



Gymnázium
České Budějovice
Jírovcova 8

MATURITNÍ PRÁCE

Neuronové sítě

Jonáš Havelka

vedoucí práce: Dr. rer. nat. Michal Kočer


Prohlášení

Prohlašuji, že jsem tuto práci vypracoval samostatně s vyznačením všech použitých pramenů.


V Českých Budějovicích dne podpis

Jonáš Havelka

Abstrakt





Neuronové sítě se dnes objevují všude, ať už jde o vyhledávání, překládání nebo třeba jen zpracovávání dat. Mnoho programovacích jazyků má své knihovny pro práci s umělou inteligencí, ale právě  Kotlin, který je mým oblíbeným programovacím jazykem a který lze použít skoro kdekoli (webové stránky, servery, mobily), takovou knihovnu postrádá. Proto jsem se rozhodl svoji práci koncipovat jako snahu o implementování takové knihovny.

Klíčová slova

Neuronové sítě, Neuron, Umělá inteligence, Aktivační funkce,  Kotlin, Multiplatformní knihovna, Java, Javascript

Poděkování

Poděkování patří hlavně mému učiteli informatiky, který je zároveň vedoucím mé práce, za skvělou výuku na hodinách a velkou trpělivost při kontrole našich prací. Také nesmím zapomenout na Alžbětu Neubauerovou, která mě celý rok podporovala a několikrát provedla korekturu mé práce.

Dále bych rád poděkoval všem komunitám, jejichž nástroje jsem používal, tj. JetBrains, v jejichž programovacím jazyce  Kotlin programuji a jejichž prostředí IntelliJ k tomu využívám,  Gradle, který používám ke kompilaci, \LaTeX , ve kterém píšu, text a dále  git a  GitHub, jež uchovávají má data, ať už text nebo knihovnu.

Obsah

I	Teoretická část	9
1	Laický náhled na neuronové sítě	10
1.1	Neuron	10
1.2	Aktivační funkce	10
1.3	Sítě	11
1.4	Dopředná propagace a zpětná propagace	12
1.5	Využití neuronových sítí	12
2	Formální náhled	15
2.1	Definice neuronu a sítě	15
2.2	Dopředná propagace	16
2.3	Chybová funkce	16
2.4	Zpětná propagace	17
2.5	Sít	18
2.5.1	Dopředná propagace	19
2.5.2	Zpětná propagace	19
2.5.3	Zakomponování biasu	20
2.6	Aktivační funkce	21
2.7	Shrnutí	25
II	Praktická část	27
3	Struktura knihovny	29
3.1	core	29
3.1.1	IActivationFunctions	29

3.1.2	ActivationFunctions	29
3.1.3	CustomFunction	30
3.1.4	INeuralNetwork	30
3.1.5	BasicNeuralNetwork	31
3.1.6	ConvolutionalNetwork	31
3.2	mnistDatabase	32
3.2.1	Databáze MNIST	32
3.2.2	Databáze EMNIST	33
4	Používání knihovny	34
4.1	Trénování sítě	34
4.2	Používání sítě	35
4.3	Nastavování hodnot	35
	Apendix	35
	Slovníček pojmů	37
	Bibliografie	40
	Seznam obrázků	41
	Přílohy	42
	Zdrojový kód knihovny	USB
	Dokumentace	USB
	Testovací dataset	USB
	Zdrojový kód ukázkového programu	USB
	Ukázkový program	USB
	Zdrojový kód práce v L ^A T _E Xu	USB
	Přehled grafů aktivačních funkcí	42
	Zdrojový kód knihovny	44
	src/commonMain/kotlin/core/ActivationFunctions.kt	44
	src/commonMain/kotlin/core/BasicNeuralNetwork.kt	47
	src/commonMain/kotlin/core/ConvolutionalNeuralNetwork.kt	50

src/commonMain/kotlin/core/CustomFunction.kt	52
src/commonMain/kotlin/core/IActivationFunctions.kt	52
src/commonMain/kotlin/core/INeuralNetwork.kt	52
src/commonMain/kotlin/mnistDatabase/loadFile.kt	53
src/commonTest/kotlin/sample/Constants.kt	55
src/commonTest/kotlin/sample/NeuralNetworkTest.kt	55
src/jsTest/kotlin/sample/ConstantsJS.kt	56
src/jvmMain/kotlin/mnistDatabase/loadFileJVM.kt	56
src/jvmTest/kotlin/sample/ConstantsJVM.kt	56
src/jvmTest/kotlin/sample/NeuralNetworkTestJVM.kt	57

Úvod

Neuronové sítě jsou v poslední době velmi skloňované téma. Nikdo pořádně neví, jak to, že fungují tak dobře. Cílem této práce však nebude zkoumat neuronové sítě, ale implementovat je v co největším rozsahu (ať už struktury bez širšího využití jako asociativní paměť, nebo často používané konvoluční sítě na rozpoznávání obrázků).

Kotlin je ideální programovací jazyk pro vývoj knihovny, protože je interoperabilní s Javou, Javascriptem i C, a tak umožňuje tuto knihovnu používat jak pro JVM, tak i v prohlížeči nebo v programech kompilovaných přímo do binárního kódu.

V textu jsou použita pojmy ze stavby biologického neuronu, objektově orientovaného programování, Kotlinu, ...Tyto pojmy jsou vysvětleny na konci práce.

Celá maturitní práce je k dispozici na GitHubu, text včetně zdrojového LaTeXu na adrese https://github.com/JoHavel/Maturitni-Seminarni-Prace/tree/my_work a knihovna samotná pak na <https://github.com/JoHavel/NeuralNetwork>.

Část I

Teoretická část

1 Laický náhled na neuronové sítě

1.1 Neuron

Počítačové neuronové sítě nejsou jen výmysl lidí, jejich základ nalezneme v nervových soustavách živočichů. Základní stavební jednotka takové soustavy (stejně tak i neuronové sítě) je *neuron*. Neuron funguje tak, že přes *dendrity* přijímá elektrické (přesněji iontové) signály od jiných neuronů a když součet signálů přeteče určitou danou *mez*, vyšle neuron signál přes *axony* dál do dalších neuronů.

Přenos signálu z axonu do dendritu se odehrává v malých prostorách mezi nimi zvaných *synapse*. Vodivost synapsí je ovlivněna jejich chemickým složením, a proto se domníváme, že proces učení probíhá měněním těchto chemických spojů [1, s. 491].

Náš umělý neuron tedy bude mít *seznam dendritů* (nesoucích informaci z jakého neuronu vedou signál a jak ho mění synapse), tzv. *aktivační funkci* (viz dále) a výstupní signál. Často navíc bude obsahovat základní hodnotu (angl. *bias*), která reprezentuje mez, při jejímž překročení začne neuron vysílat signál. Jinak řečeno posouvá aktivační funkci ve směru osy x .

Neuron (hlavně ten umělý) ilustruje obrázek 2.1 nacházející se v další kapitole. Podrobněji o souvislosti biologických a umělých neuronových sítích pojednává [2].

1.2 Aktivační funkce

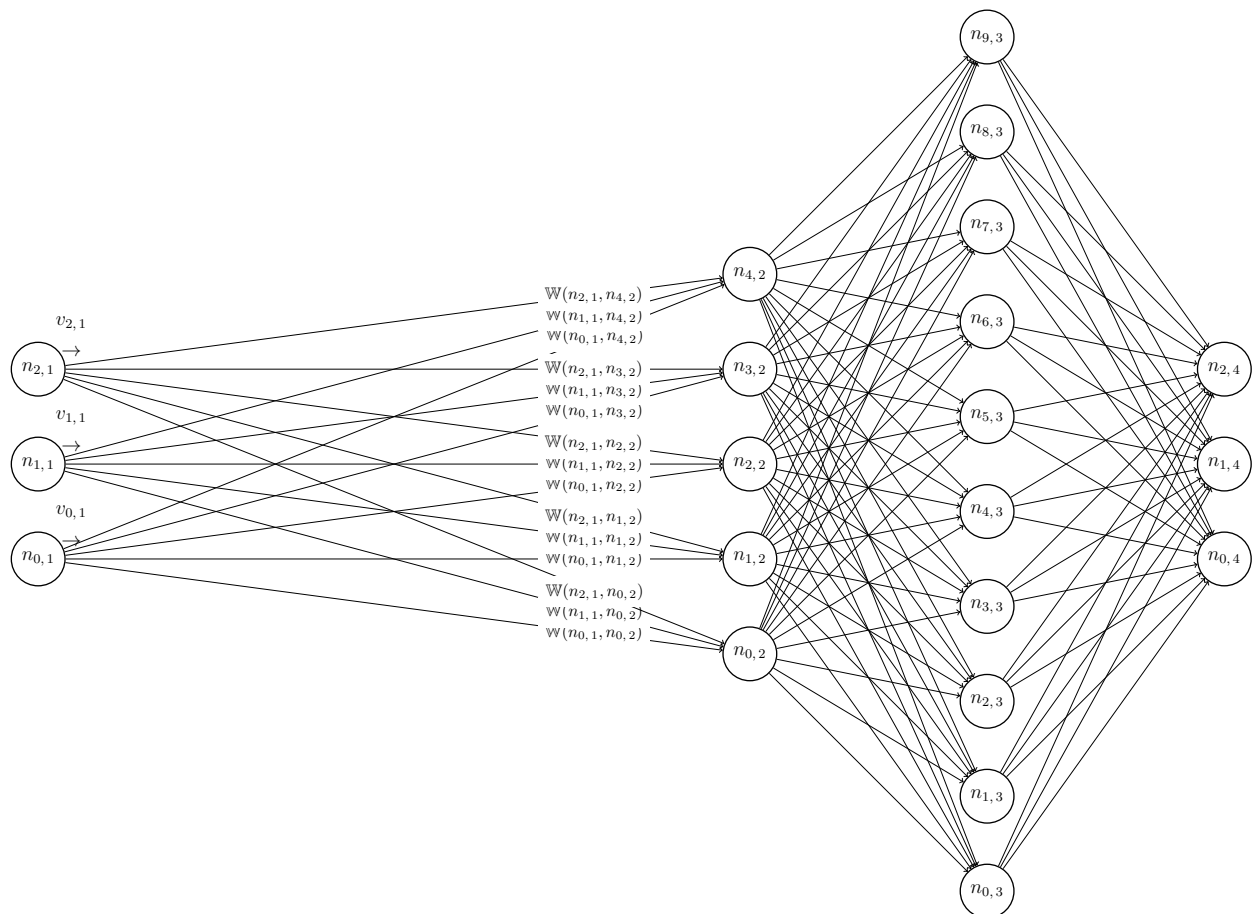
Jak už bylo zmíněno, přírodní neuron funguje na principu toho, že když součet vstupních signálů nepřekračuje určitou mez, nevysílá neuron žádný (nebo téměř žádný) signál. Když je však tato mez překonána, neuron vyšle signál. V podstatě tedy vysílá buď 0 nebo 1. Pro účely umělého neuronu je 0 a 1 nedostačující, jelikož při procesu učení potřebujeme měnit hodnoty jemně, abychom nerozbili již naučené znalosti.

Proto se jako aktivační funkce (tedy to, co určuje jaký má být výstup v závislosti na

součtu vstupů, v případě přírody tedy funkce zobrazující interval $-\infty$ až mez (bias) na 0 a zbylá čísla na 1 viz *binární krok* v sekci 2.6) používají funkce co nejvíce podobné právě tomuto binárnímu kroku, které jsou ale spojité a mají co „nejhezčí“ derivace (protože při zpětné propagaci právě podle derivace určíme, jak moc daný neuron ovlivňuje výsledek).

1.3 Sítě

Jelikož „nahodilé neurony“ by se těžko udržovaly v paměti a operace na nich by byly velmi pomalé, potřebujeme síť nějak uspořádat. Nejjednodušším uspořádáním jsou *vrstvy*. Každý neuron z jedné vrstvy má dendrity ze všech neuronů z vrstvy minulé. Tak se předejde cyklům, které jsou složité na výpočty, a navíc si nemusíme u každého neuronu pamatovat, ze kterých neuronů do něj vede signál.



Obr. 1.1: Běžná neuronová síť (W jsou váhy, n neurony a v je výstupní signál, viz kapitola 2 a sekce 2.5)

Velmi využívanými strukturami jsou také konvoluční neuronové sítě, kde nejdříve aplikujeme filtry¹ na části vstupních dat a teprve výstupy z těchto filtrů jsou vstupem do neuronové

¹Často malé neuronové sítě, které sami vytvoříme. Síť používané jako filtry se nemusí učit². Další možný filtr je třeba

sítě. O konvolučních sítích se můžete dočíst v [4] nebo v [5].

I tak se „nahodilé neurony“ občas používají, jelikož při malém množství neuronů a hlavně při malém množství synapsí je přepočítání samotných neuronů efektivnější než počítání celých vrstev. Ukázkou takové malé sítě je asociativní paměť, kde neuronům přiřadíme objekty, které si tato síť má „pamatovat“. Když chceme zjistit, co je v paměti asociováno s daným objektem, vybudíme (v umělé síti to znamená nastavíme výstupní signál na 1) neuron odpovídající tomuto objektu a následně sledujeme, které další neurony jsou vybudeny. Takto funguje i lidská paměť, pamatujeme si právě asociace. Umělou asociativní paměť zmiňuje [1].

1.4 Dopředná propagace a zpětná propagace

Dopředná propagace (častěji se používá anglický výraz *forward propagation*) je jednoduše spočítání signálů ve všech neuronech. Tedy u každého neuronu se sečtou vstupní signály (popř. přičte *bias*) a spočítá se funkční hodnota aktivační funkce v tomto bodě.

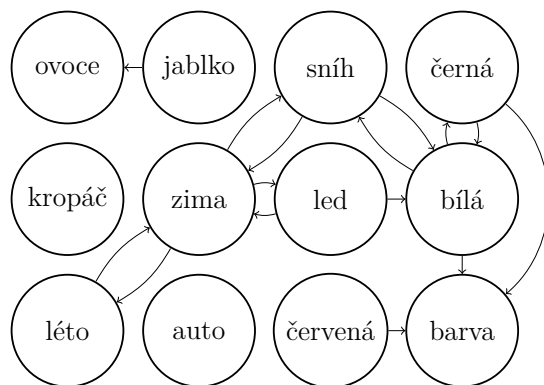
Naopak zpětná propagace (častěji se používá anglický výraz *backward propagation* či *backpropagation*) je na základě chyby, kterou spočítáme z výstupu neuronové sítě a předpokládaného výstupu, určení, které proměnné hodnoty (*synapse* a *biasy*) se na ní nejvíce podílejí. Potom tyto hodnoty posuneme odpovídajícím způsobem (stejně jako příroda mění chemické vlastnosti *synapse*). Z matematického pohledu se hodnoty posunou proti směru gradientu chyby, jelikož právě gradient udává, kterým směrem máme souřadnice (tj. váhy a *biasy*) posunout, aby funkce (tj. chyba) vzrostla.

1.5 Využití neuronových sítí

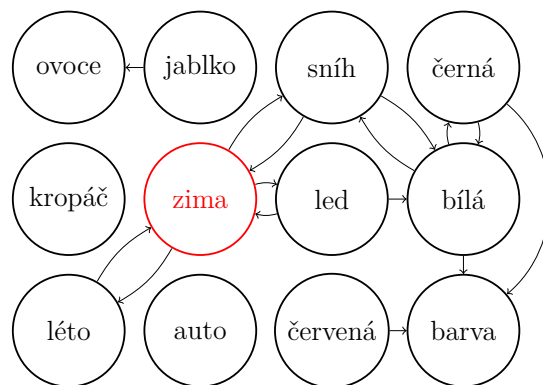
Než se pustíme do matematiky, která stojí za fungováním neuronových sítí, ještě si řekneme, kde a jaké neuronové sítě využíváme. Jedno z nejviditelnějších využití je rozpoznávání obrázků, protože takovou úlohu jen stěží zvládnou běžné algoritmy. Mezi rozpoznávání obrázku patří jak strojové čtení textů, tak třeba rozpoznávání tváře nebo klasifikace, zda je na obrázku morče, nebo slon. K tomu se používají hlavně konvoluční sítě, jelikož filtr rozezná hrany a různé útvary a neuronová síť podle toho určí dané rozřazení (znak, člověka, zvíře...).

Další oblastí je překlad. Překládat slova zvládneme jednoduše podle slovníků, ale aby věta dávala smysl a slovo bylo přeloženo v kontextu věty, potřebujeme něco více. Pro to se

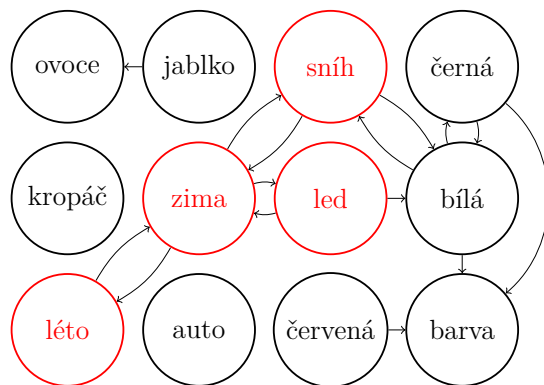
Fourierova transformace viz [3], kterou se však dále zabývat nebudeme (tato možnost není ani implementována v knihovně).



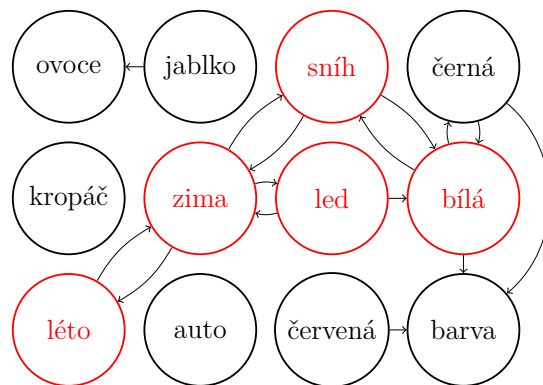
(a) Vybuzení 1. neuronu (zimy)



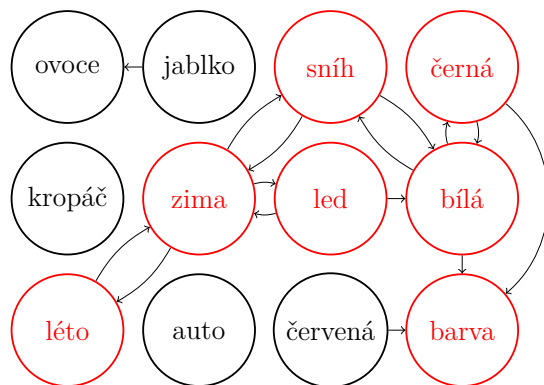
(b) 1 krok po vybuzení 1. neuronu



(c) 2 kroky po vybuzení 1. neuronu



(d) 3 kroky po vybuzení 1. neuronu



(e) 4 kroky po vybuzení 1. neuronu

Obr. 1.2: Asociativní paměť, červeně jsou vybuzené neurony



Obr. 1.3: Generovaná tvář [7]

používá vektorový prostor slov, tedy všem slovům přiřadíme určitý vektor (to musíme udělat vždy, protože neuronová síť nemá jiný vstup) a poté na vzorovém textu učíme neuronovou síť odhadovat slovo podle několika okolních slov. Při tom ale neupravujeme jen hodnoty neuronové sítě, ale i vektorů slov. Tím dostaneme vektorový prostor slov, na kterém se překládající neuronová síť (jiná než ta, co vyrobila vektorový prostor) naučí překládat velmi lidsky. Stejný vektorový prostor se dá použít i na neuronovou síť generující text.

Když už bylo zmíněno generování, umělé neuronové sítě jsou schopny i generovat obrázky, hudbu, atd.³ K tomu se používá systém GAN (tj. Generative adversarial network) [6], což jsou dvě sítě, jedna generuje a druhá dostane dvojici objekt vytvořený člověkem (resp. skutečností v případě fotek) a objekt vygenerovaný první sítí a má za úkol určit, který je který. Tyto sítě se učí spolu a výsledkem jsou relativně pěkná díla viz obrázek 1.3.

³Stále je to však na základě nějakého datasetu obrázků nebo hudby.

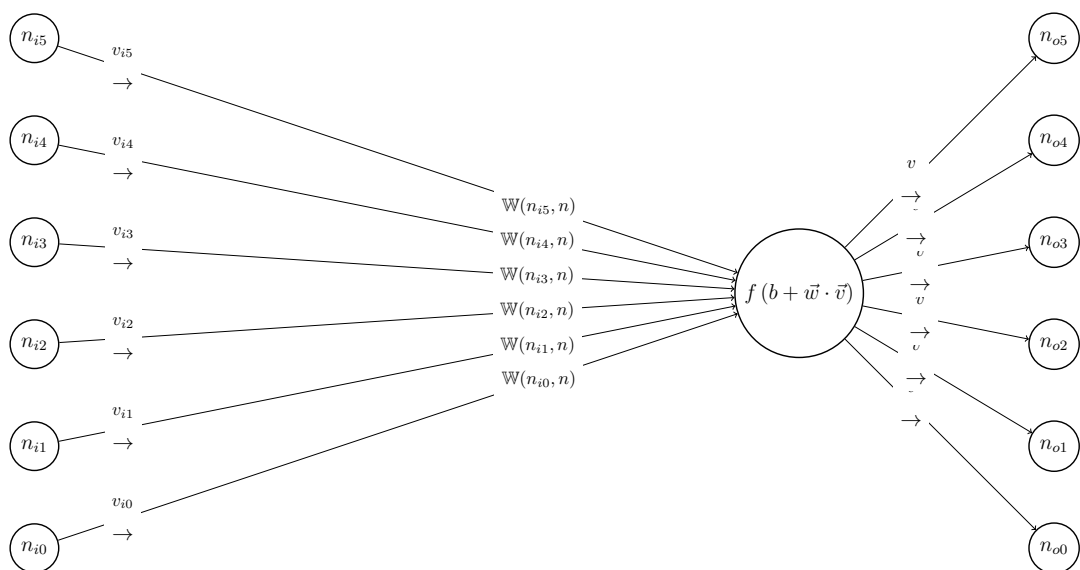
2 Formální náhled

Matematikou za neuronovými sítěmi a její implementací v Pythonu se zabývají videa [8]. Kniha zabývající se touto problematikou je např. [9].

V dalším textu $\vec{x} \cdot \vec{y}$ značí skalární součin¹ vektorů \vec{x} a \vec{y} . Vektory jsou uvedeny horizontálně, ale chápeme to jako by byly vertikálně².

2.1 Definice neuronu a sítě

Vstupní neurony: N_{in} Axony vstupních neuronů Synapse z n_{i*} do n Dendrity Vyšetřovaný neuron: n Výstupní neurony: N_{out}



Obr. 2.1: Neuron

Označme $\nu = (N, W, F)$ neuronovou síť, N je množina všech jejích neuronů, $W : N \times N \rightarrow \mathbb{R}$ jsou váhy (angl. weights) udávající sílu synapse mezi dvěma neurony (v případě, že mezi neurony synapse není, je W rovno 0) a $F : \mathbb{R}^{|N_v|} \rightarrow \mathbb{R}$ je chybová funkce udávající velikost chyby podle rozdílu reálných hodnot od chtěných hodnot výstupních neuronů (N_v).

¹To jest to samé jako $\vec{x}^T \vec{y}$.

²Mohli bychom doplnit za každou definici vektoru T , třeba 2.2 přepíšeme jako $\vec{v} = (v_1, v_2, \dots)^T$

Nechť $n \in N$, $n = (N_{in}, N_{out}, f, b, v, \varepsilon)$ je neuron, kde $N_{in} = \{n_x \in N | W(n_x, n) \neq 0\}$ je množina neuronů, které vysílají signál do n , $N_{out} = \{n_x \in N | W(n, n_x) \neq 0\}$ je množina neuronů, které přijímají signál od n , $f: \mathbb{R} \rightarrow \mathbb{R}$ je aktivační funkce, $b \in \mathbb{R}$ je bias, $v \in \mathbb{R}$ je signál vycházející z n a ε je chyba (parciální derivace chybové funkce podle $f^{-1}(v)$ ³).

2.2 Dopředná propagace

Potom dopředná propagace (tedy spočítání v) vypadá takto⁴:

$$v = f \left(b + \sum_{n_x \in N_{in}, v_x \in n_x} v_x \cdot W(n_x, n) \right) \quad (2.1)$$

To lze při označení

$$\vec{v} = (v_1, v_2, \dots) \quad (2.2)$$

$$\vec{w} = (w_1, w_2, \dots) \quad (2.3)$$

$$(\forall n_x \in N_{in}) (\exists i \in \mathbb{N}) (v_i \in n_x \wedge w_i = W(n_x, n)) \quad (2.4)$$

zapsat vektorově jako:

$$v = f(b + \vec{w} \cdot \vec{v}) \quad (2.5)$$

Případně můžeme do vektorů „zakomponovat“ i bias⁵:

$$\vec{v} = (1, v_1, v_2, \dots) \quad (2.6)$$

$$\vec{w} = (b, w_1, w_2, \dots) \quad (2.7)$$

$$(\forall n_x \in N_{in}) (\exists i \in \mathbb{N}) (v_i \in n_x \wedge w_i = W(n_x, n)) \quad (2.8)$$

$$v = f(\vec{w} \cdot \vec{v}) \quad (2.9)$$

2.3 Chybová funkce

Anglicky loss function nebo někdy také cost function. Udává, nakolik se neuronová síť strefila do správného výstupu. Většinou nás ale nezajímá její hodnota (rozlišujeme pouze, zda síť odpověděla dobře, nebo ne), používáme ji jen jako pomyslné hodnocení ve zpětné propagaci.

³Derivace aktivačních funkcí se často snadno spočítá z funkční hodnoty, proto uvádím, že hledám derivaci v bodě, kde je daná funkční hodnota, značím přitom $f^{-1}(y) = x \Leftrightarrow f(x) = y$.

⁴ $v_x \in n_x$ značí, že v_x je signál neuronu n_x , obdobně u ostatních informací v neuronu.

⁵To v knihovně není použito z důvodu netriviálního přidávání prvku do vektoru.

Její gradient, tedy derivace podle všech proměnných (vah a biasů) v neuronové síti, totiž udává, jak poupravit hodnoty, aby neuronová síť odpovídala lépe.

Pro naše potřeby stačí pouze jediná chybová funkce

$$E(x) = 0,5 \sum_{n_o \in O} (v_{od} - v_o) \quad (2.10)$$

, kde O je množina výstupních neuronů, v_o jsou jejich výstupní signály a v_{od} jsou odpovídající chtěné výstupní signály. Tato funkce má výhodu, že její derivace podle libovolného v_o je

$$\frac{\delta E}{\delta v_o} = v_{od} - v_o \quad (2.11)$$

, tedy ε výstupních neuronů spočítáme pouze jako rozdíl chtěných a reálných výstupů.

2.4 Zpětná propagace

Při zpětné propagaci je důležitý vzorec pro derivaci složené funkce, někdy také znám jako „řetízkové pravidlo“ (pro funkci jedné proměnné platí (2.12), pro více pak (2.13))⁶

$$\frac{dy}{dx} = \frac{dz}{dx} \frac{dy}{dz} \quad (2.12)$$

$$\frac{\delta y}{\delta x} = \sum_z \frac{\delta z}{\delta x} \frac{\delta y}{\delta z} \quad (2.13)$$

Díky tomu můžeme ε neuronu spočítat pomocí

$$f_x^{-1}(v_x) = \sum_{n_y \in N_{out, x}, v_y \in n_y} v_y \cdot W(n_y, n_x) \quad (2.14)$$

tj.

$$\frac{\delta f_x^{-1}(v_x)}{\delta v_y} = W(n_y, n_x) \quad (2.15)$$

takto:

$$\varepsilon = \frac{\delta E}{\delta f^{-1}(v)} = \sum_{n_x \in N_{out}, f_x \in n_x, v_x \in n_x} \frac{\delta E}{\delta f_x^{-1}(v_x)} \cdot \frac{\delta f_x^{-1}(v_x)}{\delta f^{-1}(v)} \quad (2.16)$$

$$\varepsilon = \sum_{n_x \in N_{out}, f_x \in n_x, v_x \in n_x} \frac{\delta E}{\delta f_x^{-1}(v_x)} \cdot \frac{\delta f_x^{-1}(v_x)}{\delta v} \cdot \frac{\delta v}{\delta f^{-1}(v)} \quad (2.17)$$

$$\varepsilon = \frac{\delta v}{\delta f^{-1}(v)} \sum_{n_x \in N_{out}, \varepsilon_x \in n_x} \varepsilon_x \cdot W(n, n_x) \quad (2.18)$$

$$\varepsilon = f'(f^{-1}(v)) \sum_{n_x \in N_{out}, \varepsilon_x \in n_x} \varepsilon_x \cdot W(n, n_x) \quad (2.19)$$

⁶Pro funkce musí platit, že mají v daných bodech derivaci, viz [10, s. 623].

ε nás dovede k tomu, o kolik musíme posunout bias. Hlavním parametrem neuronové sítě jsou ale váhy (funkce W). Derivaci chybové funkce podle váhy určíme za pomoci 2.14, tj.

$$\frac{\delta f_y^{-1}(v_y)}{\delta W(n, n_y)} = v \quad (2.20)$$

a z rovnice 2.2:

$$\frac{\delta E}{\delta W(n, n_y)} = \frac{\delta E}{\delta f_y^{-1}(v_y)} \cdot \frac{\delta f_y^{-1}(v_y)}{\delta W(n, n_y)} = e_y \cdot v \quad (2.21)$$

Obdobně jako v předchozím případě definujeme vektory⁷:

$$\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots) \quad (2.22)$$

$$\vec{w} = (w_1, w_2, \dots) \quad (2.23)$$

$$\frac{\delta E}{\delta \vec{w}} = \left(\frac{\delta E}{\delta w_1}, \frac{\delta E}{\delta w_2}, \dots \right) \quad (2.24)$$

$$(\forall n_x \in N_{out}) (\exists i \in \mathbb{N}) (\varepsilon_i \in n_x \wedge w_i = W(n, n_x)) \quad (2.25)$$

$$\varepsilon = f'(f^{-1}(v)) \cdot (\vec{w} \cdot \vec{\varepsilon}) \quad (2.26)$$

$$\frac{\delta E}{\delta \vec{w}} = \vec{\varepsilon} \cdot v \quad (2.27)$$

Vektor $\frac{\delta E}{\delta \vec{w}}$ už stačí jen přičíst k \vec{w} , abychom upravili hodnoty $W(n, n_x)$.

Pomocí tohoto můžeme spočítat všechno kromě ε na výstupních neuronech. To můžeme z rovnice 2.11 ($n_o \in O$ jsou výstupní neurony, $\varepsilon_o \in n_o$, $f_o \in n_o$ a $v_o \in n_o$):

$$\varepsilon_o = \frac{\delta E}{\delta f_o^{-1}(v_o)} = \frac{\delta E}{\delta v_o} \cdot \frac{\delta v_o}{\delta f_o^{-1}(v_o)} = (v_{od} - v_o) f'_o(f_o^{-1}(v_o)) \quad (2.28)$$

2.5 Síť

V 1.3 jsme se bavili o tom, že nejpoužívanější sítě mají neurony seřazené do vrstev. Nechtě jsou tudíž neurony uspořádány ve vrstvách číslovaných přirozenými čísly od 1 a nechtě jsou navíc i neurony v každé vrstvě zvlášť očíslovány přirozenými čísly od 1 (tj. vrstva je vlastně vektor neuronů). Potom značme L_x vrstvu s indexem x a $n_{x,y}$ neuron s indexem y příslušící do L_x . To znamená, že pokud $N_{in} \in n_{x,y}$, tak $N_{in} = L_{x-1}$, a pokud $N_{out} \in n_{x,y}$, tak $N_{out} = L_{x+1}$. Následně zavedme vektory ($v_{x,i} \in n_{x,i}$, $b_{x,i} \in n_{x,i}$, $f_{x,i} \in n_{x,i}$ a $\varepsilon_{x,i} \in n_{x,i}$):

$$\vec{v}_x = (v_{x,1}, v_{x,2}, \dots) \quad (2.29)$$

⁷Značení $\frac{\delta E}{\delta \vec{w}}$ a $\frac{\delta E}{\delta W}$ z rovnic 2.24 a 2.43 neznačí derivace podle vektoru a matice, ale je to symbolické značení pro vektor a matici derivací podle jednotlivých složek daného tensoru.

$$\vec{w}_{x,i} = (W(n_{x,1}, n_{x+1,i}), W(n_{x,2}, n_{x+1,i}), \dots) \quad (2.30)$$

$$\vec{w}'_{x,i} = (W(n_{x,i}, n_{x+1,1}), W(n_{x,i}, n_{x+1,2}), \dots) \quad (2.31)$$

$$\vec{b}_x = (b_{x,1}, b_{x,2}, \dots) \quad (2.32)$$

$$\vec{f}_x = (f_{x,1}, f_{x,2}, \dots) \quad (2.33)$$

$$\vec{\varepsilon}_x = (\varepsilon_{x,1}, \varepsilon_{x,2}, \dots) \quad (2.34)$$

$$\frac{\delta E}{\delta \vec{w}'_{x,i}} = \left(\frac{\delta E}{\delta W(n_{x,y}, n_{x+1,1})}, \frac{\delta E}{\delta W(n_{x,y}, n_{x+1,2})}, \dots \right) \quad (2.35)$$

2.5.1 Dopředná propagace

Přepíšeme rovnici dopředné propagace 2.5:

$$v_{x,i} = f_{x,i}(b_{x,i} + \vec{w}_{x-1,i} \cdot \vec{v}_{x-1}) \quad (2.36)$$

Můžeme využít matici vah a maticové násobení (aplikaci vektoru funkcí $\vec{f}(\vec{x})$ chápejme tak, že na každou složku \vec{x} se aplikuje odpovídající složka \vec{f}):

$$\mathbb{W}_x = \begin{pmatrix} w_{x,1} \\ w_{x,2} \\ \vdots \end{pmatrix} = \begin{pmatrix} W(n_{x,1}, n_{x+1,1}) & W(n_{x,2}, n_{x+1,1}) & \dots \\ W(n_{x,1}, n_{x+1,2}) & W(n_{x,2}, n_{x+1,2}) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (2.37)$$

$$\vec{v}_x = \vec{f}_x(\vec{b}_x + \mathbb{W}_{x-1} \cdot \vec{v}_{x-1}) \quad (2.38)$$

2.5.2 Zpětná propagace

Nyní přepíšeme rovnice 2.26 a 2.27 zpětné propagace:

$$\varepsilon_{x,i} = f'_{x,i}(f_{x,i}^{-1}(v_{x,i})) \odot (\vec{w}'_{x,i} \cdot \vec{\varepsilon}_{x+1}) \quad (2.39)$$

$$\frac{\delta E}{\delta \vec{w}'_{x,i}} = \vec{\varepsilon}_{x+1} \cdot v_{x,i} \quad (2.40)$$

Rovnici 2.39 můžeme převést hned do maticového tvaru (\mathbb{W}^T značí transponovanou matici \mathbb{W} , $\vec{f}^{-1}(x)$ a $\vec{f}'(x)$ značí aplikaci inverzní funkce a derivace funkce podobně jako v 2.38, \odot značí násobení po složkách⁸):

$$\vec{\varepsilon}_x = \vec{f}'_x(\vec{f}_x^{-1}(\vec{v}_x)) \odot (\mathbb{W}_x^T \cdot \vec{\varepsilon}_{x+1}) \quad (2.41)$$

⁸Násobením vektorů $\vec{x} = (x_1, x_2, \dots)$ a $\vec{y} = (y_1, y_2, \dots)$ tzv. po složkách získáme vektor $\vec{x} \odot \vec{y} = (x_1 \cdot y_1, x_2 \cdot y_2, \dots)$.

Pro rovnici 2.40 potřebujeme spojit definice 2.24 (definice vektoru derivací), kterou přepíšeme do tvaru vrstev:

$$\frac{\delta E}{\delta \vec{w}_{x,i}} = \left(\frac{\delta E}{\delta W(n_{x,i}, n_{x+1,1})}, \frac{\delta E}{\delta W(n_{x,i}, n_{x+1,2})}, \dots \right) \quad (2.42)$$

a 2.37 (definici matice vah):

$$\frac{\delta E}{\delta \mathbb{W}_x} = \begin{pmatrix} \frac{\delta E}{\delta w_{x,1}} \\ \frac{\delta E}{\delta w_{x,2}} \\ \vdots \end{pmatrix} = \begin{pmatrix} \frac{\delta E}{\delta W(n_{x,1}, n_{x+1,1})} & \frac{\delta E}{\delta W(n_{x,1}, n_{x+1,2})} & \dots \\ \frac{\delta E}{\delta W(n_{x,2}, n_{x+1,1})} & \frac{\delta E}{\delta W(n_{x,2}, n_{x+1,2})} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (2.43)$$

Nyní jsme již schopni zapsat 2.40 maticově:

$$\frac{\delta E}{\delta \mathbb{W}_x} = \vec{\varepsilon}_{x+1} \vec{v}_x^T \quad (2.44)$$

I spočítání ε u poslední vrstvy (tj. neuronů v O , značme ji L_o) lze zapsat vektorově (\vec{v}_{od} zde značí vektor předpokládaných výsledků):

$$\vec{\varepsilon}_o = \vec{f}_x \left(\vec{f}_x^{-1}(\vec{v}_x) \right) \odot (\vec{v}_{od} - \vec{v}_o) \quad (2.45)$$

2.5.3 Zakomponování biasu

Nejdříve musíme upravit vektory a matice:

$$\vec{v}_x = (1, v_{x,1}, v_{x,2}, \dots) \quad (2.46)$$

$$\vec{f}_x = (1, f_{x,1}, f_{x,2}, \dots) \quad (2.47)$$

$$\vec{\varepsilon}_x = (0, \varepsilon_{x,1}, \varepsilon_{x,2}, \dots) \quad (2.48)$$

$$\mathbb{W}_x = \begin{pmatrix} 1 & 0 & 0 & \dots \\ b_{x+1,1} & W(n_{x,1}, n_{x+1,1}) & W(n_{x,2}, n_{x+1,1}) & \dots \\ b_{x+1,2} & W(n_{x,1}, n_{x+1,2}) & W(n_{x,2}, n_{x+1,2}) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (2.49)$$

$$\frac{\delta E}{\delta \mathbb{W}_x} = \begin{pmatrix} 0 & 0 & 0 & \dots \\ \frac{\delta E}{\delta b_{x+1,1}} & \frac{\delta E}{\delta W(n_{x,1}, n_{x+1,1})} & \frac{\delta E}{\delta W(n_{x,1}, n_{x+1,2})} & \dots \\ \frac{\delta E}{\delta b_{x+1,2}} & \frac{\delta E}{\delta W(n_{x,2}, n_{x+1,1})} & \frac{\delta E}{\delta W(n_{x,2}, n_{x+1,2})} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (2.50)$$

Rovnice 2.38 (samozřejmě bez biasu:

$$\vec{v}_x = \vec{f}_x(W_{x-1} \cdot \vec{v}_{x-1}) \quad (2.51)$$

), 2.41 a 2.44 poté fungují pořád stejně. Rovnice 2.45 funguje také shodně, jelikož prostě řekneme, že první člen odhadu vyšel tak, jak má, tj. $\vec{\varepsilon} = (0, \dots)$

2.6 Aktivační funkce

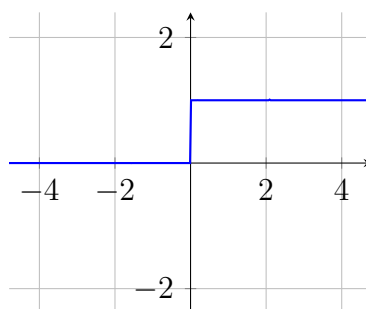
Jelikož neurony mají bias, není nutné udávat aktivační funkce obecně, stačí je jen udat tak, že $x = 0$ odpovídá mezi v pomyslném biologickém neuronu. Mezi aktivační funkce⁹ patří:

- *Binary step*

$$f(x) = \begin{cases} 0, & \text{když } x < 0 \\ 1, & \text{když } x \geq 0 \end{cases} \quad (2.52)$$

$$f'(x) = \begin{cases} 0, & \text{když } x \neq 0 \\ +\infty, & \text{když } x = 0 \end{cases} \quad (2.53)$$

(česky *binární krok*), již zmíněná funkce, jež odpovídá reálnému neuronu, ale není použitelná pro učení na základě gradientu, jelikož má derivaci 0 všude kromě bodu $x = 0$, kde je nespojitá.



Obr. 2.2: Binární krok

- *Identity*

$$f(x) = x \quad (2.54)$$

$$f'(x) = 1 \quad (2.55)$$

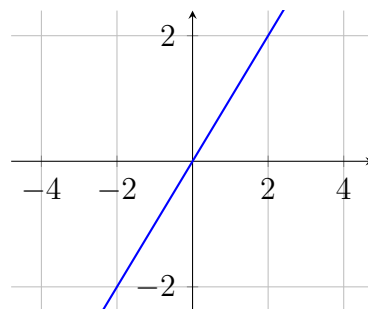
(česky *identita*) odpovídá stavu, jako kdyby tam žádná funkce nebyla. Její derivace je 1, tedy se velmi snadno určí v libovolném bodě.

- *Sigmoid* [12] (značí se σ)

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.56)$$

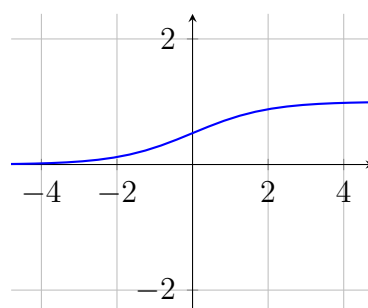
$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} \left(1 - \frac{1}{1 + e^{-x}}\right) = \sigma(x) \cdot (1 - \sigma(x)) \quad (2.57)$$

⁹Funkce jsem čerpal převážně z [11].



Obr. 2.3: Identita

je jedna z nejznámějších aktivačních funkcí. Je to vlastně takový hladký přechod mezi 0 a 1. Také je na σ dobře vidět, proč se často počítá derivace z funkční hodnoty, místo počítání exponenciální funkce a dělení si vystačíme s násobením a odčítáním.



Obr. 2.4: σ

- Nesmíme zapomenout na *sigmoidě* podobnou a také často používanou funkci *hyperbolický tangens* (\tanh) [12] [13]:

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{2}{1 + e^{-2x}} - 1 = 2 \cdot \sigma(2x) - 1 \quad (2.58)$$

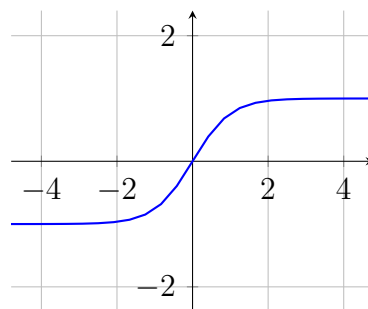
$$\tanh'(x) = \frac{1}{\cosh^2(x)} = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)} = 1 - \tanh^2(x) \quad (2.59)$$

$$\tanh'(x) = 4 \cdot \sigma'(2x) = 4 \cdot \sigma(2x) \cdot (1 - \sigma(2x)) \quad (2.60)$$

Největší rozdíl oproti σ je, že může nabývat i záporných hodnot, což sice moc neodpovídá přírodnímu neuronu, ale když si rozmyslíme, že stačí zvětšit biasy u neuronů, do kterých neuron s aktivační funkcí \tanh vysílá signál, dospějeme k výsledku, že tato funkce také funguje.

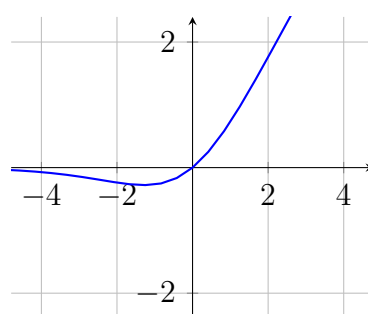
- Další funkce s vazbou na *sigmoidu* je funkce *swish* [12]:

$$f(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \quad (2.61)$$



Obr. 2.5: Hyperbolický tangens

Nepodařilo se mi ale najít derivaci za pomoci funkční hodnoty. *Sigmoida* se také používá ve spojení s ostatními funkcemi, většinou $\sigma(x)$ pro kladné a druhá funkce pro záporné.



Obr. 2.6: Swish

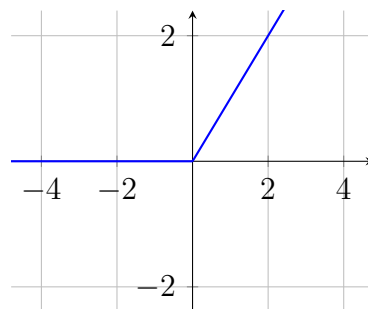
- Ukazuje se, že identita jako taková se v podstatě použít nedá, ale hojně využívaná je její „upravená“ verze *rectified linear unit* [12] (česky něco jako *napravená přímá úměrnost*), která záporná čísla převádí na nulu a v kladných se chová jako *identita*:

$$f(x) = \begin{cases} 0, & \text{když } x < 0 \\ x, & \text{když } x \geq 0 \end{cases} \quad (2.62)$$

$$f'(x) = \begin{cases} 0, & \text{když } x < 0 \\ 1, & \text{když } x > 0 \\ \text{neexistuje,} & \text{když } x = 0 \end{cases} \quad (2.63)$$

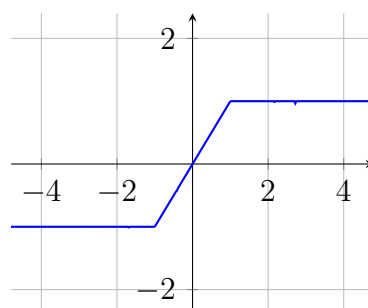
Trochu připomíná biologický neuron, protože pro záporné hodnoty nevysílá, ale na rozdíl od něj má variabilní hodnotu vysílaného signálu. Často se například používá ve filtrech, jelikož chceme detekovat, zda je někde hrana, ale nechceme vysílat záporný signál, když někde hrana není, protože může být o pixel vedle.

- Kromě této verze je v knihovně ještě *leaky* (děravá či prosakující) *rectified linear unit*



Obr. 2.7: Rectified linear unit

[12], která v záporných hodnotách nedává nulu, ale *přímou úměrnost*. K těmto funkcím můžeme přiřadit i *hard hyperbolic function*, která je *identitou* pouze na intervalu $(-1, 1)$, tedy odpovídá biologickému neuronu asi nejvíce z těchto „lineárních funkcí“.



Obr. 2.8: Hard hyperbolic function

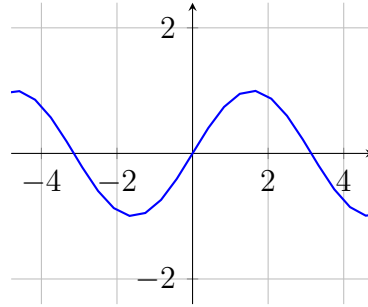
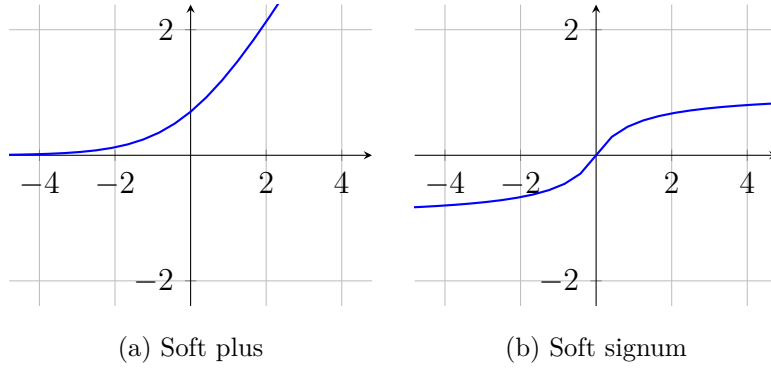
- *Rectified unit* není hladká (nemá derivaci v bodě nula), ale to lze napravit, když použijeme funkci *soft plus* [12] ($\ln(1 + e^x)$). Podobnou úpravu lze udělat i s funkcí *signum* (*znaménko*, často se značí *sign*), což je téměř *binární krok*¹⁰, akorát v záporných hodnotách nabývá funkční hodnoty -1 místo 0. *Signum* se dá zapsat jako podíl x a $|x|$, tudíž tato úprava (*soft sign* [12]) vypadá následovně:

$$f(x) = \frac{x}{|x| + 1} \quad (2.64)$$

- Jednou skupinou funkcí, se kterými se sice experimentuje, ale stěží najdete nějaké využití, jsou ty, které nejsou monotonní¹¹, jako sinus, kosinus, *Gaussova funkce* (e^{-x^2}), apod. Vzhledem k jejich mizivému využití je implementován pouze *sinus*.

¹⁰Z důvodu téhle podobnosti není ani implementována.

¹¹Můžeme si všimnout, že téměř všechny předchozí funkce jsou neklesající, většina dokonce rostoucí na celém definičním oboru.



Obr. 2.10: Sinus

2.7 Shrnutí

Rovnice:

- Dopředné propagace, tj. 2.5 resp. 2.9 nebo 2.38 resp. 2.51:

$$v = f(b + \vec{w} \cdot \vec{v})$$

$$v = f(\vec{w} \cdot \vec{v})$$

$$\vec{v}_x = \vec{f}_x(\vec{b}_x + W_{x-1} \cdot \vec{v}_{x-1})$$

$$\vec{v}_x = \vec{f}_x(W_{x-1} \cdot \vec{v}_{x-1})$$

- Zpětné propagace, tj. 2.19 a 2.27 nebo 2.41 a 2.44:

$$\varepsilon = f'(f^{-1}(v)) \sum_{n_x \in N_{out}, \varepsilon_x \in n_x} \varepsilon_x \cdot W(n, n_x)$$

$$\frac{\delta E}{\delta \vec{w}} = \vec{\varepsilon} \cdot v$$

$$\vec{\varepsilon}_x = \vec{f}_x'(\vec{f}_x^{-1}(\vec{v}_x)) \odot (\mathbb{W}_x^T \cdot \vec{\varepsilon}_{x+1})$$

$$\frac{\delta E}{\delta \mathbb{W}_x} = \vec{\varepsilon}_{x+1} \vec{v}_x^T$$

- Prvotní části zpětné propagace, tj. 2.28 nebo 2.45

$$\varepsilon_o = (v_{od} - v_o) f'_o \left(f_o^{-1}(v_o) \right)$$

$$\vec{\varepsilon}_o = \vec{f}_x \left(\vec{f}_x^{-1}(\vec{v}_x) \right) \odot (\vec{v}_{od} - \vec{v}_o)$$

nám popisují matematiku stojící za fungováním neuronových sítí, tedy naším cílem bude je implementovat. Navíc k implementování těchto rovnic potřebujeme naprogramovat samotný neuron, který jsme si definovali v 2.1 jako:

$$n = (N_{in}, N_{out}, f, b, v, \varepsilon)$$

Také často používáme aktivační funkce, proto by v naší knihovně neměly chybět.

Část II

Praktická část

Cílem této práce je knihovna, která nám umožní používat neuronové sítě v Kotlinu. Jak bylo řečeno na konci minulé kapitoly, musí obsahovat aktivační funkce, nejlépe všechny uvedené v 2.6, zároveň však umožnit uživateli definovat si funkce vlastní. Poté se zaměříme hlavně na neuronovou síť obsahující vrstvy, její implementace bude zároveň zahrnovat jak implementaci neuronu, tak implementaci dopředné a zpětné propagace.

Konvoluční síť pro jednoduchost naprogramujeme za použití sítí s vrstvami, tedy jedinou věc, kterou potřebujeme implementovat je způsob používání filtru. Nakonec se podíváme i na asociativní paměť, ale tu z nedostatku času implementuji pouze částečně.

Aby nemuselo být v knihovně implementováno maticové násobení, použil jsem knihovnu `koma` (celým názvem `Kotlin math`), která implementuje základny lineární algebry v Kotlinu. Knihovna je pro JVM, Javascript i pro binární kód, avšak ve Windows ji nelze zkompileovat, proto naše knihovna funguje pouze pro JVM a Javascript. [14]

3 Struktura knihovny

Knihovna je rozdělena do dvou částí:

- První, a ta hlavní, je `core` (česky jádro), které obsahuje definice neuronových sítí (tj. konvoluční neuronovou síť, obyčejnou neuronovou síť, asociativní paměť) a definice pro ně potřebné (například aktivační funkce).
- Druhá je `mnistDatabase`, která se stará o učení neuronových sítí na datech z databázi ve formátu MNIST.

3.1 core

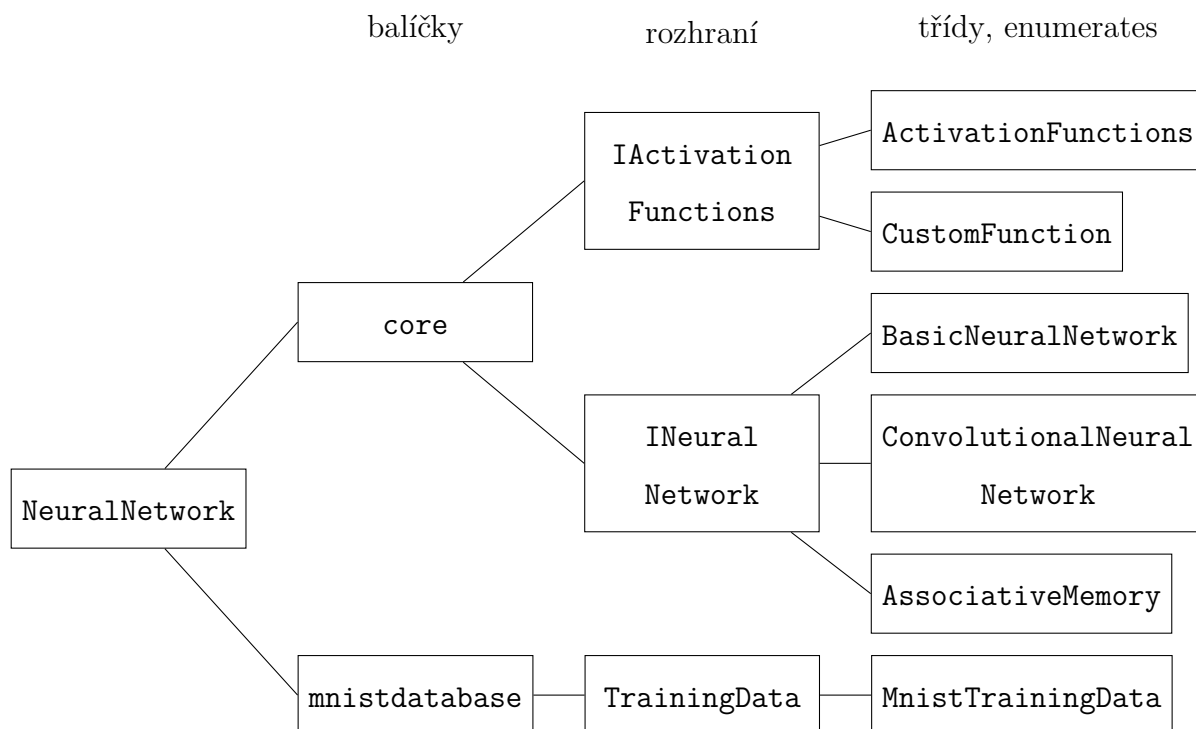
3.1.1 IActivationFunctions

Rozhraní, které zahrnuje `ActivationFunction` a `CustomFunction`. Jeho instance se používají jako aktivační funkce. Funkce lze zavolat s parametrem typu `Double`, což nám dá hodnotu funkce v tomto bodě, popřípadě lze obdobně zavolat jejich dvě metody `xD` a `yD` udávající v pořadí hodnotu derivace v bodě `x` a v bodě, kde je funkční hodnota rovna parametru.

3.1.2 ActivationFunctions

Enumerate častých funkcí, jež se používají jako aktivační funkce v neuronech. Některé jsou označeny jako překonané (anglicky `deprecated`), jelikož u funkcí, které nejsou všude hladké, neexistuje všude derivace. Taktéž u funkcí, jež nejsou prosté, nelze vždy určit derivaci podle funkční hodnoty.

Implementovány jsou všechny funkce uvedené v sekci 2.6



Obr. 3.1: Struktura knihovny

3.1.3 CustomFunction

Poskytuje možnost implementovat si vlastní aktivační funkci, má stejné metody (zde jsou to vlastnosti typu `() -> Unit`) jako `ActivationFunctions`.

3.1.4 INeuralNetwork

Rozhraní, které implementuje základní funkce neuronových sítí, které mají jako vstup i výstup vektor `Double`. Obsahuje funkce:

- `run(vstupní vektor)`, která je koncipována tak, aby ze vstupního vektoru spočítala vektor výstupní (tedy většinou udělala dopřednou propagaci). Jako vstupní vektor lze dát jak `Matrix<Double>` z knihovny `koma`, tak `DoubleArray`, které je převedeno na `Matrix<Double>`, následně se zavolá funkce `run` s tímto typem a výstup se převede zpět na `DoubleArray`.¹

Navíc (hlavně kvůli konvolučním neuronovým sítím) může být vstup i dvourozměrný, v tomto případě je pak nutno u `DoubleArray` uvést i šířku řádku.

¹`DoubleArray` je použito, protože je to typ Kotlinu samotného, ale jelikož matematika v neuronových sítích je implementována pomocí `Matrix<Double>`, musí se převést mezi typy.

- `train(vstupní vektor, chtěný výstupní vektor)` resp. `train(vstupní vektory, chtěné výstupní vektory)`, která je koncipována tak, aby nejdříve provedla `run(vstupní vektor)`, výsledek porovnála s chtěným a přepočítala váhy v neuronové síti tak, aby se výstup `run(vstupní vektor)` přiblížil (zmenšila se velikost jejich rozdílu) chtěnému výstupnímu vektoru. Kromě verze s parametry typu `DoubleArray` je funkce implementována i pro typ `Array<DoubleArray>`, tedy trénovací vstupy a výstupy lze vložit i všechny najednou.

3.1.5 BasicNeuralNetwork

Tato třída rozhraní `INeuralNetwork` implementuje nejčastěji používanou neuronovou síť, kde jsou neurony uspořádány do vrstev a ovlivňují se pouze jedním směrem. Parametry, které lze nastavit, jsou:

- `numberOfHiddenLayers`, neboli počet skrytých vrstev (tj. ty, jež jsou mezi vstupní a výstupní vrstvou). Čím více vrstev je nastaveno, tím hůře se síť učí, většinou je proto třeba nastavit pouze jednu skrytou vrstvu nebo nastavit velmi malou hodnotu `learning rate` (proměnná, jež není v konstruktoru, která udává rychlost změn vah).
- `activationFunctions`, česky aktivační funkce, musí být vybrána z třídy `ActivationFunction`. Při použití funkcí, které nejsou hladké, se neurony mohou chovat nepředvídatelným způsobem.

3.1.6 ConvolutionalNetwork

Tato třída rozhraní `INeuralNetwork` implementuje konvoluční neuronové sítě. Její konstruktor přijímá dva parametry typu `BasicNeuralNetwork`, první je filtr, druhá je samotná neuronová síť. Dalším parametrem je logická hodnota, zda se má i filtr učit (to se ale téměř nepoužívá, takže je tato hodnota při neuvedení nastavena na `false`).

Companion object této třídy navíc obsahuje příklad takového filtru (jednoduchý filtr detekující hrany viz obrázek 3.2)

0	1	1	1	1	0	0	-1	-1	-1	-1	-1	0	1	1
-1	0	1	1	0	-1	1	0	-1	-1	0	-1	0	1	1
-1	-1	0	0	-1	-1	1	1	0	1	1	0	1	1	1

1	1	1	-1	-1	-1	-1	0	1	1	1	1	0	0	-1
0	0	0	0	0	0	-1	0	1	-1	0	1	-1	1	-1
-1	-1	-1	1	1	1	-1	0	1	1	0	-1	1	0	-1

Obr. 3.2: Ilustrace filtru z třídy `ConvolutionalNetwork`, vstupem je matice 3×3 pixely, ta se po složkách násobí vždy s 1 z 8 matic výše, sečtou se všechny prvky výsledné matice, aplikuje se *rectified linear unit* a každé z výsledných 8 čísel pak udává, jak moc je v původní matici hrana odpovídající dané matici výše (tzn. jak moc je pixel násobený 1 bílý a pixel násobený -1 černý)

3.2 mnistDatabase

Pro otestování knihovny je potřeba nějaký dataset. K tomuto účelu je v knihovně implementována třída `TrainingDataMnist`, která umí přečíst data z databáze MNIST a EMNIST. Poté poskytuje vždy jedno zadání (obrázek číslice / písmena) a jeho řešení (ve formě vektoru, kde pouze na správném místě je 1, jinak je všude 0).

Jako parametry přijímá řetězec (`String`) s názvem souboru s obrázkem a řetězec s názvem souboru s daty (identifikací toho, co je na obrázcích). Zároveň nastavením parametru `inverse` na `true` lze převrátit osy obrázku (viz 3.2.2). Tato třída zatím funguje pouze v JVM, jelikož používá funkci na načtení souboru a tuto funkci jsem zatím v Javascriptu neimplementoval (soubor je většinou uložen někde na serveru, takže je obtížnější ho načíst).

Dále tento balíček rozšiřuje rozhraní `INeuralNetwork` o funkci `train` s parametrem typu `TrainData`, což je pouze `typealias` (tzn. jiný název pro typ v Kotlinu) za `Sequence<Pair<DoubleArray, DoubleArray>>`, jež je implementován výše zmíněnou třídou `TrainingDataMnist`.

3.2.1 Databáze MNIST

„Dataset MNIST, dataset ručně psaných číslic dostupná na stránkách <http://yann.lecun>.



Obr. 3.3: Příklad obrázků z datasetu MNIST [16]

com/exdb/mnist/ obsahuje 60 000 tréninkových a 10 000 ověřovacích příkladů. MNIST vychází z databáze spravované NIST (National Institute of Standards and Technology). Číslice mají normalizovanou velikost a jsou vycentrované v obrázcích shodné velikosti.“ [15, přeloženo] Ukázku takových obrázků vidíme na obrázku 3.3.

Tuto databázi jsem použil pro první testování své BasicNeuralNetwork, jelikož má pro první testování dostačující velikost. Pro pozdější testování využívám převážně EMNIST.

3.2.2 Databáze EMNIST

„Databáze MNIST se stala standardem pro učení umělého vidění. Databáze MNIST je odvozená z databáze NIST Special Database 19, která obsahuje ručně psané číslice a velká i malá písmena. EMNIST (Extended MNIST), varianta celé databáze NIST, přebírá uspořádání z databáze MNIST².“ [17, přeloženo]

Tato databáze obsahuje více příkladů než MNIST, navíc obsahuje i sety s písmeny, proto jsem po prvních pokusech s MNIST přešel na tuto databázi.

²Má však prohozené řádky a sloupce pixelů v obrázcích.

4 Používání knihovny

4.1 Trénování sítě

Příklad takového tréninku je v souboru `NeuralNetworkTestJVM` funkce `mnist()`. Takové trénování ale trvá více než deset minut (konkrétně tato funkce běží asi tři čtvrtě hodiny), tudíž ho nelze zahrnout do testů. V testech je pouze trénování malinké sítě, aby fungovala jako xor.

Nejprve musíme neuronovou síť natrénovat a uložit. Trénování neuronové sítě probíhá za pomoci funkce `train`. Té musíme poskytovat tréninkové vstupy s odpovídajícími výstupy, což můžeme udělat tak, že funkci `train` budeme volat z cyklu, který bude tato data postupně načítat. Dalším způsobem je předat rovnou celý `Array` vstupů a výstupů, to ale často znamená načíst miliony objektů třídy `Double`, proto to může výrazně zpomalit učení. Poslední možností (pokud máme data ve formátu MNIST) je využít třídy `TrainingData`, které poskytneme soubory s daty a ona vytvoří příslušné objekty typu `Double` až ve chvíli, kdy dojde na danou dvojici vstup – výstup.

Dobré je také během učení pomalu snižovat `learningRate`, jelikož nejdříve se neuronová síť vlastně učí hlavně konkrétní obrázky (v této fázi nejlépe poznává obrázky, které dostala v tréninku naposledy), poté ale umí čím dál více věcí a nechceme, aby se přepisovali již nabyté vědomosti. Já jsem například trénoval síť desetkrát na stejných datech (to není úplně vhodné, data by se měla měnit, aby se co nejméně naučila konkrétní obrázky¹, ale pro jednoduchost to stačí) s tím, že pokaždé jsem `learningRate` vydělil 1,5.

Poté už můžeme síť hned používat (například ji otestovat), ale většinou ji chceme používat víckrát a třeba i v rámci jiného programu. Proto mají třídy rozhraní `INeuralNetwork` funkci `save`, která vrátí data neuronové sítě jako řetězec (takový „osekaný“ JSON), který je pak možno uložit. V JVM je přímo definována funkce `saveFile(název souboru, data)`.

¹Při opakování malého datasetu se může stát, že neuronová síť bude umět rozpoznat jen obrázek, který je na pixel přesně shodný s tréninkovými obrázky.

4.2 Používání sítě

Ukázka načítání sítě je v programu `JSTest2` řádek 43 až 45 a ukázka výpočtu je parametr funkce `evaluateButton.addEventListener`. Můžete si všimnout, že použití je v rámci jednotek řádků kódu, zbytek se pouze stará o uživatelský vstup (program funguje jak za pomoci klikání myší, tak v mobilu pomocí dotyku).

Jakmile máme nějakou síť natrénovanou a uloženou v řetězci, můžeme ji znovu nahrát pomocí funkce `load(data)` nacházející se v companion objectu třídy `BasicNeuralNetwork` nebo `ConvolutionalNeuralNetwork`. Návrátovou hodnotou této funkce je samotná neuronová síť, takže ji stačí uložit do proměnné, na které pak zavoláme funkci `run` s vstupním vektorem jako parametrem a tím získáme výstupní vektor, který stačí už jen zpracovat (např. při rozpoznávání číslic to znamená zjistit, který z výsledných 10 neuronů vysílá největší výstupní signál).

4.3 Nastavování hodnot

Neuronová síť má mnoho hodnot, které lze nastavit. Knihovnu jsem zkoušel na rozpoznávání čísel v databázích MNIST a EMNIST a zjistil jsem, že vhodné nastavení hodnot je asi:

- Learning rate je třeba nastavit na cca 0.1 a samozřejmě snižovat.
- Počet skrytých vrstev musí být právě jedna (dvě už se nenaučí propojit vstup s výstupem a bez skryté vrstvy vůbec nefunguje). Pokud byste potřebovali učit síť s více skrytými vrstvami, musíte nastavit learning rate na daleko nižší hodnotu.
- Počet neuronů ve skryté vrstvě je hodně variabilní, ideálně mezi hodnotami 100 a 300.
- Jako aktivační funkce stačí třeba sigmoida, jiné jsem nepoužíval.

Závěr

Cílem mé práce bylo implementovat neuronovou síť, což se mi podařilo dokonce do takové míry, že v programu, kde zabírá pár řádků, je schopna rozeznávat číslíce (ukázka je na stránkách moznabude.cz, nebo na přiloženém USB). Největším přínosem je asi třída `BasicNeuralNetwork`, která implementuje velkou část matematiky obtížnou na rozmyšlení a stojící za téměř všemi neuronovými sítěmi, o niž se programátor v Kotlinu díky mojí knihovně už nemusí starat.

Zároveň jsem si díky rozdělení do balíčků a využití možností objektově orientovaného programování připravil dobrý podklad pro rozšiřování knihovny. Dále bych mohl pokračovat například implementováním lepšího ukládání do souboru (ukládání typu `Double` jako textového řetězce není moc efektivní), implementování některých genetických algoritmů, či naprogramování konvoluční sítě tak, aby filtry mohly pracovat n rozměrně.

Pro mě samotného byl asi největší přínos, že jsem si poprvé zkusil napsat formálnější kód a to jak v Kotlinu, tak i v LaTeXu. Navíc, už jen rozmyšlení si, co má tento text obsahovat byla pro mě velká životní zkušenost.

Slovníček pojmů

Array typ v Kotlinu odpovídající tzv. polím či vektorům v jiných programovacích jazycích, uchovává uspořádanou množinu objektů. 31, 34, 37

DoubleArray Array pro typ `Double`. 30–32

Double implementace 64bitových čísel s plovoucí desetinnou čárkou v Kotlinu. 30, 34, 36, 37

Pair typ v Kotlinu obsahující dvě vlastnosti `first` a `second`, dva objekty libovolného typu. 32

false opak `true`, většinou reprezentován 0. 31, 37

true hodnota typu `Boolean` (typ nabývající hodnot `true` a `false`) udávající pravdu, většinou reprezentován 1. 32, 37

axon výběžek vedoucí signál z neuronu. 10, 38

balíček anglicky `package` je něco jako složka, používá se k izolování proměnných, funkcí, tříd a rozhraní, které se dají nastavit na použití pouze v daném balíčku, a zároveň také udává samostatné části programu nebo knihovny, které jsou na sobě téměř nezávislé. 32

companion object tzv. statická část třídy v Kotlinu odpovídající modifikátoru `static` v Javě, obsahuje funkce a vlastnosti, které má třída i bez instance. 31, 35

cyklus pojem z teorie grafů, cyklus je posloupnost vrcholů (zde neuronů), přičemž z každého vrcholu do dalšího a z posledního do prvního vede hrana (zde `axon` → `dendrit`), tzn. pokud graf nemá cykly, nemůžeme se do vrcholu dostat vícekrát (zde nemusíme ho počítat vícekrát)

pojem z programování, používá se pro to, aby počítač opakoval kód. 11, 34

dendrit výběžek vedoucí signál do neuronu. 10, 11, 38

enumerate česky výčet, typický prvek Javy či Kotlinu, třída, která má přesně definované instance (např. dny v týdnu by se implementovali jako enumerate). 29, 30, 38

gradient vektor derivací funkce podle jednotlivých proměnných, v našem světě si ho můžeme představit jako vodorovnou šipku (v každém bodě světa), která ukazuje, kterým směrem a jak moc jde krajina nejvíce do kopce z tohoto bodu (proměnné jsou pro tento příklad vodorovné souřadnice, funkcí je výška třeba nad mořem). 12, 17, 21

JSON zkratka JavaScript Object Notation, lidsky čitelný formát ukládání Javascriptových objektů, každý parametr objektu se uloží jako „navez”: hodnota“ a celý objekt je obalený „{ }“. 34

JVM Java Virtual Machine je virtuální stroj, který umožňuje běh Java Bytecodu, kódu, do kterého se překládá Java a Kotlin. 8, 28, 32, 34

Kotlin programovací jazyk vyvíjený firmou JetBrains, založen na Javě. 3, 4, 8, 28, 30, 32, 36–38

rozhraní anglicky interface je v objektově orientovaném programování zabalení funkcí a vlastností třídy, které by měla každá třída z nějaké skupiny mít (např. každá fronta by měla mít funkci pro přidání a odebrání prvku a jedna z jejích vlastností je velikost). 29–31, 34, 37, 38

synapse spojení (mezera) mezi axonem a dendritem, jež podle svých chemických vlastností zesílí nebo zeslabí signál předávaný z axonu do dendritu. 10, 12

typ třída nebo rozhraní, jehož instancí je daný objekt. 29–32, 34, 36, 37

třída anglicky class je základní prvek objektově orientovaného programování. Obsahuje funkce a vlastnosti, které bude mít objekt, který se vytvoří z dané třídy (popřípadě třídy, jež budou z této třídy dědit). 30–32, 34–38, 41

xor tzv. výlučné nebo, neboli binární (tj. přijímá dvě hodnoty / tvrzení) logická funkce, která je pravda právě tehdy, když jedno tvrzení je pravdivé a jedno nepravdivé. 34

Bibliografie

- [1] J. Glenn Brookshear, David T. Smith a Dennis Brylow. *Informatika*. cs. Přel. en Jakub Goner. 1. vyd. Brno, CZ: Computer Press, 2013, s. 608. ISBN: 978-80-251-3805-2.
- [2] Kevin Gurney (University of Sheffield, UK). *An Introduction to Neural Networks*. Taylor & Francis Ltd, 5. srp. 1997, s. 234. ISBN: 1857285034. URL: https://www.ebook.de/de/product/3243601/kevin_university_of_sheffield_uk_gurney_an_introduction_to_neural_networks.html.
- [3] Harry Pratt et al. *FCNN: Fourier Convolutional Neural Networks*. en. Tech. zpr. University of Liverpool, Liverpool, L69 3BX, UK. URL: <http://ecmlpkdd2017.ijs.si/papers/paperID11.pdf> (cit. 30.01.2020).
- [4] Fei-Fei Li, Andrej Karpathy a Justin Johnson. *Lecture 7: Convolutional Neural Networks*. en. Online. Presentation. Stanford University, 27. led. 2016. URL: http://cs231n.stanford.edu/slides/2016/winter1516_lecture7.pdf (cit. 30.01.2020).
- [5] David Stutz. *Understanding Convolutional Neural Networks*. en. semreport. Fakultät für Mathematik, Informatik und Naturwissenschaften, 30. srp. 2014. URL: <https://davidstutz.de/wordpress/wp-content/uploads/2014/07/seminar.pdf> (cit. 30.01.2020).
- [6] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: (10. červ. 2014). arXiv: <http://arxiv.org/abs/1406.2661v1> [stat.ML].
- [7] Inc. Generated Media. *Generated Photos*. en. 2019. URL: <https://generated.photos/> (cit. 29.01.2020).
- [8] Daniel Shiffman. *Neural Networks - The Nature of Code*. en. YouTube. 26. červ. 2017. URL: https://www.youtube.com/user/shiffman/playlists?view_as=subscriber&shelf_id=6&view=50&sort=dd (cit. 30.01.2020).

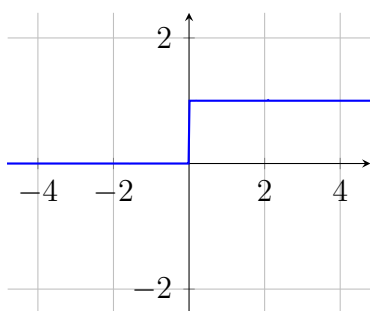
- [9] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015. URL: <http://neuralnetworksanddeeplearning.com/> (cit. 30.01.2020).
- [10] L. Pick et al. *Matematická analýza 1. (velmi předběžná verze)*. 3. dub. 2019.
- [11] Wikipedia contributors. *Activation function — Wikipedia, The Free Encyclopedia*. [Online]. 2019. URL: https://en.wikipedia.org/w/index.php?title=Activation_function&oldid=933057521 (cit. 09.01.2020).
- [12] Chigozie Nwankpa et al. “Activation Functions: Comparison of trends in Practice and Research for Deep Learning”. In: (8. lis. 2018). arXiv: <http://arxiv.org/abs/1811.03378v1> [cs.LG].
- [13] Farnoush Farhadi. *Learning activation functions in deep neural networks*. Université De Montréal (école Polytechnique De Montréal), 2017.
- [14] Kyle Kauffman. *Koma*. en. 2016. URL: <http://koma.kyonifer.com/> (cit. 30.01.2020).
- [15] Yann LeCun, Corinna Cortes a Christopher J.C. Burges. *THE MNIST DATABASE of handwritten digits*. en. 1998. URL: <http://yann.lecun.com/exdb/mnist/> (cit. 15.12.2019).
- [16] Wikimedia Commons. *File:MnistExamples.png — Wikimedia Commons, the free media repository*. [Online]. 2020. URL: <https://commons.wikimedia.org/w/index.php?title=File:MnistExamples.png&oldid=390556927> (cit. 07.02.2020).
- [17] Gregory Cohen et al. “EMNIST: an extension of MNIST to handwritten letters”. In: (17. ún. 2017). arXiv: <http://arxiv.org/abs/1702.05373v2> [cs.CV].

Seznam obrázků

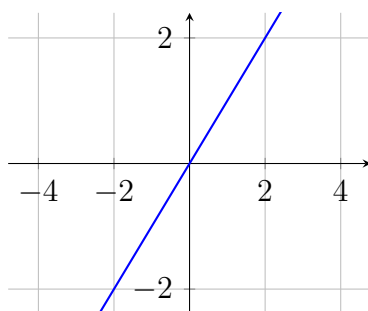
1.1	Běžná neuronová síť (W jsou váhy, n neurony a v je výstupní signál, viz kapitola 2 a sekce 2.5)	11
1.2	Asociativní paměť, červeně jsou vybuzené neurony	13
1.3	Generovaná tvář [7]	14
2.1	Neuron	15
2.2	Binární krok	21
2.3	Identita	22
2.4	σ	22
2.5	Hyperbolický tangens	23
2.6	Swish	23
2.7	Rectified linear unit	24
2.8	Hard hyperbolic function	24
2.10	Sinus	25
3.1	Struktura knihovny	30
3.2	Ilustrace filtru z třídy ConvolutionalNetwork, vstupem je matice 3×3 pixely, ta se po složkách násobí vždy s 1 z 8 matic výše, sečtou se všechny prvky výsledné matice, aplikuje se <i>rectified linear unit</i> a každé z výsledných 8 čísel pak udává, jak moc je v původní matici hrana odpovídající dané matici výše (tzn. jak moc je pixel násobený 1 bílý a pixel násobený -1 černý)	32
3.3	Příklad obrázků z datasetu MNIST [16]	33

Přílohy

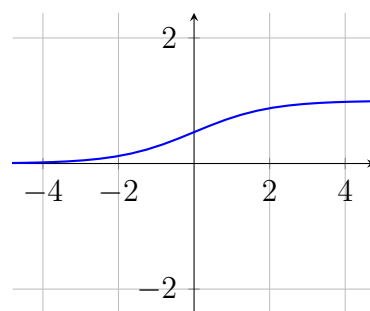
1. Zdrojový kód knihovny (složka NeuralNetwork)
2. Dokumentace (složka Dokumentace)
3. Testovací dataset (složka MNIST)
4. Zdrojový kód ukázkového programu (složka JSTest2)
5. Ukázkový program (soubor JSTest2/Main.html)
6. Zdrojový kód práce v \LaTeX u (složka LaTeX)
7. Přehled grafů aktivačních funkcí (následující stránky)
8. Zdrojový kód knihovny (následující stránky)



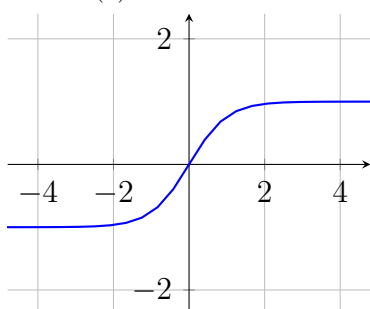
(a) Binární krok



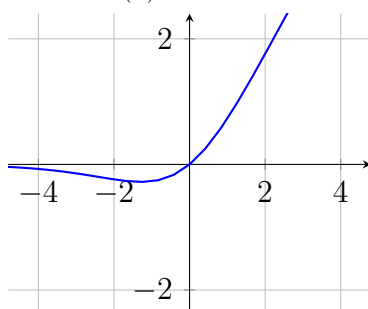
(b) Identita



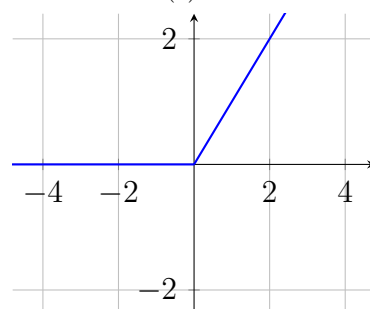
(c) σ



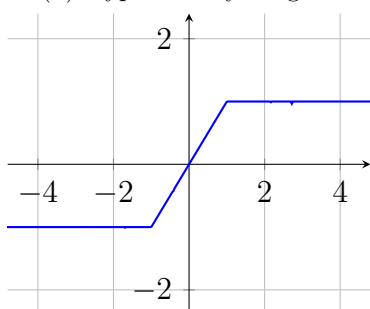
(d) Hyperbolický tangens



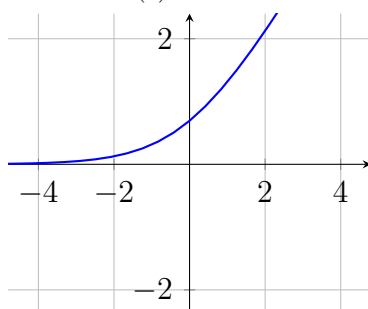
(e) Swish



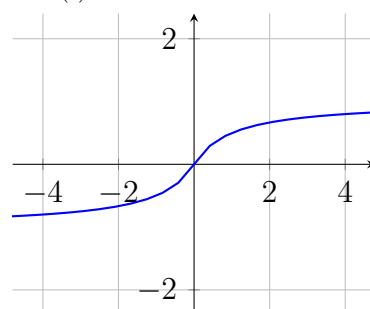
(f) Rectified linear unit



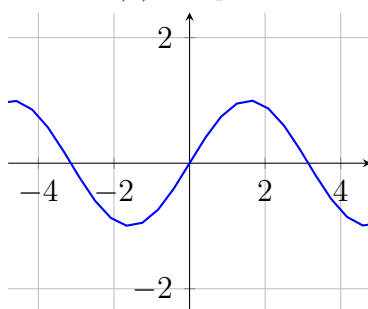
(g) Hard hyperbolic function



(h) Soft plus



(i) Soft signum



(j) Sinus

Přehled grafů aktivačních funkcí

Listing 1: src/commonMain/kotlin/core/ActivationFunctions.kt

```

1  /*
2   * Licensed under the MIT License. See LICENSE file in the project root for full license information.
3   */
4  package core
5
6  import kotlin.math.*
7
8  /**
9   * Enumerate of many common functions ([invoke] returns f(x)), with it's derivation (f'(x)) for 2 cases: when we have x - [xD] or
10   * ↪ when we have f(x) - [yD]
11   *
12   * @param[xD] Derivation (f'(x)) when we have x value
13   * @param[yD] Derivation (f'(x)) when we have y = f(x) value
14   */
15  enum class ActivationFunctions(
16      private val function: (Double) -> Double,
17      override val xD: (Double) -> Double,
18      override val yD: (Double) -> Double
19  ) : IActivationFunctions {
20      /**
21       * Zero for negative values, one for others.
22       */
23      @Deprecated("This function isn't smooth", level = DeprecationLevel.WARNING)
24      BinaryStep({
25          if (it < 0) {
26              0.0
27          } else {
28              1.0
29          }
30      }, {
31          if (it == 0.0) {
32              Double.POSITIVE_INFINITY
33          } else {
34              0.0
35          }
36      }, { 0.0 })),
37      /**
38       * Identity for -1 <= x <= 1, -1 for x < -1 and 1 for x > 1
39       */
40      @Deprecated("This function isn't smooth", level = DeprecationLevel.WARNING)
41      HardHyperbolicFunction({
42          when {
43              it < -1 -> -1.0
44              it > 1 -> 1.0
45              else -> it
46          }
47      }, {
48          when {
49              it < -1 || it > 1 -> {
50                  0.0
51              }
52              it == -1.0 || it == 1.0 -> {
53                  Double.NaN
54              }
55              else -> {
56                  1.0
57              }
58          }
59      }, {
60          if (it == -1.0 || it == 1.0) {
61              0.0
62          } else {
63              1.0
64          }
65      })

```

```

65     }},
66
67     /**
68      * Zero for negative x, identity for positive x
69      */
70     @Deprecated("This function isn't smooth", level = DeprecationLevel.WARNING)
71     RectifiedLinearUnit({ max(0.0, it) }, {
72         when {
73             it < 0 -> {
74                 0.0
75             }
76             it == 0.0 -> {
77                 Double.NaN
78             }
79             else -> {
80                 1.0
81             }
82         }
83     }, {
84         if (it == 0.0) {
85             0.0
86         } else {
87             1.0
88         }
89     })),
90
91     /**
92      * Identity for positive x, scaled identity ([ALPHA] * x) for negative x
93      */
94     @Deprecated("This function isn't smooth", level = DeprecationLevel.WARNING)
95     LeakyRectifiedLinearUnit({
96         if (it < 0) {
97             ALPHA * it
98         } else {
99             it
100         }
101     }, {
102         when {
103             it < 0 -> {
104                 ALPHA
105             }
106             it == 0.0 -> {
107                 Double.NaN
108             }
109             else -> {
110                 1.0
111             }
112         }
113     }, {
114         if (it < 0.0) {
115             ALPHA
116         } else {
117             1.0
118         }
119     })),
120
121     /**
122      *  $f(x) = x$ 
123      */
124     Identity({
125         it
126     }, {
127         1.0
128     }, {
129         1.0
130     })),
131

```

```

132  /**
133   * Smooth step:  $f(x) = 1 / (1 + e^{-x})$ 
134   */
135  Sigmoid({
136      1 / (1 + exp(-it))
137  }, {
138      val expIt = exp(-it)
139      expIt / (1 + expIt).pow(2)
140  }, {
141      it * (1 - it)
142  }),
143
144  /**
145   * Hyperbolic tangents
146   */
147  Tanh({
148      tanh(it)
149  }, {
150      1 / cosh(it).pow(2)
151  }, {
152      1 - it.pow(2)
153  }),
154
155  /**
156   * Sign with smoothing ( $x / (|x| + 1)$ )
157   */
158  Softsign({
159      it / (abs(it) + 1)
160  }, {
161      1 / (1 + abs(it)).pow(2)
162  }, {
163      (1 - abs(it)).pow(2)
164  }),
165
166  /**
167   * [RectifiedLinearUnit] with smoothing ( $\ln(1 + \exp(x))$ )
168   */
169  Softplus({
170      ln(1 + exp(it))
171  }, {
172      1 / (1 + exp(-it))
173  }, {
174      1 / (2 - exp(it))
175  }),
176
177  /**
178   * Identity for positive x, scaled exponential ( $[ALPHA] * \exp(x) - 1$ ) for negative x
179   */
180  ExponentialLinearUnit({
181      if (it > 0) {
182          it
183      } else (ALPHA * exp(it) - 1)
184  }, {
185      if (it > 0) {
186          1.0
187      } else (ALPHA * (exp(it) + 1) - 1)
188  }, {
189      if (it > 0) {
190          1.0
191      } else (it + ALPHA)
192  }),
193
194  /**
195   *  $x * [Sigmoid] (x / (1 + \exp(-x)))$ 
196   */
197  Swish({
198      it / (1 + exp(-it)) // it * Sigmoid(it)

```

```

199     }, {
200         val expIt = exp(-it)
201         1 / (1 + expIt) + it * expIt / (1 + expIt).pow(2)
202     }, {
203         TODO("WTF?")
204     }},
205
206     /**
207      * [Sigmoid] for negative x, [ExponentialLinearUnit] for positive x and 0
208      */
209     ExponentialLinearSquashing({
210         if (it < 0) (Sigmoid(it)) else (ExponentialLinearUnit(it))
211     }, {
212         if (it < 0) (Sigmoid.xD(it)) else (ExponentialLinearUnit.xD(it))
213     }, {
214         if (it < 0) (Sigmoid.yD(it)) else (ExponentialLinearUnit.yD(it))
215     }},
216
217     /**
218      * ??? for negative x, [ExponentialLinearUnit] for positive x and 0
219      */
220     HardExponentialLinearSquashing({
221         if (it < 0) ((exp(it) - 1) * max(0.0, min(1.0, (it + 1) / 2))) else (ExponentialLinearUnit(it))
222     }, {
223         if (it < 0) (Sigmoid.xD(it)) else (ExponentialLinearUnit.xD(it))
224     }, {
225         TODO("WTF~2")
226     }},
227
228     /**
229      * Simply sinus
230      */
231     Sinus({
232         sin(it)
233     }, {
234         cos(it)
235     }, {
236         sqrt(1 - it.pow(2))
237     })
238     ;
239
240     override operator fun invoke(double: Double) = function(double)
241
242     companion object {
243         const val ALPHA = 1.0
244     }
245 }

```

Listing 2: src/commonMain/kotlin/core/BasicNeuralNetwork.kt

```

1  /*
2   * Licensed under the MIT License. See LICENSE file in the project root for full license information.
3   */
4  package core
5
6  import koma.create
7  import koma.extensions.map
8  import koma.matrix.Matrix
9  import koma.rand
10 import koma.zeros
11 //import kotlin.math.abs
12 import kotlin.math.sqrt
13
14 /**
15  * Basic Neural Network consisted only of some layers of neurons.
16  *

```

```

17  * @constructor creates new [BasicNeuralNetwork] with
18  * * [numberOfHiddenLayers] hidden layers, which have sizes generated by [sizes]
19  * * neurons activating given by [activationFunction]
20  * * input for [inputLayerSize] [Double]s
21  * * answering with [outputLayerSize] [Double]s
22  *
23  * @param[numberOfHiddenLayers] number of hidden layers (without input and output layers)
24  * @param[activationFunction] how neurons are activated
25  * @param[sizes] sizes of hidden layers
26  * @param[inputLayerSize] size of input
27  * @param[outputLayerSize] size of output
28  * @param[weights] list of matrices, which state weights of connections between neurons in previous layer and neurons in next one
29  */
30  class BasicNeuralNetwork(
31      private val numberOfHiddenLayers: Int,
32      val activationFunction: IActivationFunction = ActivationFunctions.Sigmoid,
33      val sizes: (Int) -> Int = { numberOfHiddenLayers },
34      val inputLayerSize: Int = numberOfHiddenLayers,
35      val outputLayerSize: Int = numberOfHiddenLayers,
36      private val weights: MutableList<Matrix<Double>> = MutableList(numberOfHiddenLayers + 1) {
37          if (numberOfHiddenLayers == 0) {
38              rand(outputLayerSize, inputLayerSize)
39          } else when (it) {
40              0 -> rand(sizes(it), inputLayerSize) * (sqrt(2.0 / (sizes(it) + inputLayerSize)))
41              numberOfHiddenLayers -> rand(
42                  outputLayerSize,
43                  sizes(it - 1)
44              ) * (sqrt(2.0 / (outputLayerSize + sizes(it - 1))))
45              else -> rand(sizes(it), sizes(it - 1)) * (sqrt(2.0 / (sizes(it) + sizes(it - 1))))
46          }
47      },
48      private val biases: MutableList<Matrix<Double>> = MutableList(numberOfHiddenLayers + 1) {
49          //rand(if (it == numberOfHiddenLayers) { outputLayerSize } else { sizes(it) }, 1)
50          zeros(
51              if (it == numberOfHiddenLayers) {
52                  outputLayerSize
53              } else {
54                  sizes(it)
55              }, 1
56          )
57      },
58      private val values: MutableList<Matrix<Double>> = MutableList(numberOfHiddenLayers + 2) {
59          zeros(
60              when (it) {
61                  0 -> inputLayerSize
62                  numberOfHiddenLayers + 1 -> outputLayerSize
63                  else -> sizes(it)
64              }, 1
65          )
66      }
67  ) : INeuralNetwork {
68
69      /**
70       * [Double] value which declares how quickly weights and biases are changing
71       */
72      var learningRate = 0.1
73
74      override fun run(input: Matrix<Double>): Matrix<Double> {
75          require(inputLayerSize == input.size) { "Wrong size of input! This NN has input size $inputLayerSize, but you offer it
76              ↪ input with size ${input.size}." }
77          values[0] = input
78          for (index in weights.indices) {
79              values[index + 1] = (weights[index] * values[index] + biases[index]).map { activationFunction(it) }
80          }
81          return values.last()
82      }

```



```

83 override fun train(input: Matrix<Double>, output: Matrix<Double>) = train(output - run(input))
84
85 fun train(er: Matrix<Double>): Matrix<Double> {
86     var error = er
87     for (i in numberOfHiddenLayers downTo 0) {
88         val derivations = values[i + 1].map { activationFunction.yD(it) }.elementTimes(error)
89         biases[i] += derivations * learningRate
90         error = weights[i].T * derivations
91         weights[i] += derivations * values[i].T * learningRate
92     }
93     return error
94 }
95
96 override fun save() =
97     when (activationFunction) {
98         is ActivationFunctions -> "$numberOfHiddenLayers;$activationFunction;${(0..numberOfHiddenLayers + 1).map(
99             sizes
100         )};$inputLayerSize;$outputLayerSize;${weights.map { it.toList() }};${biases.map { it.toList() }}"
101         else -> TODO("It's hard to save unknown function")
102     }
103
104 companion object {
105     /**
106      * Load [BasicNeuralNetwork] from [data]
107      */
108     fun load(data: String): BasicNeuralNetwork {
109
110         val dataList = data.split(";")
111         val numberOfHiddenLayers = dataList[0].toInt()
112         val sizeList = dataList[2].removePrefix("[").removeSuffix("]").split(", ").map { it.toInt() }
113         val sizes: (Int) -> Int = { sizeList[it] }
114         val inputLayerSize = dataList[3].toInt()
115         val outputLayerSize = dataList[4].toInt()
116         return BasicNeuralNetwork(
117             numberOfHiddenLayers,
118             try {
119                 ActivationFunctions.valueOf(dataList[1])
120             } catch (e: Exception) {
121                 TODO("It's hard to save unknown function")
122             },
123             sizes,
124             inputLayerSize,
125             outputLayerSize,
126             dataList[5].removePrefix("[[").removeSuffix("]]").split("[", "]").mapIndexed
127             { index, it ->
128                 if (numberOfHiddenLayers == 0) {
129                     create(
130                         it.split(", ").map { str -> str.toDouble() }.toDoubleArray(),
131                         outputLayerSize,
132                         inputLayerSize
133                     )
134                 } else when (index) {
135                     0 -> create(
136                         it.split(", ").map { str -> str.toDouble() }.toDoubleArray(),
137                         sizes(index),
138                         inputLayerSize
139                     )
140                     numberOfHiddenLayers -> create(
141                         it.split(", ").map { str -> str.toDouble() }.toDoubleArray(),
142                         outputLayerSize,
143                         sizes(index - 1)
144                     )
145                     else -> create(
146                         it.split(", ").map { str -> str.toDouble() }.toDoubleArray(),
147                         sizes(index),
148                         sizes(index - 1)
149                     )
150                 }
151             }
152         )
153     }
154 }

```

```

150         }
151     }.toMutableList(),
152     dataList[6].removePrefix("[[").removeSuffix("]").split(", ").mapIndexed
153     { index, it ->
154         create(
155             it.split(", ").map { str -> str.toDouble() }.toDoubleArray(),
156             if (index == numberOfHiddenLayers) {
157                 outputLayerSize
158             } else {
159                 sizes(index)
160             },
161             1
162         )
163     }.toMutableList()
164 )
165 }
166 }
167 }

```

Listing 3: src/commonMain/kotlin/core/ConvolutionalNeuralNetwork.kt

```

1 package core
2
3 import koma.create
4 import koma.extensions.*
5 import koma.matrix.Matrix
6 import koma.sqrt
7
8 /**
9  * Convolutional Neural Network consisted only of two [BasicNeuralNetwork].
10  *
11  * @constructor creates new [ConvolutionalNeuralNetwork] with
12  * * [filter] as small [BasicNeuralNetwork] that applies on every part of image before [neuralNetwork]
13  * * [neuralNetwork] as the main network
14  *
15  * @param [filter] small main network
16  * @param [neuralNetwork] main neural network
17  * @param [trainBoth] if filter should be trained
18  */
19 class ConvolutionalNeuralNetwork(
20     private val filter: BasicNeuralNetwork,
21     private val neuralNetwork: BasicNeuralNetwork,
22     private val trainBoth: Boolean = false
23 ) :
24     INeuralNetwork {
25
26     /**
27      * Size of one side of [filter]
28      */
29     private val filterSizeSqrt: Int
30
31     /**
32      * [Double] value which declares how quickly weights and biases are changing
33      */
34     var learningRate = 0.1
35     set(value) {
36         field = value
37         filter.learningRate = value
38         neuralNetwork.learningRate = value
39     }
40
41     init {
42         val s = sqrt(filter.inputLayerSize)
43         filterSizeSqrt = s.toInt()
44         require(s == filterSizeSqrt.toDouble()) { "Filter is not square" }
45         require(neuralNetwork.inputLayerSize % filter.outputLayerSize == 0) { "Filter is not for this neural network" }

```

```

46     }
47
48     /**
49     * Applies filter on every square of [input]
50     */
51     private fun runFilter(input: Matrix<Double>): Matrix<Double> {
52         val output = Matrix(
53             (input.numRows() - filterSizeSqrt + 1) * (input.numCols() - filterSizeSqrt + 1) * filter.outputLayerSize,
54             1
55         ) { _, _ -> 0.0 }
56         var offset = 0
57         for (i in 0 until input.numRows() - filterSizeSqrt + 1) {
58             for (j in 0 until input.numCols() - filterSizeSqrt + 1) {
59                 val output1 = filter.run(input[i until i + filterSizeSqrt, j until j + filterSizeSqrt].toDoubleArray())
60                 output1.forEachIndexed { it, ele ->
61                     output[offset + it] = ele
62                 }
63                 offset += output1.size
64             }
65         }
66         return output
67     }
68
69     override fun run(input: Matrix<Double>): Matrix<Double> {
70         val input2 = runFilter(input)
71         require(input2.size == neuralNetwork.inputLayerSize) { "Invalid input matrix size for neural network" }
72         return neuralNetwork.run(input2)
73     }
74
75     override fun train(input: Matrix<Double>, output: Matrix<Double>): Matrix<Double> {
76         val input2 = runFilter(input)
77         val error = neuralNetwork.train(input2, output)
78         return if (trainBoth) {
79             val error2 = Matrix(filter.outputLayerSize, 1) { _, _ -> 0.0 }
80             error.forEachIndexed { idx: Int, ele: Double -> error[idx % filter.outputLayerSize] += ele }
81             filter.train(error2.map { it * filter.outputLayerSize / error.size })
82         } else create(DoubleArray(0))
83     }
84
85     override fun save() = filter.save() + ";;" + neuralNetwork.save()
86
87     companion object {
88         /**
89         * Load [ConvolutionalNeuralNetwork] from [data]
90         */
91         fun load(data: String): ConvolutionalNeuralNetwork {
92             val nns = data.split(";;")
93             return ConvolutionalNeuralNetwork(BasicNeuralNetwork.load(nns[0]), BasicNeuralNetwork.load(nns[1]))
94         }
95
96         /**
97         * Data of [Matrix] for [edgeFilter]
98         */
99         private val edgeFilterData = mutableListOf(
100             mutableListOf(1.0, 1.0, 1.0, 0.0, 0.0, 0.0, -1.0, -1.0, -1.0),
101             mutableListOf(1.0, 0.0, -1.0, 1.0, 0.0, -1.0, 1.0, 0.0, -1.0),
102             mutableListOf(-1.0, -1.0, -1.0, 0.0, 0.0, 0.0, 1.0, 1.0, 1.0),
103             mutableListOf(-1.0, 0.0, 1.0, -1.0, 0.0, 1.0, -1.0, 0.0, 1.0),
104             mutableListOf(1.0, 1.0, 0.0, 1.0, 0.0, -1.0, 0.0, -1.0, -1.0),
105             mutableListOf(-1.0, -1.0, 0.0, -1.0, 0.0, 1.0, 0.0, 1.0, 1.0),
106             mutableListOf(0.0, 1.0, 1.0, 1.0, 0.0, -1.0, -1.0, -1.0, 0.0),
107             mutableListOf(0.0, -1.0, -1.0, -1.0, 0.0, 1.0, 1.0, 1.0, 0.0)
108         )
109
110         /**
111         * Example filter, detects edges
112         */

```

```

113     val edgeFilter: BasicNeuralNetwork
114     get() = BasicNeuralNetwork(
115         0, ActivationFunctions.RectifiedLinearUnit, { 0 }, 9, 8,
116         mutableListOf(Matrix(8, 9) { row: Int, cols: Int ->
117             edgeFilterData[row][cols]
118         })
119     )
120 }
121
122 }

```

Listing 4: src/commonMain/kotlin/core/CustomFunction.kt

```

1 package core
2
3 class CustomFunction(
4     private val function: (Double) -> Double,
5     override val xD: (Double) -> Double,
6     override val yD: (Double) -> Double
7 ) : IActivationFunctions {
8     override fun invoke(double: Double): Double = function(double)
9 }

```

Listing 5: src/commonMain/kotlin/core/IActivationFunctions.kt

```

1 package core
2
3 /**
4  * Interface for activation functions for neural networks
5  */
6
7 interface IActivationFunctions {
8     /**
9      * Derivation (f'(x)) when we have x value
10     */
11     val xD: (Double) -> Double
12
13     /**
14      * Derivation (f'(x)) when we have y = f(x) value
15     */
16     val yD: (Double) -> Double
17
18     /**
19      * Returns functional value (f([double]))
20     */
21     operator fun invoke(double: Double): Double
22 }

```

Listing 6: src/commonMain/kotlin/core/INeuralNetwork.kt

```

1 /**
2  * Licensed under the MIT License. See LICENSE file in the project root for full license information.
3  */
4 package core
5
6 import koma.create
7 import koma.extensions.toDoubleArray
8 import koma.matrix.Matrix
9
10 /**
11  * Neural Network Interface
12  *
13  * Basic usage is train it by [train] and then use it by [run]
14  */

```

```

15 interface INeuralNetwork {
16
17     /**
18      * Takes input, process it throw neural network and returns Matrix vector of [Double] outputs
19      */
20     fun run(input: Matrix<Double>): Matrix<Double>
21
22     fun run(input: DoubleArray, numCols: Int = 1) = run(create(input, input.size / numCols, numCols))
23
24     /**
25      * Takes input and desired output, compute estimated output and apply backpropagation
26      */
27     fun train(input: Matrix<Double>, output: Matrix<Double>): Matrix<Double>
28
29     fun train(input: DoubleArray, output: DoubleArray, inNumCols: Int = 1, outNumCols: Int = 1) =
30         train(
31             create(input, input.size / inNumCols, inNumCols),
32             create(output, output.size / outNumCols, outNumCols)
33         ).toDoubleArray()
34
35     fun train(input: Array<DoubleArray>, output: Array<DoubleArray>) {
36         require(input.size == output.size) { "Wrong training sets! Size of input is ${input.size}, size of output is ${output.size}
37             ↪ }." }
38         for (i in input.indices) {
39             train(input[i], output[i])
40         }
41     }
42
43     /**
44      * Returns save of NN in [String]
45      */
46     fun save(): String

```

Listing 7: src/commonMain/kotlin/mnistDatabase/loadFile.kt

```

1 package mnistDatabase
2
3 import core.INeuralNetwork
4
5 expect fun loadFile(file: String): ByteArray
6 expect fun loadFileString(file: String): String
7 expect fun saveFile(file: String, text: String)
8
9 private fun Byte.toUInt() = (this.toInt() + 256) % 256
10 private fun Byte.toUNDouble() = ((this.toDouble() + 256.0) % 256.0) / 256
11
12 private fun List<Byte>.toIntArray(): IntArray {
13     require(size % 4 == 0)
14     val result = IntArray(size / 4)
15     for (i in indices) {
16         result[i / 4] += (this[i].toUInt() shl 8 * (when (i % 4) {
17             0 -> 3
18             1 -> 2
19             2 -> 1
20             3 -> 0
21             else -> 4
22         }))
23     }
24     return result
25 }
26
27 typealias TrainingData = Sequence<Pair<DoubleArray, DoubleArray>>
28
29 class MnistTrainingData(imageFile: String, numberFile: String, val inverse: Boolean) : TrainingData {
30

```

```

31 private val imageBytes = loadFile(imageFile.removeSuffix(".idx3-ubyte") + ".idx3-ubyte")
32 private val imageFirstInts = imageBytes.slice(4 until 16).toIntArray()
33 private val numberOfImages = imageFirstInts[0]
34 private val numberOfRows = imageFirstInts[1]
35 private val numberOfColumns = imageFirstInts[2]
36 private val sizeOfImage = numberOfColumns * numberOfRows
37
38 private val numberBytes = loadFile(numberFile.removeSuffix(".idx1-ubyte") + ".idx1-ubyte")
39
40 init {
41     require(numberOfImages == numberBytes.slice(4 until 8).toIntArray().first()) { "Error" }
42 }
43
44 override fun iterator(): Iterator<Pair<DoubleArray, DoubleArray>> {
45     return object : Iterator<Pair<DoubleArray, DoubleArray>> {
46         val data = this@MnistTrainingData
47         val indexes = (0 until numberOfImages).shuffled()
48         var index = 0
49         override fun hasNext() = index < numberOfImages
50
51         override fun next(): Pair<DoubleArray, DoubleArray> {
52
53             var image =
54                 data.imageBytes.slice(16 + indexes[index] * sizeOfImage until 16 + (indexes[index] + 1) * sizeOfImage)
55                 .map { byte -> byte.toUNDouble() }.toDoubleArray()
56
57             if (inverse) {
58                 val newimage = DoubleArray(sizeOfImage)
59                 for (i in 0 until numberOfRows) {
60                     for (j in 0 until numberOfColumns) {
61                         newimage[i * 28 + j] = image[j * 28 + i]
62                     }
63                 }
64                 image = newimage
65             }
66
67             val position = numberBytes[8 + indexes[index]].toUInt()
68             val number = DoubleArray(10) {
69                 if (it == position) {
70                     1.0
71                 } else {
72                     0.0
73                 }
74             }
75
76             index++
77             return image to number
78         }
79     }
80 }
81 }
82
83 fun INeuralNetwork.train(data: TrainingData, numCols: Int = 1) {
84     for ((input, output) in data) {
85         train(input, output, numCols)
86     }
87 }

```

Listing 8: src/commonTest/kotlin/sample/Constants.kt

```

1 package sample
2
3 import koma.matrix.Matrix
4 import koma.matrix.MatrixFactory
5
6 const val wrongInputLayerSize = 1

```

```

7  const val numberOfHiddenLayers = 1
8  const val numberOfDigits = 10
9  const val imageWidth = 28
10 const val imageHeight = 28
11 const val blackFrom = 0.5
12 const val learningRateEpochDecrease = 1.5
13
14 val input: DoubleArray = DoubleArray(2) { 1.0 }
15 //get() = DoubleArray(2) { 1.0 }
16 val output: DoubleArray = input
17 //get() = input
18 val inputTest = input.copyOf()
19 val outputTest = output.copyOf()
20
21 expect val defaultDoubleMatrixFactory: MatrixFactory<Matrix<Double>>

```

Listing 9: src/commonTest/kotlin/sample/NeuralNetworkTest.kt

```

1  package sample
2
3  import core.BasicNeuralNetwork
4  import koma.create
5  import koma.matrix.Matrix
6  import kotlin.test.Test
7  import kotlin.test.assertFailsWith
8  import kotlin.test.assertTrue
9
10 class NeuralNetworkTest {
11
12     init {
13         Matrix.doubleFactory = defaultDoubleMatrixFactory
14     }
15
16     // @Test
17     fun inputs() {
18         assertFailsWith<IllegalArgumentException>("Wrong size of input! This NN has input size $wrongInputLayerSize, but you offer
19             ↪ it input with size ${input.size}." ) {
20             val nn = BasicNeuralNetwork(numberOfHiddenLayers, inputLayerSize = wrongInputLayerSize)
21             nn.run(input)
22         }
23
24     // @Test
25     fun learning() {
26         val nn = BasicNeuralNetwork(numberOfHiddenLayers, inputLayerSize = input.size, outputLayerSize = output.size)
27         repeat(1000) {
28             nn.train(input, output)
29         }
30         assertTrue("Error of simple memory is bigger than 0.1") {
31             (nn.run(input) - create(
32                 output,
33                 output.size,
34                 1
35             )).elementSum() <= 0.1
36         }
37         assertTrue("Input changed (from $inputTest to $input)") { input.contentEquals(inputTest) }
38         assertTrue("Output changed (from $outputTest to $output)") { output.contentEquals(outputTest) }
39     }
40
41     // @Test
42     fun xor() {
43         val dataset = setOf(
44             DoubleArray(2) { listOf(0.0, 0.0)[it] } to DoubleArray(1) { 0.0 },
45             DoubleArray(2) { listOf(1.0, 0.0)[it] } to DoubleArray(1) { 1.0 },
46             DoubleArray(2) { listOf(0.0, 1.0)[it] } to DoubleArray(1) { 1.0 },
47             DoubleArray(2) { listOf(1.0, 1.0)[it] } to DoubleArray(1) { 0.0 }

```

```

48     )
49     val nn = BasicNeuralNetwork(
50         numberOfHiddenLayers,
51         inputLayerSize = dataset.random().first.size,
52         outputLayerSize = dataset.random().second.size,
53         sizes = { 2 })
54     repeat(50000) {
55         val (input, output) = dataset.random()
56         nn.train(input, output)
57     }
58     dataset.forEach {
59         println(it.first.toList())
60         println(it.second.toList())
61         println(nn.run(it.first).toList())
62     }
63 }
64 }

```

Listing 10: src/jsTest/kotlin/sample/ConstantsJS.kt

```

1 package sample
2
3 import koma.internal.default.generated.matrix.DefaultDoubleMatrixFactory
4 import koma.matrix.Matrix
5 import koma.matrix.MatrixFactory
6
7 actual val defaultDoubleMatrixFactory: MatrixFactory<Matrix<Double>> = DefaultDoubleMatrixFactory()

```

Listing 11: src/jvmMain/kotlin/mnistDatabase/loadFileJVM.kt

```

1 package mnistDatabase
2
3 import java.io.*
4
5 actual fun loadFile(file: String) = File(file).readBytes()
6 actual fun saveFile(file: String, text: String) {
7     val f = File(file)
8     f.createNewFile()
9     val bw = BufferedWriter(Writer(f))
10    bw.append(text)
11    bw.close()
12 }
13
14 actual fun loadFileString(file: String): String = BufferedReader(FileReader(File(file))).readLine()

```

Listing 12: src/jvmTest/kotlin/sample/ConstantsJVM.kt

```

1 package sample
2
3 import koma.internal.default.generated.matrix.DefaultDoubleMatrixFactory
4 import koma.matrix.Matrix
5 import koma.matrix.MatrixFactory
6 import mnistDatabase.MnistTrainingData
7
8 actual val defaultDoubleMatrixFactory: MatrixFactory<Matrix<Double>> = DefaultDoubleMatrixFactory()
9
10 val mnistDigitTrainingDataset = MnistTrainingData("train-images", "train-labels", false)
11 val emnistDigitTrainingDataset = MnistTrainingData("emnist-digits-train-images", "emnist-digits-train-labels", true)

```

Listing 13: src/jvmTest/kotlin/sample/NeuralNetworkTestJVM.kt

```

1 package sample
2

```



```

3 import core.BasicNeuralNetwork
4 import core.ConvolutionalNeuralNetwork
5 import mnistDatabase.loadFileString
6 import mnistDatabase.saveFile
7 import mnistDatabase.train
8 import org.junit.Test
9
10 class NeuralNetworkTestJVM {
11     @Test
12     fun serialization() {
13         val nn = BasicNeuralNetwork(numberOfHiddenLayers, inputLayerSize = input.size, outputLayerSize = output.size)
14         val saved = nn.save()
15         println(nn.save())
16         val nn2 = BasicNeuralNetwork.load(saved)
17         println(nn2.save())
18         nn2.run(input)
19         nn2.train(input, output)
20     }
21
22     // @Test
23     fun mnist() {
24         val nn = BasicNeuralNetwork(
25             numberOfHiddenLayers,
26             inputLayerSize = imageWidth * imageHeight,
27             outputLayerSize = numberOfDigits,
28             sizes = { 100 })
29         repeat(10) {
30             nn.train(mnistDigitTrainingDataset)
31             nn.train(emnistDigitTrainingDataset)
32             nn.learningRate /= learningRateEpochDecrease
33
34             val data = mnistDigitTrainingDataset.iterator().next()
35             println(nn.run(data.first).toList())
36             println(data.second.toList())
37             saveFile("output.txt", nn.save())
38         }
39     }
40
41     // @Test
42     fun savedNN() {
43         var error = 0
44         repeat(100) {
45             val data = mnistDigitTrainingDataset.iterator().next()
46             val answer =
47                 BasicNeuralNetwork.load(loadFileString("output.txt")).run(data.first).toList()
48             if (answer.indexOf(answer.maxBy { it }) != data.second.indexOf(1.0)) {
49                 error++
50             }
51         }
52         println(error)
53     }
54
55     // @Test
56     fun mnistC() {
57         val nn = ConvolutionalNeuralNetwork(
58             ConvolutionalNeuralNetwork.edgeFilter,
59             BasicNeuralNetwork(
60                 1,
61                 inputLayerSize = (imageWidth - 2) * (imageHeight - 2) * 8,
62                 outputLayerSize = numberOfDigits,
63                 sizes = { 100 })
64         )
65         repeat(10) {
66             nn.train(mnistDigitTrainingDataset, imageWidth)
67             nn.train(emnistDigitTrainingDataset, imageWidth)
68             nn.learningRate /= learningRateEpochDecrease
69

```

```

70         saveFile("outputC.txt", nn.save())
71         savedNNC()
72     }
73 }
74
75 // @Test
76 fun savedNNC() {
77     var error = 0
78     repeat(100) {
79         val data = mnistDigitTrainingDataset.iterator().next()
80         val answer =
81             ConvolutionalNeuralNetwork.load(loadFileString("outputC.txt")).run(data.first, imageWidth).toList()
82         if (answer.indexOf(answer.maxBy { it }) != data.second.indexOf(1.0)) {
83             error++
84         }
85     }
86     println(error)
87 }
88
89 // @Test
90 fun print() {
91     fun Pair<DoubleArray, DoubleArray>.print() {
92         for (i in 0 until imageHeight) {
93             for (j in 0 until imageWidth) {
94                 print(
95                     if (first[j + i * imageWidth] < blackFrom) {
96                         "."
97                     } else {
98                         "#"
99                     }
100                 )
101             }
102             println()
103         }
104         println(second.indexOf(1.0))
105     }
106
107     val data = mnistDigitTrainingDataset.iterator().next()
108     data.print()
109 }
110 }

```