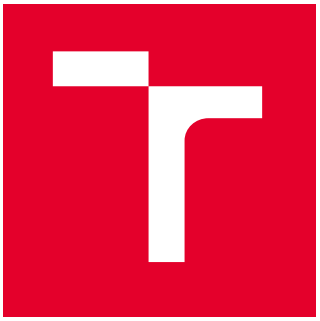


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

MOŽNOSTI NEURONOVÝCH SÍTÍ VYUŽÍVAJÍCÍCH TRANSFORMERY PRO ZPRACOVÁNÍ MEDICÍNSKÝCH OBRAZŮ

POTENTIAL OF NEURAL NETWORKS USING TRANSFORMERS FOR MEDICAL IMAGE PROCESSING

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. Tomáš Valík

AUTHOR

VEDOUCÍ PRÁCE — Ing. Jiří Chmelík, Ph.D.

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Bioinženýrství**

Ústav biomedicínského inženýrství

Student: Bc. Tomáš Valík ***ID:*** 217744

Ročník: 2 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Možnosti neuronových sítí využívajících transformery pro zpracování medicínských obrazů

POKYNY PRO VYPRACOVÁNÍ:

1) Nastudujte problematiku neuronových sítí se zaměřením na tzv. transformery a proveďte rešerši metod jejich využití pro zpracování medicínských obrazů. 2) Nalezněte alespoň dvě veřejně dostupné obrazové databáze (jednu obecnou a jednu medicínskou) vhodné pro aplikaci transformerů a formulujte problém k řešení. 3) Navrhněte a implementujte algoritmus pro řešení daného problému s využitím standardních neuronových sítí. 4) K řešení stejného problému navrhněte a implementujte algoritmus využívající transformerů. 5) Oba implementované algoritmy optimalizujte na vybraných datasetech. 6) Proveďte objektivní statistické vyhodnocení metod a dosažené výsledky vhodně diskutujte.

DOPORUČENÁ LITERATURA:

[1] SHAMSHAD, Fahad; KHAN, Salman; ZAMIR, Syed Waqas; KHAN, Muhammad Haris; HAYAT, Munawar et al., 2023. Transformers in medical imaging: A survey. Online. Medical Image Analysis. Roč. 88. ISSN 13618415. Dostupné z: https://doi.org/10.1016/j.media.2023.102802. [cit. 2023-08-14].

[2] LIN, Tianyang; WANG, Yuxin; LIU, Xiangyang a QIU, Xipeng, 2022. A survey of transformers. Online. AI Open. Roč. 3, s. 111-132. ISSN 26666510. Dostupné z: https://doi.org/10.1016/j.aiopen.2022.10.001. [cit. 2023-09-05].

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 22.5.2024

Vedoucí práce: Ing. Jiří Chmelík, Ph.D.

doc. Ing. Radim Kolář, Ph.D.
předseda rady studijního programu

ABSTRAKT

Tato diplomová práce se zabývá možnostmi využití neuronových sítí založených na architektuře transformerů pro zpracování medicínských obrazů. Hlavním cílem bylo porovnat výkonnost modelů ResNet18 a Vision Transformer (ViT-B-16) na dvou odlišných datasetech, konkrétně Intel Image Classification a ChestXray. Modely byly optimalizovány pomocí frameworku Optuna a nakonec byl každý z nich trénován desetkrát pro zajištění robustnosti výsledků. Ty ukazují, že modely využívající Vision Transformery dosahují vyšších hodnot váženého F1 skóre ve srovnání s modely ResNet18. Konkrétně dosáhl model ViT-B-16 nejvyššího F1 skóre 0,939 na datasetu Intel Image a 0,907 na datasetu ChestXray, zatímco ResNet18 dosáhl hodnot 0,883, respektive 0,885. Statistické analýzy pomocí Wilcoxonova testu potvrdily, že rozdíly ve výkonnosti mezi modely jsou statisticky signifikantní, což naznačuje výhodu použití Vision Transformerů pro tyto úlohy. Uveden je také rozbor výpočetní náročnosti, která je pro ViT mnohem vyšší.

KLÍČOVÁ SLOVA

strojové učení, transformer, neuronová síť, zpracování medicínských obrazů, self-attention, vision transformer

ABSTRACT

This thesis explores the potential of neural networks based on transformer architecture for medical image processing. The main objective was to compare the performance of ResNet18 and Vision Transformer (ViT-B-16) models on two distinct datasets, specifically Intel Image Classification and ChestXray. The models were optimized using the Optuna framework and subsequently trained ten times each to ensure robustness of the results. These results indicate that models utilizing Vision Transformers achieve higher weighted F1 scores compared to ResNet18 models. Specifically, the ViT-B-16 model achieved the highest F1 score of 0.939 on the Intel Image dataset and 0.907 on the ChestXray dataset, whereas ResNet18 achieved scores of 0.883 and 0.885, respectively. Statistical analyses using the Wilcoxon test confirmed that the differences in performance between the models are statistically significant, suggesting an advantage of using Vision Transformers for these tasks. An analysis of computational complexity is also provided, highlighting that ViT requires significantly higher computational resources.

KEYWORDS

machine learning, transformer, neural network, medical image processing, self-attention, vision transformer

VALÍK, Tomáš. *Možnosti neuronových sítí využívajících transformery pro zpracování medicínských obrazů*. Diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024. Vedoucí práce: Ing. Jiří Chmelík, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Bc. Tomáš Valík
VUT ID autora:	217744
Typ práce:	Diplomová práce
Akademický rok:	2023/24
Téma závěrečné práce:	Možnosti neuronových sítí využívajících transformery pro zpracování medicínských obrazů

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu semestrální práce panu Ing. Jiřímu Chmelíkovi, Ph.D. za odborné vedení, konzultace a podnětné návrhy k práci.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

V rychle se vyvíjejícím oboru umělé inteligence a strojového učení představuje architektura transformeru nový způsob, jakým přistupujeme ke složitým výpočetním úlohám. Tato práce se zaměřuje na možnosti využití neuronových sítí, založených na transformerech, pro zpracování medicínských obrazů.

Transformery byly poprvé představeny Vaswanim a spol. v roce 2017. [34] Tento přístup se výrazně odlišuje od tradičních rekurentních neuronových sítí (RNN) a konvolučních neuronových sítí (CNN) díky své schopnosti efektivně zachytit globální kontext pomocí mechanismu self-attention. Tento mechanismus umožňuje modelu paralelně zpracovávat různé části vstupu, což vede k rychlejšímu trénování a lepší škálovatelnosti. V kontextu medicínských obrazů může tento přístup poskytnout významné výhody při analýze složitých obrazových dat.

V první části této práce je proveden detailní rozbor základních komponent architektury transformeru, včetně enkodéru a dekodéru, a je vysvětlen koncept self-attention. Dále se práce zaměřuje na adaptaci transformerů pro zpracování obrazů, což zahrnuje výzvy spojené s jejich aplikací, jako jsou vysoké nároky na tréninková data a absence tzv. inductive bias, který je charakteristický pro konvoluční neuronové sítě.

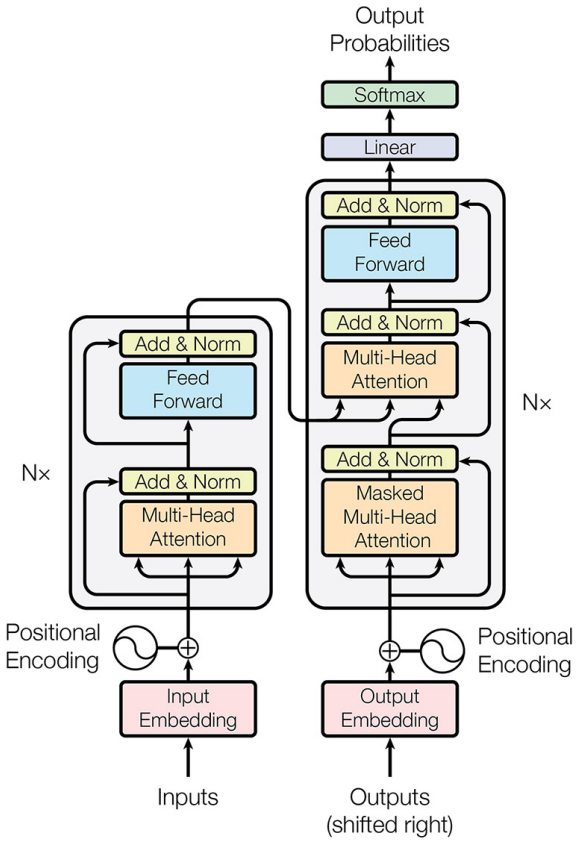
Praktická část této práce je zaměřena na implementaci a porovnání standardních konvolučních sítí s Vision Transformery při řešení klasifikačních úloh na obecném datasetu Intel Image Classification a medicínském datasetu ChestXray. Konkrétní zvolené architektury byly ResNet18 a ViT-B-16, které byly optimalizovány pomocí frameworku Optuna, který umožňuje efektivní hledání nejlepších hyperparametrů. Výsledky experimentů jsou statisticky vyhodnoceny a diskutovány. Byla provedena také analýza výpočetní náročnosti jednotlivých modelů podle vícero metrik.

Tato práce tedy poskytuje nejen teoretický přehled a adaptaci architektury transformerů, ale také praktické implementace a srovnání s konvolučními neuronovými sítěmi, což přispívá k lepšímu pochopení jejich potenciálu v oblasti medicínského zobrazování.

1 Literární rešerše

1.1 Architektura transformeru

Originální architekturu transformeru, kterou představil Vaswani et al. v roce 2017 [34], je zobrazena na Obrázku 1.1. Byla navrhnuta pro zpracování přirozeného jazyka, konkrétně pro překlad textu. Skládá se z enkodéru, který zpracovává text v překládaném jazyce a kontextuální informace přenáší dekodéru, který se na jejich základě pokouší předpovědět správný překlad. Pro porozumění textu využívá koncept self-attention, který bude dále spolu s jednotlivými částmi této architektury popsán.



Obr. 1.1: Architektura transformeru, převzato z [34]

1.1.1 Attention

Attention je mechanismus v neuronových sítích, který umožňuje modelu zaměřit se na konkrétní části vstupu při vytváření predikcí. V tradičních neuronových sítích je celý vstup zpracován k vytvoření jednoho výstupu, zde však může docházet k záměnám ve významu slov či shodě podmětu s přísudkem. Díky attention modulu tomu

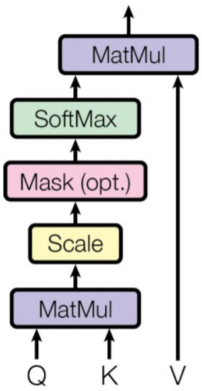
lze předejít, jelikož každému vstupnímu slovu je přiřazena váha, která indikuje jeho relativní důležitost. V konečném důsledku je díky tomu zachycen globální kontext. [34]

Scaled dot-product attention

Typ attention, který použil Vaswani et al. je *Scaled dot-product attention*. [34] Jedná se o operaci, která pracuje se třemi různými vstupními vektory, a to *query*, *keys* a *values*. Schéma je zobrazeno na Obrázku 1.2 a matematicky vyjádřeno na Rovnici (1.1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1.1}$$

Query je vektor představující aktuální pozici nebo prvek ve vstupní sekvenci, na který se model zaměřuje. Následně je vypočítána jeho podobnost s keys, které představují všechny ostatní pozice v sekvenci. Tato podobnost je vyjádřena ze skalárního součinu, který je poté upraven na základě délky vstupního vektoru keys d_k. Tuto úpravu je třeba provést, jelikož dosažené hodnoty při použití velkého vektoru keys mohou být vysoké. [34]



Obr. 1.2: Schéma scaled dot-product attention, převzato z [34]

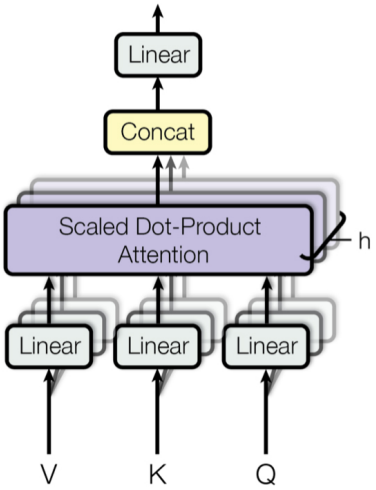
Následně dochází k maskování, které je volitelné podle toho, kde je modul využit. V enkodéru se nepoužívá, naopak v dekodéru se jím zajišťuje to, že model dokáže vnímat jen slova, která byla ve větě již obsažena, nikoli ta, která předpovídá. Tím se zajišťuje autoregresivní generování výstupní sekvence. [34]

Použití funkce *softmax*, transformuje reálné hodnoty vah do rozdělení pravděpodobnosti, tedy rozsahu od 0 do 1 a zároveň tak, že jejich součet bude roven 1. Tyto získané hodnoty jsou váhy attention, které popisují vztahy mezi jednotlivými

slovy sekvence a po vynásobení s *values* nám obohacují vstupní hodnoty o kontext získaný ze sekvence. [34]

Multi-head attention

Aby bylo možné lépe zachytit závislosti mezi jinými prvky vstupní sekvence, je implementován blok *Multi-head attention*, zobrazený na Obrázku 1.3 a vyjádřený Rovnicí (1.2). Díky rozdělení vstupních vektorů a následné lineární transformaci jsme schopni funkci *Scaled dot-product attention* paralelizovat. Zatímco výpočetní náročnost zůstane podobná tomu, kdyby byla použita pouze jedna "hlava", dosáhne se společného sledování různých subsetů na různých pozicích.



Obr. 1.3: Schéma multi-head attention, převzato z [34]

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \cdot W^O \tag{1.2}$$

Získané hodnoty z jednotlivých hlav se poté spojí a jsou lineárně transformovány naučenou maticí vah pro získání finálního výstupu bloku *Multi-head attention*. Dosahujeme tím tak více komplexní kombinace výstupů než pouze z jedné hlavy. [34]

1.1.2 Poziční kódování

Rekurentní či konvoluční neuronové sítě zpracovávají sekvenčně a pořadí prvků je tak zachyceno v jejich skrytém stavu. Jelikož se architektura transformerů spoléhá čistě na self-attention, nemá sekvenční pořadí vstupních prvků sama jak zachytit. Proto bylo představeno poziční kódování, které je přidáno k vstupním embeddingům před

vstupem do bloků enkodéru a dekodéru a má stejné rozměry jako tyto embeddingy, takže je k nim přičteno.

Poziční kódování je generováno pomocí funkcí sinus a cosinus s různými frekvencemi podle Rovnice (1.3) a (1.4):

$$\mathrm{PE}_{\mathrm{pos},2i} = \sin\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{1.3}$$

$$\mathrm{PE}_{\mathrm{pos},2i+1} = \cos\left(\frac{\mathrm{pos}}{10000^{2i/d_{\mathrm{model}}}}\right) \tag{1.4}$$

Zde *pos* reprezentuje pozici v sekvenci, i index dimenze a d_{model} délku embeddingu. Tato podoba pomáhá, jelikož hodnoty jsou limitovány v rozsahu od -1 do 1, jsou periodické a dají se ze znalosti funkce extrapolovat. Je tak možné efektivně zachytit závislosti mezi různými prvky ve vstupní sekvenci, i přesto, že transformer nemá sekvenční povahu rekurentních či konvolučních sítí. [34]

1.2 Využití pro zpracování obrazů

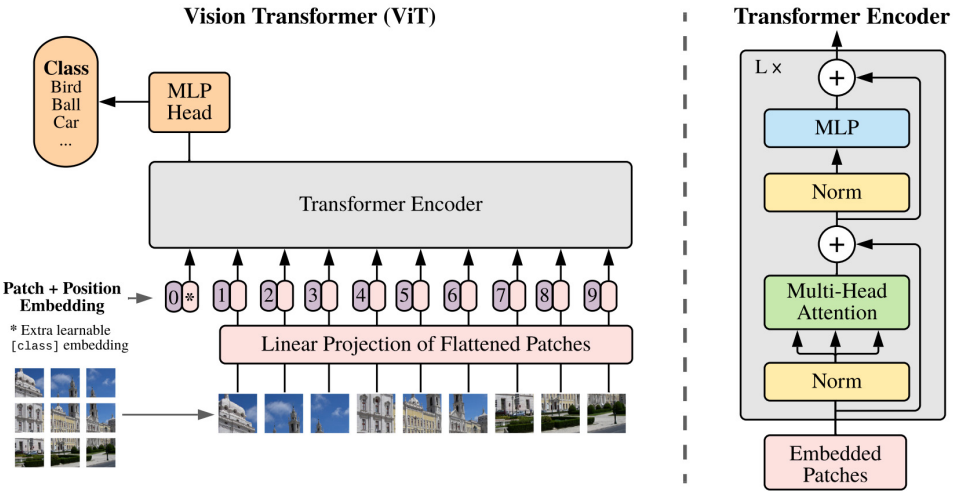
Zatímco se brzy po představení staly transformery state-of-the-art metodou v mnoha úlohách pro zpracování přirozeného jazyka, u zpracování obrazů k jejich plné implementaci došlo později.

V práci *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* [9] z roku 2020 se Dosovitskiy et al. věnoval vytvoření architektury Vision Transformeru (ViT) pro zpracování obrazů tak, aby byla co nejpodobnější originálnímu návrhu. Na rozdíl od předchozích přístupů, které využívaly self-attention v kombinaci s konvolučními neuronovými sítěmi, se obrázek rozloží na menší díly s pozičním kódováním, které vstupují do transformeru, a ten s nimi zachází podobně jako se slovy u aplikace při zpracováním přirozeného jazyka. [9]

Když byly tyto modely trénovány na středně velkých datasetech, jako je např. ImageNet, dosahovaly přesnosti o pár procent nižší než ResNet o srovnatelné velikosti. V případě, kdy byly natrénovány na mnohem větším interním datasetu (300 milionů obrazů), se jejich výsledky velmi zlepšily. Na základě této práce tedy vyplývá, že vision transformery mohou dosahovat lepších výsledků než nejmodernější konvoluční neuronové sítě, potřebují však ke svému tréninku obrovské množství dat, aby dokázaly generalizovat tak dobře, jako např. ResNet. [9]

1.2.1 Optimalizace transformerových architektur

Ze znalosti z minulých kapitol lze říci, že transformery dosahují srovnatelných, mnohdy i lepších výsledků než tradiční konvoluční neuronové sítě. Největší nevý-



Obr. 1.4: Schéma architektury Vision Transformeru, převzato z [9]

hodu, kterou představují, je ale množství dat, které k tréninku potřebují. Pro skvělou schopnost zachovat globální závislosti v obraze potřebují větší množství dat, než které jsou biomedicínské datasety schopny nabídnout. A proto již zanedlouho, od představení originálního Vision Transformeru, kterému došlo v říjnu 2020, přicházeli výzkumníci s optimalizacemi v jeho architektuře.

Data-efficient transformer

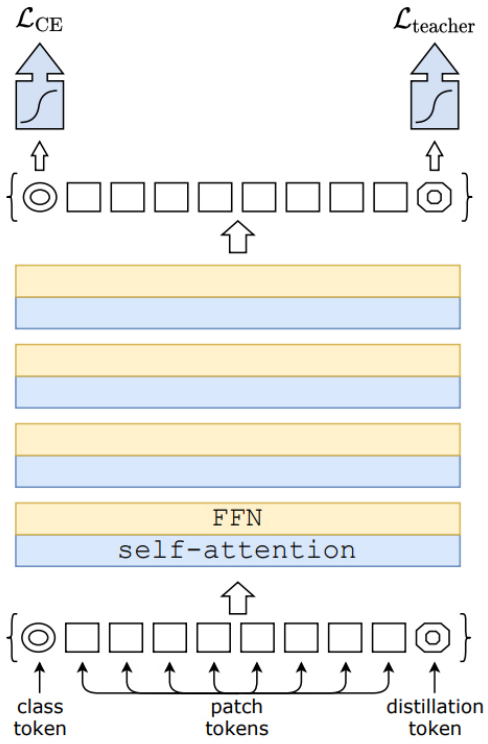
V práci *Training data-efficient image transformers & distillation through attention* [33] představuje Touvron et al. metodu pro trénování Vision Transformerů na omezeném množství dat, pro jejich zpřístupnění k použití na více aplikací. DeiT (Data-efficient Image Transformer) zahrnuje nový přístup učení učitel-student, který se zakládá na distilačním (distillation) tokenu, díky kterému se zajistí, aby se student učil od učitele skrze attention. [33]

Knowledge distillation je technika komprese modelů, při které se vnášejí znalosti z komplexnějšího modelu, učitele, na menší a rychlejší model, studenta, se snahou udržet přesnost výsledků. [14] Proces při využití této techniky je tedy natrénování učitelského modelu, kterým může být např. konvoluční neuronová síť nebo jiný klasifikátor, na rozsáhlém datasetu. Následně se natrénuje student s využitím učitelského modelu jako správného rozdělení tříd pomocí jedné z následujících metod.

Soft distillation je metoda, která hledá minimální hodnotu Kullback-Leiblerovy divergence mezi výsledkem softmax aktivační funkce učitele a studenta. Tato hodnota je míra relativní entropie, která značí rozdíl mezi odhadovanou třídou a skutečnou třídou [5] a k dosažení správného výsledku je třeba balancovat její ztrátu s vzájemnou entropií na ground truth datech. Oproti tomu u hard distillation jsou

brány třídy učitele jako správné hodnoty a cíl je zde balancovat vzájemnou entropii predikovaných tříd studenta a učitele. Na základě experimentů bylo zjištěno, že hard distillation dosahuje na každé z variant DeiT lepších výsledků než použití soft distillation či žádné z uvedených metod. [33]

Hlavním přínosem této práce je představení distillation tokenu, který slouží ke zvýšení úspěšnosti studentského modelu. Do transformeru vstupuje spolu s ostatními embeddingy a dále se transformuje v self-attention vrstvách. Na výstupu je jeho úkolem předpovědět třídu učitele, nikoli pravdivou třídu. Díky tomu dosahuje model s tímto přidaným tokenem ještě lepších výsledků než čistě použití hard distillation u učitele. Tato funkce pomáhá při učení na augmentovaných datech, kdy se může stát, že subjekt, který chceme klasifikovat, je uříznut ze snímku, čemuž se učitel dokáže přizpůsobit, zatímco pravdivá třída bude nezměněna. [33]



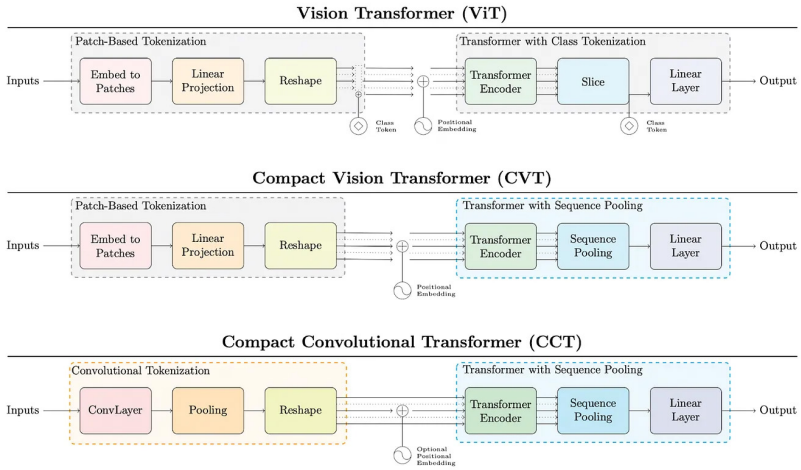
Obr. 1.5: Schéma DeiT využívající distillation tokenu, převzato z [33]

Bylo porovnáno také použití jiných modelů jako učitelů a výsledky ukazují, že model studenta používající distillation více koreluje s použitou konvoluční neuronovou sítí než transformerem, který je učen od začátku. Další analýza ukazuje, že klasifikátor založený pouze na distillation tokenu má blíže ke konvoluční neuronové síti než ten, který obsahuje pouze class token. Tento fakt ukazuje důležitost inductive biasu na trénovaném modelu. [33] Tento dopad knowledge distillation je potvrzen v dalších pracích. [1, 27]

Kompaktní transformery

Jiný přístup, který se snaží snížit množství trénovacích dat, je představen v práci *Escaping the Big Data Paradigm with Compact Transformers* [12]. Autoři představují a porovnávají 3 nové modifikace ViT. Konkrétně jsou jimi ViT-Lite, Compact Vision Transformer (CVT) a Compact Convolutional Transformer (CCT).

ViT-Lite je kompaktnější verze originálního Vision Transformeru, která byla upravena pro učení na menších datech a snížení výpočetních nákladů. CVT také vychází z původního ViT, oproti němu však využívá Sequence Pooling (SeqPool), který nahrazuje běžnou tokenizaci. Na tom dále staví CCT model, který využívá konvoluce pro vytvoření vstupních dílů oproti klasickému rozdělení obrazu. Díky tomu ponechává lokální informaci a vztahy mezi jednotlivými díly. [12]



Obr. 1.6: Porovnání upravených verzí Vision Transformeru, převzato z [12]

Konvoluční tokenizace (Convolutional Tokenization) nahrazuje tradiční tokenizaci, aby se přidal do modelu inductive bias, který čisté transformerové architektury postrádají. Skládá se z jednoduché konvoluce, ReLU aktivace a max poolingu. Tímto obraz rozdělí na nepřekrývající se díly, které jsou převedeny na vektory a transformovány do latentního prostoru. Konvoluce a max pooling se můžou překrývat, čímž model zvyšuje svou přesnost přidáváním inductive biasu, díky kterému model zachovává lokální prostorovou informaci. [12]

SeqPool, představený v této práci, je metoda založená na pozornosti, určená k použití na výstupní sekvenci tokenů v transformerových klasifikátorech. Cílem je přidat různé váhy sekvenčnímu embeddingu latentního prostoru, který je vytvořen enkóderem trasformeru a korelovat tato data se vstupními. Touto metodou generujeme váhy důležitosti pro každý vstupní token, které se pak aplikují na výstupní sekvenci, která poté může být předána klasifikátoru. Motivací za SeqPoolem je to,

že výstupní sekvence obsahuje relevantní informace napříč různými částmi vstupního obrazu a zachování těchto informací může zlepšit výkon bez přidání dalších parametrů k učení. [12]

Díky modelu CCT dosahují autoři v této práci lepších výsledků než tradiční Vision Transformer, ale také state-of-the-art konvoluční sítě jako ResNet. Hlavním přínosem je přiblížení používání transformerů i v oblastech, kde se pracuje pouze s malými datasety, což může být právě u zpracování medicínských dat. [12]

1.3 Aplikace na medicínských obrazech

1.3.1 Klasifikační problémy

Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification

Pro detekci přítomnosti Alzheimerovy choroby u pacientů vytvořili Li et al. model Trans-ResNet. [20] Ten pracuje se skeny magnetické rezonance, konkrétně T1 váhovanými obrazy. Již zmiňovaný nedostatek transformerů a jejich potřebu pro využití velkých datasetů pro natrénování, řeší autoři předtrénováním modelu na jiném úkolu, predikce věku, a poté dotrénováním na hlavním úkolu, detekce Alzheimerovy choroby. Objemová data jsou zpracována 3D konvoluční neuronovou sítí ResNet-18, následně rozdělena do dílků, převedena na sekvence a vstupují do transformerového enkodéru. Na jeho výstupu je použit global average pooling a vícevrstvý perceptron, pro klasifikaci.

Pořadí toku dat ze sítě ResNet do transformerového enkodéru není náhodné. V práci byl proveden pokus, kdy byly vytvořeny další dvě architektury. Jedna, kde bylo pořadí opačné a první zpracování prováděl transformer, který následně předával data do ResNet, a druhá, kde pracovaly tyto dva enkodéry paralelně a jejich výstupy byly poté spojeny a vyhodnoceny. Právě metoda, kdy je první použita konvoluční neuronová síť, dosahovala nejvyšší přesnosti (92.26%), následována paralelní (91.55%) a reversní metodou (90.68%). Tento výsledek podporuje tvrzení, že konvoluční neuronové sítě dokáží dobře extrahovat lokální závislosti, zatímco transformer globální kontext. [20]

Úspěšnost modelu byla porovnávána na dvou různých datasetech, ADNI a AIBL, na kterých byl zároveň dotrénováván, a oproti jiným architekturám jako ResNet či DeiT-Tiny dosahoval Trans-ResNet vždy lepších výsledků, jak je zobrazeno v Tabulce 1.1.

Dataset	Model	Přenost (%)	AUC
ADNI	ResNet	92,61	0,964
	DeiT-Tiny	78,25	0,846
	Trans-ResNet	**93,85**	**0,968**
AIBL	ResNet	93,32	0,943
	DeiT-Tiny	89,13	0,872
	Trans-ResNet	**93,94**	**0,957**

Tab. 1.1: Klasifikační úspěšnost různých modelů, převzato a upraveno z [20]

TransSLC: Skin Lesion Classification in Dermatoscopic Images Using Transformers

Sarker et al. řeší problém klasifikace lézí z dermatoskopických snímků pomocí transformerů. [30] Ke zpracování využívají obousměrný enkodér BEiT, který k předtrénování využívá masked image modelling (MIM), systém učení pod vlastním dohledem. Hlavním dílem této architektury je enkóder Vision Transfomeru, jehož výstupní vektory vedou do klasifikátoru, kde se rozhodne pro jednu ze sedmi tříd. Úspěšnost je porovnána s 5 architekturami založenými na konvolučních neuronových sítích (ResNet-101, Inception-V3, Inception-ResNet-V2, Xception, EfficientNet-B7), ze kterých TransSLC dosahuje jako jediný nad 90% úspěšnost. [30]

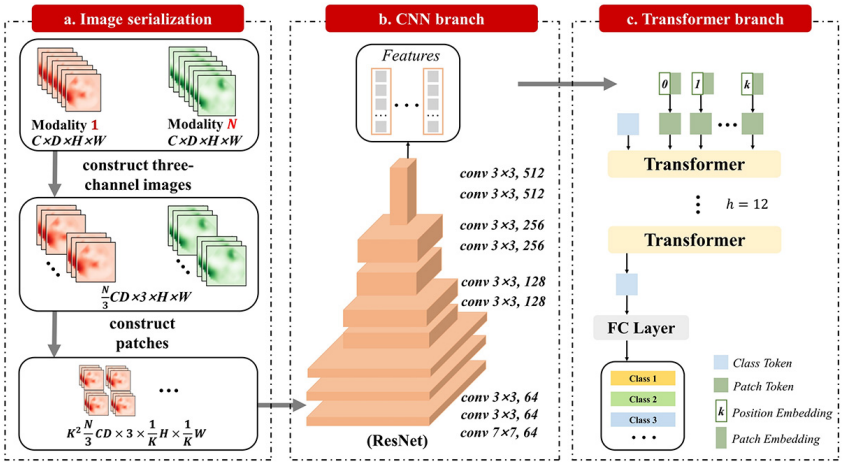
Multi-Label Retinal Disease Classification Using Transformers

Pro potřeby této práce si vytvořili Rodríguez et al. svůj dataset MuReD [28], který kombinuje více dostupných datasetů, ze kterých jsou převzaty snímky tak, aby bylo dobré rozdělení distribuce tříd a dosahovaly dostatečné kvality. Pro klasifikaci dat využívají model C-Tran, který je přímo určen pro klasifikaci dat do více tříd. Ze snímku se získají příznaky konvoluční neuronovou sítí, které vstupují do enkodéru transformeru spolu s label a state embeddingy. Ty pomáhají modelu k učení pouze s částečnou informací. Pro dosažení nejlepších výsledků porovnávají také různé extraktory příznaků, velikosti vstupních snímků či dílů. I díky tomu dosahují v porovnání s jinými modely nejlepších výsledků. [28]

TransMed: Transformers Advance Multi-Modal Medical Image Classification

Multimodální medicínské obrazy obsahují oproti přirozeným obrazům dlouhodobé závislosti, se kterými mají standardní konvoluční sítě problém. V této práci se snaží Dai et al. využít transformerů pro zachycení těchto závislostí, které však potřebují obrovské datasety, aby dosáhly lepších výsledků. Dai et al. představují TransMed, který kombinuje výhody konvolučních neuronových sítí a transformerů, aby efektivně

vytáhnul příznaky nízké úrovně z obrazu a díky tomu vytvořil dlouhodobé závislosti mezi obrazy. V době zveřejnění článku se jednalo o první aplikaci transformerů na multimodální obrazy. [6]



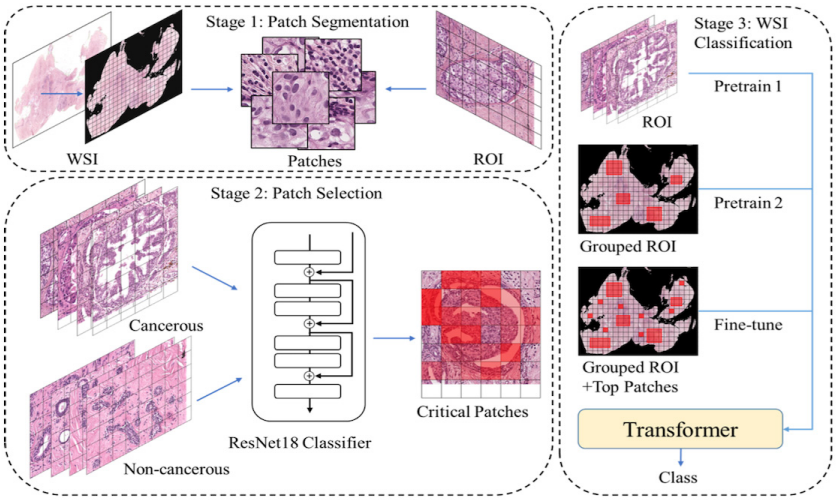
Obr. 1.7: Schéma modelu TransMed, převzato z [6]

Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers

Práce se zabývá využitím ViT na klasifikaci rakoviny tlustého střeva z histologických obrazů. Využívají dataset CRC, který obsahuje 5000 obrázků s osmi kategoriemi tkání, a pro klasifikaci používají dva druhy modelů, klasický Vision transformer a také Compact Convolutional Transformer, které dosahují přesnosti 93,3% a 95%. [38]

Breast Tumor Image Classification in Bright Challenge VIA Multiple Instance Learning and Deep Transformers

Zhan et al. představuje nový model hlubokého učení MIL-Transfer-Transformer Network (MTTN) pro klasifikaci nádorů prsu z histologických snímků. V první části dochází k předzpracování, kdy je obraz rozdělen na menší dílky i za pomoci oblasti zájmu, která pomáhá vyřadit dílky obsahující pozadí. V druhé části je vytvořena binární klasifikace modelem ResNet18, který vybírá ty nejvíce reprezentativní dílky a získává odpovídající příznaky. Ve třetí fázi je využit transformer, do kterého vstupují právě tyto příznaky a na jeho výstupu je získána odpovídající klasifikace. Výsledky ukazují účinnost navrhované metody, která překonává výsledky referenčního modelu pro obě úlohy. [39]



Obr. 1.8: Schéma navrhované klasifikační metody MTTN, převzato z [39]

1.3.2 Segmentační problémy

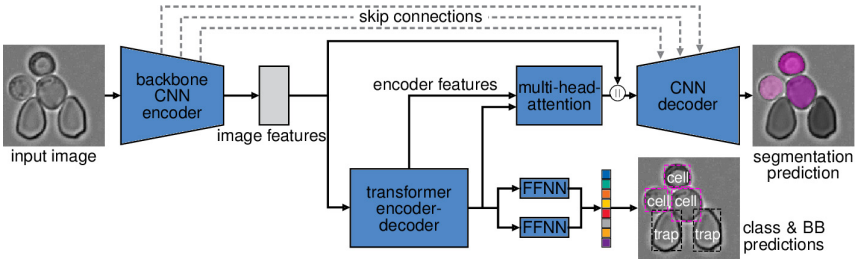
Attention-Based Transformers for Instance Segmentation of Cells in Microstructures

Inovaci, kterou Prangemeier et al. v této práci představuje, je Cell-DETR [26], model založený na základech DETR (detection transformer) [4], který je optimalizovaný pro instanční segmentaci na biomedicínských obrazech. Data, se kterými pracují, jsou fluorescenční mikroskopické snímky kvasinek, mezi kterými se nachází mikrofluidní pasti. Ty se zde nachází, jelikož slouží pro dlouhodobou kultivaci kvasinek v ohniskové rovině mikroskopu. Cílem architektury je přiřadit každému pixelu jednu ze tří tříd, buňka, past a pozadí. [26]

Architektura, která je zobrazena na Obrázku 1.9, je složena z dvou hlavních částí, podpůrné konvoluční sítě a transformerové sítě. Pro získání příznaků je využita síť ResNet-50 předtrénovaná na datasetu ImageNet, dále pak transfomerová síť založená na architektuře DETR. Představují nové poziční schéma pro enkoding, které zaznamenává prostorové informace buněk ve vstupním snímku. Dekodér je navíc modifikován, aby vracel vyznačenou masku instance místo bounding boxu. V porovnání s jinými segmentačními modely dosahuje Cell-DETR podobných výsledků, to však s výrazně méně parametry a za o 30% kratší čas. [26]

Multi-compound Transformer for Accurate Biomedical Image Segmentation

Pro dosažení lepších výsledků při segmentaci biomedicínských obrazů představili Ji et al. více složkový model MCTrans. [17] Zavádí pojem proxy embedding, který je důležitý pro modelování vzájemných vazeb mezi třídami. Bloky self-attention a



Obr. 1.9: Schéma modelu Cell-DETR, převzato z [26]

cross-attention umožňují modelu lépe rozlišit rozdíly mezi třídami a konzistenci v rámci jedné třídy. Úspěšnost segmentace byla testována na šesti různých datasetech o třech typech, segmentace buněk, polypu a kožních lézí, kde MCTrans překonal výsledky jiných předních segmentačních metod. Celkově tento model demonstruje účinnost a generalizaci na různých datech. [17]

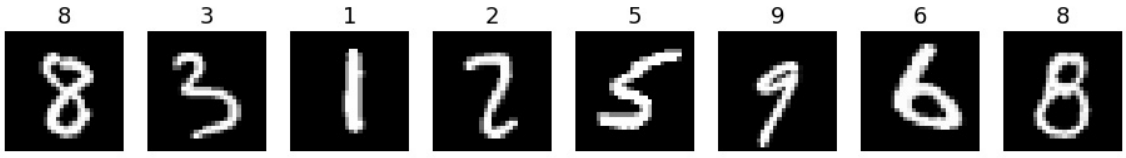
2 Obrazové databáze

V oblasti zpracování medicínských obrazů vykazují transformery slibné výsledky v již popsaných aplikacích, jako je segmentace či klasifikace. Jejich úspěch v těchto úlohách lze přičíst jejich schopnosti učit se na komplexních vztazích ve vstupních datech, kde konvoluční sítě naopak lépe extrahují lokální závislosti těchto dat. [22] Současné modely založené na transformerech musí čelit výzvám, jako je nedostatek dat či absence inductive biasů. S nedostatky dat, ale také zavedení inductive biasů, se snaží různé týmy vypořádat jinými způsoby, kterými může být např. knowledge distillation [33], sequence pooling [12] či předtrénování na jiném datasetu s odlišným cílem a následném dotrénování. [20]

2.1 Obecné datasety

MNIST

Dataset MNIST (Modified National Institute of Standards and Technologies) je kolekce ručně psaných číslic, často využívaná pro výzkum a vývoj algoritmů, pro rozpoznávání psaného textu. Prochází z původní databáze NIST a obsahuje celkem 70 000 obrázků, kdy 60 000 je pro trénování a 10 000 pro testování. Každý obrázek obsahuje ručně psanou číslici, která byla předzpracována, včetně segmentace a normalizace. Černobílé číslice jsou normalizovány podle velikosti a umístěny do středu obrazu o rozměrech 28×28 pixelů. V důsledků toho je dimenzionalita každého vektoru vzorku obrazu 784 a každý prvek je binární. [21]

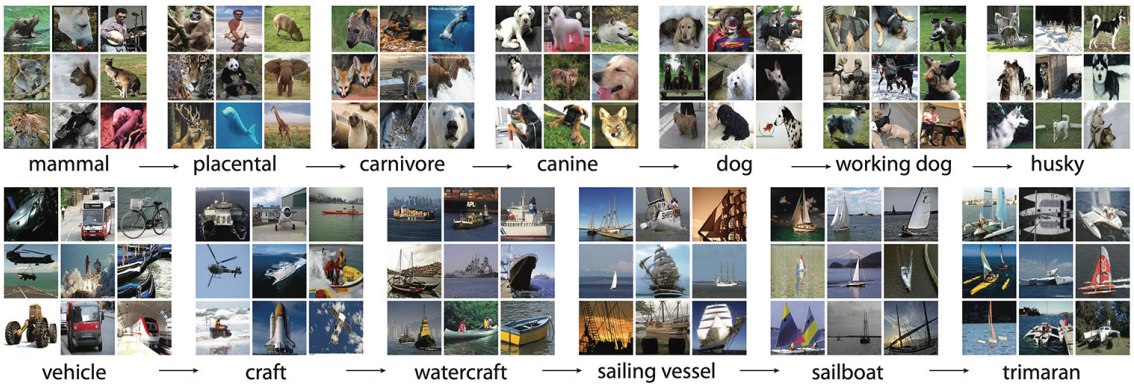


Obr. 2.1: Příklad osmi vzorků z datasetu MNIST a jejich označení

ImageNet

ImageNet je velká databáze obrazů organizovaná podle hierarchie WordNet, obsahující více než 100 000 synsetů, a to především podstatných jmen. Tento dataset má za cíl poskytovat výzkumníkům obrazová data pro účely trénování a porovnávání v oblasti počítačového vidění a hlubokého učení. [7]

Od roku 2010 do roku 2017 byla data používána pro ILSVRC (ImageNet Large Scale Visual Recognition Challenge). Kompletní dataset obsahuje 14 milionů obrázků, ze kterých soutěž využívala jeho, nyní veřejně dostupnou, podmnožinu obsahující 1 281 167 obrázků pro trénink, 50 000 pro validaci a 100 000 pro testování. Soutěž se stala měřítkem v oblasti klasifikace obrazu a detekce objektů. [29]



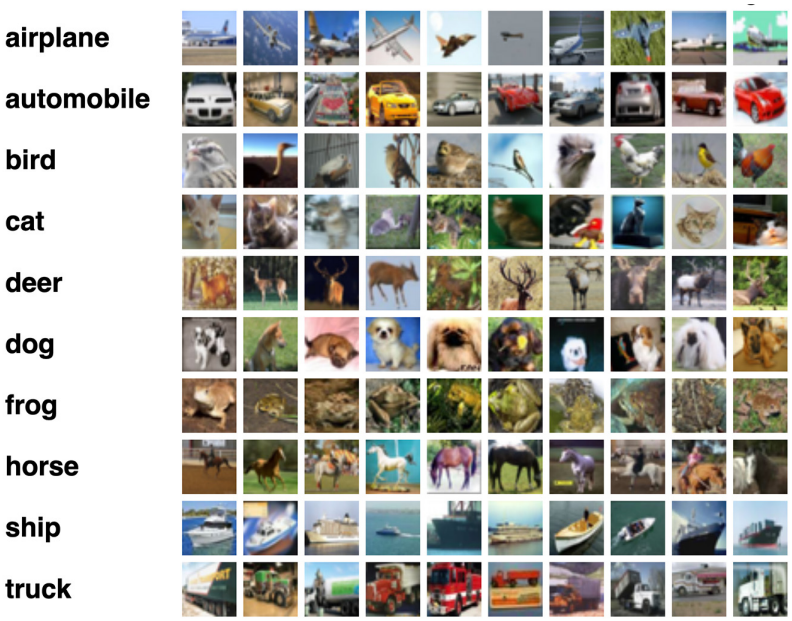
Obr. 2.2: Příklad hierarchie obrázků v datasetu ImageNet, převzato z [7]

CIFAR

Datasety CIFAR-10 a CIFAR-100 jsou kolekcemi malých barevných obrázků o velikosti 32x32 pixelů používaných pro experimenty s rozpoznáváním a klasifikací objektů, pojmenované po Canadian Institute for Advanced Research, který projekt financoval. Dataset CIFAR-10 obsahuje 6 000 příkladů pro každou z 10 tříd, kterými jsou letadlo, automobil (ale ne nákladní auto ani dodávka), pták, kočka, jelen, pes, žába, kůň, loď a nákladní auto. Třídy CIFAR-100 obsahují 600 obrázků pro každou ze 100 tříd, které se nepřekrývají s třídami CIFAR-10 a mohou tak být použity jako negativní příklady pro CIFAR-10. Například CIFAR-10 obsahuje třídy automobil a nákladní auto, ale žádná z těchto tříd neobsahuje obrázky dodávek, kterou naopak obsahuje CIFAR-100. Třídy CIFAR-100 jsou navíc rozděleny do 20 nadřazených tříd po pěti každá. [19]

Intel Image Classification

Obrazová databáze Intel Image Classification byla původně vytvořena společností Intel pro Analytics Vidhya, jejichž výzvou byla klasifikace do více tříd. Obsahuje přibližně 25 000 obrázků rozdělených do šesti různých tříd: budovy, les, ledovec, hory, moře a ulice. Slouží jako cenný zdroj pro výzkumníky v oblasti strojového učení a počítačového vidění, nabízí rozmanitou škálu přírodních a městských scén pro vývoj a benchmarking algoritmů klasifikace obrázků. Každý prvek datasetu je obrázek ve



Obr. 2.3: Příklad obrázků ke každé třídě v datasetu CIFAR-10, převzato z [19]

formátu PNG s rozlišením 150x150 pixelů a třemi kanály RGB. Celá databáze je rozdělena do tří podmnožin: trénovací, která má 14 034 obrázků, testovací obsahující 3 000 obrázků a predikční se 7 302 obrázky.

Obr. 2.4: Příklady obrázků pro každou třídu z datasetu Intel Image Classification

2.2 Medicínské datasety

MedMNIST

MedMNIST vznikl za cílem zjednodušit porovnávání algoritmů strojového učení na medicínských datech, kdy se snaží tyto odlišnosti odstranit jednotnou standardizací a úpravou obrazů. Jeho současná verze obsahuje 12 2D a 6 3D datasetů, kdy každý prvek má velikost 28^2, respektive 28^3. Předností tohoto souboru je to, že obsahuje různorodé datasety velikosti od stovek po statisíce vzorků, u kterých se snaží dosáhnout různých úkolů jako binární klasifikace, klasifikace do více tříd či ordinální regrese. [37]

Pro sjednocení vyhodnocení na testovacích datech jsou také předem rozděleny na trénovací, validační, a právě testovací data. Pokud měl zdrojový soubor dat, ze kterých byl MedMNIST sestavován, oficiální rozdělení na trénovací a validační, autoři považují oficiální validační soubor za testovací a trénovací berou 10% pro validaci. Pokud neměla zdrojová data oficiální rozdělení, jsou rozdělena v poměru 7:1:2 z celkového množství dat na trénovací, validační a testovací. Použitá data jsou různých modalit, která jsou spolu s dalšími parametry uvedeny v Tabulce 2.1.

MedMNIST2D	Modalita dat	Úkol (počet tříd)	Množství dat
PathMNIST	patologie tlustého střeva	více tříd (9)	107,180
ChestMNIST	rentgen hrudi	více rozdělení (14), dvě třídy (2)	112,120
DermalMNIST	dermatoskopie	více tříd (7)	10,015
OCTMNIST	OCT sítnice	více tříd (4)	109,309
PneumonialMNIST	rentgen hrudi	dvě třídy (2)	5,856
RetinaMNIST	fundus kamera	ordinální regrese (5)	1,600
BreastMNIST	ultrazvuk prsu	dvě třídy (2)	780
BloodMNIST	mikroskop krevních buněk	více tříd (8)	17,092
TissueMNIST	mikroskop kůry ledvin	více tříd (8)	236,386
OrganAMNIST	CT břišní dutiny	více tříd (11)	58,850
OrganCMNIST	CT břišní dutiny	více tříd (11)	23,660
OrganSMNIST	CT břišní dutiny	více tříd (11)	25,221

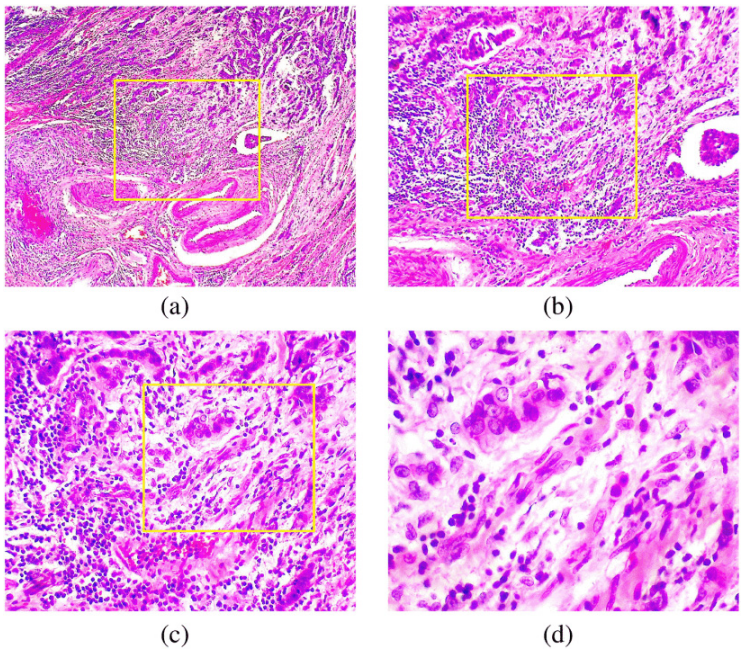
Tab. 2.1: Parametry 2D datasetů MedMNIST, převzato z [37]

BreakHis

Dataset BreakHis je sbírkou mikroskopických bioptických obrazů z klinické studie, která byla realizována v průběhu roku 2014 v Brazílii. Zúčastnilo se jí 82 pacientů, u kterých bylo podezření na rakovinu prsu, ze kterých 24 mělo benigní a 58 maligní nádory. Díky zaznamenávání obrazů se zvětšovacími faktory 40×, 100×, 200× a 400× se dataset skládá z celkem 7 909 obrázků. Dataset kategorizuje nádory do čtyř histologicky odlišných benigních typů - adenosis (A), fibroadenoma (F), phyllodes

Obr. 2.5: Příklad vzorků z datasetu OrganAMNIST a jejich označení

tumor (PT) a tubular adenoma (TA) - a čtyř maligních typů - ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) a papillary carcinoma (PC). [31]

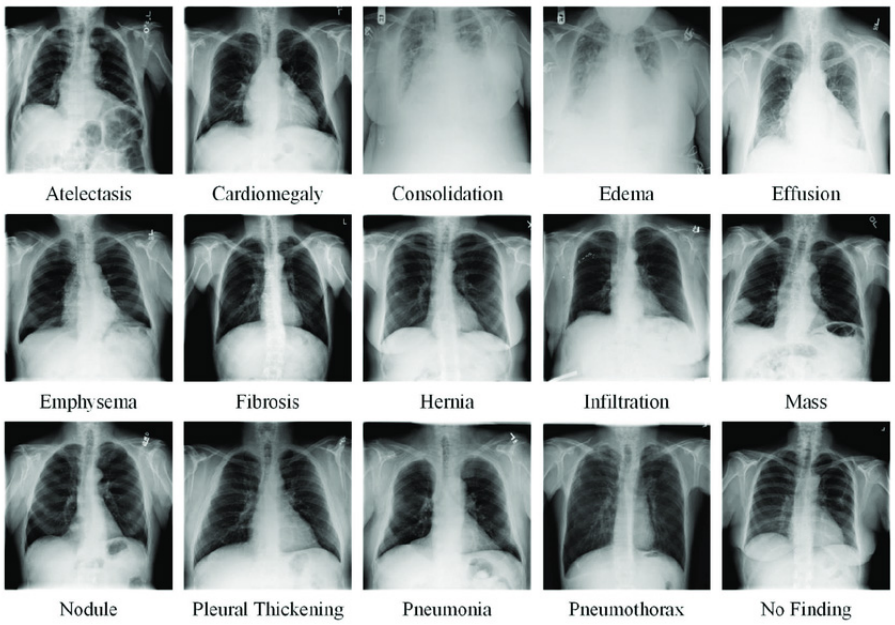


Obr. 2.6: Preparát zhoubného nádoru prsu pod různými zvětšeními (a)40×, (b)100×, (c)200×, (d)400×. Žlutě zvýrazněná oblast je oblast zájmu pro následné zvětšení, převzato z [31]

NIH Chest X-ray dataset

Jedná se o soubor 112 120 frontálních rentgenových snímků hrudi od 30 805 různých pacientů, které byly pořízeny od roku 1992 do roku 2015. Oproti jiným datasetům, kdy jsou anotace tvořeny odborníky, vznikaly pro tento dataset za pomocí algoritmu zpracování přirozeného jazyka, který přiřazoval jednotlivým rentgenům jednu či více ze 14 různých patologií na základě lékařských zpráv. Z tohoto postupu tak vychází, že k jednomu rentgenu může patřit více patologií a rozložení tříd není rovnoměrné.

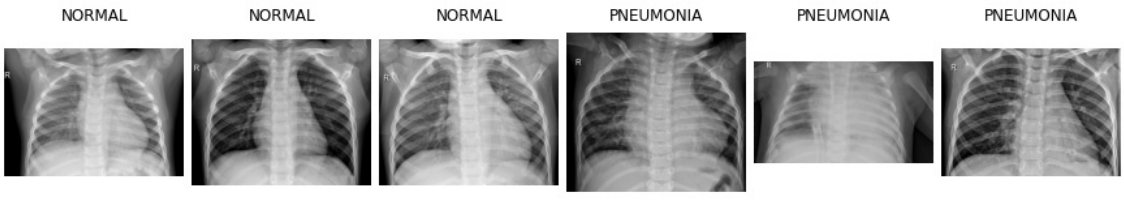
Většina snímků je bez patologie a zbytek odpovídá přibližnému rozložení v populaci. Tato verze, zvaná také jako ChestX-ray14, je rozšířením originálního datasetu ChestX-ray8, který obsahoval pouze 8 patologií a méně snímků. [35]



Obr. 2.7: Snímky reprezentující 14 různých patologií hrudníku a jeden bez nálezu, převzato z [15]

ChestXray

Dataset ChestXray je derivován z obsáhlejší sady ChestX-ray8, která je původně zdrojována z klinického centra NIH. Tento subset obsahuje 5 856 frontálních rentgenových skenů klasifikovaných do 2 tříd: *NORMAL* a *PNEUMONIA*. Všechny obrazy v datasetu jsou ve formátu JPEG a šedotónové, což odráží standardní výstup pro medicínské rentgenové snímání. Jednotlivé obrazy se liší v rozměrech, kdy nejmenší dosahuje rozměru 384x127 pixelů a největší 2916x2583 pixelů. Data jsou rozdělena na 624 testovacích a 5 232 trénovacích vzorků snímaných na odlišných pacientech.



Obr. 2.8: Příklady obrázků z datasetu Chest Xray

3 Metodologie

3.1 Implementace

Architektury byly implementovány za využití programovacího jazyka Python 3.10 v prostředí Spyder. Klíčovou použitou knihovnou byl PyTorch [3] [24], který je optimalizovanou tenzorovou knihovnou pro hluboké učení pomocí GPU a CPU a jeho část Torchvision nabízí adaptované architektury, jako je ResNet18 či ViT-B-16, pro využití v praxi. Pro samotné trénování jsem využil knihovnu Pytorch Lightning [10], která slouží pro zefektivnění práce při trénování modelů a získávání výsledků. Optimalizaci hyperparametrů jsem provedl pomocí frameworku Optuna [2]. Pro debugování a pomoc s vizualizací dat jsem použil nástroj ChatGPT firmy OpenAI.

Pro potřeby tréninku samostatných modelů byl využit počítač Ústavu biomedicínského inženýrství FEKT VUT, který disponuje grafickou kartou Nvidia Titan Xp s 12 GB paměti typu GDDR5, procesorem Intel Xeon E5-2603v4, který má 6 jader běžících na frekvenci 1,7 GHz.

3.2 Použité architektury

Pro provedení porovnání výsledků mezi Vision Transformerem a standardní neuronovou sítí jsem si vybral jako konvoluční neuronovou síť architekturu ResNet18, představenou v roce 2015. [13] Průlomové je u této metody využití reziduálních spojení, které řeší problém s trénováním velmi hlubokých sítí, kdy může docházet ke *zmizení* či *explozi* gradientu. Jejich základní myšlenkou je to, že namísto toho, aby se vrstvy učily přímé mapování vstupů na výstupy, se učí rozdíl mezi vstupem a výstupem, taky označovaný jako reziduum. [13] Například, pokud je požadovaným výstupem bloku $H(x)$, kde (x je vstup), vstvy v bloku se snaží naučit reziduální funkci $F(x) = H(x) - x$. Výstup každého reziduálního bloku je pak $F(x) + x$, což je naučené reziduum přidané zpět ke vstupu. Tímto přístupem se pomáhá bojovat s problémem mizejících gradientů, protože se umožňuje přímo procházet sítí skrze skoková spojení během tréninku. [13]

Představeno bylo několik verzí ResNetu odlišujících se v počtu použitých vrstev, ale také jejich stavbě. ResNet18 a ResNet34 používají základní konfiguraci bloků, kdy množství vrstev odpovídá jejich číselnému označení. Každý blok v této konfiguraci se skládá ze standardních konvolučních vrstev s batch normalizací a aktivacemi ReLU. U architektur ResNet50, ResNet101 a ResNet152 se mění design na využití bottlenecku pro každý reziduální blok. Využívají na jeden blok tři vrstvy místo dvou (jako je tomu u ResNet18 a ResNet34), které zahrnují konvoluci 1x1, poté 3x3 a pak zpět 1x1 pro obnovení rozměrů. Tímto uspořádáním se docílí snížení počtu

parametrů a výpočetní náročnosti ve srovnání s přímým navýšením počtu vrstev. [13]

Vision Transformery, jak byly již popsány v kapitole 1.2, představil v roce 2020 Dosovitskiy et al. [9], a to v několika konfiguracích, které se odlišují především svou velikostí a složitostí. Parametry jednotlivých variant jsou zobrazeny v Tabulce 3.1. Pro moje potřeby jsem si zvolil základní variantu Vision Transformeru, *ViT-Base*, podobně jako jsem zvolil základní variantu u architektury ResNet. Obě rozhodnutí byla ovlivněna také výpočetní náročností, která u složitějších variant roste několikanásobně. Další možnou volbu představuje velikost vstupního výřezu obrázku, který může být buď 16x16 nebo 32x32 pixelů. Menší velikost vede k delší vstupní sekvenci, což může zvýšit výpočetní náročnost, ale také vylepšit schopnost zachytit jemné detaily v obraze. [9]

Model	# vrstev	velikost skryté vrstvy	velikost MLP	# hlav	# parametrů
ViT-Base	12	768	3072	12	86 mil.
ViT-Large	24	1024	4096	16	307 mil.
ViT-Huge	36	1280	5120	16	632 mil.

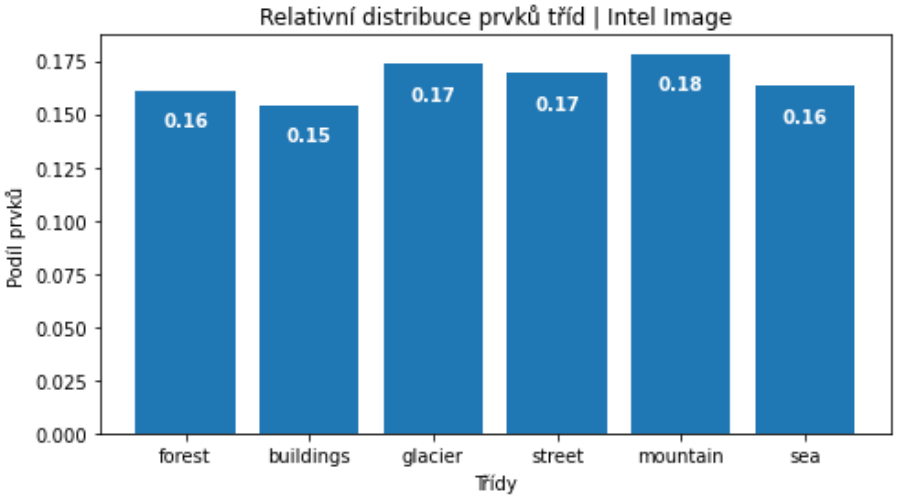
Tab. 3.1: Parametry variant Vision Transformerů, převzato z [9]

Jak již bylo zmíněno dříve v této práci, trénování Vision Transformerů od nuly trvá příliš dlouhou dobu. Konkrétní hodnoty, které uvádí Dosovitskiy et. al. [9], jsou vázány k předtrénování na interním datasetu firmy Google JFT-300M datasetu, který má 300 milionů obrázků, a dosahují 2,5k *TPUv3-core-days* pro největší ViT-H-14 a 0,68k *TPUv3-core-days* pro menší ViT-L-16. Tyto jednotky referují k použití třetí generace TPU (Tensor Processing Unit) firmy Google, které jsou specializované pro urychlení procesů a použití ve strojovém učení. [18] Pro vyjádření doby předtrénování pro ViT-H-14 to znamená, že by se musel trénovat po dobu 2 500 dní, kdyby byla použita jedna jednotka TPUv3. Pro ViT-L-16 by to pak bylo 680 dní.

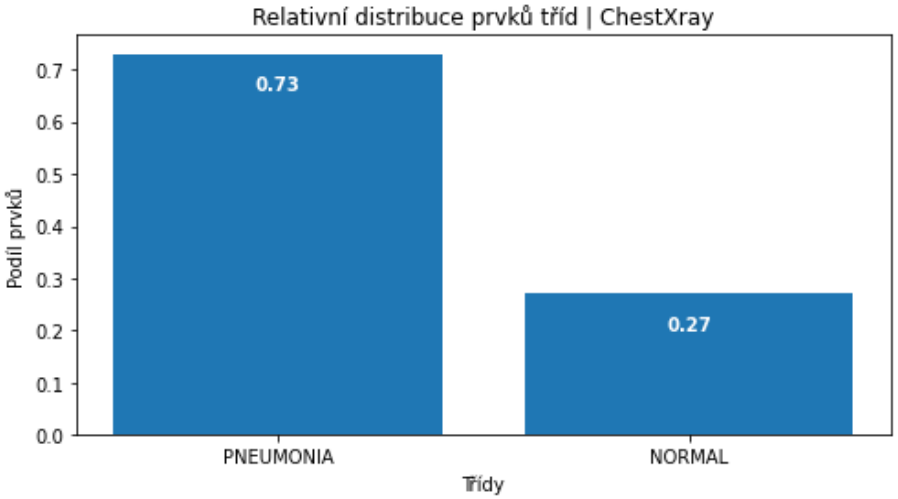
Tento problém se podařil vyřešit použitím předtrénovaných vah, které použitá knihovna *torchvision* nabízí. Tyto váhy byly předtrénovány na obecném datasetu ImageNet-1k, který obsahuje 1 000 různých tříd. Na něm dosáhl předtrénovaný model top1 úspěšnosti 81,072% a top5 úspěšnosti 95,318%. Jelikož knihovna nabízela předtrénované váhy na stejném datasetu i pro ResNet18, využil jsem jich také pro urychlení mého učení.

Datasety, které jsem používal v mé práci, byly již předzpracované a nebylo třeba provádět tolik úprav. U obecného, Intel Image Classification, bylo klíčové to, že jsem nevyužil všechny obrázky, které byly k dispozici, a to z toho důvodu, že v predikční složce sice bylo 7 302 vzorků, žádné ale nebyly označeny a pro účely trénování nebo

testování tak neposloužily. Samotné obrázky jsem před vstupem do obou architektur převedl na šedotónové a změnil jejich velikost z původních 150x150 na 224x224 pixelů. Medicínský dataset ChestXray obsahoval obrázky ve velkých rozměrech, což způsobovalo podstatné zdržení práce s daty. Protože jsem si stanovil, že do obou architektur budou obrázky vstupovat v pevné velikosti 224x224 pixelů, vytvořil jsem si kopii tohoto datasetu, kde byla změněna velikost veškerých obrazů na tento rozměr, což velmi urychlilo práci s daty. Na Obrázku 3.1 a 3.2 jsou zobrazeny relativní distribuce prvků v jednotlivých datasetech. U datasetu Intel Image Classification je vidět, že jsou data poměrně vyrovnaná, zatímco u medicínského, ChestXray, je méně pacientů ve zdravém stavu.



Obr. 3.1: Relativní distribuce prvků v datasetu Intel Image Classification



Obr. 3.2: Relativní distribuce prvků v datasetu ChestXray

3.3 Optimalizace

Pro dosažení co nejlepších výsledků bylo mým cílem provést prvotní optimalizaci se širokým rozsahem parametrů a tu poté zúžit na základě výsledků pro efektivnější výběr. Hyperparametry, které jsem volil u mých modelů, byl krok učení, batch size, optimalizační algoritmus a jeho weight decay. Jako ztrátová funkce byla zvolena vzájemná entropie.

Krok učení jsem vybíral z rozsahu 10^{-5} až 10^{-1}, který pokrývá široké spektrum všech možných hodnot pro učení. Batch size byl zvolen odlišný u konvoluční neuronové sítě a transformeru z důvodu nedostatečné paměti počítače, na kterou jsem narážel během tréninku, a byly tak stanoveny hodnoty od 32 do 86 pro ResNet a od 32 do 64 pro Vision Transformer. Optimalizační algoritmy byly vybrány tři, a to RMSprop, Adam a Adamax. Pro předcházení přetrénování byla zavedena u jednotlivých metod také L2 regularizace v podobě hyperparametru *weight decay*, který byl nastaven v rozsahu od 0 do 0,1.

3.3.1 Framework Optuna

Optimalizace hyperparametrů byla provedena za pomocí aplikačního prostředí Optuna, které využívá několik sofistikovaných algoritmů pro vzorkování, které efektivně navigují prostor hyperparametrů, a to konkrétně *Tree-structure Parzen Estimator (TPE)* a *Covariance Matrix Adaptation Evolution Strategy (CMA-ES)*. [2]

Tree-structure Parzen Estimator (TPE) funguje tak, že sestavuje modely pro odhad výkonnosti hyperparametrů na základě již použitých dat. Toho dosahuje modelováním dvou odlišných hustot rozdělení pravděpodobnosti (anglicky *Probability Density Function*) - jednu pro hyperparametry vedoucí k lepšímu výkonu (označované jako "dobré") a jednu pro ty, které vedou k horšímu výkonu (označované jako "špatné"). Tyto distribuce jsou obvykle odvozeny pomocí jádrového odhadu distribuční funkce (anglicky *Kernel Density Estimation*), neparametrického způsobu odhadu hustoty rozdělení pravděpodobnosti náhodné proměnné. Při rozhodování pro výběr nové sady hyperparametrů využívá TPE akviziční funkci odvozenou z poměru obou hustot rozdělení pravděpodobnosti. Cílem je vybírat hyperparametry tak, aby maximalizovaly tento poměr, který odpovídá oblastem prostoru, kde je pravděpodobnost optimalizace vyšší. Akviziční funkce používaná v TPE je podobná očekávanému zlepšení (anglicky *Expected Improvement*), které je používáno v bayesovské optimalizaci a pomáhá vyvažovat průzkum nových oblastí proti využití známých dobrých oblastí. [36]

Covariance Matrix Adaptation Evolution Strategy (CMA-ES) začíná generováním souboru potenciálních řešení, kde každé řešení je definováno sadou

parametrů. Ty jsou vzorkovány z vícerozměrné normální distribuce, jejíž střed je umístěn na aktuálně nejlepším odhadu řešení. Kovariační matice tohoto rozdělení řídí vzorkování a podporuje průzkum prostoru ve směrech, kde bylo v minulosti pozorováno zlepšení. Jádrem této metody je její adaptační mechanismus pro kovariační matici, který zaznamenává strategii pochopení prohledávaného prostoru. Adaptace je založena na konceptu tzv. evolučních cest, které akumulují informace o úspěšných směrech hledání napříč několika generacemi. Tyto údaje se používají pro úpravu kovarianční matice, čímž se zvýrazňují směry, které v minulosti vedly ke zlepšení výsledků, a modifikuje se tvar distribuce pro vyhledávání. [11]

Tyto zmíněné metody mohou být v Optuně kombinovány také s metodami pruningu, které dále zlepšují optimalizační proces. Jejich integrace umožňuje brzké ukončení méně slibných pokusů a šetří tím výpočetní zdroje a soustředí úsilí na co nejlepší konfigurace hyperparametrů.

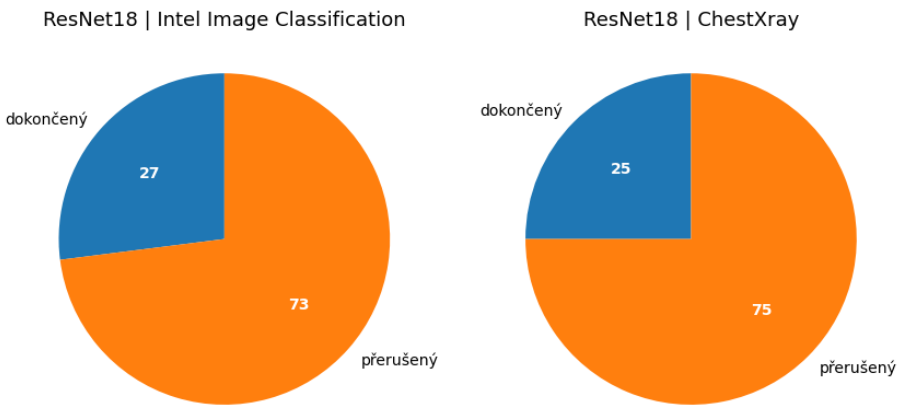
Popsané aplikační prostředí Optuna jsem využil pro vytvoření systému optimalizace hyperparametrů, které mi dovolovalo nastavit zmíněné rozsahy parametrů, ukládat výsledky trénování a také jejich následné porovnávání. Postup jsem měl nastaven tak, že jsem prováděl čtyři různá učení, a to jako kombinace pro dvě architektury, konvoluční neuronovou síť a vision transformer, a dva datasety, obecný a medicínský.

3.3.2 Prvotní optimalizace podle vzájemné entropie

Pro optimalizaci hyperparametrů achitektury ResNet na obecném a medicínském datasetu jsem provedl celkem 100 pokusů ukládaných do databáze formátu *sqlite3*, která byla jednoduše zobrazitelná v nástroji knihovny Optuna, Optuna Dasboard. K porovnávání a posuzování úspěšnosti pokusu docházelo na základě hodnot ztrátové funkce vzájemné entropie na validačních datech po dobu 15 epoch. V průběhu jsem zde tedy sledoval konečnou hodnotu po uplynutí tréninku a také její průběh po dobu učení. Zajímalo mě, jestli hodnota opravdu postupně konvergovala či jen neoscilovala mezi nízkými a vysokými hodnotami ztráty.

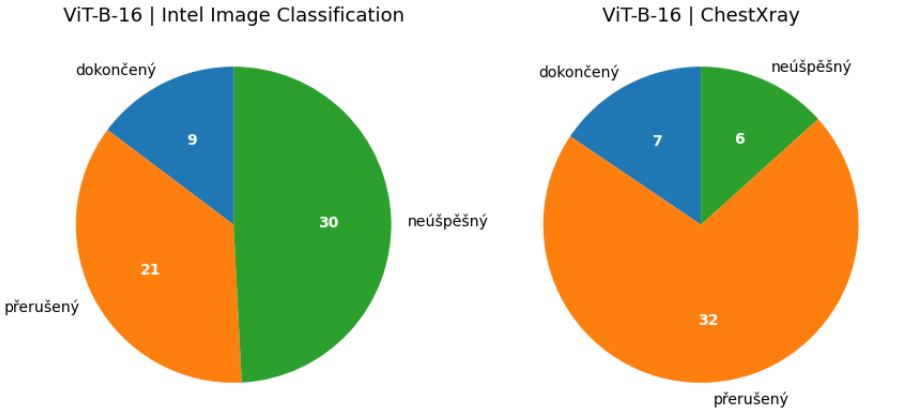
Aby nedocházelo ke zbytečnému učení pokusů, které už od první epochy nedosahují dobrých výsledků, byla implementována již popsaná metoda pruningu. Díky té jsem dokázal vyzkoušet více kombinací hyperparametrů v kratším čase a ze 100 pokusů, které byly provedeny na každém datasetu pro konvoluční síť, došlo k pruningu u 73 pokusů u obecného a 75 u medicínského datasetu, jak je zobrazeno na Obrázku 3.3.

Podle stejného protokolu jsem postupoval také u Vision Transformerů, tedy hodnocení úspěšnosti trénování po 15 epochách podle vzájemné entropie na validačních datech. V tomto případě trval však trénink podstatně déle a provedl jsem tak pouze



Obr. 3.3: Výsledek jednotlivých pokusů pro ResNet architekturu na dvou datasetech

61 a 45 pokusů na jednotlivých datasetech. U této architektury docházelo k zastavování pokusů nikoli na základě pruningu, ale selhání, kdy hodnoty vzájemné entropie dosahovaly hodnot *NaN*. Na základě zpětné analýzy, kterou jsem provedl, to bylo zapříčiněno příliš velkými kroky učení, které způsobovaly explodující gradienty a příliš velké hodnoty ztráty vzájemné entropie, které poté dosahovaly nedefinovaných hodnot. V Obrázku 3.4 je zobrazeno rozložení úspěšnosti pokusů v tomto kroku optimalizace, které ukazuje nutnost přizpůsobit výběr rozsahu pro hyperparametry nejen z důvodu omezení výše zmíněného problému, ale také optimalizace učení a dosáhnutí lepších výsledků.

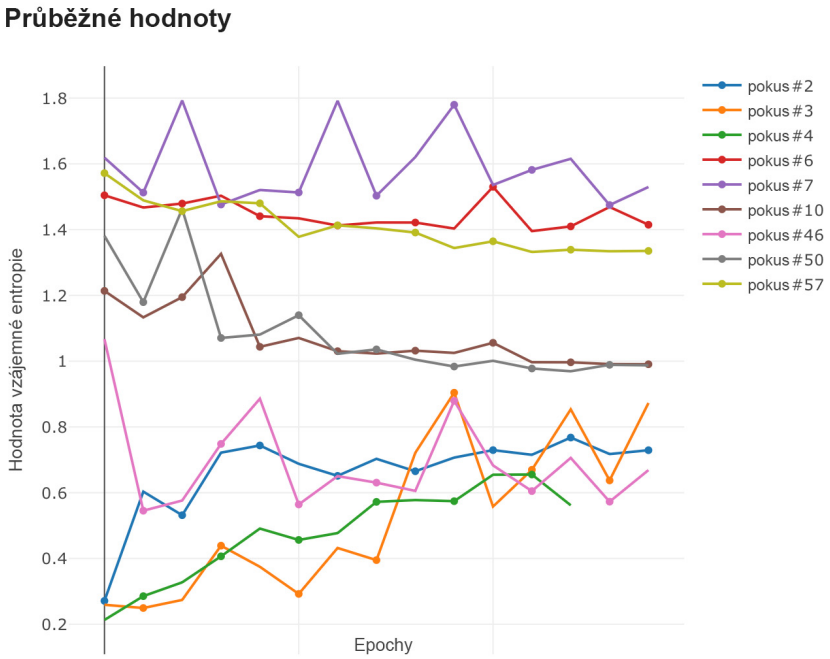


Obr. 3.4: Výsledek jednotlivých pokusů pro ViT architekturu na dvou datasetech

3.3.3 Optimalizace podle F1 skóre

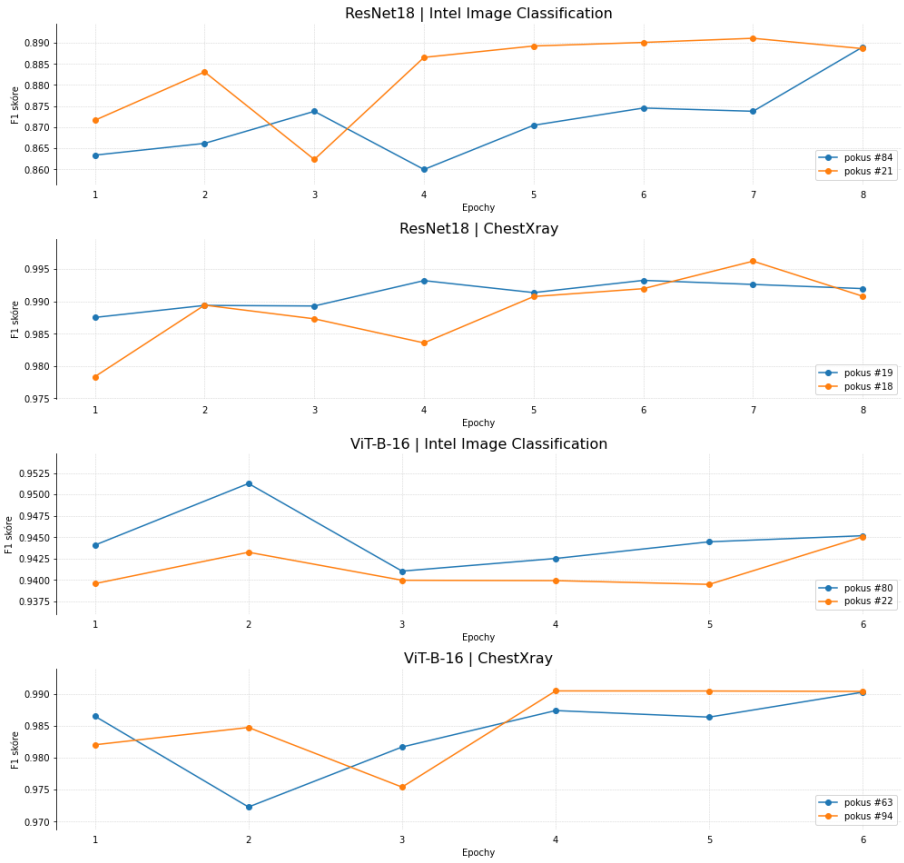
Mimo limitaci rozsahů pro následnou optimalizaci jsem se rozhodl také změnit metriku pro hodnocení úspěšnosti pokusu. Jak je zobrazeno na Obrázku 3.5, průběžné

hodnoty vzájemné entropie u pokusů na obecném datasetu s architekturou Vision Transformeru nekonvergovaly správně a naopak rostly. Z toho důvodu jsem se přiklonil k použití jiné kriteriální metriky, a to F1 skóre na validačních datech, které oproti úspěšnosti bere v potaz i nevyváženosti datasetů a dobře reprezentuje kvalitu modelu.



Obr. 3.5: Průběžné hodnoty vzájemné entropie při pokusech u Vision Transformeru na obecném datasetu

Na základě znalostí z průběhu učení u prvního testování jsem upravil kromě kriteriální metriky také množství epoch, po které se modely učily, a to bylo stanoveno na 8 u konvoluční neuronové sítě a 6 u transformerové architektury. Pokusů pak bylo provedeno 100 pro každou kombinaci architektury a datasetu, načež jsem získal hyperparametry pro použití na trénink svých modelů. Z pokusů jsem volil vždy ten s nejvyšší hodnotou F1 skóre, ale přihlížel jsem také na to, jak křivka konverguje. Nejlepší dva pokusy u každé varianty modelu jsou zobrazeny na Obrázku 3.6 a modře je zvýrazněn ten, který jsem vybral jako finální. Varianta, kde je použit Vision Transformer na datasetu ChestXray, je jediná, kde jsem nezvolil nejlepší pokus podle finální hodnoty validačního F1 skóre, ale až ten druhý, a to na základě minimální odchylky a také lepšího průběhu učení. V Tabulce 3.2 jsou zobrazeny zvolené hyperparametry pro trénink jednotlivých modelů na základě provedené optimalizace.



Obr. 3.6: Průběžné hodnoty validačního F1 skóre během tréninku u architektur na různých datasetech

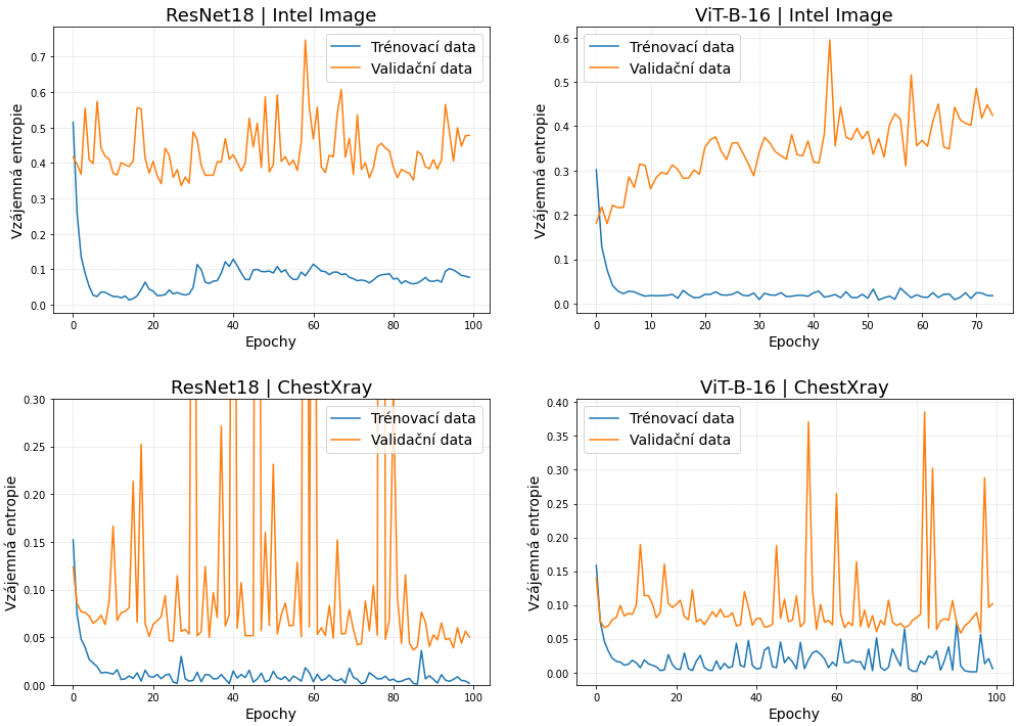
Architektura	Dataset	Krok učení	Batch size	Weight decay	Optimalizační algoritmus
ResNet18	Intel Image	$2,442 \cdot 10^{-4}$	61	$2,649 \cdot 10^{-2}$	Adamax
ResNet18	ChestXray	$1,798 \cdot 10^{-5}$	52	$7,314 \cdot 10^{-2}$	RMSprop
ViT-B-16	Intel Image	$1,267 \cdot 10^{-5}$	32	$7,585 \cdot 10^{-4}$	Adam
ViT-B-16	ChestXray	$1,959 \cdot 10^{-5}$	64	$1,636 \cdot 10^{-2}$	Adamax

Tab. 3.2: Výsledný výběr hyperparametrů pro trénink modelů

3.4 Trénink modelů

Aby bylo vidět, jak se modely trénují v dlouhodobém měřítku, nechal jsem je nejprve trénovat po dobu 100 epoch. Průběh učení je zobrazen na Obrázku 3.7, kde můžeme pozorovat rychlý pokles hodnot vzájemné entropie na trénovacích datech, a to jak u konvoluční neuronové sítě, tak u Vision Transformeru. Na základě postupu validačních ztrát jsem se rozhodl trénovat modely na méně epochách, které byly stanoveny pro ResNet18 na 27 epoch a pro ViT-B-16 na 18 v případě datasetu Intel Image. U první architektury jsem se pro to rozhodl konkrétně kvůli poklesu hodnoty vzájemné entropie na validačních datech, ke které dochází v rozmezí od 20. po 30.

epochu, a také následnému nárůstu ztrát na trénovacích datech. U druhé architektury, tedy Vision Transformeru, dochází u vzájemné entropie na validačních datech k postupnému nárůstu, který se ale u výsledné hodnoty F1 skóre tolik neprojevuje. I přesto jsem nechtěl trénovat tento model příliš dlouho a hranice tak byla nastavena na 18 epoch kvůli nárůstu ztráty, ke které dochází poté okolo 20. epochy. Pro modely trénované na medicínském datasetu jsem postupoval obdobně. U konvoluční neuronové sítě ResNet18 lze pozorovat fluktuace vzájemné entropie validačních dat, a proto jsem zvolil pro ukončení trénování epochu 25, po které dochází k nejvyšším výkyvům. U Vision Transformeru jsem podobně jako na obecných datech zvolil kratší dobu trénování, tj. 16 epoch, jelikož dále vzájemná entropie také kolísá.



Obr. 3.7: Průběh učení po dobu 100 epoch pro modely ResNet18 a ViT-B-16 vytvořené na datasetech Intel Image a ChestXray

4 Výsledky a diskuze

4.1 Statistická analýza

V této části prezentuji analýzu výkonu modelů ResNet18 a ViT-B-16 na dvou různých datasetech. Výkon je měřen pomocí váženého F1 skóre vypočítaného funkcí $classification\_report$ z knihovny $sklearn.metrics$ [25] a na tuto metriku se budu v práci odkazovat i dále. Každý model byl trénován desetkrát, aby se zohlednila jeho variabilita při tréninku. Jelikož jsou počáteční váhy využívány již z předtrénovaých modelů, vychází variabilita převážně z náhodného výběru subsetu pro validaci dat.

Tabulka 4.1 poskytuje přehled popisné statistiky vážených F1 skóre získaných během tréninku. Lze zde pozorovat dosažení vyšších středních hodnot i maximálních hodnot pro vážené F1 skóre u modelů využívající Vision Transformery. Zároveň se ukazuje také nižší směrodatná odchylka pro obecný dataset Intel Image, obzvláště pak právě také pro transformerovou architekturu.

Architektura	Model	Střední hodnota F1 skóre	Směrodatná odchylka	Maximální F1 skóre
ResNet18	Intel Image	0,868	0,015	0,883
ViT-B-16	Intel Image	**0,930**	0,011	**0,939**
ResNet18	ChestXray	0,843	0,025	0,885
ViT-B-16	ChestXray	**0,881**	0,025	**0,907**

Tab. 4.1: Popisná statistika vážených F1 skóre pro modely ResNet18 a ViT-B-16 na dvou datasetech

Aby bylo možné výsledky zhodnotit také objektivně a určit, zda je rozdíl mezi výsledky statisticky signifikantní, provedl jsem sérii statistických testů. Pro ověření normálního rozdělení jsem využil Shapiro-Wilkův test, který slouží nejlépe pro použití s malým množstvím vzorků. [23] Pouze výsledky jednoho modelu, ResNet18 na datasetu ChestXray, vykazovaly normalitu dat, a to s p hodnotou 0,93. Ostatní modely se pohybovaly pod hodnotou 0,05, tj. 0,0012 pro ResNet na Intel Image, 0,0064 pro ViT na Intel Image a 0,038 pro ViT na ChestXray – a pro porovnání byl tedy využit neparametrický párový Wilcoxonův test.

Párové porovnání bylo provedeno mezi modely trénovanými na stejném datasetu. Pro varianty vázající se k obecnému datasetu Intel Image byla získána p hodnota Wilcoxonova testu $1,95 \cdot 10^{-3}$, která naznačuje statisticky významný rozdíl mezi výslednými hodnotami natrénovaných modelů ResNet18 a ViT-B-16. Pro medicínský dataset ChestXray dosahovaly výsledky neparametrického Wilcoxonova testu p hodnoty $1,95 \cdot 10^{-2}$ a taktéž statisticky signifikantní rozdíl mezi jednotlivými modely zdůrazňující vyšší účinnost architektury ViT-B-16.

4.2 Porovnání výsledků

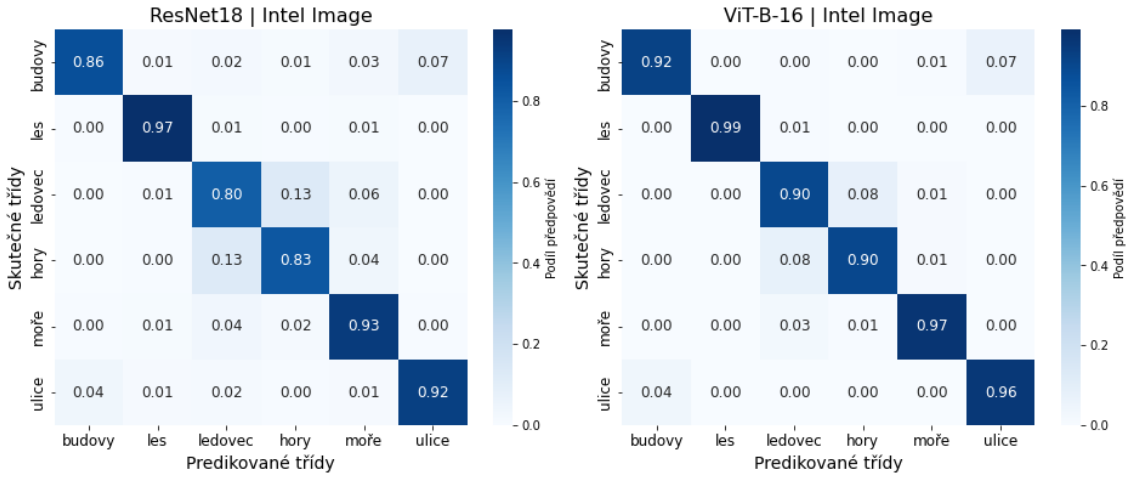
Nejlepší model pro každou variantu jsem vybral na základě nejvyšší hodnoty váženého F1 skóre a jejich výsledné výkonnostní metriky jsou zobrazeny v Tabulce 4.2. Absolutní rozdíl mezi modely je přibližně 0,056 u aplikace na datasetu Intel Image a 0,022 pro medicínský dataset ChestXray. Když dojde k porovnání hodnot konvolučních neuronových sítí, dosahuje nyní model ResNet18 na datasetu ChestXray lepších výsledků než na obecných datech. Na základě hodnot zobrazených v Tabulce 4.1 lze ale říci, že stabilnější výsledky bude poskytovat model ResNet18 Intel Image, který při provedení více tréninkových cyklů dosahoval výsledků s menší směrodatnou odchylkou.

Architektura	Model	F1 skóre	Precision	Recall
ResNet18	Intel Image	0,883	0,884	0,883
ViT-B-16	Intel Image	**0,939**	**0,939**	**0,939**
ResNet18	ChestXray	0,885	0,904	0,889
ViT-B-16	ChestXray	**0,907**	**0,920**	**0,910**

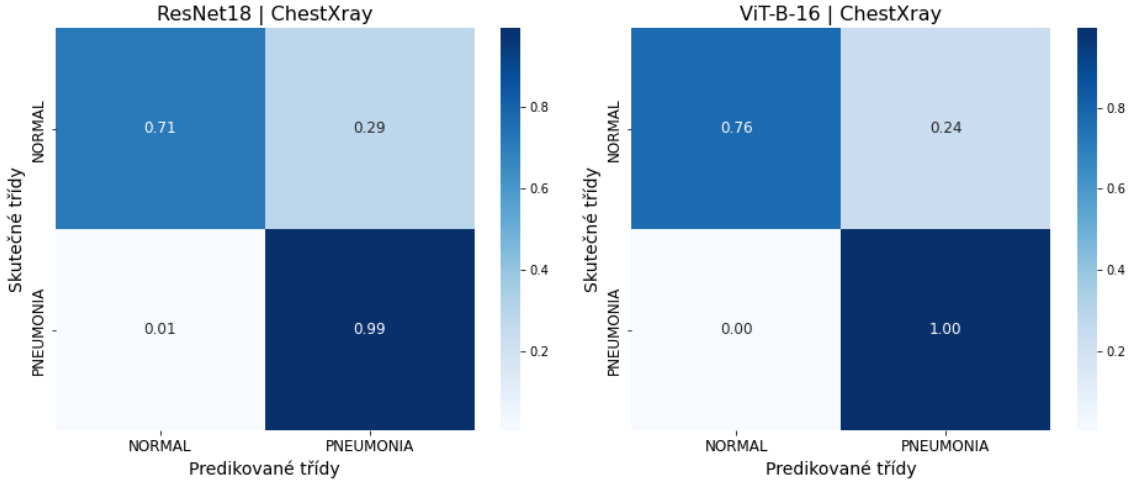
Tab. 4.2: Výsledné výkonnostní metriky pro jednotlivé modely

Na Obrázku 4.1 je zobrazena normalizovaná matice záměn pro oba modely trénované na datasetu Intel Image Classification. V obou případech lze hovořit o dobré klasifikaci pro třídu lesa, která dosahuje hodnot 0,97 pro ResNet18 a 0,99 pro Intel Image. Podstatný rozdíl lze ale pozorovat u klasifikace hor a ledovce. Ke zvýšené záměně dochází i u modelu ViT-B-16, to ale menší než u ResNet18. Důvod pro záměnu samotnou může být přisouzen podobnosti vizuálních rysů těchto tříd, kdy ledovce mohou obsahovat v obrázku také hory a stejně tak i moře, k jehož záměně dochází převážně u ResNetu také. K další chybě, pravděpodobně ze stejného důvodu, dochází i mezi klasifikací budov a ulice. Obecně lze ale říci, že i když u modelu ViT-B-16 k těmto chybám dochází také, vyvarovává se ve větší míře chybné klasifikaci mezi jinými třídami oproti modelu ResNet18.

Na Obrázku 4.2 jsou zobrazeny normalizované matice záměn pro oba modely trénované na medicínském datasetu ChestXray. Zde se na výsledku podílí nevyváženost datasetu a hlavním rozdílem mezi jednotlivými modely je to, že u ViT-B-16 dochází k nižší míře záměny mezi třídami. Takto natrénovaný model vede ke zvýšené míře falešně pozitivně klasifikovaných zdravých pacientů, ale na druhou stranu limituje falešně negativní detekce, které by pacientům mohly uškodit.



Obr. 4.1: Normalizovaná matice záměn pro modely na obecném datasetu Intel Image Classification



Obr. 4.2: Normalizovaná matice záměn pro modely na medicínském datasetu ChestXray

4.3 Výpočetní náročnost

Přestože dosahují modely Vision Transformerů lepších výsledků než standardní konvoluční síť, jak již bylo představeno, jejich nevýhodou je výpočetní náročnost, která bude následně demonstrována několika metrikami.

4.3.1 Velikost modelů

Počet parametrů v modelu ukazuje jeho schopnost učit se z dat a obecně platí, že čím více parametrů, tím složitější souvislosti jsou schopny modely zachytit. S tím je

ale také spojeno větší riziko přeučení. Jak ukazuje Tabulka 4.3, modely Vision Transformeru mají výrazně vyšší počet parametrů, přibližně 85,8 milionu, ve srovnání s modely ResNet18, které mají přibližně 11,7 milionu parametrů.

Velikost modelu v megabytech může být také relevantní pro určitá použití v prostředích s omezenými úložnými kapacitami. Na základě dat v Tabulce 4.3 jsou modely ViT výrazně větší, s velikostmi kolem 982 MB, zatímco modely ResNet18 jsou menší, s velikostmi 127,99 MB pro aplikaci na datasetu Intel Image Classification a 85,32 MB pro ChestXray. Zatímco modely ViT mohou nabízet lepší výkon díky vyšší složitosti, vyžadují také výrazně více úložného prostoru, což nemusí být ve všech scénářích použití proveditelné.

Architektura	Dataset	Počet parametrů (mil.)	Velikost
ResNet18	Intel Image	11,173	127,99 MB
ViT-B-16	Intel Image	85,803	982,12 MB
ResNet18	ChestXray	11,171	85,32 MB
ViT-B-16	ChestXray	85,800	982,08 MB

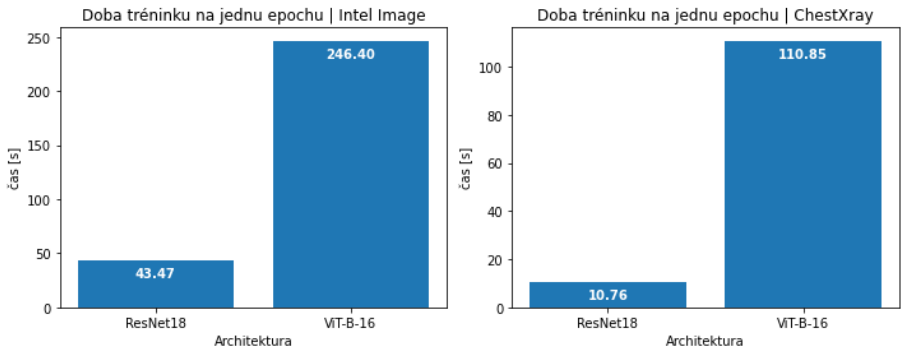
Tab. 4.3: Počet parametrů a velikost jednotlivých modelů

4.3.2 Čas trénování

Přestože dosahuje architektura Vision Transformeru lepších výsledků než standardní konvoluční síť, jejich hlavní nevýhodou, mimo zmíněnou velikost modelů, je výpočetní náročnost. Tu jsem vyjádřil na jednu epochu, ve které je zahrnuto trénování na tréninkových datech, ale také validace na validačním subsetu. Data jsem získal z tréninku na 100 epochách a následného přepočtu pouze na jednu. Při tréninku na hardwaru zmíněném v Kapitole 3.1, který disponuje grafickou kartou, dosáhl čas tréninku na jednu epochu pro ViT na datasetu Intel Image přibližně 246 vteřin, což je řádově šestkrát více než pro trénink modelu ResNet18. Pro medicínský dataset, který je menší a jedná se pouze o binární klasifikaci, dosahovaly hodnoty doby tréninku pro ViT téměř 111 vteřin oproti pouze 10,76 vteřinám pro ResNet18, kde se poměr zvýšil na jedenáctinásobek hodnoty.

4.3.3 FLOPs

Jiným způsobem, jak lze objektivně vyjádřit náročnost modelu, je metrika zvaná FLOPs (Floating Point Operations), která udává počet operací s pohyblivou řádovou čárkou pro inferenci jednoho vzorku z datasetu. Vyšší počet FLOPs znamená potřebu vyššího výpočetního výkonu, ale také vyšší energetické nároky a delší dobu



Obr. 4.3: Porovnání doby tréninku na jednu epochu mezi jednotlivými kombinacemi datasetů a architektur

inference. Počet FLOPs je považován za spolehlivý ukazatel skutečné latence a spotřeby energie, přičemž je přesnější než počet parametrů modelu při hodnocení energetické náročnosti a latence. [32]

Tabulka 4.4 porovnává právě tyto hodnoty pro ResNet18 a ViT-B-16 na dvou použitých datasetech. Pro každou kombinaci je uvedena hodnota v milionech, ale také celková, aby nebyly zanedbané malé rozdíly, ke kterým dochází mezi variantami na obecném a medicínském datasetu, které jsou v řádech tisíců. Tyto malé rozdíly by měly být ovlivněny počtem klasifikovaných tříd, jelikož obecný dataset Intel Image je rozdělen do šesti tříd, zatímco u medicínského ChestXray se jedná o binární klasifikaci.

Hlavní porovnání, které je pro tuto práci důležité, je mezi architekturami ResNet18 a ViT-B-16. Zde dosahuje Vision Transformer hodnot 16,85 GFLOPs oproti 1,74 GFLOPs pro konvoluční neuronovou síť.

Architektura	Dataset	FLOPs	FLOPs (G)
ResNet18	Intel Image	1 737 372 672	1,74
ViT-B-16	Intel Image	16 851 519 793	16,85
ResNet18	ChestXray	1 737 370 624	1,74
ViT-B-16	ChestXray	16 851 516 721	16,85

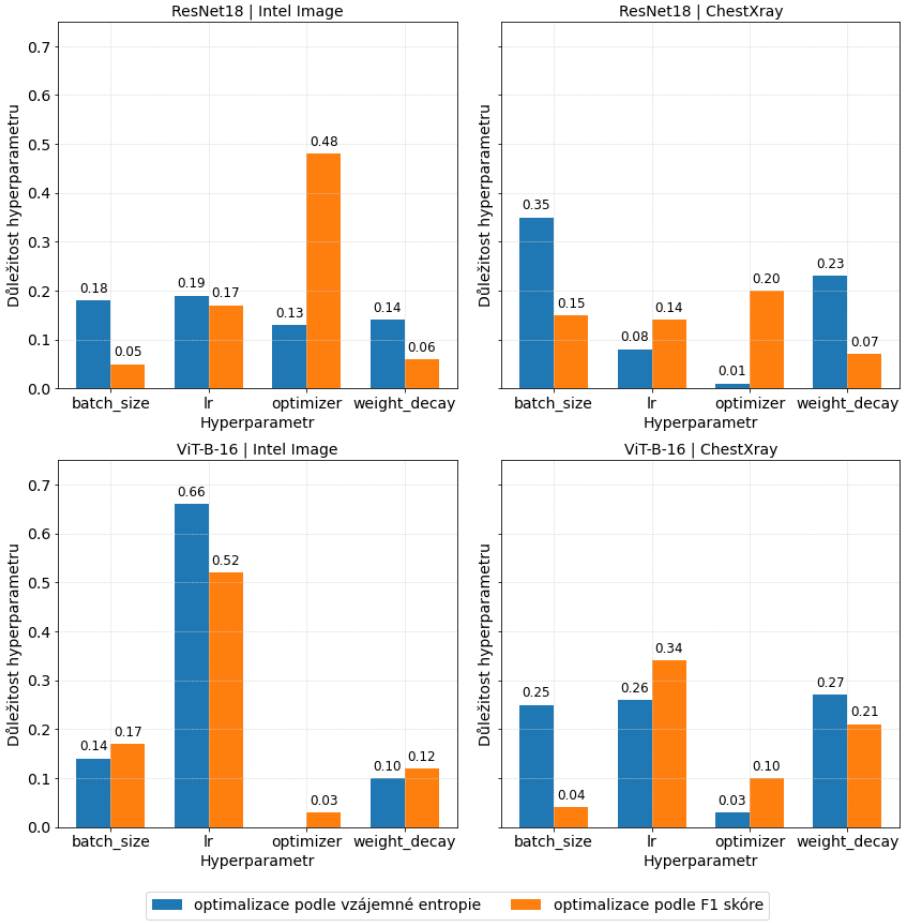
Tab. 4.4: FLOPs hodnoty pro jednotlivé modely

4.4 Důležitost hyperparametrů

Během optimalizace vypočítává framework Optuna důležitost hyperparametrů. K tomu využívá Funkční analýzu rozptylu (fANOVA) k posouzení příspěvku každého

hyperparametru k celkové variabilitě výkonu modelu. Proces začíná fitováním modelu náhodného lesa na data shromážděná během optimalizačního procesu založeného na Bayesovských metodách a rozděluje hyperparametrový prostor na oblasti s konstantními predikčními hodnotami. [16]. Na základě těch vypočítává marginální efekt každého hyperparametru tím, že zprůměruje výkon modelu přes všechna možná nastavení ostatních hyperparametrů. Výsledná důležitost je kvantifikována jako podíl celkového rozptylu.

Tímto přístupem umožňuje detailní rozbor toho, jak jednotlivé hyperparametry a jejich interakce přispívají k variabilitě výkonu. Analyzováním dat z náhodného lesa identifikuje, které hyperparametry jsou nejdůležitější pro ladění a jak spolu interagují. [16] Znalosti důležitosti hyperparametrů jsem při samotné práci nevyužil, díky tomuto pohledu lze ale získat větší porozumění optimalizačnímu procesu.



Obr. 4.4: Porovnání důležitosti parametrů pro jednotlivé modely a fáze optimalizace

Na Obrázku 4.4 je zobrazeno porovnání důležitostí během optimalizace na základě vzájemné entropie a následné optimalizace podle F1 skóre na všech čtyřech modelech. U kombinace architektury ResNet18 a datasetu Intel Image Classification

vystupuje oproti ostatním u druhé optimalizace hodnota důležitosti optimalizačního algoritmu. Zde docházelo k výběru mezi metodami Adamax a RMSprop a ovlivněno to může být tím, že 14 nejlepších dokončených pokusů využilo algoritmus Adamax a až pro patnáctý nejlepší výsledek dle validačního F1 skóre byla použita metoda RMSprop. Po limitaci rozsahů lze pozorovat také pokles důležitosti parametrů *batch_size* a *weight_decay*, který může být zapříčiněn právě optimalizovaným rozsahem, ale také nárůstem důležitosti jiných parametrů. Volba kroku učení se potvrzuje jako zásadní pro trénink Vision Transformerů, kdy naopak optimalizační algoritmus dosahuje hodnot velmi malých oproti architekturám ResNet18.

Závěr

Tato diplomová práce se zaměřila na možnosti využití neuronových sítí založených na architektuře transformerů pro zpracování medicínských obrazů. Byl představen jejich původ pro zpracování přirozeného jazyka i jejich adaptace na obrazová data. Uvedeny byly také varianty použité na praktických příkladech, kde dochází např. ke kombinaci se standardními konvolučními sítěmi.

Bylo provedeno srovnání výkonnosti modelů ResNet18 a Vision Transformeru (ViT-B-16) na dvou odlišných datasetech, konkrétně obecném Intel Image Classification a medicínském ChestXray. Výsledky ukázaly, že modely využívající transformerovou architekturu dosahují vyšších hodnot váženého F1 skóre ve srovnání s modely ResNet18. Konkrétně dosáhl model ViT-B-16 nejvyššího F1 skóre 0,939 na datasetu Intel Image a 0,907 na datasetu ChestXray, zatímco pro ResNet18 to bylo 0,883, respektive 0,885. Statistická analýza pomocí Wilcoxonova testu potvrdila, že rozdíly ve výkonnosti modelů jsou statisticky signifikantní a lze tak říci, že použitím Vision Transformeru dosáhneme lepších výsledků.

Vyšší úspěšnost modelu s sebou však nese vyšší výpočetní nároky, a to nejen při inferenci, ale také trénovacím procesu a optimalizaci. Čas tréninku na jednu epochu dosahoval pro ViT-B-16 u obecného datasetu Intel Image téměř šestinásobek času oproti ResNet18 a pro medicínský dataset ChestXray se jednalo skoro o jedenáctinásobek. Tento fakt může být klíčovým faktorem při rozhodování použití jedné či druhé varianty v praxi.

Získané výsledky v této práci poskytují pohled do problematiky použití Vision Transformerů v oblasti zpracování medicínských obrazů a porovnávají je v praxi se standardními konvolučními sítěmi. Tím ukazují, že mají transformery značný potenciál k využití pro dosáhnutí větší přesnosti a spolehlivosti diagnostických systémů.

Literatura

[1] ABNAR, Samira; DEHGHANI, Mostafa a ZUIDEMA, Willem, 2020. Transferring Inductive Biases through Knowledge Distillation. Online. Dostupné z: https://arxiv.org/abs/2006.00555. [cit. 2024-01-03].

[2] AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru a KOYAMA, Masanori, 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. Online. Dostupné z: https://arxiv.org/abs/1907.10902. [cit. 2024-05-13].

[3] ANSEL, Jason; YANG, Edward; HE, Horace; GIMELSHEIN, Natalia; JAIN, Animesh et al., 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. Online. In: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. New York, NY, USA: ACM, 2024-04-27, s. 929-947. ISBN 9798400703850. Dostupné z: https://doi.org/10.1145/3620665.3640366. [cit. 2024-05-22].

[4] CARION, Nicolas; MASSA, Francisco; SYNNAEVE, Gabriel; USUNIER, Nicolas; KIRILLOV, Alexander et al., 2020. End-to-End Object Detection with Transformers. Online. Dostupné z: https://arxiv.org/abs/2005.12872. [cit. 2024-01-03].

[5] CUI, Jiequan; TIAN, Zhuotao; ZHONG, Zhisheng; QI, Xiaojuan; YU, Bei et al., 2023. Decoupled Kullback-Leibler Divergence Loss. Online. Dostupné z: https://arxiv.org/abs/2305.13948. [cit. 2024-01-03].

[6] DAI, Yin; GAO, Yifan a LIU, Fayu, 2021. TransMed: Transformers Advance Multi-Modal Medical Image Classification. Online. *Diagnostics*. Roč. 11, č. 8, s. 1-15. ISSN 2075-4418. Dostupné z: https://doi.org/10.3390/diagnostics11081384. [cit. 2023-11-24].

[7] DENG, Jia; DONG, Wei; SOCHER, Richard; LI, Li-Jia; KAI LI et al., 2009. ImageNet: A large-scale hierarchical image database. Online. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, s. 248-255. ISBN 978-1-4244-3992-8. Dostupné z: https://doi.org/10.1109/CVPR.2009.5206848. [cit. 2024-01-02].

[8] DING, Meidan; QU, Aiping; ZHONG, Haiqin a LIANG, Hao, 2021. A Transformer-based Network for Pathology Image Classification. Online. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine*

(BIBM). IEEE, 2021-12-9, s. 2028-2034. ISBN 978-1-6654-0126-5. Dostupné z: https://doi.org/10.1109/BIBM52615.2021.9669476. [cit. 2023-11-24].

[9] DOSOVITSKIY, Alexey; BEYER, Lucas; KOLESNIKOV, Alexander; WEISSENBORN, Dirk; ZHAI, Xiaohua et al., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Online. Dostupné z: https://arxiv.org/abs/2010.11929. [cit. 2024-01-03].

[10] FALCON, William a THE PYTORCH LIGHTNING TEAM, 2019. PyTorch Lightning (Version 1.4). Online. In: . Dostupné z: https://doi.org/https://doi.org/10.5281/zenodo.3828935. [cit. 2024-05-22].

[11] HANSEN, Nikolaus, 2023. The CMA Evolution Strategy: A Tutorial. Online. Dostupné z: https://doi.org/https://doi.org/10.48550/arXiv.1604.0077. [cit. 2024-05-13].

[12] HASSANI, Ali; WALTON, Steven; SHAH, Nikhil; ABUDUWEILI, Abulikemu; LI, Jiachen et al., 2021. Escaping the Big Data Paradigm with Compact Transformers. Online. Dostupné z: https://arxiv.org/abs/2104.05704. [cit. 2024-01-03].

[13] HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing a SUN, Jian, 2015. Deep Residual Learning for Image Recognition. Online. Dostupné z: https://doi.org/10.48550/arXiv.1512.03385. [cit. 2024-05-14].

[14] HINTON, Geoffrey; VINYALS, Oriol a DEAN, Jeff, 2015. Distilling the Knowledge in a Neural Network. Online. Dostupné z: https://arxiv.org/abs/1503.02531. [cit. 2024-01-03].

[15] HUANG, Zhiwei; LIN, Jinzhao; XU, Liming; WANG, Huiqian; BAI, Tong et al., 2020. Fusion High-Resolution Network for Diagnosing ChestX-ray Images. Online. *Electronics*. Roč. 9, č. 1. ISSN 2079-9292. Dostupné z: https://doi.org/10.3390/electronics9010190. [cit. 2024-01-03].

[16] HUTTER, Frank; HOOS, Holger a LEYTON-BROWN, Kevin, 2014. An Efficient Approach for Assessing Hyperparameter Importance. Online. *Proceedings of the 31st International Conference on Machine Learning*. Roč. 2014, č. 32, s. 754–762. Dostupné z: https://proceedings.mlr.press/v32/hutter14.html. [cit. 2024-05-17].

[17] JI, Yuanfeng; ZHANG, Ruimao; WANG, Huijie; LI, Zhen; WU, Lingyun et al., 2021. Multi-compound Transformer for Accurate Biomedical Image Segmentation. Online. In: DE BRUIJNE, Marleen; CATTIN, Philippe C.; COTIN, Stéphane; PADOY, Nicolas; SPEIDEL, Stefanie et al. (ed.). *Medical*

Image Computing and Computer Assisted Intervention — MICCAI 2021. Lecture Notes in Computer Science. Cham: Springer International Publishing, s. 326-336. ISBN 978-3-030-87192-5. Dostupné z: https://doi.org/10.1007/978-3-030-87193-2_31. [cit. 2023-11-24].

[18] JOUPPI, Norman P.; YOON, Doe Hyun; KURIAN, George; LI, Sheng; PATIL, Nishant et al., 2020. A domain-specific supercomputer for training deep neural networks. Online. *Communications of the ACM*. 2020-06-18, roč. 63, č. 7, s. 67-78. ISSN 0001-0782. Dostupné z: https://doi.org/10.1145/3360307. [cit. 2024-05-16].

[19] KRIZHEVSKY, Alex, 2009. Learning Multiple Layers of Features from Tiny Images. Online. *Tech. Report (University of Toronto)*. Dostupné z: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. [cit. 2024-01-03].

[20] LI, Chao; CUI, Yue; LUO, Na; LIU, Yong; BOURGEAT, Pierrick et al., 2022. Trans-ResNet: Integrating Transformers and CNNs for Alzheimer-s disease classification. Online. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022-3-28, s. 1-5. ISBN 978-1-6654-2923-8. Dostupné z: https://doi.org/10.1109/ISBI52829.2022.9761549. [cit. 2023-11-24].

[21] LI DENG, 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. Online. *IEEE Signal Processing Magazine*. Roč. 29, č. 6, s. 141-142. ISSN 1053-5888. Dostupné z: https://doi.org/10.1109/MSP.2012.2211477. [cit. 2024-01-02].

[22] LIN, Tianyang; WANG, Yuxin; LIU, Xiangyang a QIU, Xipeng, 2022. A survey of transformers. Online. *AI Open*. Roč. 3, s. 111-132. ISSN 26666510. Dostupné z: https://doi.org/10.1016/j.aiopen.2022.10.001. [cit. 2024-01-02].

[23] NORNADIAH, Mohd Razali a YAP, Bee, 2011. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. Online. *J. Stat. Model. Analytics*. Roč. 2011, č. 2. Dostupné z: https://www.researchgate.net/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests. [cit. 2024-05-20].

[24] PASZKE, Adam; GROSS, Sam; MASSA, Francisco; LERER, Adam; BRADBURY, James et al., WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; D'ALCHÉ-BUC, F.; FOX, E. et al. (ed.), 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Online. *Advances in Neural Information Processing Systems*. Roč. 2019, č. 32, s. 8024-8035. Dostupné z: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. [cit. 2024-05-21].

[25] PEDREGOSA, F.; VAROQUAUX, G; GRAMFORT, A.; MICHEL, V.; THIRION, B. et al., 2011. Scikit-learn: Machine Learning in Python. Online. *Journal of Machine Learning Research*. Roč. 2011, č. 12, s. 2825–2830. [cit. 2024-05-22].

[26] PRANGEMEIER, Tim; REICH, Christoph a KOEPPL, Heinz, 2020. Attention-Based Transformers for Instance Segmentation of Cells in Microstructures. Online. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020-12-16, s. 700-707. ISBN 978-1-7281-6215-7. Dostupné z: https://doi.org/10.1109/BIBM49941.2020.9313305. [cit. 2023-11-23].

[27] REN, Sucheng; GAO, Zhengqi; HUA, Tianyu; XUE, Zihui; TIAN, Yonglong et al., 2022. Co-advise: Cross Inductive Bias Distillation. Online. In: *2022 IEEE/-CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, s. 16752-16761. ISBN 978-1-6654-6946-3. Dostupné z: https://doi.org/10.1109/CVPR52688.2022.01627. [cit. 2023-11-30].

[28] RODRÍGUEZ, Manuel Alejandro; ALMARZOUQI, Hasan a LIATSIS, Panos, 2023. Multi-Label Retinal Disease Classification Using Transformers. Online. *IEEE Journal of Biomedical and Health Informatics*. Roč. 27, č. 6, s. 2739-2750. ISSN 2168-2194. Dostupné z: https://doi.org/10.1109/JBHI.2022.3214086. [cit. 2023-11-24].

[29] RUSSAKOVSKY, Olga; DENG, Jia; SU, Hao; KRAUSE, Jonathan; SATHEESH, Sanjeev et al., 2015. ImageNet Large Scale Visual Recognition Challenge. Online. *International Journal of Computer Vision*. Roč. 115, č. 3, s. 211-252. ISSN 0920-5691. Dostupné z: https://doi.org/10.1007/s11263-015-0816-y. [cit. 2024-01-02].

[30] SARKER, Md Mostafa Kamal; MORENO-GARCÍA, Carlos Francisco; REN, Jinchang a ELYAN, Eyad, 2022. TransSLC: Skin Lesion Classification in Dermatoscopic Images Using Transformers. Online. In: YANG, Guang; AVILES-RIVERO, Angelica; ROBERTS, Michael a SCHÖNLIEB, Carola-Bibiane (ed.). *Medical Image Understanding and Analysis*. Lecture Notes in Computer

Science. Cham: Springer International Publishing, s. 651-660. ISBN 978-3-031-12052-7. Dostupné z: https://doi.org/10.1007/978-3-031-12053-4_48. [cit. 2023-11-24].

[31] SPANHOL, Fabio A.; OLIVEIRA, Luiz S.; PETITJEAN, Caroline a HEUTTE, Laurent, 2016. A Dataset for Breast Cancer Histopathological Image Classification. Online. *IEEE Transactions on Biomedical Engineering*. Roč. 63, č. 7, s. 1455-1462. ISSN 0018-9294. Dostupné z: https://doi.org/10.1109/TBME.2015.2496264. [cit. 2024-01-02].

[32] TANG, Raphael; ADHIKARI, Ashutosh a LIN, Jimmy, 2018. FLOPs as a Direct Optimization Objective for Learning Sparse Neural Networks. Online. Dostupné z: https://arxiv.org/abs/1811.03060v2. [cit. 2024-05-22].

[33] TOUVRON, Hugo; CORD, Matthieu; DOUZE, Matthijs; MASSA, Francisco; SABLAYROLLES, Alexandre et al., 2021. Training data-efficient image transformers & distillation through attention. Online. Dostupné z: https://arxiv.org/abs/2012.12877. [cit. 2024-01-03].

[34] VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion et al., 2017. Attention Is All You Need. Online. *Advances in Neural Information Processing Systems*. Dostupné z: https://arxiv.org/abs/1706.03762. [cit. 2024-01-03].

[35] WANG, Xiaosong; PENG, Yifan; LU, Le; LU, Zhiyong; BAGHERI, Mohammadhadi et al., 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Online. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, s. 3462-3471. ISBN 978-1-5386-0457-1. Dostupné z: https://doi.org/10.1109/CVPR.2017.369. [cit. 2024-01-03].

[36] WATANABE, Shuhei, 2023. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. Online. Dostupné z: https://doi.org/10.48550/arXiv.2304.11127. [cit. 2024-05-13].

[37] YANG, Jiancheng; SHI, Rui; WEI, Donglai; LIU, Zequan; ZHAO, Lin et al., 2023. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Online. *Scientific Data*. Roč. 10, č. 1. ISSN 2052-4463. Dostupné z: https://doi.org/10.1038/s41597-022-01721-8. [cit. 2024-01-02].

[38] ZEID, Magdy Abd-Elghany; EL-BAHNASY, Khaled a ABO-YOUSSEF, S. E., 2021. Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers. Online. In: *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2021-12-5, s. 224-230. ISBN 978-1-6654-4076-9. Dostupné z: https://doi.org/10.1109/ICICIS52592.2021.9694125. [cit. 2023-11-24].

[39] ZHAN, Yangen; BIAN, Hao; CHEN, Yang; LI, Xiu a ZHANG, Yongbing, 2022. Breast Tumor Image Classification in Bright Challenge VIA Multiple Instance Learning and Deep Transformers. Online. In: *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE, 2022-3-28, s. 1-5. ISBN 978-1-6654-5172-7. Dostupné z: https://doi.org/10.1109/ISBIC56247.2022.9854733. [cit. 2023-11-24].

Seznam symbolů a zkratek

Adam	Adaptivní odhad momentů
Adamax	Variace optimalizačního algoritmu Adam
BEiT	Dvousměrový enkodér z reprezentací obrazových transformerů
CCT	Kompaktní konvoluční transformer
CIFAR-10	Dataset Canadian Institute for Advanced Research s 10 třídami
CIFAR-100	Dataset Canadian Institute for Advanced Research se 100 třídami
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CNN	Konvoluční neuronová síť
CVT	Kompaktní vision transformer
DeiT	Data-efficient Image Transformer
fANOVA	Funkční analýza rozptylu
FLOPs	Počet operací s plovoucí desetinnou čárkou
GPU	Grafická procesorová jednotka
MB	Megabyte
mil.	Milion
MLP	Vícevrstvý perceptron
MNIST	Modified National Institute of Standards and Technology dataset
PE	Poziční kódování
PNG	Portable Network Graphics
ReLU	Rectified Linear Unit
ResNet	Reziduální síť
ResNet18	Reziduální síť s 18 vrstvami
RGB	Model červená-zelená-modrá
RNN	Rekurentní neuronová síť

TPE	Tree-structured Parzen Estimator
TPU	Jednotka pro zpracování tensorů
ViT	Vision Transformer
ViT-B-16	Základní Vision Transformer s 16x16 vstupními výřezy