Arhitekture dubokog učenja u područjima primjene

Dubinska analiza podataka

11. predavanje

Pripremio: izv. prof. dr. sc. Alan Jović

Ak. god. 2023./2024.

Sadržaj

- Klasifikacija vremenskih nizova
  - ROCKET, Hydra+MultiROCKET, InceptionTime
- Klasifikacija slika
  - ResNet, YOLO, ViT
- Vizualizacija konvolucijskih mreža i transformera
- Obrada prirodnog jezika
  - BERT, RoBERTa, GPT-4, namjenske arhitekture
- Generiranje slika iz teksta
  - Stabilna difuzija

Analiza vremenskih nizova

ROCKET

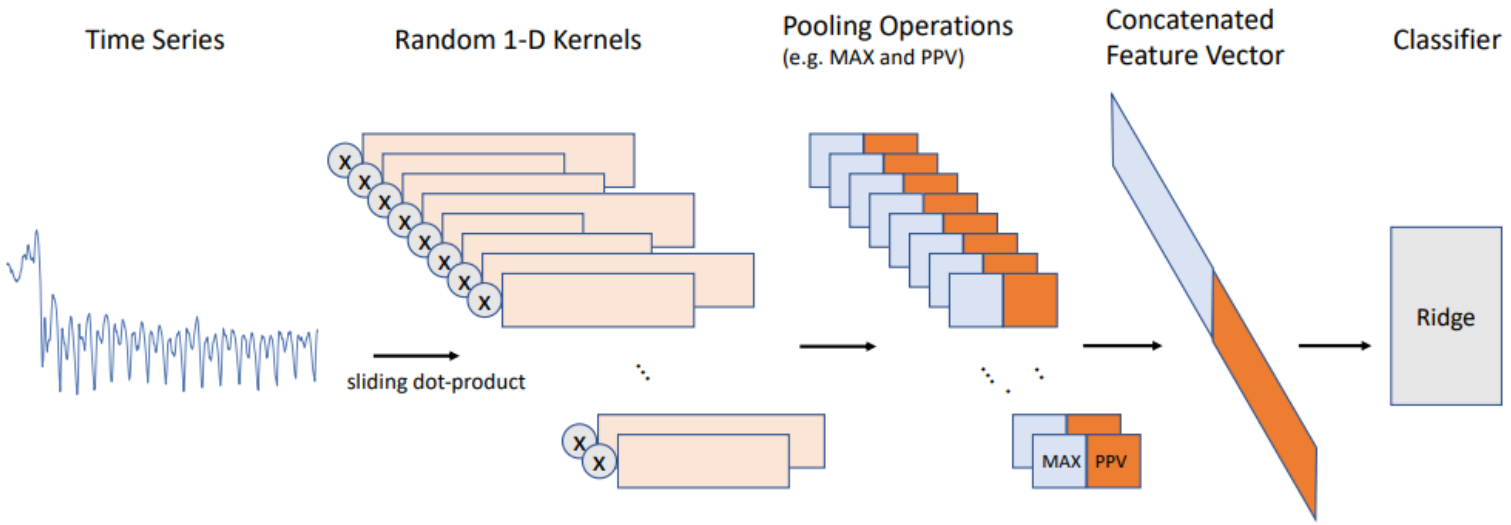
- ***R**and**O**m **C**onvolutional **KE**rnel **T**ransform*, ROCKET, *A. Dempster et al. 2020*
- *State-of-the-art* algoritam za **klasifikaciju vremenskih nizova** (univarijatnih ili multivarijatnih) po pogledu točnosti i brzine izgradnje modela na skupu za učenje
- Iako nije neuronska mreža, izgradnja modela je **motivirana uspjehom dubokih konvolucijskih neuronskih mreži** u klasifikaciji slika i drugih vrsta podataka, može se smatrati jednoslojnom konvolucijskom mrežom
- ROCKET **transformira** vremenski niz s pomakom 1 koristeći veliki broj slučajnih konvolucijskih jezgri (kernela)
  - Jezgre imaju **slučajnu**: duljinu vektora (7, 9 ili 11), težine (norm. razdioba [0,1]), *bias* (unif. razdioba [-1,1]), dilataciju (široka skala, detalji u radu) i dopunu (slučajan izbor ima li ili ne neki broj nula na početku i kraju vremenskog niza)
  - Jedini hiperparametar ROCKET-a je **broj slučajnih konvolucijskih kernela**, pri čemu ima *defaultnu* vrijednost *k* =10 000 i uglavnom ju nije potrebno mijenjati
- Transformirane značajke se koriste za učenje jednostavnog klasifikatora (*ridge* regresija – L2 regularizacija ili logistička regresija, ovisno o veličini – za probleme s manje primjeraka od broja značajki koristi se *ridge* regresija)

A. Dempster, F. Petitjean, G. I. Webb. 2020. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery , Vol. 34, 5 (2020), 1454--1495

ROCKET

- Četiri stvari razlikuju ROCKET od konvolucijskih slojeva u konvolucijskim neuronskim mrežama i od prethodnih pristupa:
  - ROCKET koristi veliki broj **slučajnih kernela i ne uči težine kernela** kao neuronske mreže
  - ROCKET koristi **raznolike kernele**, što je različito od tipičnih konvolucijskih neuronskih mreža gdje je uobičajeno da grupe kernela (po slojevima) imaju istu veličinu, dilataciju i dopunu
  - ROCKET naglašeno koristi **dilataciju kernela**, uzorkuje se slučajno za svaki kernel – proizvodi velik raspon dilatacija, čime se obuhvaća informacija s različitim frekvencijama i vremenskim skalama, što je važno za perfomanse metode
  - ROCKET koristi **dvije značajke na transformiranim nizovima**:
    1) maksimalnu vrijednost transformiranog niza (MAX) (slično *max poolingu*),
    2) značajku proporcije pozitivnih vrijednosti (PPV), koja omogućuje klasifikatoru da uoči dani obrazac u nizu (slično metodama vremenskih nizova zasnovanima na rječnicima – npr. shapeleti i sl.)

ROCKET



Ridge regresija:

Funkcija gubitka

$$\min_{\mathbf{w},b_0} \left(\|\mathbf{y} - (\mathbf{X}\mathbf{w} + b_0)\|^2 + \lambda \|\mathbf{w}\|^2 \right)$$

Rješenje za težine na skupu za učenje:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{y} - b_0)$$

- Za multivarijatne nizove, dimenzija kernela se slučajno generira, radi se višestruki skalarni produkt s ulaznim vektorom i računa ukupni max i ppv
- Ridge regresija se može koristiti za klasifikacijske probleme ako odluku o klasi donosimo na temelju toga je li dobiveni *y* manji (klasa 0) ili veći (klasa 1) od 0.5, a za višeklasne probleme po principu jedan protiv svih

Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: a review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery, 1-74.

Hydra+MultiROCKET

- Kombinira klasifikatore Hydra i MultiROCKET u jedan ansambl, *A. Dempster et al. 2023*
- **Hydra** – koristi slučajno postavljene konvolucijske kernele za transformaciju vremenskog niza kao i ROCKET, ali podijeljene u ***g* grupa po *k* kernela**
  - U svakoj od *n* točaka vremenskog niza, mjeri se aktivacija svakog kernela (duljine 9 uzoraka, normalna razdioba između 0 i 1) pojedine grupe s vremenskim nizom i određuje se kernel koji se najbolje poklapa s nizom (ima najveći rezultat konvolucije), njegov se brojač uveća za 1 – postoji *k*-dimenzionalni vektor brojanja za svaku grupu, ukupno *g* x *k* značajki
  - Primijenjuje *d* dilatacija (2^0, 2^1, 2^2, … , *n*), za svaku dilataciju računa *g* x *k* značajki
  - Primijenjuje se na izvorni vremenski niz i na vremenski niz prvih razlika vrijednosti izvornog niza
  - Hiperparametri su *g* (default je 64) i *k* (default je 8)
  - Na izlazu za 2 x *d* x *g* x *k* koristi linearni klasifikator

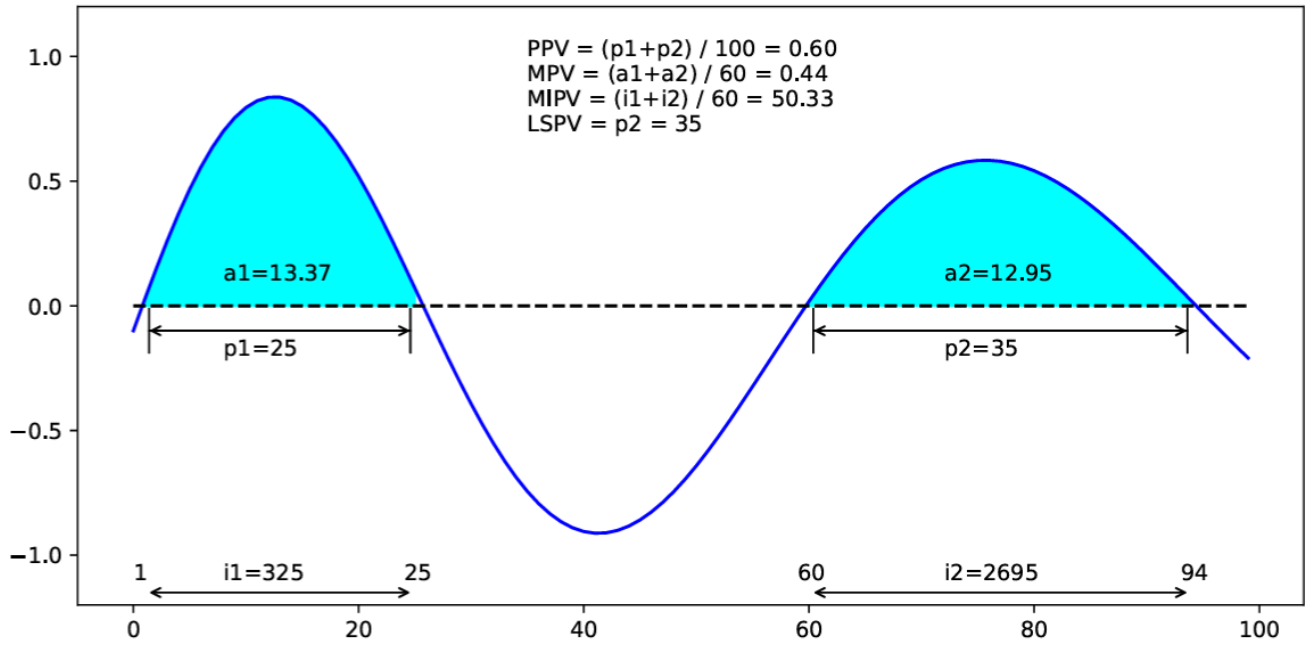
Dempster, A., Schmidt, D.F. & Webb, G.I. Hydra: competing convolutional kernels for fast and accurate time series classification. *Data Min Knowl Disc* **37**, 1779–1805 (2023). https://doi.org/10.1007/s10618-023-00939-3

Hydra+MultiROCKET

- **MultiROCKET** (*Tan et al. 2022*)
  - Koristi kernele duljine 9 (šestero vrijednosti iznose fiksno -1, a troje 2) za transformaciju vremenskog niza, što rezultira u ukupno mogućih različitih $k = 84$ kernela
  - Generira veći broj dilatacija (fiksno od 1 do veličine niza) i *biasa* (kvartili iz izlaza konvolucije za slučajni uzorak za učenje), rezultirajući u ukupno 6.216 konvolucijskih kernela za svaki niz duljine *n;* pomak pri transformaciji niza iznosi 1
  - Računa 4 značajke po kernelu: PPV, srednja vrijednost pozitivnih vrijednosti (MPV), srednja vrijednost indeksa pozitivnih vrijednosti (MIPV) i najduži potez pozitivnih vrijednosti (LSPV)
  - Radi transformacije na izvornom nizu i nizu prvih razlika
  - Ukupnih 6.216 x 4 x 2 ≈ 50.000 značajki ulaze u linearni klasifikator

Hydra+MultiROCKET

- Vizualizacija (ilustracija) računanja značajki za MultiROCKET:

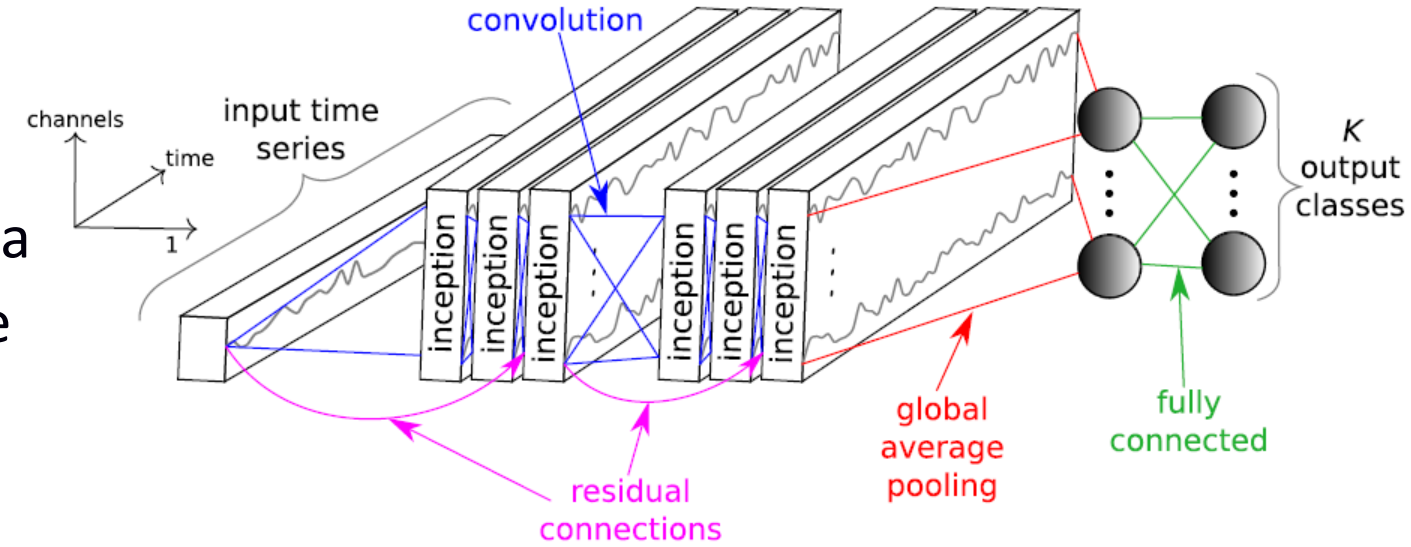


Izvor: Tan, C.W., Dempster, A., Bergmeir, C. et al. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Min Knowl Disc* 36, 1623–1646 (2022). https://doi.org/10.1007/s10618-022-00844-1

Neka je izlaz konvolucije jedne jezgre duljine Z = 100 (duljina vrem. niza je isto 100). p1 i p2 su brojevi pozitivnih vrijednosti izlaza konvolucije, a1 i a2 su sume pozitivnih vrijednosti, i1 i i2 su sume indeksa pozitivnih vrijednosti. Budući da je p2 dulji niz pozitivnih vrijednosi, LSPV = p2

InceptionTime

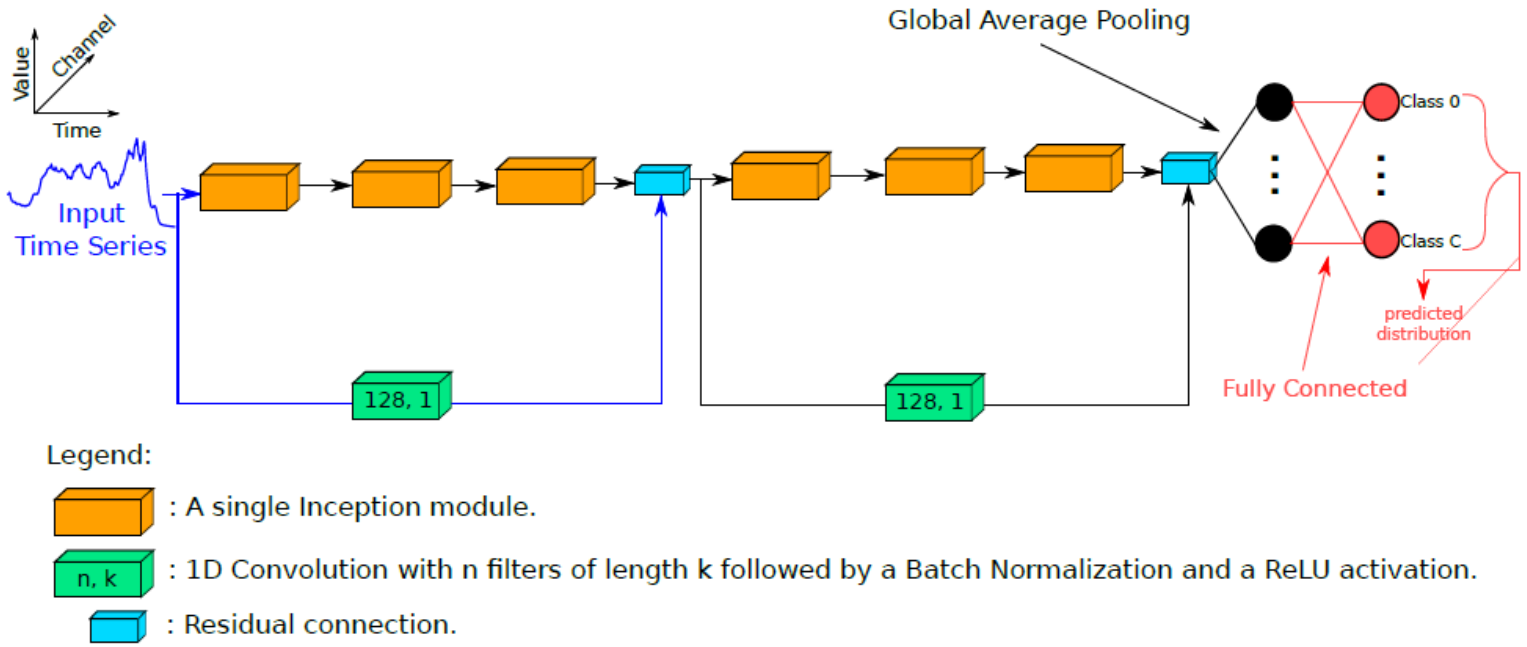
- InceptionTime (*Fawaz et al. 2020*) je trenutačno najtočnija arhitektura dubokih neuronskih mreža za klasifikaciju općenitih vremenskih nizova (točnija od “običnih” potpuno konvolucijskih mreža i ResNeta)
- Ansambl koji se sastoji od 5 konvolucijskih mreža s istom arhitekturom, ali inicijaliziranih na slučajan način, koje su zasnovane na modulima “inception”
- Većinski glasa o klasi



Ismail Fawaz, H., Lucas, B., Forestier, G. *et al.* InceptionTime: Finding AlexNet for time series classification. *Data Min Knowl Disc* **34**, 1936–1962 (2020). https://doi.org/10.1007/s10618-020-00710-y

InceptionTime

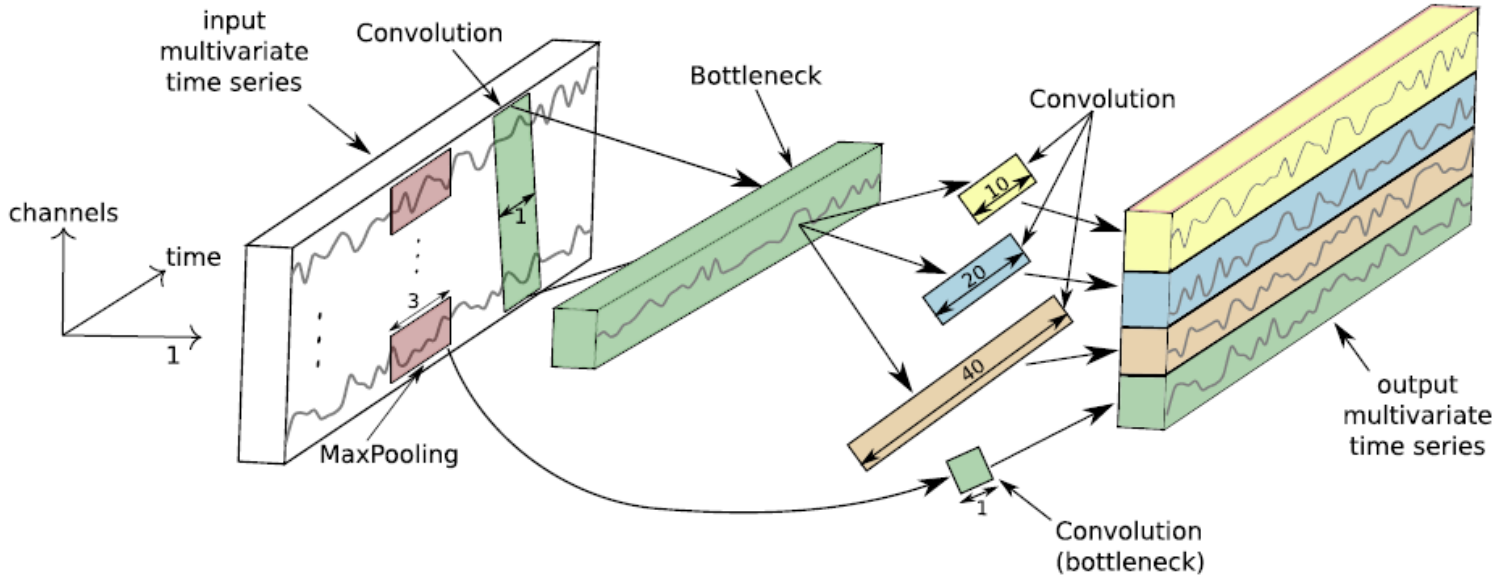
Izvor: Ismail-Fawaz A, Devanne M, Weber J, Forestier G (2022) Deep learning for time series classication using new hand-crafted convolution lters. In: 2022 IEEE International Conference on Big Data (IEEE BigData 2022), pp 972-981



- Mreža se sastoji od dva slijedna rezidualna bloka, svaki s tri modula *inception*, a rezidualne veze koriste se za rješavanje problema nestajućeg gradijenta
- Modul *Inception* najprije primjenjuje sloj koda (*bottleneck*) da bi smanjio dimenzionalnost multivarijatnog vremenskog niza (s *M* dimenzija na m << *M*) (ako je broj dimenzija *M* malen, *bottleneck* može i povećati dimenziju, jer je krajnji cilj izvući korisne informacije)
- Potom primjenjuje veći broj konvolucijskih filtara različitih veličina tako da uhvati vremenske značajke na različitim skalama

InceptionTime

Izvor: Ismail Fawaz, H., Lucas, B., Forestier, G. et al. InceptionTime: Finding AlexNet for time series classification. Data Min Knowl Disc 34, 1936–1962 (2020). https://doi.org/10.1007/s10618-020-00710-y



- *Bottleneck* je sloj koji se dobiva pomicanjem konvolucije s *m* filtara duljine 1 i korakom duljine 1 po ulaznim nizovima
- Primjena *m* filtara duljine 10, 20 i 40 (težine filtra su različito inicijalizirane), korak duljine 1 na sloju *bottleneck*
- Dodatno sažimanje najvećom vrijednošću (širine 3) na izvornom nizu + *bottleneck* (konvolucija s *m* filtara duljine 1)
- Izlaz iz modula *Inception* dobiva se **konkatenacijom** rezultata primjene filtara i sažimanja s najvećom vrijednošću
- Svaki modul *Inception* slučajno inicijalizira 4 x *m* filtara (za tri duljine filtara + za sažimanje), filtri se potom uče

Klasifikacijski algoritmi – performance

- Usporedba najboljih algoritama za TSC

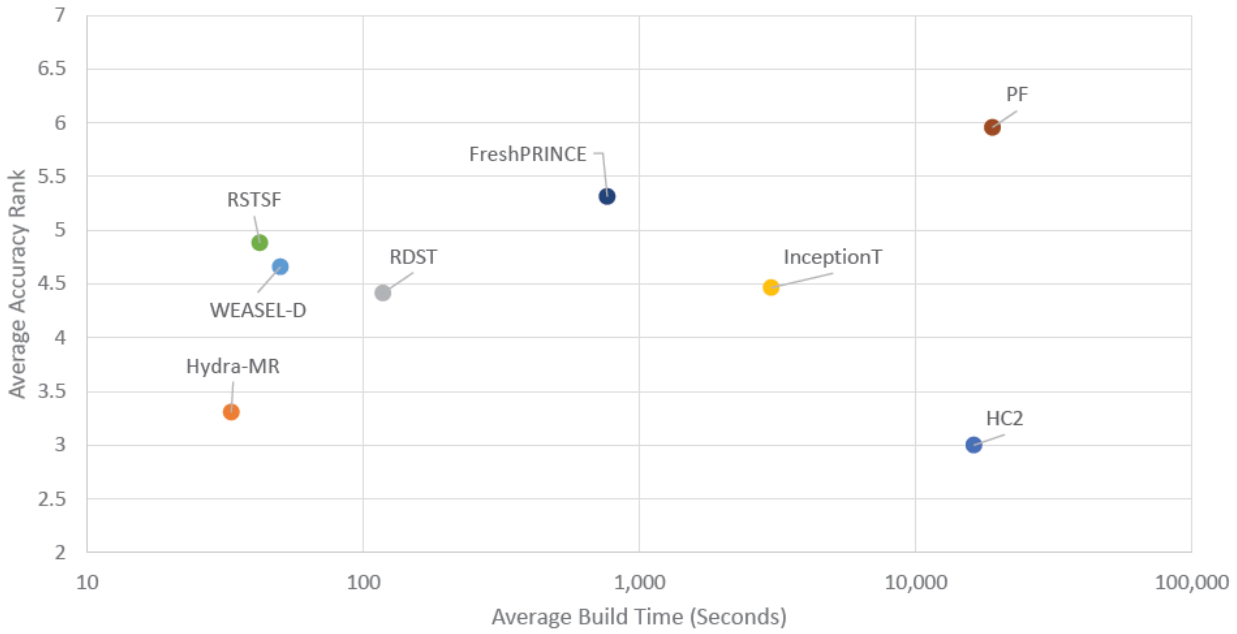


Table 18 Average rank of classifiers on 30 resamples of 142 TSC problems split by problem type.

	DEVICE (11)	ECG (7)	IMAGE (34)	MOTION (27)
HC2	**2.000 (1)**	**2.143 (1)**	3.441 (2)	**2.759 (1)**
Hydra-MR	2.455 (2)	2.571 (2)	**2.912(1)**	3.056 (2)
InceptionT	4.455 (4)	4.286 (3)	4.676 (5)	3.833 (4)
RDST	3.909 (3)	4.571 (5)	3.971 (4)	3.722 (3)
WEASEL-D	5.364 (7)	4.286 (4)	3.706 (3)	4.741 (5)
RSTSF	5.091 (6)	5.429 (6)	5.324 (6)	5.926 (6)
FreshPRINCE	4.909 (5)	5.571 (7)	5.676 (7)	6.019 (8)
PF	7.818 (8)	7.143 (8)	6.294 (8)	5.944 (7)
	SENSOR (35)	SIMULATED (12)	SPECTRO (12)	
HC2	**3.414 (1)**	3.333 (2)	**2.167 (1)**	
Hydra-MR	3.957 (2)	**3.083 (1)**	3.792 (3)	
InceptionT	4.200 (4)	3.917 (4)	5.958 (8)	
RDST	4.871 (5)	5.667 (7)	5.083 (5)	
WEASEL-D	4.900 (6)	6.833 (8)	4.083 (4)	
RSTSF	4.071 (3)	5.167 (6)	3.417 (2)	
FreshPRINCE	5.171 (7)	3.750 (3)	5.667 (6)	
PF	5.414 (8)	4.250 (5)	5.833 (7)	

Izvor: Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: a review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery, 1-74.

Klasifikacija slika

Primjenska područja za klasifikaciju slike

- **Medicinska dijagnostika**
  - Detekcije i klasifikacije anomalija: rendgen, CT, MRI, ultrazvuk, histologija, slike kože...
- **Prepoznavanje, detekcija i identifikacija objekata i osoba**
  - Detekcija objekata i klasa objekata: autonomna vožnja, sigurnosni sustavi, robotika
- **Kontrola kvalitete**
  - Detekcija i klasifikacija kvarova i nekonzistentnosti proizvoda u industriji i proizvodnji
- **Nadgledanje okoliša**
  - Detekcija i klasifikacija promjena u okolišu: agrikultura, šumarstvo, arhitektura
- **Umjetnost**
  - Identificiranje starosti, autora, stila različitih umjetničkih djela (slike, kipovi)

Razlozi zašto su konvolucijske mreže dobre za slike

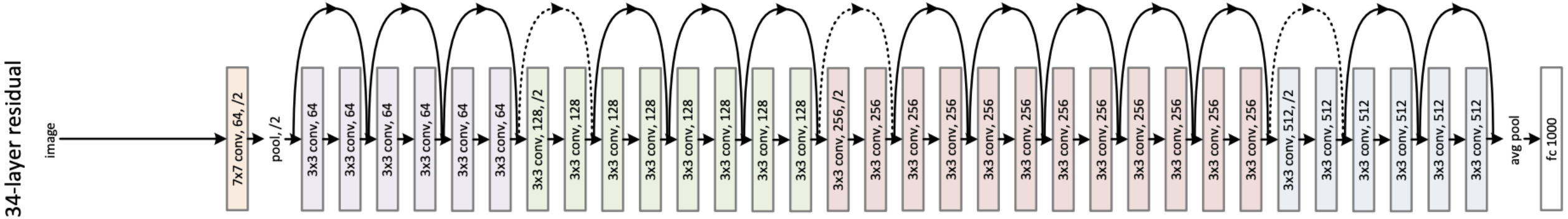
- Konvolucijske mreže su specifično oblikovane da obrađuju **prostornu strukturu slikovnih podataka**, za razliku od višeslojnog perceptrona koji tretira sve ulazne vrijednosti kao neovisne varijable
- Konvolucijske mreže mogu naučiti **hijerarhijsku reprezentaciju ulaznih podataka** – uče sve kompleksnije značajke kroz unutarnje slojeve, što je korisno za klasifikaciju slika kod kojih se objekti sastoje od više manjih dijelova
- Konvolucijske mreže koriste **slojeve sažimanja** da bi poduzorkovale reprezentaciju slike, što smanjuje dimenzionalnost podataka i čini arhitekturu računski učinkovitom
- Konvolucijske mreže su **prilagodljive problemu** i danas su **prvi izbor za klasifikaciju objekata na** slikama za većinu primjena, no imaju i nedostatke koji se pokušavaju riješiti alternativnim arhitekturama

ResNet

- *Deep Residual Learning for Image Recognition*, ResNet, *K. He et al., 2015*
- Obitelj arhitektura dubokih konvolucijskih neuronskih mreža s mnogo slojeva – ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152...
- Najvažnija primjena – klasifikacija slika (tu je slabija samo u odnosu na najbolje transformerske modele), ali se i kombinira s drugim metodama za detekciju i segmentaciju objekata
- Arhitektura je zasnovana na **rezidualnim blokovima**
  - Sastoje se od nekoliko (2-3) konvolucijskih slojeva s ReLU aktivacijom i normalizacijom podskupa podataka
  - Povezani su s drugim blokovima **rezidualnim** (**skip**) vezama
  - **Izlaz iz svakog bloka može se dodati ulaznim podacima** (koji su preskočili blok/ove putem rezidualne veze) kao ulaz sljedećeg rezidualnog bloka
  - Tijekom propagacije pogreške unazad, rezidualne veze služe da se **gradijent direktno propagira do svakog bloka**, olakšavajući učenje i otežavajući pojavu nestajućeg gradijenta
- Obično se fino podešava (engl. *fine-tuning*) za dani problem, ali može se učiti i od početka (engl. *from scratch*)

K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. Tech Report, 2015, https://doi.org/10.48550/arXiv.1512.03385

ResNet



- Dublje arhitekture, s većim brojem slojeva, obično daju bolje rezultate, ali učenje zahtijeva više podataka i jače resurse, postoje arhitekturne razlike između pojedinih članova obitelji ResNet
- Dobro se kombiniraju s drugim metodama pogodnima za detekciju i segmentaciju objekata, npr.
  - **Brža regijska konvolucijska neuronska mreža** (engl. *Faster Region Convolutional Neural Network*, *Faster-RCNN*)
  - **Konvolucijske mreže s mogućnošću deformacije jezgre**

S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS, vol. 28, 2015.

ResNet – arhitekturne razlike članova obitelji

- Postoje manje ili veće arhitekturne razlike između pojedinih članova obitelji ResNet po pitanju veličine filtra, broja filtara i broja podslojeva u jednom sloju:

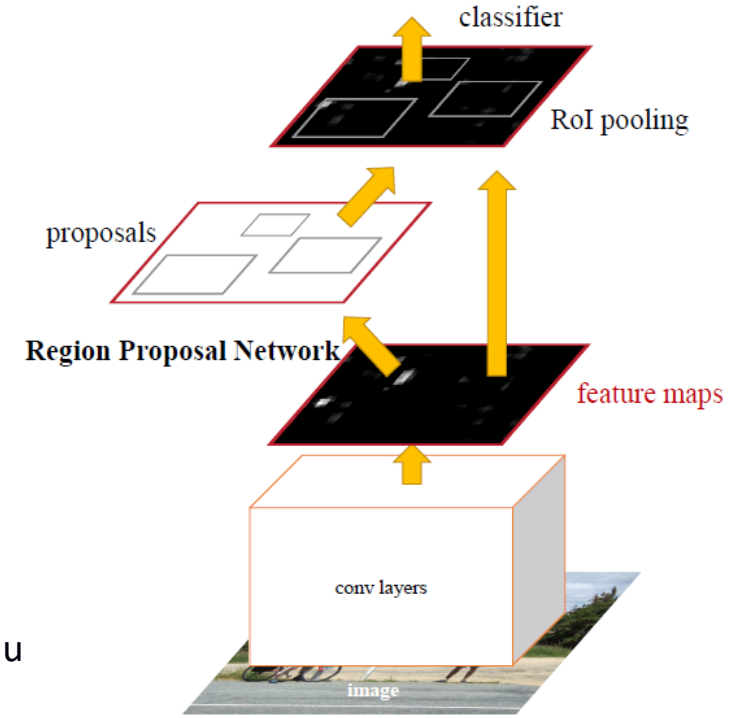
layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

fc – fully connected

https://datagen.tech/guides/computer-vision/resnet-50/

Brža regijska konvolucijska neuronska mreža

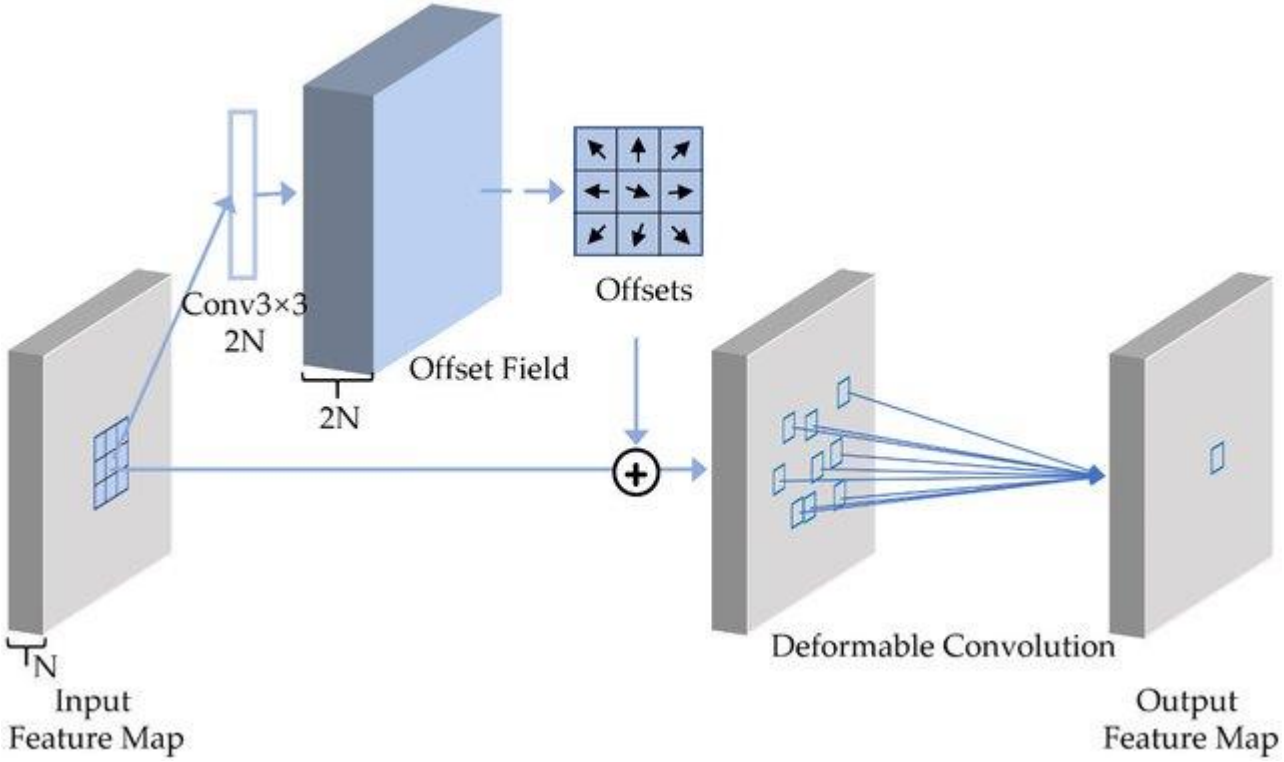
- *Faster Region Convolutional Neural Networks*, Faster-RCNN, Ren et al. 2015.
- Rješenje *end-to-end* za učinkovitu detekciju i klasifikaciju objekata u slikama
- Kombinacija dvije glavne komponente: **mreže za prijedlog regije** i **mreže za detekciju**
- Iz slike se **značajke na početku izlučuju nekom dubokom konvolucijskom mrežom** (npr. ResNet-101) – pritom se ResNet tu naziva *backbone* mrežom
- Mreža za prijedlog regije – CNN koja generira potencijalne kandidate okvira objekata na temelju značajki (koordinate vrhova)
  - Koristi predefinirane sidrene okvire (engl. *anchor boxes*) različitih skala i omjera stranica na slučajnim mjestima koji služe kao predložak
  - Tijekom učenja pomiče prozor po slici, uči pomake sidrenih okvira i mjeri koliko je izgledno da je u regiji objekt
- Mreža za detekciju – potpuno povezani slojevi koji rade klasifikaciju objekata u predloženim okvirima s određenom vjerojatnosti
  - Klasifikator koristi značajke fiksne širine izlučene iz okvira različitih dimenzija putem RoI – sloja sažimanja



Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, Extended tech report, 2016, https://arxiv.org/abs/1506.01497

Konvolucijske mreže s mogućnošću deformacije jezgre

- *Deformable Convolutional Networks*, 2017.
- Konvolucijske jezgre **adaptivno mijenjaju mjesta u receptivnom polju** na koja djeluju
- Uči se **skup pomaka** (engl. *offsets*) za svaku prostornu lokaciju unutar jezgre
- Operacija konvolucije je uobičajena, ali s definiranim pomacima
- Omogućuje bolju detekciju objekata koje imaju varijacije ili deformacije
- Kombinira se s ResNetom u njegovim zadnjim slojevima da bi se napravio detektor objekata



J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable Convolutional Networks, 2017, https://doi.org/10.48550/arXiv.1703.06211
S. Wang et al. Automatic Detection and Classification of Steel Surface Defect Using Deep Convolutional Neural Networks. Metals. 11. 388. 10.3390/met11030388.

YOLO

- *You Only Look Once*, YOLO, Redmon et al. 2016
- Brza i točna neuronska mreža koja u **jednom prolazu** detektira objekte u stvarnom vremenu
- Slika se dijeli na mrežu nepreklapajućih ćelija i u svakoj ćeliji predviđa se postojanje objekta kao i koordinate *bounding boxa*
- *Bounding boxovi* se uče na temelju **odmaka u dimenzijama** (x, y, širina, visina) od većeg broja referentnih sidrenih okvira (*anchors*) za svaku ćeliju
- Više objekata može biti detektirano u pojedinoj ćeliji
- Ako je objekt veći od pojedinih ćelija, onda se objekt detektira unutar one ćelije koja je u centru objekta
- Detekcija objekata se formulira **kao regresijski problem** u kontekstu predviđanja x, y, širine i visine **odmaka** od sidrenih okvira
  - Klasifikacijski gubitak računa se prema odmaku vjerojatnosti klasa od stvarnih objekata (gubitak unakrsne entropije)
  - Tijekom učenja, teži se minimizirati regresijski i klasifikacijski gubitak

- YOLOv3 model, introduced by Redmon et al. in 2018
- YOLOv4 model, released by Bochkovskiy et al. in 2020,
- YOLOv4-tiny model, research published in 2021
- YOLOR (You Only Learn One Representation) model, published in 2021
- YOLOX model, published in 2021
- NanoDet-Plus model, published in 2021
- PP-YOLOE, an industrial object detector, published in 2022
- YOLOv5 model v6.1 published by Ultralytics in 2022
- YOLOv7, published in 2022

J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016 pp. 779-788. doi: 10.1109/CVPR.2016.91
https://viso.ai/deep-learning/yolov7-guide/
https://ijicis.journals.ekb.eg/article_233993_4f08100cd6662d26912e245c0861001b.pdf

YOLOv7

- Ideja mreže YOLOv7 je poskupiti učenje korištenjem **većeg broja heuristika i optimizacija** (*bag-of-freebies*) za veću točnost, ali zato postići vrlo brzo zaključivanje
- Fokus na:
  - Učinkovitijem načinu učenja mreže – razmatra se detaljno kako se gradijent treba propagirati između slojeva i blokova mreže
  - Robusnijoj funkciji gubitka (*deep supervision*) – koristi se i pomoćna funkcija gubitka unutar mreže u plićim slojevima
  - Dodatnim optimizacijama (npr. *batch normalization* nakon svakog konv. sloja unutar sloja RepConv i sl.)
    - RepConv (*Receptive Field Block Convolutional Layer*) – specifičan način kombiniranja više konvolucija 1x1, 3x3, 1x1 u blok za dobivanje informacije na više skala
  - I dr.
- Danas je najnovija mreža YOLOv8 – nešto brža i točnija u detekciji manjih objekata od YOLOv7

- YOLOv3 model, introduced by Redmon et al. in 2018
- YOLOv4 model, released by Bochkovskiy et al. in 2020,
- YOLOv4-tiny model, research published in 2021
- YOLOR (You Only Learn One Representation) model, published in 2021
- YOLOX model, published in 2021
- NanoDet-Plus model, published in 2021
- PP-YOLOE, an industrial object detector, published in 2022
- YOLOv5 model v6.1 published by Ultralytics in 2022
- YOLOv7, published in 2022

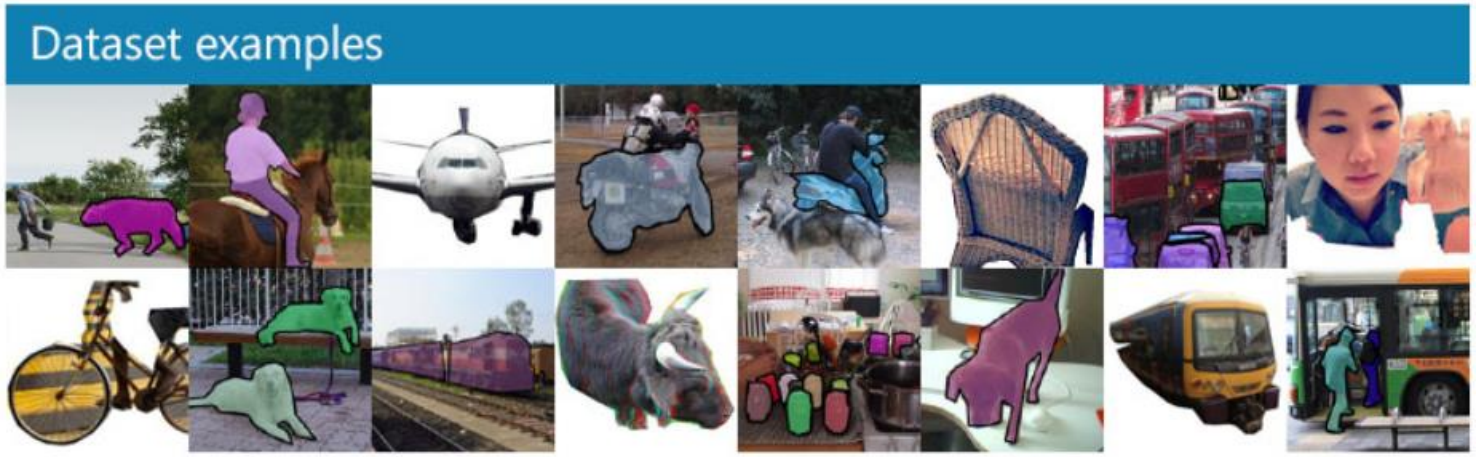
Chien-Yao Wang and Alexey Bochkovskiy and Hong-Yuan Mark Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2207.02696, 2022., https://arxiv.org/pdf/2207.02696.pdf
https://viso.ai/deep-learning/yolov7-guide/
https://ijicis.journals.ekb.eg/article_233993_4f08100cd6662d26912e245c0861001b.pdf

Skup podataka COCO

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints



- Suradnja većeg broja sveučilišta i tvrtki u SAD-u (Caltech, CMU, Google...)
- Koristi se za izgradnju modela za segmentaciju i detekciju objekata, npr. **YOLOv7** je učen isključivo na COCO-u

https://cocodataset.org/#home

Transformerske arhitekture za klasifikaciju slika

- *Vision Transformers*, ViT, Google Research 2020, 2021
- Kod ViT-a, slika se tretira kao **slijed fragmenata** (engl. *patches*) fiksne veličine bez preklapanja i transformerska arhitektura se koristi za obradu ovih fragmenata za klasifikaciju – koristi se samo **koderski** dio
  - 2D fragmenti se izravnaju u vektor vrijednosti
  - Vektorima fragmenata se doda informacija o poziciji u vidu vektora s kodiranom pozicijom
  - Ovakvi vektori se koriste kao **tokeni za transformer**
  - Mehanizam samopozornosti uči odnose između fragmenata (slično kontekstu kod teksta)
  - Nakon samopozornosti koristi se sloj sažimanja za agregaciju informacija iz više fragmenata
  - Na kraju koderskog dijela koristi se potpuno povezani sloj i unakrsna entropija za klasifikaciju
- Pristup je dosta dobar ako se koriste vrlo veliki skupovi podataka za učenje, na manjim skupovima podacima rezultati mogu biti slabiji od CNN-a, pristup se može fino podešavati na manjim skupovima s visokom uspješnosti
- Najbolji rezultati za većinu skala dobivaju se kombiniranjem transformera i CNN-ova, vidjeti npr. CvT

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 9th International Conference on Learning Representations, ICLR 2021.
Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang; CvT: Introducing Convolutions to Vision Transformers, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 22-31

Vizualizacija konvolucijskih mreža i transformera

Objašnjavanje dubokih modela

- Primjenjuju se ***post-hoc* metode** za objašnjenje nakon što je mreža već izgrađena
  - Najčešće korisno inženjerima za debugiranje modela i pronalazak objašnjenja zašto model radi onako kako radi, rjeđe se koristi za objašnjenje rezultata modela krajnjem korisniku
  - Globalne metode objašnjavanju model u cjelini (za sve primjerke), lokalne objašnjavaju model za pojedini primjerak (češće su)
- Metode ciljaju na vizualizaciju nekog dijela (sloja) mreže kako bi se ustanovila povezanost dijelova reprezentacije – značajki (dijela slike, teksta) s ciljnom predviđenom labelom (klasom)
- Pridruživanje predviđene labele nekom dijelu slike ili teksta naziva se **atribucija** (engl. *attribution*)

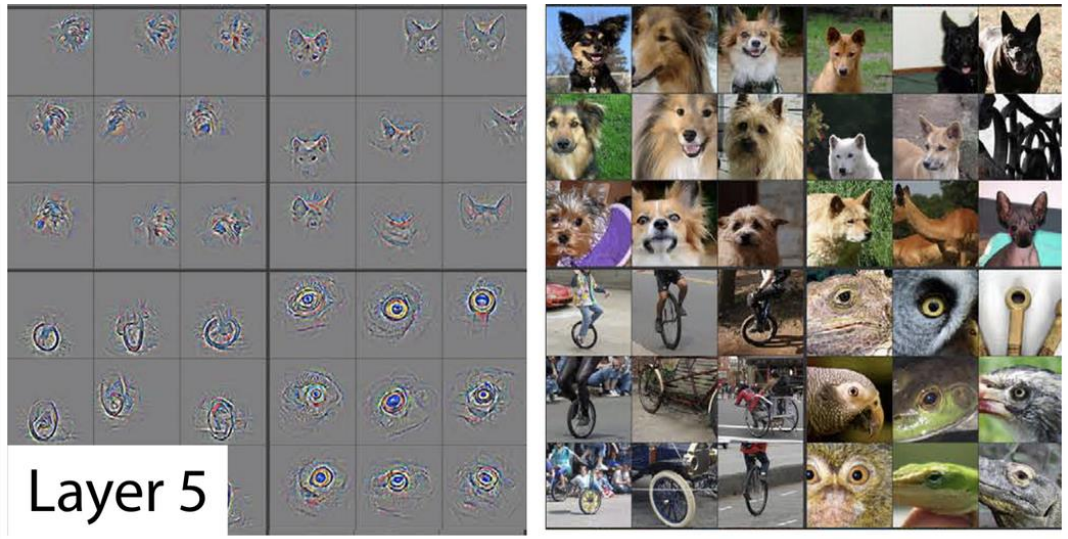
Klasifikacija metoda za vizualizaciju (objašnjenje) modela

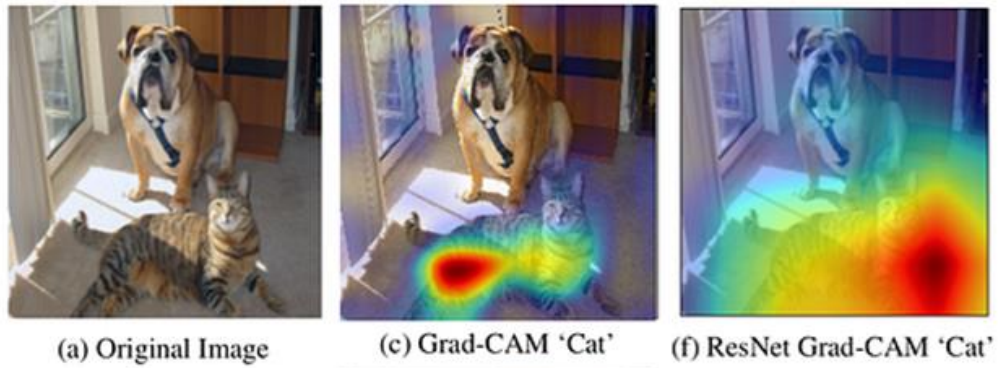
- Vizualizacija konvolucijskih mreža
  - **Atribucija aktivacije na razini sloja** (engl. *layer activation attribution*)
    - Vizualizacija aktivacije nakon određenog konvolucijskog sloja
  - **Atribucija zasnovana na gradijentima** (engl. *gradient-based attribution*, *sensitivity maps*)
    - Mape istaknutosti (engl. *saliency maps*), Grad-CAM, integrirani gradijenti (engl. *integrated gradients*)
  - **Atribucija zasnovana na perturbaciji** (engl. *perturbation-based attribution*)
    - Ablacija značajki (engl. *feature ablation*), osjetljivost na okluziju, uzorkovanje Shapleyjevih vrijednosti (engl. *Shapley value sampling*), KernelSHAP
  - **Ostale atribucijske metode i hibridne atribucijske metode** – npr. DeepLift, DeepSHAP
- Vizualizacija transformera
  - **Vizualizacija mehanizma pažnje** (engl. *attention visualization*)

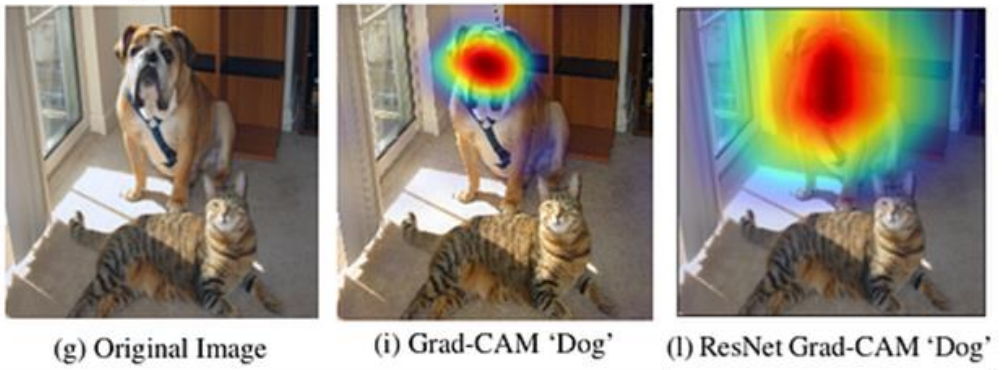
Atribucija zasnovana na gradijentima

- Ove metode koriste **gradijente u prolasku kroz mrežu unatrag** da bi računale atribucije i sve su lokalne jer funkcioniraju na razini objašnjenja jednog primjerka
- **Mape istaknutosti** (tzv. vanila gradijent)
  - Koriste apsolutne vrijednosti gradijenata (pozitivni gradijent ukazuje na povećanje vjerojatnosti predviđene klase, a negativni smanjenje) s idejom pronalaska onih piksela koje je potrebno najmanje promijeniti da bi se izlazna klasa promijenila
- **Grad-CAM** (gradijent preslikavanja aktivacije klasa)
  - Koristi gradijente skorova po klasama (prije *softmaxa*) na aktivacijskoj mapi nekog konvolucijskog sloja da bi se dobila važnost težina neurona, potom računa utežanu linearnu kombinaciju aktivacijskih mapa s težinama neurona, primijeni ReLU (da se zadrži samo pozitivan utjecaj značajki) i napravi *upsampling* na dimenziju slike za vizualizaciju

Atribucija zasnovana na gradijentima







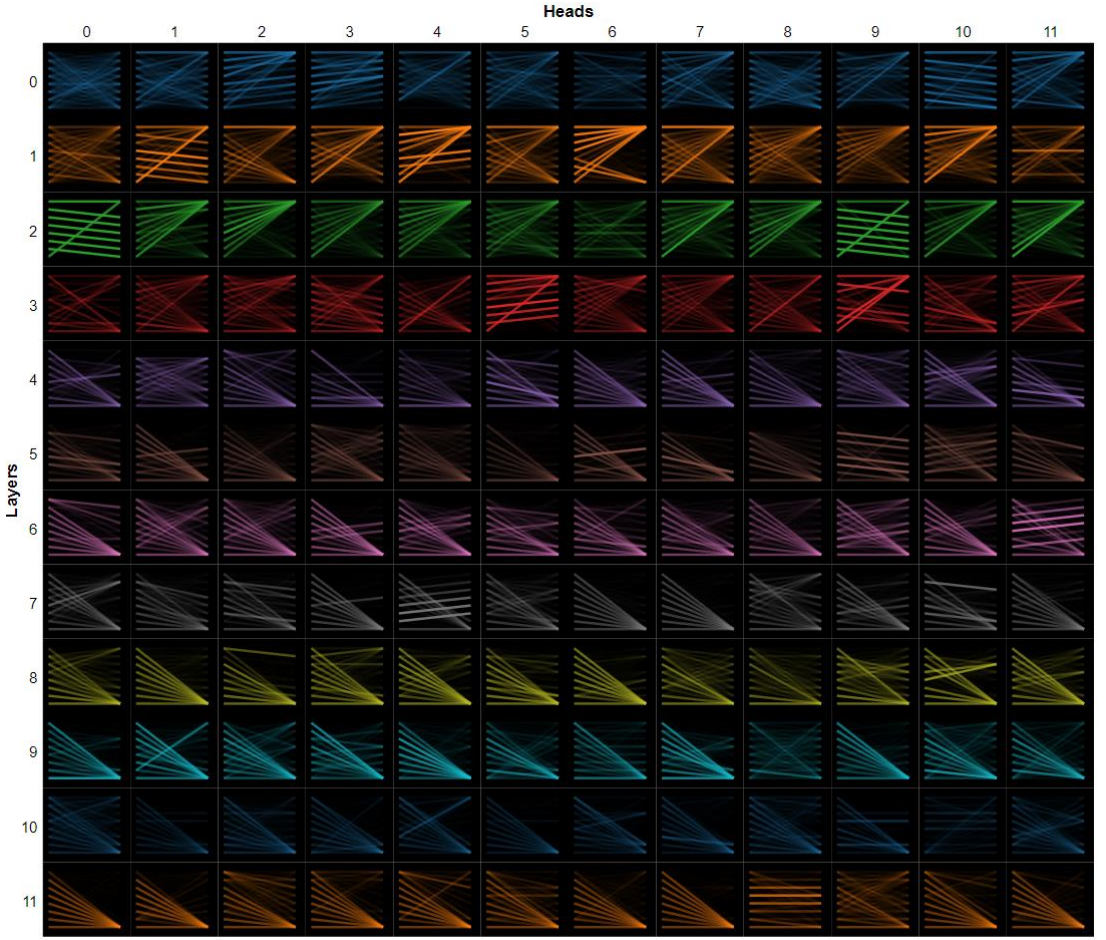
Izvor (prilagođeno): https://mrsalehi.medium.com/a-review-of-different-interpretation-methods-in-deep-learning-part-1-saliency-map-cam-grad-cam-3a34476bc24d

Atribucija zasnovana na perturbaciji

- Ove metode koriste permutacije vrijednosti piksela za procjenu koliko pojedini pikseli (ili grupe piksela pomoću primjene maski) utječu na klasifikaciju
- **Ablacija značajki**
  - Mijenja dijelove ulazne slike postavljajući vrijednosti grupa piksela na nulu, promatra razliku u predikciji klase u odnosu na izvornu sliku i označava bitnije grupe piksela svjetlijom bojom
- **Uzorkovanje Shapleyevih vrijednosti**
  - Istovremeno se permutiraju vrijednosti određenog broja piksela i razmatra se koliko se pritom izmijenila klasifikacija, broj permutacija po grupi piksela je hiperparametar

Vizualizacija mehanizma pažnje

- Služi za vizualizaciju težina u mehanizmu pažnje
- Korisno kod ustanovljavanja kako koji sloj transformera izvlači informacije iz npr. teksta (od jednostavne sintaksne povezanosti pa sve do globalne semantičke povezanosti) kako bi konačno napravio ciljni zadatak (npr. klasifikaciju teksta u više klasa)
- Vizualizacija omogućuje uočavanje jačih ili slabijih veza među tokenima u pojedinom sloju i detaljnije na pojedinoj glavi transformerskog modela
- Vidjeti BertViz

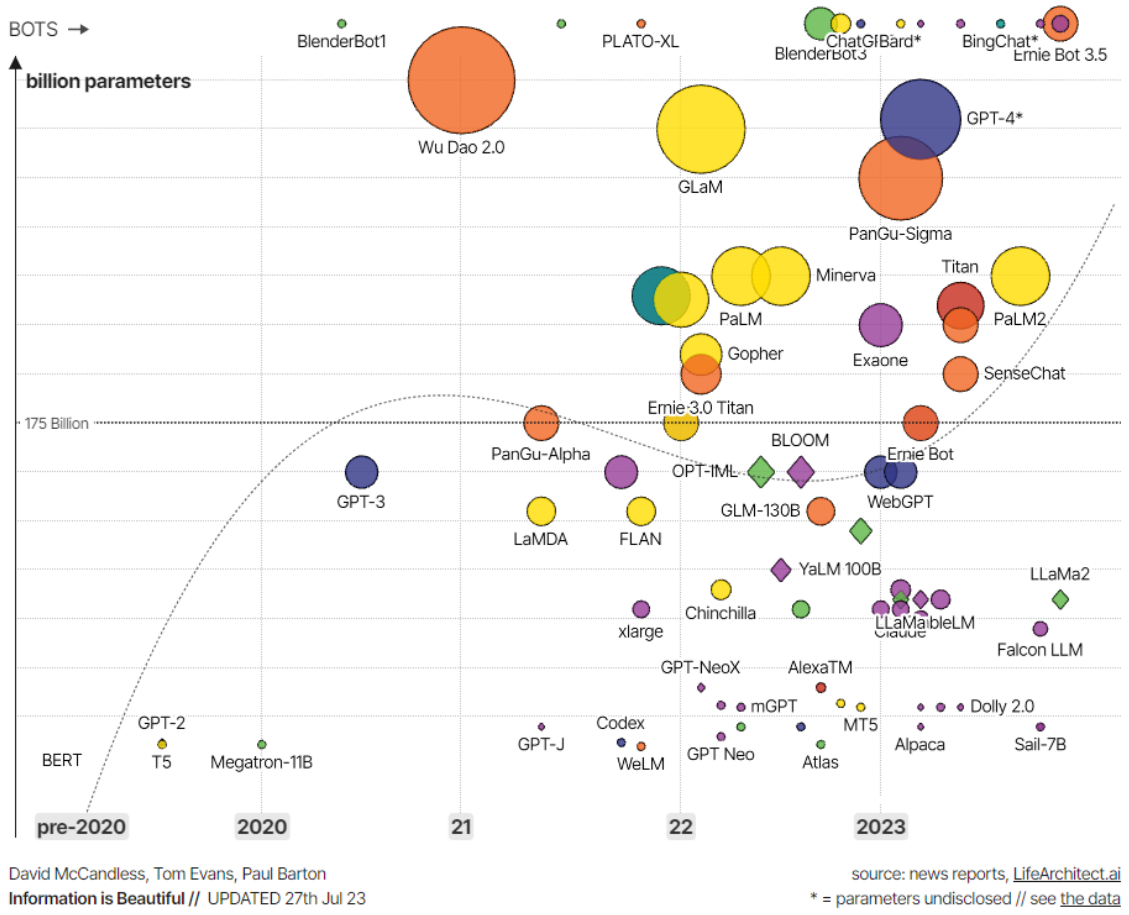


Izvor: https://github.com/jessevig/bertviz

Obrada prirodnog jezika

Obrada prirodnog jezika

- Veći broj parametara dubokog modela je bolji (u pravilu) za razumijevanje teksta
- Obrada prirodnog jezika je danas većinom svedena na varijacije **transformerske arhitekture i velike jezične modele** (bilo generativna bilo diskriminativna primjena)
- Usporedba nekih jezičnih modela u 2024.:
  - https://www.multimodal.dev/post/best-large-language-models-of-2024
- Manje arhitekture su predložene za specifične namjene (npr. mobilne platforme, ugradbene sustave i sl.)
  - **Destilacija modela**, vidjeti npr.: https://neptune.ai/blog/knowledge-distillation



Izvor: https://medium.com/@adria.cabello/the-evolution-of-language-models-a-journey-through-time-3179f72ae7eb

BERT

- *Bidirectional Encoder Representations from Transformers,* BERT, *Devlin et al. 2019*
- Prednaučeni jezični model zasnovan na **transformerskoj** arhitekturi od 12 kodera i 12 dekodera, često se koristi uz fino podešavanje, danas se koristi samo koderski dio arhitekture
- Primjena danas: klasifikacija teksta, analiza sentimenta, prepoznavanje imenovanih entiteta (NER)
- Tijekom procesa predučenja, BERT ima cilj: **modeliranje maskiranog jezika** (engl. *Masked Language Modeling*, MLM)
- BERT na ulazu prima konkatenaciju dva segmenta duljina *M* i *N* (od jedne ili više rečenice prirodnog jezika) ukupne duljine manje od hiperparametra *T* (duljina sekvence), po *defaultu* maksimum 512

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL).

BERT

- **MLM**:
  - Odabere se slučajno 15% tokena u ulaznoj sekvenci, od toga ih se 80% zamijeni s posebnim tokenom [MASKA], 10% ostaje nepromijenjeno, a 10% se zamijeni nasumično odabranim tokenom
  - MLM koristi funkciju gubitka unakrsne entropije za predviđanje vrijednosti maskiranih tokena samo na temelju njihovog konteksta
  - Slučajno maskiranje i zamjena se izvedu jednom na početku i pohrane se za trajanje cijelog procesa učenja
- Optimizator za proces učenja BERT-a je Adam (hiperparametri beta1 = 0.9, beta2 = 0.999, eps = 1e-6 i L2 *weight decay* = 0.01), uz povećanje stope učenja za prvih 10 000 iteracija do vrijednosti 1e-4 i onda linearno smanjenje
- BERT uči s *dropoutom* od 0.1 na svim slojevima i težinama sloja pozornosti, koristi aktivacijsku funkciju GELU (*Gaussian Error Linear Unit*)
- Model se naučio s podskupovima od 256 sekvenci, svakom najveće duljine 512 tokena (ali većinom 128 tokena), uz 1 milijun iteracija (otprilike 40 epoha učenja na **3.3 milijarde riječi**)
- Danas se BERT obično fino podešava za konkretni klasifikacijski zadatak

Više o Adam:u https://medium.com/ai%C2%B3-theory-practice-business/adam-optimization-algorithm-in-deep-learning-9b775dacbc9f

RoBERTa

- *A **Ro**bustly optimized **BERT A**pproach*, RoBERTa, *Liu et al. 2019*
- Jedan od poboljšanih modela u odnosu na BERT, postiže u pravilu bolje rezultate od BERT-a na različitih primjenama NLP-a
- Modifikacija postupka učenja BERT-a na više načina:
  - Veći skup podataka korišten za učenje (sa 16 GB na 160 GB) – BookCorpus, English Wikipedia, CC News, OpenWebText, Stories
  - Povećanje veličine slučajnog podskupa podataka (*batch size*) s 256 na čak 8000
  - Učenje na duljim sekvencama podataka – do 512 tokena u jednom podatku (primjerku za učenje), što čini jednu ili više kontinuiranih rečenica, BERT češće s manje tokena
  - Korištenje većeg broja različitih tokena ispod razine riječi (engl. *sub-word units*), od 30k tokena na 50k
  - Dinamična promjena parametra maskiranja primijenjena tijekom učenja, za razliku od prije početka učenja kod BERT-a
    - 10 različitih maski tijekom 40 epoha, svaki primjerak vidi 4 puta istu masku

Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," CoRR, vol. abs/1907.11692, 2019. arXiv: 1907 . 11692. [Online]. Available: http://arxiv.org/abs/1907.11692.

BERT i RoBERTa

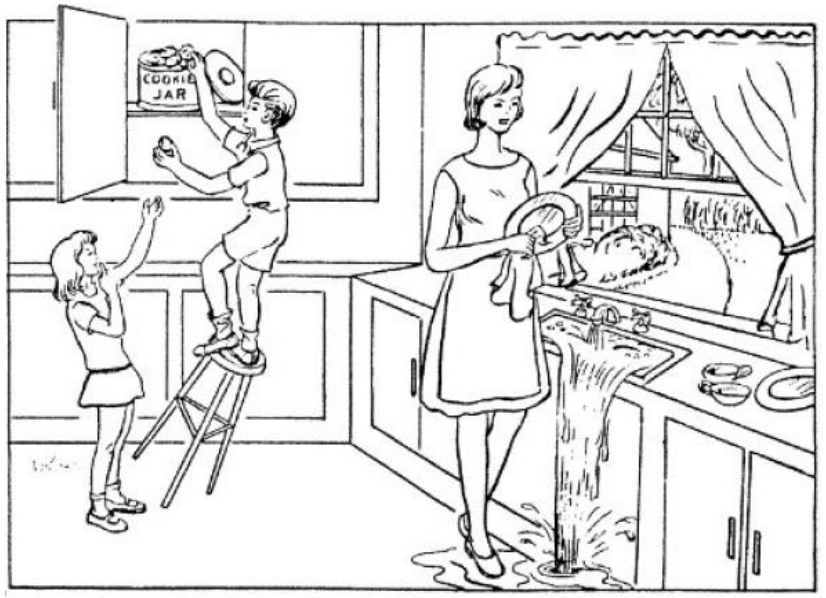
- Primjena RoBERTe: Dijagnostika demencije iz transkripata opisa slike

DATASET STATISTICS

Number of participants	Number of transcripts	Average number of transcripts per patient	Average number of characters per transcript	Average number of words per transcript
275	509	1.85	521.42	107.09

TRANSCRIPT EXAMPLE FOR DIFFERENT PREPROCESSING STEPS IN EACH OF THE EXPERIMENTS. BLUE TEXT, WHICH INDICATES REPEATED, RETRACED OR REFORMULATED SPEECH WAS USED IN THE SECOND AND THE THIRD EXPERIMENT, RED TEXT, WHICH INDICATES FILLED PAUSES WAS USED ONLY IN THE THIRD EXPERIMENT.

Transcript	Label
mhm oh I see a part of the whole kitchen is that all the kitchen or isn't it uh oh I can't read a lady a mother were in her kitchen in her kitchen doing some work I suppose and the uh there's another woman there sharing their pleasures or whatever oh have you have you checked heard of that new game that they started to play after christmas did you is a well it looks like I'd say this is well let's see it looks like oh my mother will beat me by my wife will beat me by a couple rows of this that's that's like the washing would say washing machine or let me see I can't oh that's the son come out of from school maybe or something that's a youngster there well that's just as though they getting ready to go to school or they're just coming out from school and right there he's uh same as back there except for down there in the bottom I think it's uh that's a little	Dementia
okay uh the child's falling off the chair he's taking cookies out of the jar the girl is standing on the floor uh asking for a cookie the door to the cabinet door is open mother is washing dishes the sink is overflowing the water's running uh I don't know if she's dryin em or washin em anyway and the kitchen window has curtains the window's open um it looks like a view of the back there are three dishes on the uh counter	Control



Best F1-score

BERT: 86.89%

RoBERTa: 90.28%

L. Matošević, A. Jović, "Accurate Detection of Dementia from Speech Transcripts Using RoBERTa Model," Proceedings of 45th Jubilee International Convention MIPRO 2022, pp. 1696-1702.

GPT-4

- *Generative Pre-trained Transformer*, *v4*, OpenAI, 03/2023
- GPT-4 je konverzacijski veliki jezični model zasnovan na transformerskoj arhitekturi, u nizu jezičnih modela tvrtke OpenAI (ChatGPT 3, 3.5, 4, 4 Turbo, 4 o)
- Multimodalan je u smislu da na ulazu prima tekst i/ili slike, a na izlazu odgovara tekstom
  - Koder transformira tekst i riječi u unutarnju reprezentaciju, a dekoder generira tekstni odgovor
- Kao i raniji modeli, učen je na zadatku predviđanja sljedeće riječi u rečenici
  - Veliki skup podataka za učenje, sličan onome jezičnog modela **Kosmos-1** (Microsoft):

Dataset Modality	Content Source
Text corpora	Consists of Text Data- i. **Internet data** : e.g. Wikipedia, OpenWebText2, Pile-CC, Realnews, etc. ; ii. **Academic data**: NIH Exporter; iii. **Prose Data**
Image-Caption Pairs	Consists of Embeded Image Data- LAION-2B, LAION-400M, COYO-700M
Interleaved Data	Consists of common crawl data ~2 Billion of web pages

https://openai.com/research/gpt-4

https://medium.com/@amol-wagh/whats-new-in-gpt-4-an-overview-of-the-gpt-4-architecture-and-capabilities-of-next-generation-ai-900c445d5ffe

GPT-4

- Model je učen na skupu podataka koji ponekad ima i netočne informacije, a i sam model nije savršen (iako je vrlo točan)
  - Moguć je odgovor koji dosta odudara od očekivanja kao i posve krivi odgovor (tzv. halucinacija)
- Model je poboljšavan u smislu očekivanog odgovora finim podešavanjem i to korištenjem podržanog učenja s ljudskim odgovorom
- Detalji arhitekture nisu otvoreno objavljeni, ali procjena je da ima preko 1 bilijun (engl. 1 *trillion*) parametara, Turbo verzija omogućuje veliki broj tokena u kontekstu (32 768)
- Konkurentni jezični modeli (i oni u razvoju): Claude 3 (Opus), Mixstral 8x7B (Mistral), Gemini 1.5 Pro (Google), Llama 3 70B (Meta), Kosmos 2, MAI-1 (Microsoft), Titan Text Premier (Amazon)

https://openai.com/research/learning-from-human-preferences

Namjenske arhitekture transformera

- Dosad je predloženo više tisuća različitih modela transformerske arhitekture, velika većina nerecenzirane arhitekture u znanstvenom smislu!
- Repozitorij modela (većinom) transformera: **Hugging Face** https://huggingface.co/models
- Podjela po primjenskih zadacima (engl. *downstream tasks*), najviše u područjima računalnog vida i obrade prirodnog jezika, drugdje bitno manje
- Primjeri: prevođenje teksta (npr. SeamlessM4T (Meta)), klasifikacija slika (npr. ViT base patch 16 (Google)), tekst u govor (npr. WhisperSpeech (Collabora Ltd and LAION)), itd.

Generiranje slika iz teksta

Uvod u generiranje slika iz teksta

- Zahtjevan problem zbog različite prirode ovih vrsta podataka
- Načelni koraci:
  - Skup za učenje treba sadržavati **parove tekstnog opisa slike i samu sliku**
  - **Tekstni opis treba se preslikati u vektorsku reprezentaciju** koristeći određenu metodu (npr. GloVe, Word2Vec...)
  - Gradi se **generativni model** koji nastoji izgraditi preslikavanje iz vektorske reprezentacije teksta u sliku – za to je potrebno ostvariti model preslikavanja u mreži
  - **Model se vrednuje** vizualnom procjenom kvalitete dobivenih slika za tekstni unos ili nekim objektivnim kriterijem vrednovanja (npr. nekom mjerom perceptualne sličnosti kao što je indeks strukturne sličnost – SSIM, koji mjeri osvjetljenje, kontrast i strukturnu sličnost između dvije slike)

https://medium.com/srm-mic/all-about-structural-similarity-index-ssim-theory-code-in-pytorch-6551b455541e

Stabilna difuzija

- Koristi se generativni model kao što je varijacijski autoenkoder i njegove varijante
- Stabilna difuzija modelira **šum ili slučajnost ε u latentnom prostoru** koji se koristi za generiranje novih uzoraka
- **Precizno osmišljeni difuzijski proces** se koristi za transformaciju inicijalne razdiobe šuma u niz **međurazdioba** šuma sve veće složenosti (broj međurazdioba je parametar algoritma)
  - Uzorci se generiraju iz modela počevši od osnovne razdiobe šuma pa sve do zadnje složene razine, čime se dobivaju sve različiti uzorci
  - Precizna kontrola i podešavanje razine šuma omogućuje generiranje uzoraka sa specifičnim značajkama i stilovima
- Proces počinje s Gaussovim šumom koji se transformira iz koraka u korak u sve složeniju razdiobu šuma
  - To se radi tako da se **doda malo šuma trenutačnoj razdiobi** nakon čega slijedi **korak difuzije** koji zaglađuje šum i čini ga kompleksnijim
- Difuzija je stabilna u smislu numeričke stabilnosti – metoda daje stabilne i pouzdane rezultate, pod uvjetom pažljivog podešavanja razina dodavanja šuma

Stabilna difuzija

- Najčešća primjena stabilne difuzije je u generiranju slika iz teksta
- Za to se može koristiti **uvjetni varijacijski autoenkoder** (engl. *conditional variational autoencoder* – cVAE)
- Inicijalna razdioba šuma je uvjetovana ulaznim tekstom – **pomoću teksta uči se latentna reprezentacija i inicijalna razdioba šuma**
- Utjecaj teksta na inicijalnu razdiobu šuma može se kontrolirati težinskim regularizacijski parametrom – **skalom vođenja** (engl. *guidance scale*)
- Difuzijski proces koristi se za **progresivno rafiniranje generiranih slika** s iterativnom transformacijom razdiobe šuma – svaka slika uzima se iz jedne od razdiobi šuma, a razdiobe mogu biti uvjetovane prema značajkama slika koje želimo istaknuti

Stabilna difuzija

- Primjer: **astronaut koji jaše konja**
- Svaki redak ima različitu skalu vođenja (1.1, 3, 7, 14) – omogućuje manje ili veće fokusiranje na karakteristike tekstnog opisa pri čemu:
  - manje fokusiranje – veći utjecaj šuma na model
  - veće fokusiranje – manji utjecaj šuma na model
  - jako fokusiranje nije nužno najbolje
- Svaki stupac generira sliku iz iste latentne razdiobe
- Knjižnica Diffusers: https://huggingface.co/docs/diffusers/index
- Jeremy Howard, Lesson 9: Deep Learning Foundations to Stable Diffusion, 2022 https://www.youtube.com/watch?v=_7rMfsA24Ls
- Slično: Midjourney v6: https://www.midjourney.com/

Zaključak

- Postoji mnogo arhitekturalnih varijacija osnovnih dubokih neuronskih mreža koje se koriste za razne primjene
- Brza klasifikacija velikog broja vremenskih nizova može se postići algoritmima kao što su ROCKET i Hydra+MultiROCKET
- Detekcija objekata i klasifikacija slika široko je područje u kojem se najviše koriste varijacije CNN-ova, a u novije vrijeme i transfomeri
- U obradi prirodnog jezika danas dominiraju različite varijante transformerskih modela, ovisno o namjeni
- Generiranje slika iz teksta novije je područje gdje se koriste različite varijante stabilne difuzije i generativnih modela
- Ne postoji jedna arhitektura koja je pogodna za rješavanje svih problema, ali se sve više radi na kombiniranju više izvora informacija (multimodalnost) u jedan model