazuje da su **ključne konture tela ribe, posebno peraja** i **oblik glave**, kao regije koje su najviše uticale na predikciju. Time se potvrđuje da model koristi semantički relevantne delove slike za donošenje odluke.

## Primer 2: Leptir (*butterfly*) → Predikcija: *monarch* (74.70%)

A butterfly with a black background

AI-generated content may be incorrect.

Na slici leptira MobileNet uspešno prepoznaje kategoriju *monarch* sa visokom sigurnošću od 74.70%.

A screen shot of a butterfly

AI-generated content may be incorrect.

IG jasno **naglašava krila** i **simetrične šare** kao dominantne atribute koji su doveli do ove odluke. Ovaj rezultat pokazuje sposobnost modela da prepozna detalje u strukturi krila, što je ključno za razlikovanje između različitih vrsta leptira.

## Primer 3: Morska zvezda (*starfish*) → Predikcija: *starfish* (82.26%)

A screenshot of a starfish

AI-generated content may be incorrect.

Kod ovog primera model postiže najvišu tačnost, od 82.26% za klasu starfish, što ima smisla jer je ovo jedna od retkih klasa iz ImageNet koja **nema podklase**, kao na primer u prethodna dva primera.

A screenshot of a cell phone

AI-generated content may be incorrect.

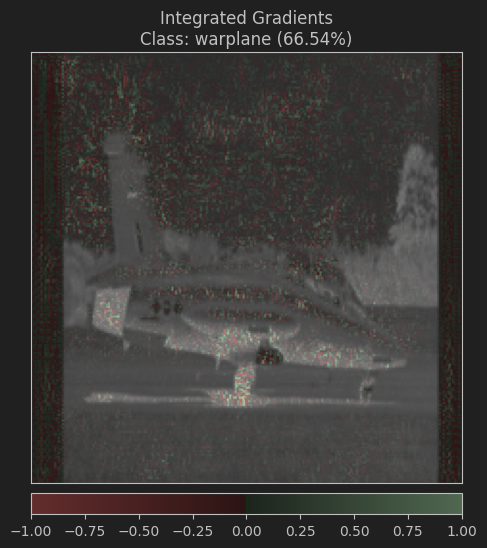
IG dodatno potvrđuje ispravnost odluke, jer ističe **konture krakova** i **teksturu površine** zvezde kao ključne faktore. Rezultat pokazuje da model pravilno koristi vizuelne atribute koji su u skladu sa tim kako čovek percipira da je nešto na slici morska zvezda.

## Primer 4: Avion (*airplane*) → Predikcija: *warplane* (66.54%)

A military jet on a runway

AI-generated content may be incorrect.

Slika aviona je klasifikovana kao *warplane* sa verovatnoćom od 66.54%. Ovo je još jedan od primera gde se ne poklapaju klase iz Caltech-ovog skupa i ImageNet, ali je rezultat blizak i ispravan s obzirom da je na slici zaista ratni avion.



Metoda IG **ističe trup** i **krila** aviona kao ključne oblasti doprinosa, dok pozadina slike ima minimalan uticaj na odluku. Ovaj rezultat ilustruje da metoda uspešno odvaja relevantne objekte od ostatka slike.

## Diskusija

Rezultati potvrđuju da i pored razlika u klasama između Caltech-101 i ImageNet skupova, MobileNetV3 daje **uverljive i interpretabilne predikcije**. IG metoda upravo **naglašava semantički značajne delove objekata**, potvrđujući da model koristi relevantne vizuelne karakteristike. U svim primerima smo mogli uočiti da metoda objašnjava predikciju na intuitivan način, što doprinosi razumevanju ponašanja modela i povećava poverenje u njegove odluke.

# Zaključak

U ovom radu demonstrirana je primena metode integrisanih gradijenata za objašnjavanje načina na koji neuronske mreže dolaze do svojih predikcija. Cilj istraživanja je bio da se ispita na koji način ova metoda omogućava bolje razumevanje ponašanja modela kroz atribuciju ulaznih karakteristika.

Kroz eksperimente na odabranim slikama iz Caltech-101 pokazano je da, iako se klase iz ovog skupa podataka ne poklapaju sa klasama ImageNet, model je i dalje **uspešno prepoznao objekte kroz semanitčki bliske kategorije**. Mnogo bitnija stvar, što je i tema ovog rada, IG je **dosledno isticao delove slike koji su intuitivno najrelevantniji** za prepoznavanje objekta.

Glavni doprinos ove metode obezbeđuje **transparentnost modelskih odluka** i **poverenje korisnika**, naročito kada su podaci kompleksni i predikcije nisu najsigurnije. Time *IG predstavlja efikasno sredstvo za evaluaciju i validaciju dubokih neuronskih mreža u realnim primerima*.

U perspektivi, dalji rad bi mogao da obuhvati poređenje IG sa drugim atribucionim metodama, kao i evaulaciju na širim i raznovrsnijim skupovima podataka. Na taj način bi se mogla dobiti bolja slika o tome u kojim uslovima metoda daje najpouzdanije objašnjenje i kako se može primeniti u domenu kritičnih sistema, gde je zapravo interpretabilnost najpotrebnija.