

Konvolúciós Neurális Hálók és Vision Transformer architektúrák összehasonlítása felhőosztályozási feladatokra

Beszámoló a Deep Learning a gyakorlatban Python és LUA alapokon c. tárgyból

Kássa Kristóf

kassak@edu.bme.hu

Barancsuk Lilla

barancsuk.lilla@vik.bme.hu



M Ű E G Y E T E M 1 7 8 2

2023. január 16.

Bevezetés

- ▶ Felhőképek osztályozása \rightsquigarrow napelemek termelésének előrejelzése
- ▶ Neurális hálók képfeldolgozásra: Konvolúciós Neurális Hálók (CNN) vs Vision Transfromerek (ViT)
- ▶ CNN
 - ▶ Konvolúciós és pooling rétegek halmaza
 - ▶ Minták keresése „csúszó” szűrők alkalmazásával
 - ▶ A kimeneti feature mapet az utolsó réteg osztályozza
 - ▶ Hagyományosan hang, kép, videó feldolgozásra
- ▶ ViT
 - ▶ A bemeneti képet pixelsorozatként kezeli
 - ▶ Pixelek közti kapcsolat \rightsquigarrow *self-attention* mechanizmus
 - ▶ Hagyományosan szövegbányászatra alkalmazták (NLP)

Célkitűzés

- ▶ Felhőkép klasszifikálás CNN és ViT modellekkel
- ▶ A két architektúra tulajdonságainak és korlátainak megismerése
- ▶ Felépítésbeli és működésbeli különbségek vizsgálata
- ▶ Elvárt (korábbi kutatások) és saját eredmények összehasonlítása új adathalmazon
 - ▶ Számítási kapacitás igényeik
 - ▶ Futási idejük
 - ▶ Pontosságuk

Korábbi kutatások

- ▶ Transzformerek alkalmazása képfeldolgozási feladatokra
- ▶ CNN memória és számításigény szempontjából hatékonyabb (konvolúciós és pooling rétegek)
- ▶ A ViT modell „adatéhesebb”, de pontosabb
- ▶ A CNN tanult jellemzőit könnyebb megérteni a szűrők vizualizálásával

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet Real	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

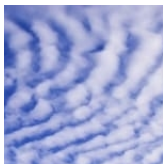
Forrás: Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale."

Felhasznált adatsor

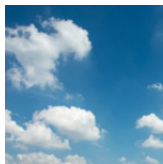
- ▶ Adathalmaz forrás: Swimcat-extend dataset (nyílt adatbázis)
- ▶ 2100 különböző kép, 6 különböző típusú égbolt, figyelembe véve a felhőzetet
- ▶ Kategóriánként 350 db kép
- ▶ A tanítás előtt a képeket normalizáltuk és 224x224 méretűre formáztuk
- ▶ Egyszerű statisztikai tulajdonságok alapján nem megkülönböztethető kategóriák



A—Clear sky



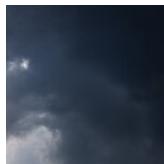
B—Patterned
clouds



C—Thin
white clouds



D—Thick
white clouds



E—Thick
dark clouds



F—Veil
clouds

A hálók felépítése

- ▶ Bemeneti réteg
- ▶ Opcionális adatdúsítás (augmentation layer)
 - ▶ Nincs adatdúsítás
 - ▶ Véletlen forgatás
 - ▶ Véletlen forgatás, tükrözés és zoom
- ▶ **Előtanított model (CNN vagy ViT, tanítható vagy rögzített súlyok)**¹
- ▶ Dropout réteg
- ▶ Kimeneti réteg

Konvolúciós neurális háló (CNN)

- ▶ EfficientNetB1²
- ▶ ResNet18D²

Vision Transformer (ViT)

- ▶ vit_tiny_patch16_224²
- ▶ vit_tiny_r_s16_p8_224_in21k²

¹ A hálókat az ImageNet-21k adatbázison előtanították

² [Tensorflow Image Models](#)

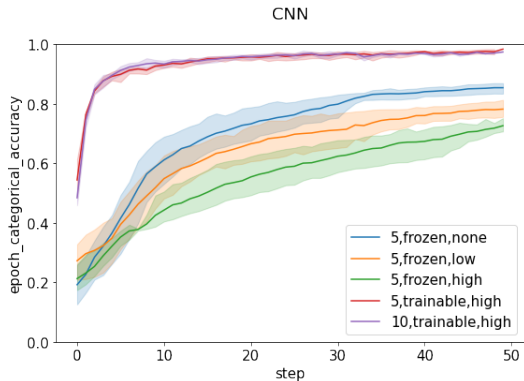
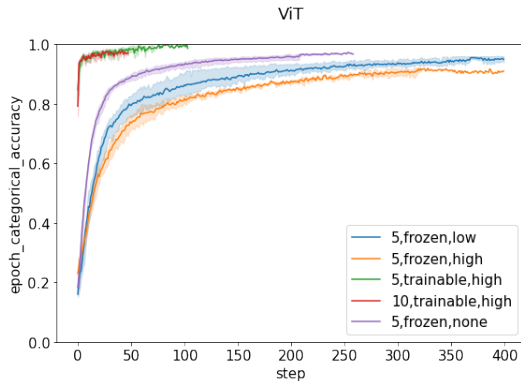
Vizsgált modellkonfigurációk, kísérleti módszertan

	Paraméterek száma	Háló taníthatósága	Adatdúsítás	CNN modell	ViT modell
1	$5 \cdot 10^6$	rögzített	nincsen	cnn_5m_f_na	vit_5m_f_na
2			alacsony	cnn_5m_f_la	vit_5m_f_la
2			magas	cnn_5m_f_ha	vit_5m_f_ha
4	$5 \cdot 10^6$	tanítható	magas	cnn_5m_t_ha	vit_5m_t_ha
5	10^7	tanítható	magas	cnn_10m_t_ha	vit_10m_t_ha

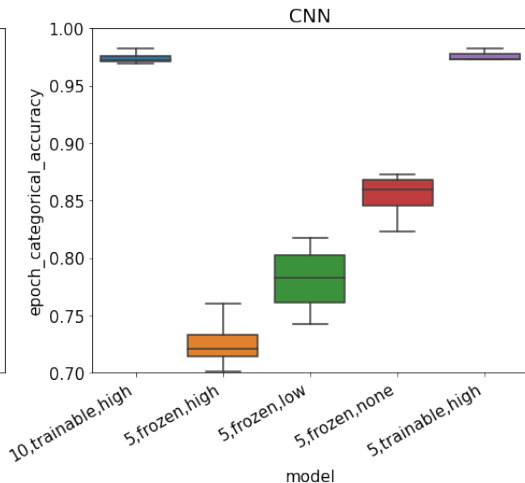
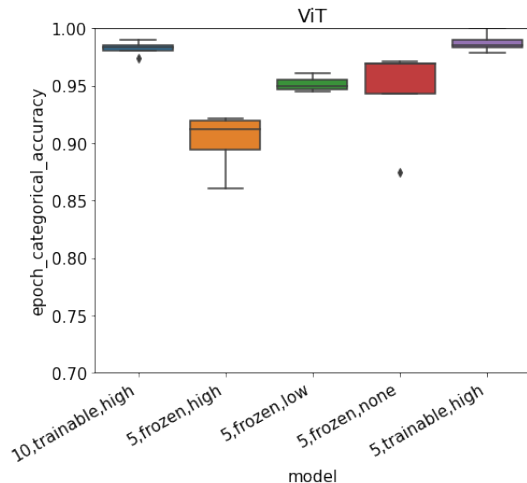
Kísérleti módszertan

- ▶ 5-szörös keresztvalidáció minden modellre
- ▶ 400 epoch tanítás + validáció
- ▶ Adam optimizer, $lr = 3 \cdot 10^{-5}$
- ▶ Early stopping 20 epoch után

Eredmények értékelése: pontosság alakulása a tanítás során



Eredmények értékelése: maximális elért validációs pontosság



Eredmények értékelése: független tesztadatsor

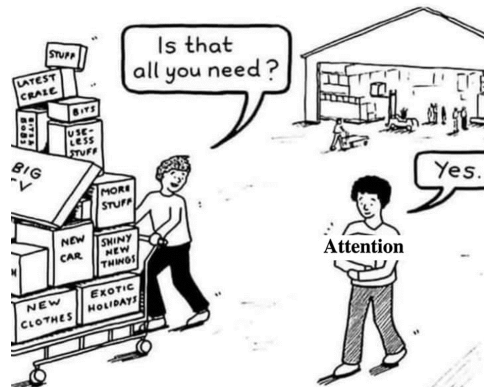
Model neve	Pontosság [%]	Veszteség
vit_10m_ha_t	77	1,52
cnn_5m_ha_t	74	1,12
vit_5m_ha_t	72	1,67
cnn_10m_ha_t	63	1,20
vit_5m_ha_f	56	1,60
vit_5m_la_f	50	2,56
vit_5m_na_f	43	3,42
cnn_5m_na_f	42	1,60
cnn_5m_ha_f	39	1,60
cnn_5m_la_f	37	1,61

Modellek performanciájának összehasonlítása

Modell	FLOPS (gigaFLOPS)	Paraméterek száma (millió)	Maximum pontosság [%]
vit_5m_ha_t	162	5,7	1,0
vit_10m_ha_t	56	10,5	0,99
cnn_5m_ha_t	75	7,8	0,98
cnn_10m_ha_t	263	11,7	0,98
vit_5m_na_f	162	5,7	0,97
vit_5m_la_f	162	5,7	0,96
vit_5m_ha_f	162	5,7	0,92
cnn_5m_na_f	75	7,8	0,87
cnn_5m_la_f	75	7,8	0,81
cnn_5m_ha_f	75	7,8	0,76

Összefoglalás

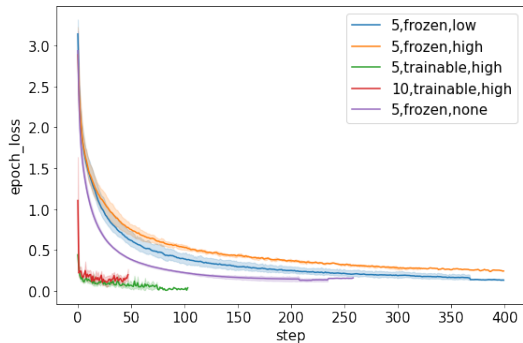
- ▶ Cél: képfeldolgozási célú neurális háló architektúrák összehasonlítása
- ▶ Probléma: felhőképek osztályozása
- ▶ Változatos paraméterű modellek
- ▶ Modellek értékelése 5-szörös keresztvalidációval
- ▶ Performancia és pontosság vizsgálata



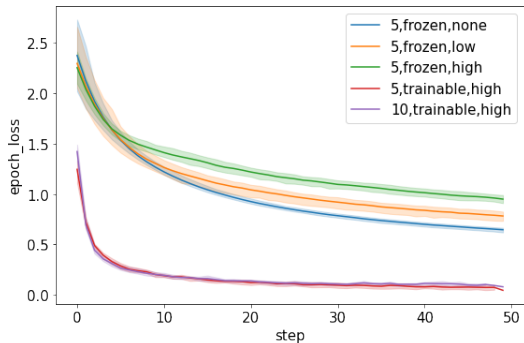
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017) - cited more than 62000 times

Veszteség alakulása a tanítás során

ViT



CNN



Különböző modellek veszteségének összehasonlítása

