# Универзитет Унион

## Рачунарски факултет

# Класификација пнеумонијалних налаза са конволуционим неуронским мрежама

Петар Арнаутоивић Београд, 2021

Ментор: проф. Др Јелена Васиљевић

# Увод

Одувек је човек покушавао да разуме шта је то интелигенција и како она фукнционише, како ми заправо размишљамо и учимо. На основу којих параметара доноси закључке и одлуке, предвиђа и тумачи одређене проблеме. То су нека питања на које и данас човечанство не може у потпуности да одговори.

Вештачка интелигенција (Artificial Intelligence) покушава не само да одговори на та питања и да их разуме, већ да опонаша људску интелигенцију. Данас, вештачка интелигенција је напредна област са много практичних делова и изазовних истраживачких тема. Интелигентни софтвери, попут препознавање говора, алгоритама за социјалне мреже, паметни асистенти (Alexa, Google Home, Apple HomePod), све су више у дневном животу присутни.

Вештачку интелигенцију први је предложио Џон Макарти 1956. Идеја, коју је предложио математичар Алан Тјуринг, била је да се направе машине које имају способност да мисли и учи као човек. Алан Тјуринг је своје идеје и хипотезе спровео у праксу, тако што је почео да тестира да ли машине могу да мисле. Након серија тестирања, метода је добила назив Тјурингов тест. Испоставило се да машине имају способност да уче слично као људи. Тјурингов тест се користи као прагматичан начин тестирања, како би се могло оценити да ли машина може да размишља као човек.

У овом раду покушаћу да објасним неке од најважнијих делова вештачке интелигенције, њене предности и мане, као и саму примену вештачке интелигенције у разним пољима науке и рада, највише о доприносу вештачке интелигенције у медицини. Описаћу проблем класификације рендгенских налаза плућа са пнеумонијом и обучити програм да сам може да класификује на основу претходних тренинга. Користићу методе дубоког учења, као што су конволуционе неуронске мреже. Користићу три различита приступа за класификацију слика: класичну конволуциону неуронску мрежу (Convolutional Neural Network), Пренешено учење (Transfer Learning), Прецизно подешавање (Fine Tuning).

# Упала плућа – пнеумонија

Пнеумонија је инфекција у једном или оба плућа узрокована бактеријама, вирусима или гљивицама. Инфекција доводи до упале у ваздушним кесама плућа, које се називају алвеоли. Алвеоле се пуне течношћу или гнојем, што отежава дисање. И вирусна и бактеријска пнеумонија су заразне. То значи да се могу ширити са особе на особу удисањем капљица у ваздуху од кашља. Упала плућа представља најопасније обољење респираторног система и може бити узрокована различитим микроорганизмима. Типичне упале плућа изазване су бактеријама.

Такође можете добити ове врсте упале плућа ако дођете у контакт са површинама или предметима који су контаминирани бактеријама или вирусима који изазивају упалу плућа.

До обољења долази када су одбрамбени механизми респираторног система поремећени или одбрамбене снаге организма смањене. Као предиспонирајући чиниоци често се наводе расхлађење организма, употреба алкохола, примена анестезије, инфекције горњих дисајних путева и хронични бронхитис.

Упала плућа може бити у распону озбиљности од благе до опасне по живот. Најозбиљније је за одојчад и малу децу, особе старије од 65 година и особе са здравственим проблемима или ослабљеним имунолошким системом.

Упала плућа је и данас чест узрок смртности, упркос савременој антибиотској терапији. Пнеумонија се односи 10 пута више живота од свих осталих заразних болести. У земљама у развоју са ниским стандардом и лошим условима живота инфекције доњих дисајних путева најчешћи су узрок умирања.

# Дубоко учење

Дубоко  учење (Deep learning) је техника обучавања рачунара  примером која је део машинског учења (Machine learning). Дубоко учење добија све већи значај у последње време. Постизањем невероватних резултата који раније нису били могући, омогућили су развој софтвера као што су: гласовна контрола, аутомобили са аутопилотом, роботика, препознавање говора, итд.

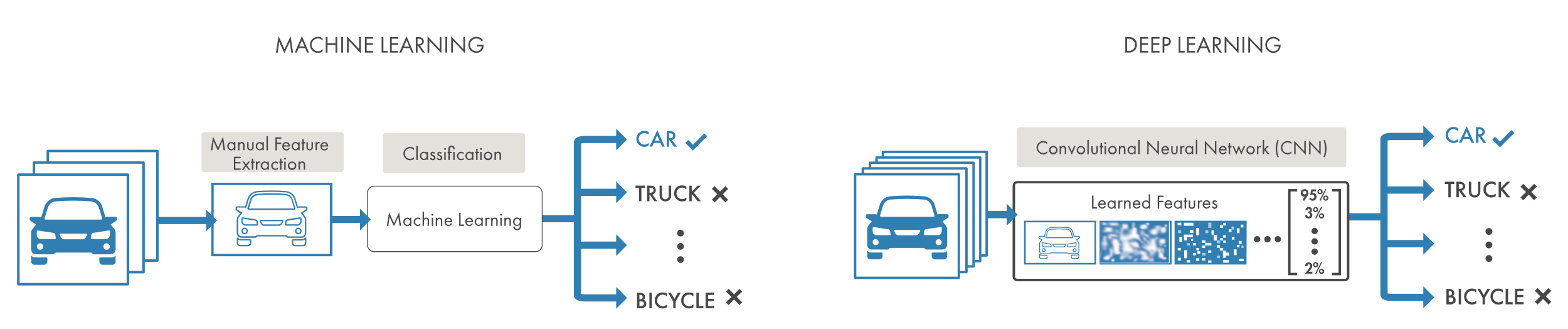
У дубоком учењу, модел учи да решава задатке као што су класификација помоћу слика, звука или текста. Такви модели, уколико су добро утренирани, могу да постигну већу тачност него човек. Да би тачност била што већа, модел мора бити добро утрениран, а то је могуће уколико користи велики скуп података и вишеслојну архитектуру неуронских мрежа.

Термин „дубоко“ се односи на већи број скривених слојева у архитектури неуронских мрежа. Основне неуронске мрежа садже два до три скривена слоја, док комплексиније архитектуре имају  знатно већи број.

## Разлика између дубоког и машинског учења

Дубоко учење је посебан облик машинског учења. Машинско учење (Machine Learning) је проучавање компјутерских алгоритама који се могу аутоматски побољшати кроз искуство и коришћење података. Процес рада машинског учења почиње тако што се релевантне функције ручно издвајају из података. Карактеристике се затим користе за креирање модела који категорише објекте на слици. У процесу рада дубоког учења, релевантне функције се аутоматски издвајају из слика. Поред тога, дубоко учење користи „end-to-end learning” (учење од краја до краја), што значи да се мрежи дају необрађени подаци и задатак који треба да изврши, уз то и учи како да се изврше аутоматски. Друга велика разлика је да се алгоритми дубоког учења скалирају са подацима, док код машинског, учење конвергира.

Када додате више примера и података у обуку, перформансе учења се заустављају на одређеном нивоу, док код дубоког учења није такав случај. Кључна предност дубоког учења је у томе што се перформансе побољшавају са повећањем броја података.



Илустрација 1:Поређење приступа машинског учења са дубоким учењем

У машинском учењу, ручно бирате карактеристике и класификатор за сортирање података. Са дубоким учењем, екстракција карактеристика и кораци моделирања су аутоматски. Maшинско учење нуди низ метода и модела које се могу ускладити на основу апликације, величине података које обрађује и врсте проблема које треба да се реши. Апликација, која користи дубоко учење захтева веома велику количину података (хиљаде слика) да би успешно обучио модел, као и графичку картицу (GPU) или графичке процесорске јединице, за ефикасну обраду података.

Уколико треба да направите избор између машинског учења и дубоког учења, размислите да ли графичка картица коју имате има високе перформансе. Ако перформансе вашег рачунара нису на одговарајућем нивоу, машинско учење би био бољи избор уместо дубоког учења. Дубоко учење је захтевније и генерално је сложеније, тако да је потребно најмање неколико хиљада слика како би добили жељене резултате. Поседовање графичке картице високих перформанси значи да ће моделу требати мање времена да анализира податке.

Diagram

Description automatically generated

Илустрација 2:Приказ односа дубоког и машинског учења

# Неуронске мреже

Неуронске мреже престављају једну од најбољих и најуспешнијих метода у машинском учењу. Први пут су представљене педесетих година прошлог века, али кроз своју историју нису биле приоритет научницима. Значајни помаци дешавају се 2006. године са појавом дубоких неуронских мрежа, а затим прави успон добијају 2012. године када дубоке неуронске мреже значајно надмашују претходне методе у решавању различитих рачунарских проблема. У неким задацима, превазишле су могућности и људских експерата. Данас представљају један од најбољих метода машинског учења са најширим доменом примене.Постоје различите врсте неуронских мрежа, развијене за специјализоване примене. Свака мрежа има одређен број елемената које називамо неуронима, по томе су препознатљиве све неуронске мреже. Као у људској анатомији, неурони прослеђују једни другима сигнале и израчунавају нове сигнале на основу претходно прослеђених сигнала. Начин повезаности неурона и на који начин они рачунају, одређују архитектуру мреже и прецизно одређују о каквој мрежи се ради. У зависности од врсте проблема, експерт покушава да дефинише најбоље одговарајућу архитектуру мреже. Да би најбоље објаснили принцип неуронских мрежа потребно је прво описати рад појединачног неурона.

# Неурон

Неурон је основна јединица израчунавања процеса и повезивањем више неурона добијамо неуронску мрежу. Рад неурона сличан је као код биолошког неурона, на улазу примају сигнале од других неурона и трансформишу добијени сигнал као свој излазни сигнал. Сви сигнали представљени су реалним бројевима, а обрада тих сигнала врши се неком математичком методом која укључује линеарну комбинацију улазних сигнала, а затим неалинеарну трасформацију те линеарне комбинације.

Дефиниција: „Неурон је функција *n* аргумената x1, x2, x3,…,xn која има општи облик:

где су 𝑤0, 𝑤1, . . . , 𝑤n параметри или тежине неурона, којима се врши линеарно комбиновање, а 𝑔 : R → R је нелинеарна активациона функција којом се та линеарна комбинација трансформише.“

Diagram

Description automatically generated

Илустрација 3:Неурон

Није одређено стриктно коју активациону фукцију треба користити, постоји више избора у зависности од архитектуре неуронске мреже. Активациона функција, у већини случајева, треба да буде монотона, непрекидна и диференцијабилна. Најчешће коришћене функције су: сигмоидна функција, хиперболички тангенс, исправљена линеарна јединица (rectified linear unit - ReLU), степенаста функција, Гаусова функција, лиенарна функција и њихове варијације. У свим слојевима мреже осим у последњим најчешће се користи нека модификација исправљене линеарне функције због успеш них резултата те функције.

Diagram

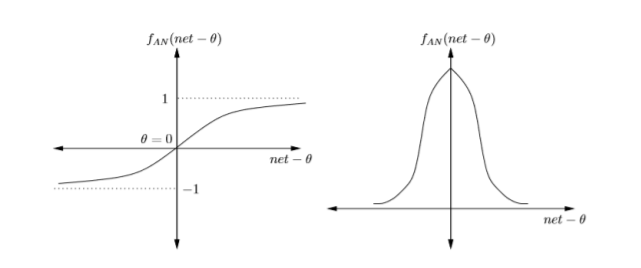
Description automatically generated

Илустрација 4: линеарна и степенаста функција

A picture containing text, antenna, day

Description automatically generated

Илустрација 5:Исправљена линеарна јединица (ReLU) и сигмоидна функција



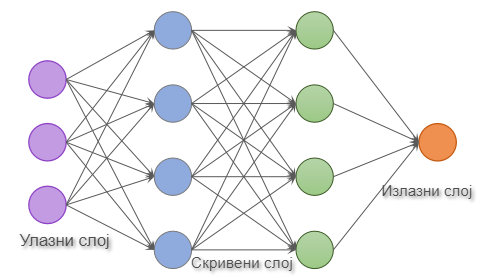
Илустрација 6: Хиперболички тангенс и Гаусова функција

Излазни сигнал је јачи, уколико је вредност линеарне комбинације већа, због моностоности активационе функције. Што су вредности улазних тежина веће, то је већи утицај тих улаза на излазни сигнал, а кад је ближа нули тај утицај се смањује.

Када се сигмоидна функција постави као активациона функција мреже, неурон постаје модел логистчке регресије, а када се постави идентититет за активациону функцију, неурон постаје модел линеарне регресије. Зато су неуронске мреже способне да извршавају класификацију и регресију.

# Потпуно повезане неуронске мреже

Потпуно повезане неуронске мреже (Fully connected neural networks) представљају традиционалан модел неуронских мрежа. То је најкоришћенији модел до сада. Као што из назива може да се претпостави, њихова структура је таква да се неурони слажу у слојеве, групе неурона, и повезане су тако да неурони једног слоја примају као улазе вредности свих неурона претхходног слоја, а вредности које се добију на том слоју прослеђују се неуронима следећег слоја.

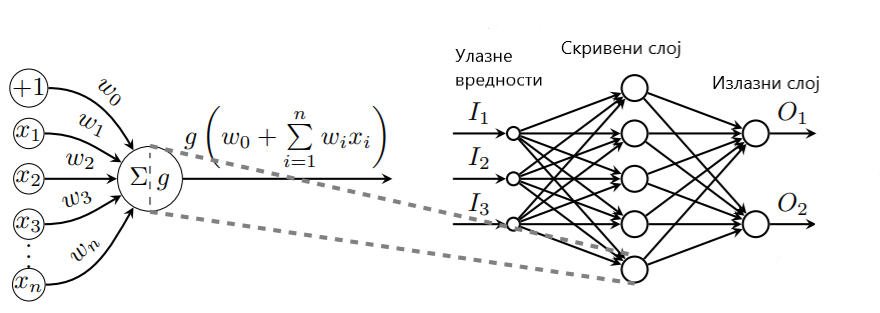


Илустрација 7: Структура потпуно повезане неуронске мреже

На слици 7 приказана је архитектура потпуно повезане неуронске мреже. На слици се могу уочити да се архитектура састоји из улазног слоја, скривених слојева и излазног слоја. Приказана су два скривена слоја неурона, који се називају тако јер се вредности тог слоја се не користе ван модела. Број скривених слојева није ограничен (1, … ,N) и због тога се називају дубоке неуронске мреже (deep neural networks). Већина мрежа које данас можемо да користимо су дубоке.

Дефиниција потпуно повезаних мрежа: „Потпуно повезана неуронска мрежа је функција ***𝑓w*** дефинисана на следећи начин:

где ***x*** представља улазни вектор, ***L*** је број слојева**, *hi***представља вектор вредности ***i-***тог слоја, ***Wi***  је матрица параметара чије врсте представљају векторе параметара неурона ***i-***тог слоја који множе улазне вредности, ***bi***  је вектор слободних коефицијената тих неурона, а ***g*** и ***g*** су активационе функције. Матрице ***Wi***  и вектори ***bi***  чине параметре тих ***W.*** “



Илустрација 8: Архитектура повезане неуронске мреже и неурона

Израчунавања која мрежа извршава састоји се од линеарних матричних трансформација и примена активационе функције. У последњем слоју мреже, могуће је да функција 𝑔˜ може да представља другачију активациону функцију од фукције 𝑔. Последњи слој има само један неурон са вредношћу оцене циљне промњиве. Случај класификације је мало другачији, последњи слој има онолико неурона колико има класа и да се инстанца класификује у класу која припада неурону са највећом вредношћу. Последњи слој користи следећу векторску активациону функцију:

Код бинарне класификације уместо два излаза и функције довољно је користи један излаз и сигмоидну функцију. Архитектура таквог типа спада у логистичку регресију над претпоследњим слојем мреже. У зависности од саме примене, потребно је поставити одговарајућу функцију грешке. На пример, случају регресије то би била квадратна грешка:

Када нам је потребна метода класификације одређеног броја класа, прво треба одредити одговарајућу репрезентацију категоричке променљиве. Обично таква презентација подразумева да је класа дефинисана вектором ***y*** димензије ***m***, који има ***m-1*** нула и једне јединице, при чему јединица на позицији ***ј*** ако и вектор ***y*** представља класу ***ј.*** Свака инстанца има један такав вектор ***y*** у зависности којој инстанци припада. Функција грешке која се највише користи у овом случају за категорилке циљне промењиве која се ослања на ту репрезентацијз је **унакрсна ентропија**.

Уколико желимо да неуронска мрежа врати жељене резулате и буде успешна, потребно је дефинисати и функцију грешке., а потом ускладити одговарајући алгоритам учења. Алгоритам који се користи за израчунавање градијента у случају неуронских мрежа назива се алгоритмом **пропагације уназад**. Алгоритам за израчунавање градијента се просто састоји од множења ивода тих активационих функција и компатибилних матрица и таблицом извода сложене функције. Проблем дубоких мрежа је у томе што њихове велике дубине могу довести до великог броја процеса, односно множења. Главна предност потпуно повезаних мрежа је у томе што су „агностичке структуре“, тј. нема посебних претпоставки о улазу. Иако структуре без претпоставке чини потпуно повезане мреже веома широко применљивим, такве мреже имају тенденцију да имају слабије перформансе од мрежа са посебном применом. Иако потпуно повезане мреже не дају никакве претпоставке о улазу, оне обично имају мање перформансе и нису добре за екстракцију карактеристика. Осим тога, имају већи број тежина за тренирање што резултира великим временом тренинга.

Код потпуно повезаних мрежа број улаза је унапред одређен и такве мреже имају ограничену примену у пракси. Највише се користе над подацима у виду вектора фиксне дужине. Из таквих разлога, потпуно повезане неуронске мреже тешко је применити над подацима попут слика различитих размера и секвенцама које се користе у биоинформатичким истраживањима. Када говоримо о подацима у табеларним формама, потпуно повезане мреже је могу дати изузетне резултате.

# Конволутивне неуронске мреже

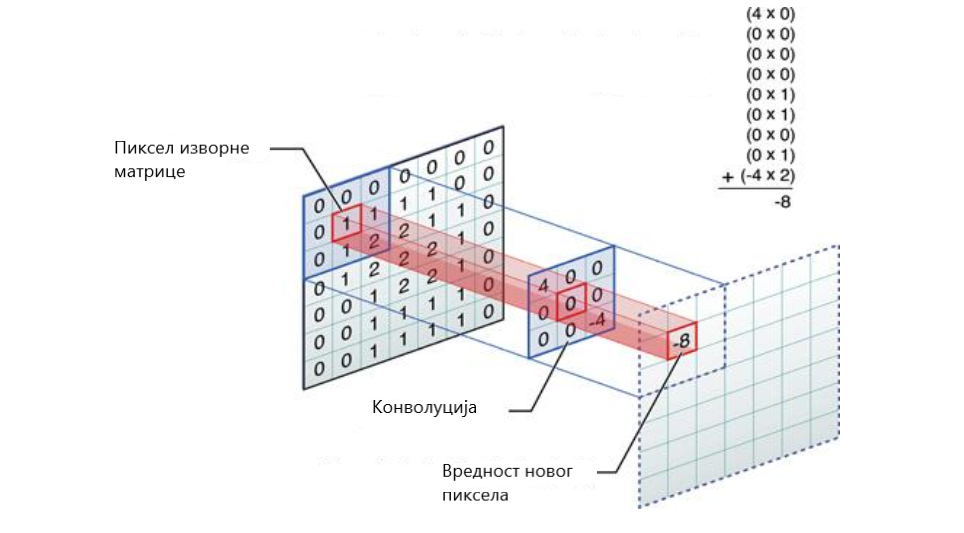
Конволутивне неуронске мреже су веома сличне порпуно повезаним неуронским мрежама: ​​састоје се од неурона који имају своје тежине и претпоставке које се могу научити. Сваки неурон прима своје улазе, изводи тачкасти производ и опционо га прати са нелинеарношћу. Цела мрежа и даље изражава једну диференцибилну функцију резултата: од необрађених пиксела слике на једном крају до резултата класе на другом. И још увек имају функцију губитка као код потпунo повезаних неуронских мрежа (нпр. Softmax) на последњем (потпуно повезаном) слоју и све што смо до сада рекли за учење потпуно повезаних мрежа и даље важи. Захваљујући оваквој архитектури, догодили су се велики помаци у сфери дубоког учења. Сама жеља за њиховим учењем дошла је из потребе за практичном применом над подацима у виду слика. На пример, за одређивање и препознавање објеката на сликама.

Потпуно повезане неуронске мреже не скалирају добро величину слике. У *CIFAR-10* скупу података, који је најчешће коришћен скуп у машинском учењу, слике су димензија 32х32х3. Што би значило да су слике слабије резолуције (32 широка, 32 висока и 3 канала у боји). Такав скуп нам дозвољава да ефикасно примењујемо различите алгоритме у зависности од проблема који треба решити. Узимајући овај случај, порпуно повезани неурон у свом скривеном слоју имао би тежину која би износила вредност од 3072 (32\*32\*3 = 3072). Јасно је да се овом тежином може управљати, али ова потпуно повезана структура се не скалира на веће слике. Уколико би слика имала димензије 200х200х3, то би значило да би неурон имао тежину од 120,000 (200\*200\*3). Са таквим начином, неурони би се преоптеретили и не би добили добре резулате. Закључак се намеће да је потпуно повезаност нескалабилна и велики број параметара би довео до преоптерећења (*overfitting*). Зато су нам неопходне методе трансформације.

Трансформације су методе које од слика производе нову слику истих или сличних димензија. То значи да можемо променити различите параметре слика као што су димензије, ивице, осветљење, контраст, изоштреност, итд. Производ таквих трансформација су стандардизоване слике, које се могу користити као својства на основу које би модел могао да учи. Различите трансформациј слика могу се вршити операцијама са две матрице. Једну матрицу би одредили као улазни сигнал, односно слику, а друга матрица би користили као филтер, који одређује процес који треба да се изврши.

Дефиниција конволуције матрица: „Конволуција матрица А ∈ R*m x n*  и B ∈ Rp x q је матрица димензија a (𝑚−𝑝+ 1)×(𝑛−𝑞 + 1) која означава производ А \* B и дефинише на следећи начин:

Конволуција је матричнa операцијa, која се састоји од језгра, матрице малих тежина, која прелази преко улазних података вршећи множење по елементима са делом улаза на коме се налази, а затим сабира резултате у излаз. У математичкој анализи, конволуција је математичка операција са две функције (f  и g) која производи трећу функцију (f \* g). Општи облик за матрицу конволуција је дат тако што, конволуција омогућава дељење тежине. То омогућава смањење броја ефективних параметара и дозвољава да се исти параметар слике открије у различитим деловима улазног простора. Оваква метода конволуције је кључна тачка конволутивних неуронских мрежа.



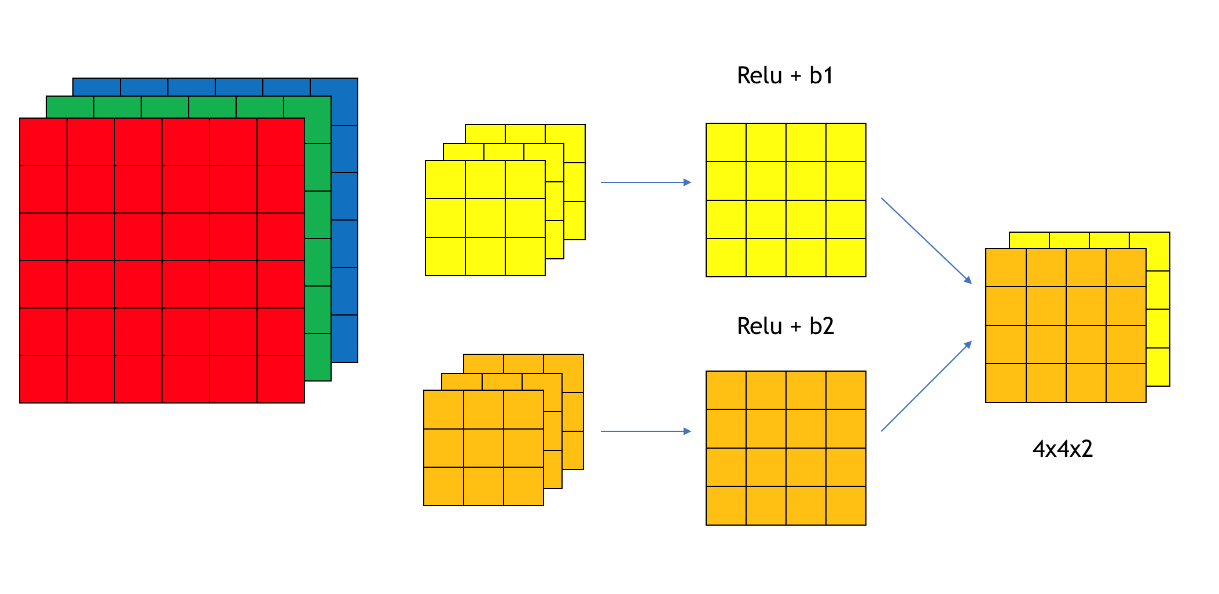
Илустрација 9: Конволутивне операције над матрицама

У општем случају матрице не морају бити квадратне. Конволуција за свако **𝑖** и **𝑗** израчунава скаларарни производ елемената матрице ***B*** са одговарајућим деловима матрице ***А*** по формули. Матрица ***В*** помера се дуж матрице  ***А*** и рачуна скаларне производе на различитим пољима и на тај начин добијамо резултат конволуције.

Често се може догодити да ни ексеперти који праве избор слика и дизајнирају релевантне трансформације, нису довољно компетентни да дефинишу довољно добре филтере за неке одређене задатке. Идеја која би могла да реши такву проблематику, имплементира се тако што би се модел могао дефинисати тако да свој саставни део научи и трансформише, а не само да донесе закључак на основу излаза. То би се могло дефинисати основна идеја конволутивних мрежа. Матрицу ***В*** посматрамо као филтер приликом израчунавања конволуције, која се креће дуж матрице ***А*** и на тај начин добијамо скаларни производ на различитим локацијама. Сваки скаларни производ могуће је узрачунати засебним неуроном. На такав начин могуће је применити процес конволуције на групи неурона. Оно што је битно, то је да сви неурони из те групе морају имати исте параметре филтера.

Конволутивна мрежа смањење грешака својих излаза на датим улазим постиже обучавањем. Филтери конволутивне мреже могу бити научени, што значи да би било потребно дефинисати архитектуру и структуру мреже која то може успешно да реши. Улази се трансформишу низом конволуција, активационим функцијама и агрегације које се међусобно преплићу, а да је производ тога порпуно повезана неуронска мрежа, која може да врши завршну класификацију улаза. Да би процес конволуције у неуронским мрежама био успешан, неопходно је имати више улазних канала. На тај начин, улаз не представља увек матрицу већ може представљати тродимензионални низ који представља неку врсту тензора. Једноставније је тродимензионални низ представити као колекцију дводимензионалних тензора тј. матрица, које називамо каналима. Сада, улазни тензор и филтер биће тродимензионални тензор, а излаз ће бити представљен као обична матрица.

Неуронска мрежа у сваком свом слоју учи одређени број матрица које представљају филтере. Слој који чини већи број филтера називамо конволутивни слој мреже. Конволутивни слој даје као излаз тензор који одговара излазу појединачног филтера. Сваки филтер има своју функцију и параметре по којима врши филтрацију, нпр. Један филтер служи да уочи вертикалне ивице, други хоризонталне и трећи косе. Нелинеарном активационом функцијом се трансформише сваки канал као и код порпуно повезаних неурноских мрежа. После једног конволутивног слоја могуће је надовезати наредни слој. Тако да виши конволутивни слојеви процесирају податке које су процесирали нижи слојеви.



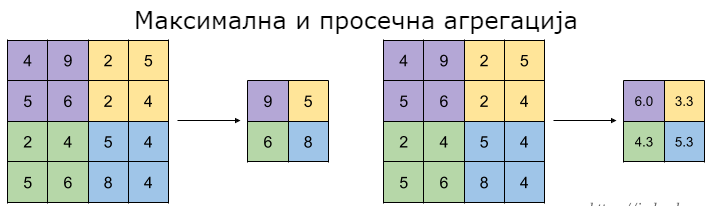
Илустрација 10: Конволутивни слој

Поред конволутивног слоја потребно је споменути и слој агрегације (*pooling layer*). Овај слој има главну улогу у удруживању параметара у циљу смањења количине израчунавања и учинили неке од детекција више робусним. Реализација решења је једноставна. Потребно је да сваки канал улазног тензора подели на делове одређених димензија који се замене једном вредношћу попут максимума вредности у том тензору. Производ високих вредности у тензору значи да је филтер који је произвео ту вредност када је нашао образац који тражи. Процесом агрегације губи се информација на ком месту се налазио максимум. Колики је то значај те информације, зависи искључиво од премене.

Сада ћемо процес агрегације објаснити преко примера. Процес је прилично једноставан. Улаз који је приказан има димензију 4х4 и на слици су приказана са процесом рачунања максимума и просека. У примеру, филтер има димензију 2х2 и помера се за два поља, због тога је потребно улазну матрицу раставити на 4 улаза са димензијом од 2х2. За добијање максималне агрегације, потребно је израчунати максималну вредност улаза из сваке под регије. Принцип за добијање просечне агрегације је исти само је потребно израчунати просечну вредност. Разлика између конволуције и агрегације је та да се максимум не рачуна на преклапајућим деловима, већ се дели на дисјунктивне подрегије који одговарају димензијама агрегације. Када улазни слој има више од једног канала, сваки слој се обрађује независно. У овом примеру је свака димензија слике дељива одговарајућом димензијом слика (*n x n*). У случају када не би била дељива димензија слике са димензијом филтера, колона или врста која представља остатак тог дељења не би била обрађена. На пример, када је димензија улазног канала 4х3, а филтер димензије 3х3, последња врста би била занемарена.

Дефиниција конволутивне мреже је функција 𝑓w која је одређена на следећи начин:

где ***х*** представља улазни тензор, ***𝐿*** је број конволутивних слојева, ***h𝑖***  за ***𝑖 = 1, . . . ,𝐿*** представља излазни тензор ***𝑖***'-тог слоја, ***W𝑖***  је тензор параметара ***𝑖***'-тог слоја***, B𝑖*** тензор слободних коефицијената који по димензијама одговара тензору ***W𝑖 \* h𝑖−1***, ***𝑎***  је функција агрегације, ***𝑔*** је активациона функција***, ˜𝑔˜w*** је потпуно повезана неуронска мрежа.



Илустрација 11: Агрегација максималне и просечне вредности

# Примењене методе у овом раду

Развоју апликације за класификацију рендгенских налаза плућа могли би смо поделити у четири фазе. Прва од четири фазе је сакупљање и обрада података, односно приказивање и анализа података кроз графике. Друга фаза се састоји од креирања и тренинга модела за класификацију слика. У трећој фази приступамо методи пренесеног учења, користећи унапред обучени модел са замрзнутим слојевима, што значи да ће тежине остати не промењене. Четврта фаза рада је прецизно подешавање. У овој фази последњи слојеви претходно обученог модела нису замрзнути. Програмски код је писан је у програмском језику Пајтон (Python) користећи Керас (Keras) и остале библиотеке.

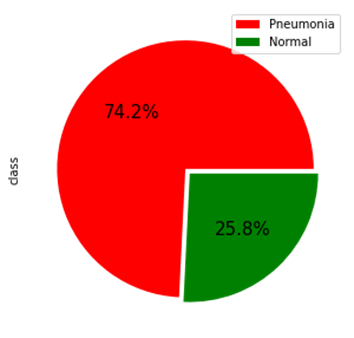
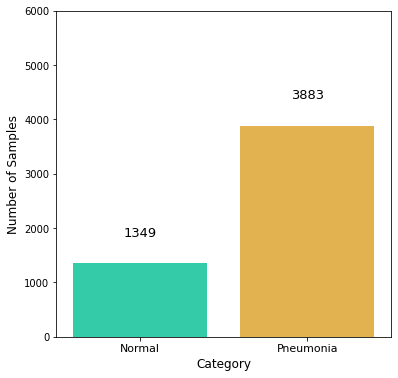
## Прикупљање података

Овај сет података садржи 5,856 рендгенских снимака у облику слика. Слике су подељене на тренинг и тест сет података од различитих пацијената. Слике у свом називу имају врсту упале као на пример: *NORMAL/BACTERIA/VIRUS,* затим насумично додељен идентификациони број пацијента и број слике пацијента.

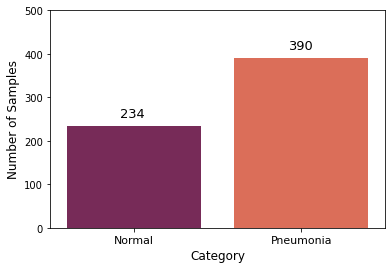
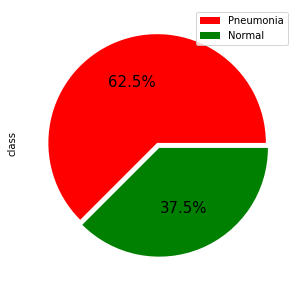
Рендгенске слике грудног коша одабране су из ретроспективних група педијатријских пацијената. Ови рендгенски снимци рађени су као део рутинске контроле и клиничке неге пацијента.

Сви рендгенски снимци су првобитно испитани и прегледани ради контроле квалитета и уклоњни су сви снимци ниског квалитета. Дијагнозе на основу слика су успоставили два стручна лекара пре него што је добијена дозвола за коришћење и обуку у интелигентним системима. Како би отклонили могућност грешке, било је потрено да провери сет података и трећи стручни лекар.

Ови подаци се користе како би се развио модел који омогућава бинарну класификацију предикцијом за свако патолошко стање. Овај сет података је распоређен у фолдере за тренирање модела који се састоји од 5,232 слике распоређене у два фолдера под називом *NORMAL* и *PNEUMONIA.* У фолдеру *NORMAL* налази се 1,349 слика које представљају плућа без упале, док у фолдеру *PNEUMONIA* се налази 3,883 слика које представљају плућа са видљивим симптомима упале плућа. Фолдер за тестирање модела садржи 624 слике које су распоређене такође у два фолдера под називом *NORMAL* и *PNEUMONIA* вођене истим принципом расподеле.

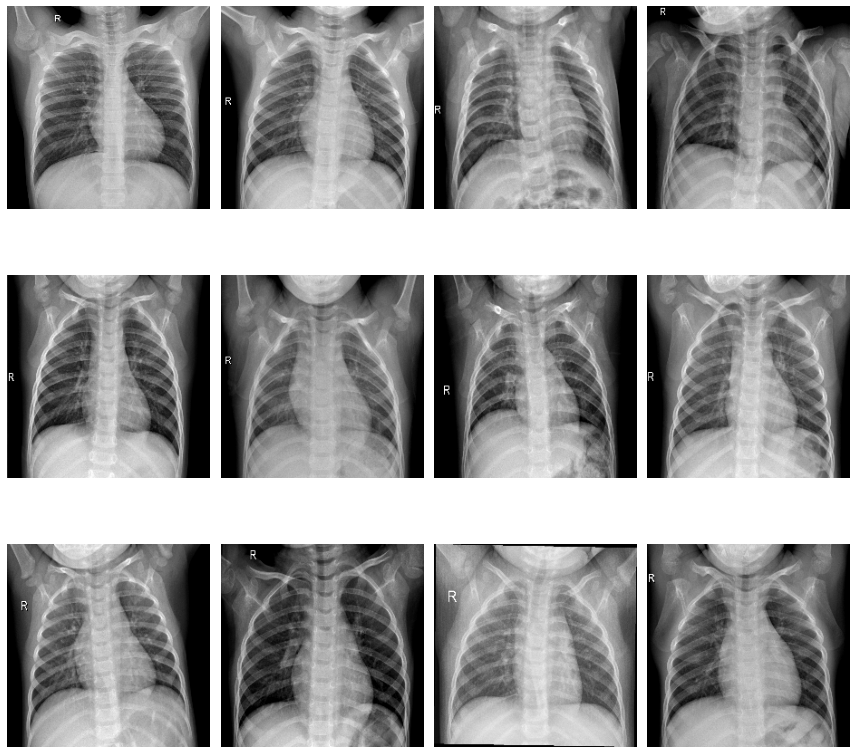


Илустрација 12 и 13: Приказ броја узорака и однос категорије NORMAL и PNEUMONIA у проценитима из фолдер за тренирање мoдела

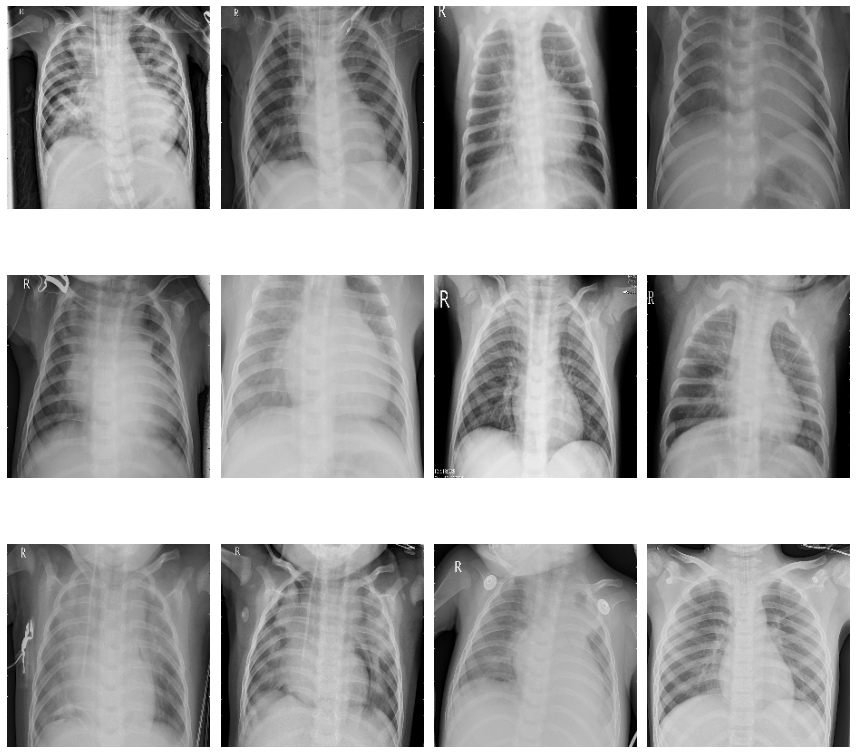
 

Илустрација 14 и 15:Приказ броја узорака и однос категорије NORMAL и PNEUMONIA у проценитима из фолдер за тестирање мoдела

На овим приказима можемо закључити да ови скупови података се мало разликују једно од других. Оба скупа су мало неуравнотежена, постоји више узорака из категорије са пнеумонијом, уз то да је скуп података за тренирање модела мало више неуравнотежен. На следећим сликама приказани су узорци из оба скупа података.



Илустрација 16: Узорци из скупа података за тренирање модела без упале



Илустрација 17: Узорци из скупа података за тренирање модела са упалом

# Припрема података

Пре креирања и тренирања модела за класификацију слика, потребно је прво формирати скуп за валидацију. Потребно је своје улазне податке расподелити у подскупове за обуку, валидацију и тестиранње како би се спречио *overfitting* у процесу учења и да би могли да ефикасно проценимо наш модел. *Оverfitting* се дешава када модел научи претерано добро до те мере да негативно утиче на перформансе модела на новим подацима. То значи да се лоши делови узорка науче као што су шум и остале нестабилности у подацима и користе као ваљани концепти у оквиру модела. Овакав начин процеса негативно утиче на способност модела да се генерализују. Вероватноћа за оваквом појавом је већа код непараметарских и нелинеарних модела који имају већу флексибилност приликом учења циљне функције. Како би смо боље схватили разлику између скупа за тренинг, валидацију и тестирање, треба обратити пажњу на одломак из текста из Риплијеве књиге „Препознавање узорака и неуронске мреже“ (Pattern Recognition and Neural Networks). Ови скупови податак су дефинисани на следећи начин:

„*– Сет за обуку: Скуп примера који се користе за учење, односно да одговарају параметрима класификатора.*

*– Скуп за валидацију: Скуп примера који се користе за подешавање параметара класификатора, на пример за избор броја скривених јединица у неуронској мрежи.*

*– Тестни скуп: Скуп примера који се користи само за процену перформанси потпуно одређеног класификатора.“ 1*

У овом случају користимо класичну поделу оригиналог скуп података за тренирање. Делимо скуп на 80% за стварну обуку и 20% за потребе валидације. Затим је потребно учитати слике из фолдера и припремити их за обраду.

Почињемо са одређивањем генератора података. Са *Keras Image Data Generator*, можемо променити и прилагодити величину пиксела и применити насумичну трансформацију за повећање података. Користимо два различита генератора, један за тренинг – *train\_imggen*, други за валидацију – *val\_imggen*.

Ове генераторе примењујемо на сваки скуп података како би све слике биле стандардизоване, помоћу функције *flow\_from\_dataframe*. Ова метода је склопу *Keras* библиотеке, али је уз то потребно и инсталирати *Pandas* библиотеку. Осим трансформација дефинисаних у сваком генератору, слике се такође мењају на основу параметара у *target\_size.* Величина слике постављена је на 224 (*image\_size* = 224), *batch* износи 32, а *seed* 42.

# Дефинисање модела

После припреме података спремни смо на следећу фазу пројекта, а то је креирање и обучавање модела за класификацију слика.

Пре креирања модела, потребно је поставити повратне позиве односно *callbacks.* Креирање модела без њих је као вожња возила без функционалних кочница. Имамо малу или никакву контролу над целим процесом који ће вероватно довести до нежељењног исхода.

Из *Keras* документације можемо дефинисати: „*Повратни позив (callback) је скуп функција које се примењују у датим фазама процеса обуке. Можете користити повратне позиве да бисте добили преглед интерних стања и статистике модела током обуке.“ 2*

Повратни поозиви (*callbacks*) највише користимо када желимо да аутоматизујемо неке процедуре након сваке обуке или епохе и помажу вам да имате контролу над целокупним процесом обуке. То значи да имате бољи увид када је неопходно зауставити тренинг када се достигне конкретна вредност за тачност или за губитак функције, затим након сваке успешне епохе чува се одређена вредност контролне тачке, прилагођава се стопа учења током процеса и још много тога. У овом пројекту користио сам *EarlyStopping* функцију. Овој функцији можете доделити различите параметре и аргументе које се могу мењати у зависности када је погодно да се процес обуке заустави. Неке од метрика које се користе су:

*monitor*: вредност коју је потребно пратити - val\_loss

*patience*: број епоха без побољшања након којих ће обука бити прекинута, у нашем случају је то 5 епоха

*min\_delta:* минимална промена у надгледаној вредности. На пример, *min\_delta* =1 значи да ће процес обуке бити заустављен ако је апсолутна промена надгледане вредности мања од 1, код нас је доста мања вредност 0.0000001

*restore\_best\_weights:* подешавамо ову метрику на Тачно (*Тrue*) ако желиmo да задржиmo најбоље тежине када се заустави процес

Други повратни позив је *ReduceLROnPlateau*. Овај повратни позив смањује стопу учења када метрика престане да се побољшава. Такође се прати количина и ако се не види напредак за одређен број епоха, стопа учења се смањује.

# Креирање модела

После успешног дефинисања модела, потребно је креирати модел за класификацију слика. У овом делу пројекта модел је пројектован од нуле. Први слој модела је улазни слој који је дефинисан и има за аргументе висину и ширину од 224 пиксела и 3 канала тј. три боје (црвена, зелена плава), односно формат слике је 224 х 224 *RGB (red, green, blue).* У овом случају је укупан број улаза 150 528 (224 х 224 х 3) за улазни слој. Може се користити и аргумент *None,* у колико димензија има променљиву величину. Вишеслојне неуронске мреже користе улазни вектор слике и трансформишу га кроз низ скривених слојева, користећи нелинеарне активационе функције. Сваки скривени слој је такође у потпуности повезан са свим неуронима у претходном слоју. Последњи слој је такође потпуно повезан и представља коначну излазну класификацију. Потребно је додати сакривене слојеве у мрежу. Постоји много типова слојева који се користе за изградњу конволутивне мреже, али они који су се користили у овом моделу подељени су у три блока и то су следећи слојеви:

* Конволутивни слој – користили смо *Keras* методу *Conv2D*(). Конволутивни слој је основни део блока конволутивне неуронске мреже. Аргументи ове методе су *filters, kernel\_size* и *padding. Filters* представља број излазних филтера у конволуцији, овом примеру је вредност 16 у првом блоку, у другом је 32 и у трећем 64. *Kernel\_size* је цео број или листа од два цела броја, који одређују висину и ширину прозора 2Д конволуције. Може бити један број за исте вредности просторне димензије, код нас је вредност 3. *Padding* може имати вредности *“valid”* и *“same”.* Вредност у овом моделу је *“valid”* што значи да нема допуњавања.
* Активациони слој – користили смо *Keras* методу *Activation()*. Као активациону функцију користили смо ‘*relu*’.
* Aгрегациони слој – дефинисали смо овај слој методом MaxPool2D(). Узима максималну вредност преко улазног прозора за сваки канал улаза. Прозор се помера корацима дуж сваке димензије. Пошто смо користили у конволутивном слоју *“valid”* за вредност *padding* онда није било потребе навести димензије овог прозора.

# Референце

1 -  Brian Ripley, страна 354, [Pattern Recognition and Neural Networks](https://amzn.to/2Y1s76G), 1996

2 - https://keras.io/api/callbacks/