

Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

DIPLOMOVÁ PRÁCE

Plzeň, 2018

Martin Majer

Prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne 25. března 2018

.....

Poděkování

Tímto bych rád poděkoval vedoucímu diplomové práce, Ing. Luboši Šmídlovi, Ph.D., za cenné rady a připomínky.

Anotace

Tato práce se zabývá klasifikací fonémů pomocí různých architektur neuronových sítí. V první části práce je představena obecná teorie dopředných a rekurentních neuronových sítí a následně metoda CTC (connectionist temporal classification). Ve druhé části je pak vyhodnocena přesnost rozpoznání šesti navržených architektur nad čtyřmi parametrizacemi pro dvě datové sady o různé velikosti. Ukázalo se, že rekurentní neuronová síť využívající dvě obousměrné LSTM vrstvy a metodu CTC dosahuje velmi vysoké přesnosti, ale pouze díky využití informace z celé nahrávky. Proto byla testována i její varianta s omezenou délkou vstupní sekvence, která pro tuto úlohu rovněž ukázala velký potenciál a mohla by být využita pro rozpoznávání v reálném čase.

Klíčová slova: rozpoznání fonémů, dopředná neuronová síť, rekurentní neuronová síť, long short-term memory, gated recurrent unit, connectionist temporal classification

Abstract

This thesis focuses on the phoneme recognition using various architectures of neural networks. The first part introduces theory of feedforward and recurrent neural networks followed by the introduction of the method CTC (connectionist temporal classification). The second part presents comparison of accuracy of recognition between six architectures on four parametrizations generated from two datasets of various size. It was shown that the recurrent neural network using two bidirectional LSTM layers combined with CTC achieved high accuracy although only due to using information from the whole recording. Therefore its alternative version which used a limited length of the input sequence was tested and also showed large potential and could be possibly used for recognition in real-time.

Keywords: phoneme recognition, feedforward neural network, recurrent neural network, long short-term memory, gated recurrent unit, connectionist temporal classification

Obsah

1	Úvod	1
2	Neuronové sítě	2
2.1	Rozpoznávání řeči pomocí neuronových sítí	2
2.2	Cenová funkce	3
2.3	Dopředné neuronové sítě	3
2.3.1	Architektura	4
2.3.2	Výstupní jednotky	5
2.3.3	Skryté jednotky	6
2.3.4	Algoritmus zpětného šíření	7
2.4	Rekurentní neuronové sítě	8
2.4.1	Algoritmus zpětného šíření časem	10
2.4.2	Obousměrné rekurentní neuronové sítě	12
2.4.3	LSTM	13
2.4.4	GRU	14
3	Optimalizace parametrů neuronových sítí	16
3.1	Stochastický gradientní sestup	16
3.2	Další modifikace gradientního sestupu	17
3.2.1	Moment	18
3.2.2	Nesterovův moment	18
3.3	ADAM	18
4	Kapacita modelu	20
4.1	Regularizace	21
4.1.1	Penalizace velikosti parametrů	21
4.1.2	Zanesení šumu	21
4.1.3	Předčasné ukončení	22
4.1.4	Dropout	22
5	CTC	24
5.1	Formální definice úlohy	24
5.2	Reprezentace výstupu	25

5.3	Dekódování	26
5.4	Dopředný a zpětný algoritmus	26
6	Klasifikace fonémů	29
6.1	Předzpracování dat	29
6.2	Architektury a trénování sítí	30
7	Experimenty a vyhodnocení	33
7.1	Experimenty nad datovou sadou ŠkodaAuto	34
7.2	Experimenty nad datovou sadou SpeechDat-E	41
7.3	Porovnání výsledku nad ŠkodaAuto a SpeechDat-E	44
8	Závěr	46
	Seznam obrázků	48
	Seznam tabulek	48
	Reference	50

1 Úvod

Neuronové sítě byly vyvinuty již v polovině minulého století, ale kvůli nedostatečné výpočetní kapacitě nemohly být plně využity pro řešení reálných problémů. Až v posledních letech, kdy došlo k významnému vývoji v oblasti hardwaru jak pro počítače, tak i pro mobilní a vestavěná zařízení, začaly být plně využívány, a to zejména pro komplexní úlohy v oblasti zpracování obrazu a řeči, kde stabilně překonávají ostatní algoritmy strojového učení. Vývoj výkonných grafických karet a optimalizovaných knihoven pak umožnil rychlé trénování těchto modelů na velkém množství dat, a díky výkonným čipům lze provádět predikci v reálném čase i na mobilních zařízeních.

Tato práce se věnuje využití neuronových sítí pro zpracování řeči, a to zejména úloze klasifikace fonémů s využitím základních metod zpracování akustického signálu představených v [1]. Ve zpracování řeči jsou běžně využívány jak sítě dopředné, tak i rekurentní, které byly vytvořeny pro klasifikaci časových řad či sekvencí a jsou schopny využívat kontextu. Pro trénování neuronových sítí je potřeba velké množství dat, aby byla zajištěna robustnost a schopnost generalizace s tím, že v úloze klasifikace fonémů jsou tato data ve formě zvukových nahrávek a odpovídajících přepisů neboli transkripcí.

Cílem práce je porovnat různé architektury neuronových sítí, které by mohly být využity pro přepis mluvené řeči do textové podoby (speech-to-text) v reálném čase. Za tímto účelem byly voleny především jednodušší sítě s nízkým počtem parametrů.

V první části práce je uvedena obecná teorie neuronových sítí a optimalizačních algoritmů. Druhá část se pak věnuje metodě trénování rekurentních neuronových sítí, která nevyžaduje předsegmentovaná trénovací data (více o segmentaci v [1]). Nakonec jsou porovnány různé architektury jak dopředných, tak i rekurentních neuronových sítí, na několika datových sadách, které vznikly využitím různých typů příznaků.

2 Neuronové sítě

Neuronová síť je algoritmus, který je schopen aproximovat i silně nelineární funkce a zároveň je schopen dosáhnout vysoké míry statistické generalizace. Tento parametrický model bývá zpravidla složen z několika vrstev reprezentovaných vektory, jejichž dimenze udává šířku modelu. Prvky těchto vektorů, jednotky či perceptrony, pracují paralelně a reprezentují funkce zobrazující vstupní vektor na skalár. Cílem je natrénovat parametry tohoto modelu tak, aby dokázal splnit zadanou úlohu s minimální chybou. Jedná se tedy o optimalizační úlohu.

Základním rozdílem mezi běžnou optimalizací a optimalizací v neuronových sítích je, že optimalizace v neuronových sítích a ve strojovém učení obecně probíhá nepřímou. To znamená, že ačkoliv optimalizujeme zvolenou metriku P , která v jistém smyslu kvantifikuje výkon algoritmu na dané úloze, minimalizujeme jinou cenovou funkci $J(\theta)$ za účelem minimalizace metriky P . V běžném optimalizačním problému bychom minimalizovali přímo cenovou funkci $J(\theta)$ za účelem její minimalizace [2–4].

2.1 Rozpoznávání řeči pomocí neuronových sítí

Ve zpracování řeči jsou neuronové sítě využívány téměř ve všech jejích oblastech, a to zejména v rozpoznávání řeči a její syntéze, kde postupně nahradily klasické statistické modely. Rozpoznávání řeči je úloha, jejímž cílem je získat přepis mluvené řeči v textové podobě. V závislosti na aplikaci systému rozpoznávání řeči a dalších omezujících podmínkách (doba odezvy, výpočetní náročnost) se pak nabízí dva přístupy rozpoznávání řeči - rozpoznávání po jednotlivých fonémech či po celých slovech.

K rozpoznávání řeči se v poslední letech začala využívat rekurentní neuronová síť, která je schopná využívat kontextu v řeči, či hluboká dopředná síť. Vhodné architektury sítí jako Seq2Seq [5] či rekurentní neuronová síť se ztrátovou funkcí CTC pak umožňují přímý přepis mluvené řeči do textové podoby bez potřeby dekodování výstupu modelu, jako tomu bylo například u skrytých Markovových modelů. V posledních letech se ovšem také začaly k rozpoznávání řeči využívat konvoluční neuronové sítě, a to buď s využitím 1D konvoluce nebo 2D konvoluce, kdy se síť učí ze spektrogramů, a jsou voleny podobné architektury jako při zpracování obrazu [6].

Mezi další zajímavá využití neuronových sítí ve zpracování řeči patří například syntéza

řeči, překlad mluvené řeči, verifikace řečníka či detekce klíčových frází. Neuronové sítě se také běžně využívají ke kompresi příznakového vektoru (může být i různé délky) do příznakového vektoru o nižší pevně dané délce. K tomu se využívá tzv. bottleneck vrstva, která má výrazně nižší počet jednotek než její sousední vrstvy a po natrénování sítě slouží jako výstupní vrstva. Tuto metodu komprese příznakového vektoru lze například využít pro rychlou detekci klíčových frází v kombinaci s algoritmem DTW.

Tato práce se věnuje rozpoznávání řeči po fonémech s využitím jak dopředných, tak rekurentních neuronových sítí s dekódováním i přímým přepisem.

2.2 Cenová funkce

Jedním ze zásadních aspektů při návrhu neuronových sítí je výběr cenové funkce. Výběr cenové funkce závisí na dané úloze a požadovaném výstupu. Cenovou funkci definujeme stejným způsobem jako u ostatních parametrických modelů, které generují hustotu pravděpodobnosti $p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$, kde \mathbf{y} je výstupní třída, \mathbf{x} je vstup modelu, a které při trénování využívají principu maximální věrohodnosti. Většinou cenovou funkci definujeme jako vzájemnou entropii mezi trénovacími daty a hustotou pravděpodobnosti predikcí modelu

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim \hat{p}_{data}} \log p_{model}(\mathbf{y} \mid \mathbf{x}). \quad (2.1)$$

Předpis cenové funkce se liší model od modelu v závislosti na tvaru $\log p_{model}$ a kromě definice ceny také může obsahovat regularizační prvky. Výhodou využití principu maximální věrohodnosti je, že specifikací modelu $p(\mathbf{y} \mid \mathbf{x})$ zároveň definujeme cenovou funkci $\log p(\mathbf{y} \mid \mathbf{x})$ [2, 4].

2.3 Dopředné neuronové sítě

Dopředné neuronové sítě, občas také nazývané vícevrstvé perceptrony, jsou základním typem neuronových sítí. Jejich cílem je aproximovat určitou funkci f^* , např. klasifikátor $\mathbf{y} = f^*(\mathbf{x})$ zobrazuje vstup \mathbf{x} na výstupní třídu \mathbf{y} . Dopředná neuronová síť definuje zobrazení $\mathbf{y} = f^*(\mathbf{x}; \boldsymbol{\theta})$ a učí se optimální hodnoty parametrů $\boldsymbol{\theta}$, při kterých by měla být dosažena nejlepší aproximace funkce.

Jak název napovídá, tok informací směřuje od vstupu \mathbf{x} přes několik po sobě jdoucích výpočetních vrstev definujících f až k výstupu \mathbf{y} . Jedná se tedy o složení několika různých

funkcí do řetězové struktury. Takto definovaný model lze také popsat jako orientovaný acyklický graf, který určuje závislosti mezi funkcemi. Mějme například síť složenou ze tří funkcí $f^{(1)}$, $f^{(2)}$ a $f^{(3)}$ spojených do řady $f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. V tomto případě nazýváme $f^{(1)}$ první skrytou vrstvou, $f^{(2)}$ druhou skrytou vrstvou a $f^{(3)}$ výstupní vrstvou. Celková délka tohoto řetězce pak udává hloubku modelu [2].

Cílem trénování sítě je, aby se naučila odpovídající zobrazení mezi $f(\mathbf{x})$ a $f^*(\mathbf{x})$. Učení probíhá na základě trénovacích dat, která poskytují zašuměná aproximovaná pozorování $f^*(\mathbf{x})$ vyhodnocena v různých bodech. Ke každému pozorování \mathbf{x} je k dispozici také skutečná hodnota $\mathbf{y} \approx f^*(\mathbf{x})$. Výstupní vrstva sítě se pro každou trénovací hodnotu \mathbf{x} snaží vyprodukovat hodnotu blízkou \mathbf{y} . Chování skrytých vrstev není přímo dáno trénovacími daty a síť se je musí naučit využívat k získání požadovaného výsledku, tj. k aproximaci $f^*(\mathbf{x})$ [2, 4].

2.3.1 Architektura

Klíčovým prvkem při návrhu neuronových sítí je jejich architektura. Architekturu sítě či modelu rozumíme celkovou strukturu sítě - kolik má mít jednotek a jak mají být tyto jednotky mezi sebou propojeny.

Jak již bylo uvedeno, neuronové sítě jsou tvořeny skupinami jednotek, které jsou uspořádány do jednotlivých vrstev. Většina architektur tyto vrstvy uspořádává do řetězové struktury, kde každá vrstva je funkcí vrstvy, která ji předchází. V této struktuře je pak první vrstva definována jako

$$\mathbf{h}^{(1)} = g^{(1)} \left(\mathbf{W}^{(1)\top} \mathbf{x} + \mathbf{b}^{(1)} \right), \quad (2.2)$$

druhá vrstva jako

$$\mathbf{h}^{(2)} = g^{(2)} \left(\mathbf{W}^{(2)\top} \mathbf{h}^{(1)} + \mathbf{b}^{(2)} \right), \quad (2.3)$$

a tak dále, kde $g^{(i)}$ je aktivační funkce vrstvy i , $\mathbf{W}^{(i)\top}$ je váhová matice vrstvy i a $\mathbf{b}^{(i)}$ je prahový vektor vrstvy i .

Pro takto definovanou řetězovou architekturu je pak hlavním problémem určení hloubky sítě a šířky každé vrstvy. Vhodnou architekturu pro danou úlohu je třeba nalézt pomocí experimentů založených na sledování chyby na validační datové sadě a apriorní znalosti o úloze a datech [2–4, 7].

2.3.2 Výstupní jednotky

Reprezentace výstupu je úzce spojena s danou úlohou a tím i s výběrem cenové funkce. Předpokládejme, že dopředná neuronová síť poskytuje výstupní vrstvě soubor skrytých příznaků $\mathbf{h} = f(\mathbf{x}; \boldsymbol{\theta})$, tj. výstup poslední skryté vrstvy. Cílem výstupní vrstvy je pak provést určitou transformaci těchto příznaků, aby síť plnila úlohu, pro kterou byla navržena. Tato transformace je provedena použitím aktivační funkce $g(\mathbf{h})$. Ačkoliv existuje mnoho aktivačních funkcí, které lze ve výstupní vrstvě využít, zde se zaměříme pouze na aktivační funkci softmax, která je využívána v experimentech provedených v rámci této práce.

Aktivační funkce softmax je využívána pro úlohy, kde je třeba reprezentovat výstup jako hustotu pravděpodobnosti diskrétní proměnné s n možnými hodnotami (výstupními třídami). K odvození aktivační funkce softmax využijeme znalosti o úloze binární klasifikace, při které predikujeme hodnotu

$$\hat{\mathbf{y}} = P(\mathbf{y} = 1 \mid \mathbf{x}), \quad (2.4)$$

kde $\hat{\mathbf{y}} \in [0, 1]$. Aby byla zajištěna numerická stabilita při optimalizaci, budeme raději predikovat hodnotu

$$\mathbf{z} = \log \tilde{P}(\mathbf{y} = 1 \mid \mathbf{x}). \quad (2.5)$$

Aplikací exponenciální funkce a následnou normalizací bychom pak dostali Bernoulliho rozložení pravděpodobnosti řízeného sigmoidální funkcí.

Pro generalizaci tohoto postupu pro diskrétní proměnnou s n hodnotami je potřeba získat vektor $\hat{\mathbf{y}}$, kde $\hat{\mathbf{y}}_i = P(\mathbf{y} = i \mid \mathbf{x})$. Pro každý prvek $\hat{\mathbf{y}}_i$ musí platit $\hat{\mathbf{y}}_i \in [0, 1]$ a zároveň $\sum_i \hat{\mathbf{y}}_i = 1$, aby bylo možné tento vektor interpretovat za hustotu pravděpodobnosti. Nejprve je potřeba provést predikci vrstvou s lineární aktivační funkcí, která predikuje ne-normalizované logaritmované pravděpodobnosti

$$\mathbf{z} = \mathbf{W}^\top \mathbf{h} + \mathbf{b}, \quad (2.6)$$

kde $\mathbf{z}_i = \log \tilde{P}(\mathbf{y} = i \mid \mathbf{x})$. Softmax funkce pak aplikuje exponenciální funkci a normalizuje

\mathbf{z} , čímž získáme požadované $\hat{\mathbf{y}}$.

$$\text{softmax}(\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i)}{\sum_j \exp(\mathbf{z}_j)} \quad (2.7)$$

Natrénováním parametrů modelu pak bude výstupní vrstva s aktivační funkcí softmax predikovat podíly počtů všech pozorovaných výsledků v trénovací datové sadě [2].

$$\text{softmax}(\mathbf{z}(\mathbf{x}; \boldsymbol{\theta}))_i \approx \frac{\sum_{j=1}^m 1_{\mathbf{y}^{(j)}=i, \mathbf{x}^{(j)}=\mathbf{x}}}{\sum_{j=1}^m 1_{\mathbf{x}^{(j)}=\mathbf{x}}} \quad (2.8)$$

2.3.3 Skryté jednotky

Jak název napovídá, skryté jednotky jsou jednotky skryté vrstvy, jejichž vstupem je vektor \mathbf{x} , který je transformován na $\mathbf{z} = \mathbf{W}^\top \mathbf{x} + \mathbf{b}$. Na takto transformovaný vstup je pak po prvcích aplikována nelineární aktivační funkce $g(\mathbf{z})$. Volba aktivačních funkcí skrytých vrstev vyžaduje mnoho experimentů a vyhodnocení přesnosti modelu na validační datové sadě. V dnešní době se pro dopředné neuronové sítě většinou volí jednotky s aktivační funkcí ReLU (z anglického "rectified linear unit") či její modifikace, pro sítě rekurentní jsou pak voleny funkce hyperbolický tangens a hard sigmoid.

- **ReLU** - tyto jednotky využívají aktivační funkci $g(\mathbf{z}) = \max\{0, \mathbf{z}\}$. Jedná se o lineární jednotky s prahem v bodě nula - levá polovina jejich definičního oboru je rovna nule. To zaručuje rychlý výpočet a vysokou hodnotu gradientu, kdykoliv je jednotka aktivní. Zásadním nedostatkem je, že tyto jednotky se pomocí gradientních metod nemohou učit z pozorování, které mají aktivační hodnotu rovnou nule. Existuje proto několik modifikací, které zajistí, že jednotky budou mít gradient všude (např. Leaky ReLU, PReLU, Maxout).
- **sigmoid** a **tanh** - tyto jednotky využívají logistickou funkci (sigmoid) $g(\mathbf{z}) = \sigma(\mathbf{z})$, resp. hyperbolický tangens $\tanh(\mathbf{z}) = 2\sigma(2\mathbf{z}) - 1$. Hlavním nedostatkem sigmoidální funkce je její citlivost a náchylnost k saturaci, kdy může dojít k tzv. explozi či vymizení gradientu a síť nebude schopná se učit.
- **hard sigmoid** - tyto jednotky využívají aktivační funkci $g(\mathbf{z}) = \max(0, \min(1, \frac{\mathbf{z}+1}{2}))$. Jedná se o lineární aproximaci funkce sigmoid a díky své výpočetní nenáročnosti jsou využívány v sítích typu LSTM [2,3].

2.3.4 Algoritmus zpětného šíření

Dopředná neuronová síť přijme na vstupu počáteční informaci o pozorování \mathbf{x} , kterou poté šíří skrz skryté jednotky ve všech skrytých vrstvách až k výstupní vrstvě, která vygeneruje odhad $\hat{\mathbf{y}}$. Informační tok tedy proudí skrz síť směrem dopředu a tomuto procesu se říká dopředné šíření. V trénovací fázi probíhá dopředné šíření tak dlouho, dokud není vygenerována skalární cena $J(\boldsymbol{\theta})$. Algoritmus zpětného šíření (anglicky backpropagation) pak umožní zpětný tok informace od ceny skrz síť za účelem výpočtu gradientu.

Analytický výpočet gradientu je poměrně přímočarý, nicméně numericky může být velice náročný. Algoritmus zpětného šíření provádí výpočet gradientu $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ pomocí jednoduché a výpočetně nenáročné procedury. Takto vypočtený gradient je pak využit při učení pomocí gradientních metod [2, 3].

Algoritmus zpětného šíření využívá k výpočtu gradientu řetězové pravidlo. Řetězové pravidlo se běžně využívá k výpočtu derivací složených funkcí, které jsou složeny z funkcí, jejichž derivace jsou známé. Mějme reálné číslo x a funkce $f : \mathbb{R} \mapsto \mathbb{R}$ a $g : \mathbb{R} \mapsto \mathbb{R}$. Dále předpokládejme, že $y = g(x)$ a $z = f(g(x)) = f(y)$. Řetězové pravidlo je pak dáno jako

$$\frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx}. \quad (2.9)$$

Toto pravidlo lze snadno rozšířit ze skalárního případu. Předpokládejme, že $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $g : \mathbb{R}^m \mapsto \mathbb{R}^n$ a $f : \mathbb{R}^n \mapsto \mathbb{R}$. Pokud $\mathbf{y} = g(\mathbf{x})$ a $z = f(\mathbf{y})$, pak

$$\frac{\partial z}{\partial \mathbf{x}_i} = \sum_j \frac{\partial z}{\partial \mathbf{y}_j} \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_i} \quad (2.10)$$

Zapišeme-li rovnici (2.10) vektorově, získáme tvar

$$\nabla_{\mathbf{x}} z = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^{\top} \nabla_{\mathbf{y}} z, \quad (2.11)$$

kde $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{n \times m}$ je Jakobián \mathbf{x} . Toto pravidlo lze dále snadno rozšířit i na tensory, jenž jsou v neuronových sítích často využívány (podrobněji v [2, 8]).

Předpokládejme síť o hloubce l , kde každé vrstvě náleží váhová matice $\mathbf{W}^{(i)}$, $i \in \{1, \dots, l\}$ a prahový vektor $\mathbf{b}^{(i)}$. Ztrátová funkce $L(\mathbf{y}, \hat{\mathbf{y}})$ závisí na skutečné hodnotě \mathbf{y} a na výstupu sítě $\hat{\mathbf{y}}$, který síť vygeneruje pro vstup \mathbf{x} . Celková cena J pro zjednodušení

odpovídá přímo ztrátové funkci L a neobsahuje žádnou regularizační složku.

```

 $\mathbf{h}^{(0)} = \mathbf{x}$ 
for  $k = 1, \dots, l$  do
     $\mathbf{a}^{(k)} = \mathbf{W}^{(k)} \mathbf{h}^{(k-1)} + \mathbf{b}^{(k)}$ 
     $\mathbf{h}^{(k)} = f(\mathbf{a}^{(k)})$ 
end
 $\hat{\mathbf{y}} = \mathbf{h}^{(l)}$ 
 $J = L(\mathbf{y}, \hat{\mathbf{y}})$ 

```

Algoritmus 1: Dopředné šíření.

Při zpětném průchodu jsou počítány gradienty aktivací $\mathbf{a}^{(k)}$ pro každou vrstvu k počínaje výstupní vrstvou až k první skryté vrstvě. Tyto gradienty indikují, jak by se měly změnit výstupy jednotlivých vrstev, aby došlo ke snížení chyby. Spočtené gradienty vah a prahů mohou být rovnou využity pro změnu parametrů pomocí optimalizačního algoritmu [2].

```

výpočet gradientu výstupní vrstvy (po dopředném průchodu sítě)
 $g \leftarrow \nabla_{\hat{\mathbf{y}}} J = \nabla_{\hat{\mathbf{y}}} L(\mathbf{y}, \hat{\mathbf{y}})$ 
for  $k = l, l-1, \dots, 1$  do
    převedení gradientu do tvaru před aplikací nelineární aktivační funkce
     $g \leftarrow \nabla_{\mathbf{a}^{(k)}} J = g \odot f'(\mathbf{a}^{(k)})$ 
    výpočet gradientů prahových vektorů a váhových matic
     $\nabla_{\mathbf{b}^{(k)}} J = g$ 
     $\nabla_{\mathbf{W}^{(k)}} J = g \mathbf{h}^{(k-1)\top}$ 
    šíření gradientu do nižší vrstvy
     $g \leftarrow \nabla_{\mathbf{h}^{(k-1)}} J = \mathbf{W}^{(k)\top} g$ 
end

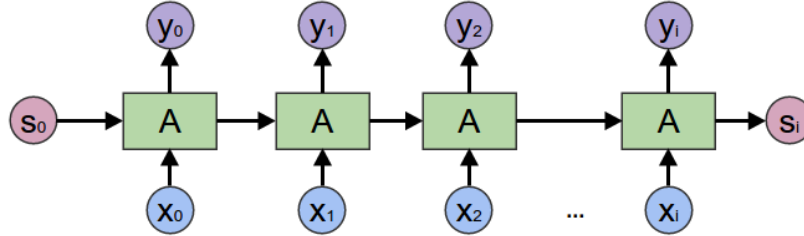
```

Algoritmus 2: Zpětné šíření.

2.4 Rekurentní neuronové sítě

Rekurentní neuronové sítě (dále jen RNN z anglického "recurrent neural networks") jsou typem neuronových sítí, které umí zpracovávat sekvenční data, tj. sekvenci hodnot $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}$, a předávat si informace mezi jednotlivými časovými kroky. Základním rozdílem oproti dopředným neuronovým sítím je zavedení zpětné smyčky a využití sdílení parametrů. RNN tedy netrénuje pro každý časový krok samostatný soubor parametrů, ale

tyto parametry jsou sdíleny přes všechny prvky sekvence. Díky tomu jsou RNN schopny generalizovat na různé délky sekvencí a na různé pozice důležitých informací v čase (např. dvě téměř stejné věty, kde jedna obsahuje datum na začátku věty a druhá na konci).



Obrázek 1: Schéma rekurentní neuronové sítě. Převzato z [9].

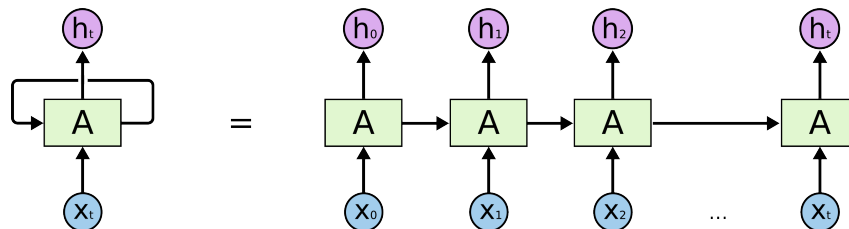
Sdílení parametrů je zajištěno "rozvinutím" rekurentního výpočtu do grafu s opakující se strukturou. Prvek sekvence v každém časovém kroku je tedy zpracováván stejnou sítí. Pro jednoduchost předpokládejme, že RNN pracuje se sekvencí obsahující vektory $\mathbf{x}^{(t)}$, $t = 1, \dots, \tau$ a model je ve tvaru

$$\mathbf{s}^{(t)} = f(\mathbf{s}^{(t-1)}; \boldsymbol{\theta}), \quad (2.12)$$

kde \mathbf{s} je stav systému. Tento rekursivní vztah je možné pro konečný počet kroků τ rozvinout τ -krát. Například pro model daný vztahem (2.12) a $\tau = 3$ získáme

$$\mathbf{s}^{(3)} = f(\mathbf{s}^{(2)}; \boldsymbol{\theta}) = f(f(\mathbf{s}^{(1)}; \boldsymbol{\theta}); \boldsymbol{\theta}). \quad (2.13)$$

Tímto rozvinutím zajistíme, že vztah neobsahuje žádnou rekurenci a je možné ho reprezentovat jako acyklický orientovaný graf [2, 10].



Obrázek 2: Ilustrace rozvinutí rekurentní neuronové sítě. Převzato z [10].

Základní rovnici RNN získáme přidáním závislosti na externím vstupu $\mathbf{x}^{(t)}$

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}), \quad (2.14)$$

kde \mathbf{h} je stav skrytých jednotek. Při trénování sítě na danou úlohu se pak síť učí využívat $\mathbf{h}^{(t)}$ jako zobrazení relevantních informací z minulých časových kroků od vstupu až po t . Toto zobrazení je ztrátové, jelikož se jedná o zobrazení sekvence o proměnné délce $(\mathbf{x}^{(t)}, \mathbf{x}^{(t-1)}, \dots, \mathbf{x}^{(2)}, \mathbf{x}^{(1)})$ na vektor pevně dané délky $\mathbf{h}^{(t)}$.

Definujme si rovnice pro dopředné šíření základní RNN s aktivační funkcí hyperbolický tangens ve skryté vrstvě. Síť je určena ke klasifikaci n tříd, kde výstup \mathbf{o} reprezentuje nenormalizované pravděpodobnosti jednotlivých tříd výstupní diskrétní proměnné a aplikací operace softmax je pak získán výsledný vektor $\hat{\mathbf{y}}$ normalizovaných pravděpodobností nad výstupem. Dopředné šíření vyžaduje specifikaci počátečního stavu $\mathbf{h}^{(0)}$, poté jsou pro každý časový krok $t = 1, \dots, \tau$ vypočteny následující rovnice

$$\mathbf{a}^{(t)} = \mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)}, \quad (2.15)$$

$$\mathbf{h}^{(t)} = \tanh(\mathbf{a}^{(t)}), \quad (2.16)$$

$$\mathbf{o}^{(t)} = \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)}, \quad (2.17)$$

$$\hat{\mathbf{y}}^{(t)} = \text{softmax}(\mathbf{o}^{(t)}), \quad (2.18)$$

kde parametry \mathbf{b} a \mathbf{c} jsou prahové vektory a váhové matice \mathbf{U} , \mathbf{V} a \mathbf{W} popořadě odpovídají skrytým spojením mezi vstupem a skrytou vrstvou, mezi skrytou vrstvou a výstupem a mezi skrytou vrstvou a předcházející skrytou vrstvou [2].

2.4.1 Algoritmus zpětného šíření časem

Pro výpočet gradientu RNN se využívá algoritmus zpětného šíření časem. Jedná se o generalizovaný algoritmus zpětného šíření nad rozvinutým výpočetním grafem a jeho časová náročnost je $O(\tau)$. Časovou náročnost tohoto algoritmu není možné snížit paralelizací, jelikož dopředných průchod je sekvenční a hodnoty v aktuálním časovém kroku jsou závislé na předchozích hodnotách.

Předpokládejme reprezentaci rozvinuté RNN ve tvaru acyklického orientovaného grafu, jehož uzly obsahují parametry \mathbf{U} , \mathbf{V} , \mathbf{W} , \mathbf{b} , \mathbf{c} , a sekvenci uzlů indexovaných časem t pro $\mathbf{x}^{(t)}$, $\mathbf{h}^{(t)}$, $\mathbf{o}^{(t)}$ a $L^{(t)}$. Pro každý uzel grafu N pak potřebujeme rekurzivně spočítat gradient $\nabla_N L$ v závislosti na uzlech grafu, které tento uzel následují. Začneme rekurzi v uzlu, který

bezprostředně předchází výslednou ztrátu L

$$L(\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(\tau)}\}, \{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(\tau)}\}) = - \sum_t \log_{p_{model}} \left(\mathbf{y}^{(t)} \mid \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t)}\} \right), \quad (2.19)$$

$$\frac{\partial L}{\partial L^{(t)}} = 1. \quad (2.20)$$

Pro výpočet gradient $\nabla_{\mathbf{o}^{(t)}} L$ nad všemi výstupy v čase t pro všechna i, t je opět využito řetězové pravidlo

$$(\nabla_{\mathbf{o}^{(t)}} L)_i = \frac{\partial L}{\partial \mathbf{o}_i^{(t)}} = \frac{\partial L}{\partial L^{(t)}} \frac{\partial L^{(t)}}{\partial \mathbf{o}_i^{(t)}} = \hat{\mathbf{y}}_i^{(t)} - \mathbf{1}_{i, \mathbf{y}^{(t)}}. \quad (2.21)$$

Výpočet gradientu dále pokračuje přes další uzly počínaje koncem sekvence. V konečném časovém kroku τ má $\mathbf{h}^{(\tau)}$ za následníka pouze $\mathbf{o}^{(\tau)}$ a gradient je

$$\nabla_{\mathbf{h}^{(\tau)}} L = \mathbf{V}^\top \nabla_{\mathbf{o}^{(\tau)}} L. \quad (2.22)$$

Gradient se dál šíří od $t = \tau - 1$ až k $t = 1$ s tím, že $\mathbf{h}^{(t)}$ (pro $t < \tau$) má jako své předchůdce $\mathbf{o}^{(t)}$ a $\mathbf{h}^{(t+1)}$. Gradient skryté vrstvy je pak

$$\begin{aligned} \nabla_{\mathbf{h}^{(t)}} L &= \left(\frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{h}^{(t)}} \right)^\top (\nabla_{\mathbf{h}^{(t+1)}} L) + \left(\frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{h}^{(t)}} \right)^\top (\nabla_{\mathbf{o}^{(t)}} L) \\ &= \mathbf{W}^\top (\nabla_{\mathbf{h}^{(t+1)}} L) \text{diag} \left(1 - (\mathbf{h}^{(t+1)})^2 \right) + \mathbf{V}^\top (\nabla_{\mathbf{o}^{(t)}} L), \end{aligned} \quad (2.23)$$

kde $\text{diag} \left(1 - (\mathbf{h}^{(t+1)})^2 \right)$ je Jakobián hyperbolického tangensu pro skryté jednotky i v čase $t + 1$. Na základě gradientů skrytých stavů lze dopočítat gradienty parametrů. Vzhledem k tomu, že parametry jsou sdíleny mezi jednotlivými časovými kroky, zavedeme nové proměnné $\mathbf{W}^{(t)}$, resp. $\mathbf{U}^{(t)}$, které jsou kopiemi váhových matic \mathbf{W} , resp. \mathbf{U} a značí, jakou

mírou tyto váhy přispívají ke gradientu v čase t [2].

$$\nabla_{\mathbf{c}} L = \sum_t \left(\frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{c}} \right)^\top \nabla_{\mathbf{o}^{(t)}} L = \sum_t \nabla_{\mathbf{o}^{(t)}} L \quad (2.24)$$

$$\nabla_{\mathbf{b}} L = \sum_t \left(\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{b}^{(t)}} \right)^\top \nabla_{\mathbf{h}^{(t)}} L = \sum_t \text{diag} \left(1 - (\mathbf{h}^{(t)})^2 \right) \nabla_{\mathbf{h}^{(t)}} L \quad (2.25)$$

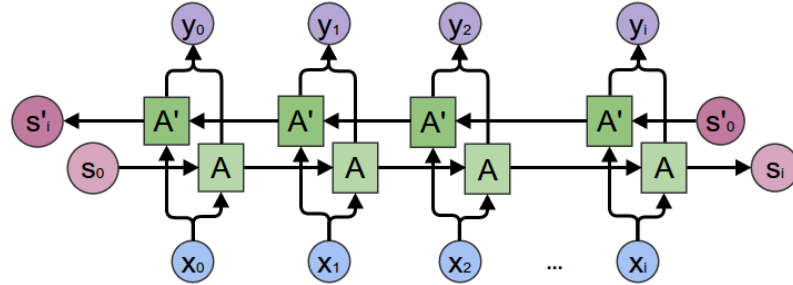
$$\nabla_{\mathbf{v}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial \mathbf{o}_i^{(t)}} \right) \nabla_{\mathbf{v}} \mathbf{o}_i^{(t)} = \sum_t (\nabla_{\mathbf{o}^{(t)}} L) \mathbf{h}^{(t)\top} \quad (2.26)$$

$$\nabla_{\mathbf{w}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial \mathbf{h}_i^{(t)}} \right) \nabla_{\mathbf{w}^{(t)}} \mathbf{h}_i^{(t)} = \sum_t \text{diag} \left(1 - (\mathbf{h}^{(t)})^2 \right) (\nabla_{\mathbf{h}^{(t)}} L) \mathbf{h}^{(t-1)\top} \quad (2.27)$$

$$\nabla_{\mathbf{u}} L = \sum_t \sum_i \left(\frac{\partial L}{\partial \mathbf{h}_i^{(t)}} \right) \nabla_{\mathbf{u}^{(t)}} \mathbf{h}_i^{(t)} = \sum_t \text{diag} \left(1 - (\mathbf{h}^{(t)})^2 \right) (\nabla_{\mathbf{h}^{(t)}} L) \mathbf{x}^{(t)\top} \quad (2.28)$$

2.4.2 Obousměrné rekurentní neuronové sítě

Rekurentní neuronové sítě zachycují do stavu v čase t pouze informaci z minulosti, tj. informace ze vstupů $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(t-1)}$ a aktuálního vstupu $\mathbf{x}^{(t)}$. V mnoha úlohách ovšem může správná predikce $\mathbf{y}^{(t)}$ záviset i na budoucích vstupech či na celé sekvenci. Tento problém řeší obousměrná rekurentní neuronová síť (dále BRNN z anglického "bidirectional recurrent neural network").



Obrázek 3: Schéma obousměrné rekurentní neuronové sítě. Převzato z [9].

BRNN kombinuje dvě RNN, kde jedna síť se pohybuje v čase kupředu od začátku sekvence, zatímco druhá se pohybuje v čase zpět od konce sekvence. Tyto dvě sítě mezi sebou sdílí skryté stavy $\mathbf{h}^{(t)}$ (síť dopředná) a $\mathbf{g}^{(t)}$ (síť zpětná), což zajistí, že výstupní jednotky $\mathbf{o}^{(t)}$ jsou závislé jak na minulosti, tak na budoucnosti. Výstupy skrytých stavů $\mathbf{h}^{(t)}$ a $\mathbf{g}^{(t)}$ nejsou mezi sebou nijak propojeny. Pro výpočet gradientů lze opět využít algoritmus zpětného šíření časem dle postupu, který je shrnut v algoritmu č. 4 [2, 11].

```

for  $t = 1, \dots, \tau$  do
    dopředné šíření dopřednou sítí od  $t = 1$  do  $t = \tau$ 
    dopředné šíření zpětnou sítí od  $t = \tau$  do  $t = 1$ 
    dopředné šíření nad výstupními jednotkami obou sítí
end
for  $t = 1, \dots, \tau$  do
    zpětné šíření nad výstupními jednotkami obou sítí
    zpětné šíření dopřednou sítí od  $t = \tau$  do  $t = 1$ 
    zpětné šíření zpětnou sítí od  $t = 1$  do  $t = \tau$ 
end
změna parametrů

```

Algoritmus 3: Postup trénování obousměrné rekurentní neuronové sítě.

2.4.3 LSTM

Rekurentní sítě jsou velice náchylné na problém vymizení či saturace gradientu. K těmto problémům dochází zejména při modelování dlouhodobých závislostí, kdy jsou při výpočtu váhy násobeny několikrát samy sebou a v závislosti na jejich řádu mohou nabývat velmi malých či velmi velkých hodnot. To má pak za následek trvalou deaktivaci některých neuronů či neschopnost sítě se učit. Tento problém řeší síť LSTM (z anglického "long short-term memory").

LSTM sítě, na rozdíl od běžných RNN, nejsou tvořeny jednotkami, ale buňkami, které kromě vnější rekurence obsahují i rekurenci vnitřní (smyčky). Každá buňka má stejný vstup a výstup jako obyčejná rekurentní jednotka, ale má více parametrů díky systému jednotek opatřených bránou, které řídí tok informací stavem buňky. Jedná se o vstupní bránu, zapomínací bránu a výstupní bránu, které určují, jaké vstupní informace mají být vpuštěny do stavu buňky, jaké informace ze stavu buňky odstranit a jaké informace ze stavu buňky mají být vypuštěny na její výstup.

Nejdůležitější komponentou LSTM je stavová jednotka $\mathbf{s}_i^{(t)}$ (pro buňku i v čase t), která je opatřena lineární smyčkou. Váha této smyčky je řízena jednotkou se zapomínací bránou $\mathbf{f}_i^{(t)}$, která díky aktivační funkci sigmoid umožňuje nastavovat tuto váhu na hodnoty mezi 0 a 1

$$\mathbf{f}_i^{(t)} = \sigma \left(\mathbf{b}_i^f + \sum_j \mathbf{U}_{i,j}^f \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j}^f h_j^{(t-1)} \right), \quad (2.29)$$

kde $\mathbf{x}^{(t)}$ je vstupní vektor v čase t , $\mathbf{h}^{(t)}$ vektor skryté vrstvy obsahující výstupy všech

LSTM buňek, \mathbf{b}^f , \mathbf{U}^f a \mathbf{W}^f jsou popořadě prahový vektor, váhová matice vstupu a rekurentní váhová matice zapomínací brány. Stav buňky se pak řídí pravidlem

$$\mathbf{s}_i^{(t)} = \mathbf{f}_i^{(t)} \mathbf{s}_i^{(t-1)} + \mathbf{g}_i^{(t)} \sigma \left(\mathbf{b}_i + \sum_j \mathbf{U}_{i,j} \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j} \mathbf{h}_j^{(t-1)} \right), \quad (2.30)$$

kde \mathbf{b} , \mathbf{U} , \mathbf{W} jsou popořadě prahový vektor, váhová matice vstupu a rekurentní váhová matice buňky. Stav buňky je také závislý na jednotce s vstupní bránou $\mathbf{g}_i^{(t)}$, která se chová podobně jako jednotka se zapomínací branou

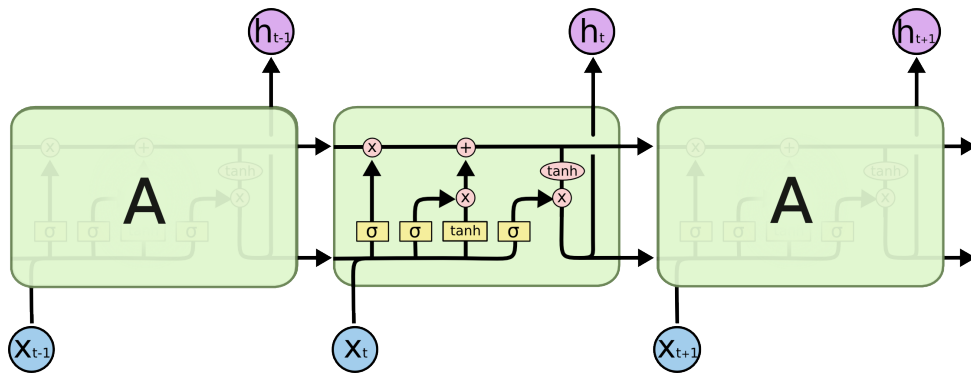
$$\mathbf{g}_i^{(t)} = \sigma \left(\mathbf{b}_i^g + \sum_j \mathbf{U}_{i,j}^g \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j}^g \mathbf{h}_j^{(t-1)} \right), \quad (2.31)$$

kde \mathbf{b}^g , \mathbf{U}^g a \mathbf{W}^g jsou popořadě prahový vektor, váhová matice vstupu a rekurentní váhová matice vstupní brány. Výstup buňky $\mathbf{h}_i^{(t)}$ je pak řízen pomocí výstupní brány $\mathbf{q}_i^{(t)}$, která opět řídí propustnost pomocí aktivační funkce sigmoid

$$\mathbf{h}_i^{(t)} = \tanh(\mathbf{s}_i^{(t)}) \mathbf{q}_i^{(t)}, \quad (2.32)$$

$$\mathbf{q}_i^{(t)} = \sigma \left(\mathbf{b}_i^o + \sum_j \mathbf{U}_{i,j}^o \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j}^o \mathbf{h}_j^{(t-1)} \right), \quad (2.33)$$

kde \mathbf{b}^o , \mathbf{U}^o a \mathbf{W}^o jsou popořadě prahový vektor, váhová matice vstupu a rekurentní váhová matice výstupní brány [2, 10, 12].



Obrázek 4: Schéma LSTM buňky. Převzato z [10].

2.4.4 GRU

Sítě GRU (z anglického "gated recurrent unit") dále zjednodušují LSTM architekturu ponecháním pouze nezbytných komponent. Výstup buňky a stav buňky je sloučen do

jednoho stavu, který je řízen jednotkou vzniklou sloučením jednotky s vstupní bránou a zapomínací bránou. Rovnice dopředného šíření má tvar

$$\mathbf{h}_i^{(t)} = \mathbf{u}_i^{(t-1)} \mathbf{h}_i^{(t-1)} + (1 - \mathbf{u}_i^{(t-1)}) \sigma \left(\mathbf{b}_i + \sum_j \mathbf{U}_{i,j} \mathbf{x}_j^{(t-1)} + \sum_j \mathbf{W}_{i,j} \mathbf{r}_j^{(t-1)} \mathbf{h}_j^{(t-1)} \right), \quad (2.34)$$

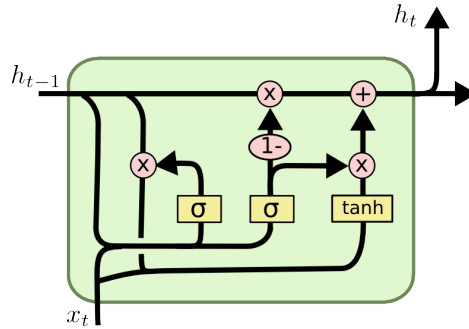
kde jednotka \mathbf{u} modifikuje stav buňky

$$\mathbf{u}_i^{(t)} = \sigma \left(\mathbf{b}_i^u + \sum_j \mathbf{U}_{i,j}^u \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j}^u \mathbf{h}_j^{(t)} \right) \quad (2.35)$$

a jednotka \mathbf{r} řídí, které informace stavu budou využity pro výpočet příštího cílového stavu

$$\mathbf{r}_i^{(t)} = \sigma \left(\mathbf{b}_i^r + \sum_j \mathbf{U}_{i,j}^r \mathbf{x}_j^{(t)} + \sum_j \mathbf{W}_{i,j}^r \mathbf{h}_j^{(t)} \right). \quad (2.36)$$

Takto definované jednotky s branami umožňují vhodně vybírat části stavového prostoru a uchovávat je ve stavu buňky. Zároveň sloučením výstupu a stavu buňky dochází k významnému snížení paměťové a výpočetní náročnosti [2, 10].



Obrázek 5: Schéma GRU buňky. Převzato z [10].

3 Optimalizace parametrů neuronových sítí

Většina učících se algoritmů využívá jistou formu optimalizace. Optimalizací rozumíme úlohu, kdy minimalizujeme či maximalizujeme předem danou funkci $f(\mathbf{x})$ změnou parametru \mathbf{x} . Hledáme tedy hodnotu \mathbf{x} takovou, aby funkce $f(\mathbf{x})$ nabývala minimální či maximální hodnoty. Většina optimalizačních metod uvažuje minimalizaci funkce $f(\mathbf{x})$ a její případná maximalizace se provádí minimalizací funkce $-f(\mathbf{x})$.

Funkce, kterou optimalizujeme, nazýváme kritérium (v terminologii strojového učení se také často objevují názvy cenová, ztrátová či chybová funkce). Tato práce se zabývá především optimalizací pro neuronové sítě, kde jsou nejčastěji využívány optimalizační metody založené na gradientu.

Základní metodou založenou na gradientu je tzv. gradientní sestup. Předpokládejme cenovou funkci $J(\boldsymbol{\theta})$ parametrizovanou souborem parametrů $\boldsymbol{\theta}$. Gradientní sestup hledá optimální soubor parametrů $\boldsymbol{\theta}^* = \operatorname{argmin} J(\boldsymbol{\theta})$ pomocí iterativního pravidla pro změnu parametrů

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \quad (3.1)$$

kde ϵ je tzv. konstanta učení, která udává velikost kroku v opačném směru největšího gradientu. Gradientní sestup pro konvexní cenové funkce vždy nalezne globální minimum a pro nekonvexní cenové funkce lokální minimum či sedlový bod [2, 3, 13].

3.1 Stochastický gradientní sestup

Ačkoliv je gradientní sestup efektivní optimalizační metoda, při optimalizaci nad velkým objemem dat může být velice pomalá a výpočetně náročná, jelikož pro jednu změnu parametrů je třeba spočítat gradient nad celou datovou sadou. Jednou z nej-používanějších modifikací gradientního sestupu, která tyto problémy řeší, je stochastický gradientní sestup.

Předpokládejme cenovou funkci ve tvaru záporného logaritmu podmíněné pravděpodobnosti

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p_{data}} L(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^m L(\mathbf{x}^i, \mathbf{y}^i, \boldsymbol{\theta}), \quad (3.2)$$

kde p_{data} je množina trénovacích dat o velikost m a $L(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = -\log p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$ je cena pro jedno pozorování v závislosti na souboru parametrů $\boldsymbol{\theta}$. Pro takto definovanou cenovou

funkci by gradientní sestup musel spočítat gradient

$$\nabla_{\theta} J(\theta) = \frac{1}{m} \sum_{i=1}^m \nabla_{\theta} L(\mathbf{x}^i, \mathbf{y}^i, \theta), \quad (3.3)$$

jehož komplexita je $O(m)$. Stochastický gradientní sestup nahlíží na gradient jako na střední hodnotu a tudíž se předpokládá, že může být aproximován menšími soubory pozorování náhodně vybíranými z datové sady, tzv. dávky. V optimalizačním kroku je náhodně vybrána dávka pozorování $\mathbf{b} = \{\mathbf{x}^1, \dots, \mathbf{x}^{m'}\}$, kde m' se volí jako malé číslo v závislosti na výpočetní kapacitě a velikosti m úplné datové sady. Při trénování modelu na grafické kartě (GPU) je vhodné volit velikost dávky jako mocninu dvou, a to v rozmezí 8 až 256, aby mohly být plně využity vektorizované operace GPU. Menší dávky mohou mít zároveň regularizační efekt díky šumu, který vnášejí do učícího se procesu. Odhad gradientu je pak vypočten s využitím dávky \mathbf{b}

$$\mathbf{g} = \frac{1}{m'} \nabla_{\theta} \sum_{i=1}^{m'} L(\mathbf{x}^i, \mathbf{y}^i, \theta). \quad (3.4)$$

Změna parametrů je dána dle pravidla

$$\theta \leftarrow \theta - \epsilon \mathbf{g}. \quad (3.5)$$

Pro pevně danou velikost modelu tedy výpočetní cena nezávisí na velikosti datové sady m [2, 13].

3.2 Další modifikace gradientního sestupu

Jedním z největších problémů při využití gradientního sestupu a gradientního stochastického sestupu je výběr konstanty učení ϵ . Vysoká hodnota ϵ může způsobit fluktuaci okolo lokálního minima nebo dokonce divergenci optimalizačního procesu a nízká hodnota ϵ může mít za následek výrazné zpomalení učení. Konstanta učení se v praxi dynamicky mění v závislosti na počtu epoch k , tedy ϵ_k . Dalším běžným problémem je uvíznutí v mělkém lokálním minimu či sedlovém bodě, což znemožní další učení. Bylo vytvořeno několik modifikací základních gradientních algoritmů, které tyto problémy do jisté míry řeší [2, 13].

3.2.1 Moment

Moment byl navržen za účelem zrychlení trénování a to především při optimalizaci ztrátových funkcí, které mají mnoho mělkých lokálních minim nebo v případech, kdy jsou gradienty značně zašuměné. Algoritmus momentu využívá plovoucí průměr minulých gradientů s exponenciálním zapomínáním a pokračuje v jejich směru.

Tento algoritmus zavádí parametr α , který udává, jakou rychlostí mají být minulé gradienty zapomínány. Pravidlo pro změnu parametrů pak vypadá následovně

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \epsilon \nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^m L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), \mathbf{y}^{(i)}) \right), \quad (3.6)$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{v}. \quad (3.7)$$

Proměnná \mathbf{v} akumuluje prvky gradientu $\nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^m L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), \mathbf{y}^{(i)}) \right)$ s tím, že čím větší je parametr α vůči konstantě učení ϵ , tím více minulé gradienty ovlivňují aktuální směr kroku. Stejně jako u konstanty učení, i parametr α může být měněn v závislosti na čase. Většinou je jako počáteční hodnota zvolena 0.5 a je navyšována až na 0.99 [2, 3, 13].

3.2.2 Nesterovův moment

Dalším variantou je Nesterovův moment, který dále rozšiřuje algoritmus momentu. Pravidlo pro změnu parametrů se změní na

$$\mathbf{v} \leftarrow \alpha \mathbf{v} - \epsilon \nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^m L(f(\mathbf{x}^{(i)}; \boldsymbol{\theta} + \alpha \mathbf{v}), \mathbf{y}^{(i)}) \right), \quad (3.8)$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{v}, \quad (3.9)$$

kde parametry α a ϵ odpovídají stejným parametrům jako u algoritmu momentu. Hlavním rozdílem oproti algoritmu momentu je, že gradient je vyhodnocen až po aplikaci minulých gradientů. Nejprve je proveden krok ve směru akumulovaných minulých gradientů a následně je provedena korekce [2, 3, 13, 14].

3.3 ADAM

Cenová funkce bývá často citlivá na určité směry v prostoru parametrů a naopak necitlivá na směry jiné. Moment tento problém do určité míry řeší, ale za cenu zavedení

nového parametru, který je třeba správně nastavit. ADAM je adaptivní metoda, která tento problém řeší tím, že zavede vlastní konstantu učení pro každý parametr, a tyto konstanty pak automaticky v průběhu učení adaptuje.

ADAM používá ke změně parametrů plovoucí průměr (první centrální moment) gradientů s exponenciálním zapomínáním \mathbf{r} a druhý necentrální moment \mathbf{v} . Počáteční hodnota obou momentů je nastavena na nulu a zejména v prvních krocích algoritmu je potřeba momenty opravit korekčním faktorem. Parametry β_1 a β_2 udávají rychlost zapomínání a δ je malá konstanta [2, 3, 13, 15].

```

inicializace
 $\mathbf{r} = 0, \mathbf{v} = 0, t = 0$ 
while zastavovací podmínka není splněna do
     $t \leftarrow t + 1$ 
    výpočet gradientu
     $g = \frac{1}{m} \nabla_{\theta} \sum_i L(f(\mathbf{x}^{(i)}; \theta), \mathbf{y}^{(i)})$ 
    změna odhadu prvního centrálního momentu
     $\mathbf{r} \leftarrow \beta_1 \mathbf{r} + (1 - \beta_1)g$ 
    změna odhadu druhého necentrálního momentu
     $\mathbf{v} \leftarrow \beta_2 \mathbf{v} + (1 - \beta_2)g \odot g$ 
    korekce odhadu prvního centrálního momentu
     $\hat{\mathbf{r}} \leftarrow \frac{\mathbf{r}}{1 - \beta_1^t}$ 
    korekce odhadu druhého necentrálního momentu
     $\hat{\mathbf{v}} \leftarrow \frac{\mathbf{v}}{1 - \beta_2^t}$ 
    změna parametrů
     $\theta \leftarrow \theta + -\epsilon \frac{\hat{\mathbf{r}}}{\sqrt{\hat{\mathbf{v}} + \delta}}$ 
end

```

Algoritmus 4: ADAM.

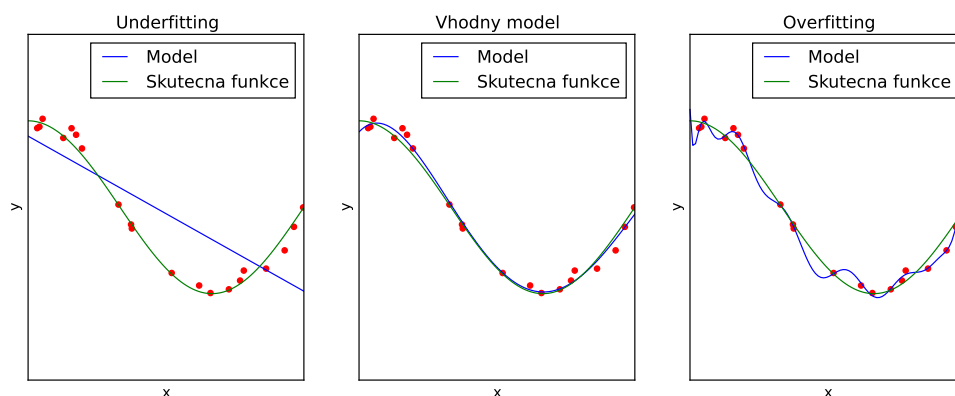
4 Kapacita modelu

Základní vlastností každého modelu by měla být generalizace. To znamená, že model musí být schopný správně klasifikovat nejen pozorování, na kterých byl natrénován, ale i většinu nových, dříve neviděných pozorování. Z tohoto důvodu se při trénování modelu datová sada běžně dělí na tři - trénovací, validační a testovací, kdy předpokládáme, že tyto sady podléhají stejnému rozložení a jednotlivá pozorování jsou nezávislá. Cílem trénování modelu je nalézt takové parametry, pro které bude mít model podobnou chybu na všech třech datových sadách. Model by měl splňovat tyto vlastnosti:

1. musí minimalizovat chybu na trénovací sadě,
2. musí minimalizovat rozdíl mezi chybou nad trénovací a testovací (popř. validační) sadou.

Během optimalizace parametrů často dochází ke dvěma nežádoucím jevům:

- **underfitting** (podtrénování) - k underfittingu dochází, pokud model není schopný dostatečně minimalizovat chybu na trénovací sadě, tj. není schopný se naučit informační souvislosti obsažené v trénovacích pozorováních;
- **overfitting** (přetrénování) - k overfittingu dochází, pokud model není schopný minimalizovat rozdíl mezi chybou nad trénovací a testovací sadou, tj. model si "zapamatuje" trénovací data místo toho, aby se naučil souvislosti obsažené v trénovacích pozorováních a na nových, dosud neviděných pozorováních selhává [2, 3, 7].



Obrázek 6: Příklad overfittingu a underfittingu.

Oba tyto jevy lze ovlivnit pomocí tzv. kapacity modelu nebo pomocí regularizačních metod. Modely s nízkou kapacitou jsou náchylné na underfitting a naopak modely s vysokou kapacitou jsou náchylné na overfitting. Kapacita modelu je většinou dána přímo zvolenou architekturou, kde snížením počtu parametrů sítě omezíme počet volných stupňů volnosti a tím lze získat vyšší generalizaci [2, 3].

4.1 Regularizace

Regularizace umožňuje navýšit generalizaci modelu bez toho, aby byla ovlivněna kapacita modelu - nevyžaduje změnu architektury modelu. Jedná se o modifikaci učícího se algoritmu, která snižuje generalizační chybu, ale zároveň nesnižuje chybu trénovací [2].

4.1.1 Penalizace velikosti parametrů

Populární metodou pro regularizaci jsou penalizační metody, a to zejména L1 a L2 penalizace. Tyto penalizační metody nepřímo omezují kapacitu modelu přidáním penalizace $\Omega(\boldsymbol{\theta})$, která odpovídá normě parametrů, k cenové funkci J . Regularizovaná cenová funkce pak vypadá následovně

$$\tilde{J}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = J(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) + \alpha \Omega(\boldsymbol{\theta}), \quad (4.1)$$

kde síla regularizace je řízena parametrem α (s vyšší hodnotou regularizace roste). Trénovací algoritmus pak minimalizuje jak původní cenovou funkci J , tak i zvolenou metriku velikosti parametrů $\Omega(\boldsymbol{\theta})$. V případě neuronových sítí jsou regularizovány pouze váhy a prahy jsou ponechány bez regularizace [2, 3].

4.1.2 Zanesení šumu

Další metodou regularizace, která neovlivňuje model samotný, je augmentace trénovacích dat ve formě aditivního šumu. Ačkoliv neuronové sítě nejsou obecně robustní vůči šumu, většinu klasifikačních i regresních úloh jsou schopny řešit i pokud jsou vstupní data zatížena malým náhodným šumem. Trénovací pozorování jsou tedy vždy před vstupem do modelu zatížena novým náhodným šumem, který vede k vyšší generalizaci výsledného modelu [2].

4.1.3 Předčasné ukončení

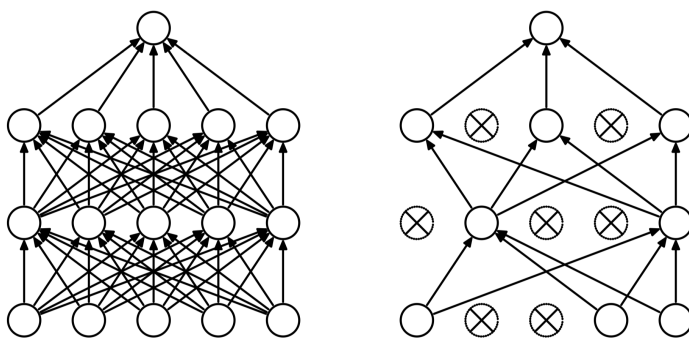
Při trénování velkých modelů s dostatečně velkou reprezentační kapacitou často dochází k jevu, kdy chyba na trénovací sadě pomalu během času klesá, zatímco chyba na validační sadě pomalu roste. Vrátime-li se k nastavení parametrů v čase s nejnižší validační chybou, můžeme získat lepší model s nižší chybou i na testovací sadě.

Algoritmus předčasného ukončení (anglicky *early stopping*) je snadno implementovatelný a výpočetně nenáročný. Pokaždé, když dojde k poklesu chyby na validační sadě, uložíme si kopii parametrů modelu. Jakmile trénování modelu skončí, vrátíme se k nejlepšímu souboru parametrů místo ponechání souboru parametrů z poslední trénovací iterace. Trénování je předčasně ukončeno, pokud žádná změna parametrů nevede k nižší chybě na validační sadě, než které bylo dosaženo s aktuálně nejlepším souborem parametrů, pro předem daný počet trénovacích iterací.

Tato forma regularizace nevyžaduje téměř žádné změny trénovacího procesu či modifikaci cenové funkce. Tento algoritmus lze využít buď samostatně nebo spolu s libovolnou regularizační metodou [2].

4.1.4 Dropout

Dropout je výpočetně nenáročnou metodou regularizace, která nejlépe funguje pro modely, které využívají dávkové učení s malými kroky. Pokaždé, kdy pozorování vstoupí do dávky, je vytvořena binární maska, která je pak aplikována na všechny vstupy skrytých jednotek v síti. Tato maska je pro každou jednotku vytvořena náhodně nezávisle na ostatních. Jednotky, jejichž odpovídající prvek v masce je roven nule, jsou pak pro dané pozorování vyloučeny z učícího se procesu.



Obrázek 7: Vlevo - před aplikací dropoutu, vpravo - po aplikaci dropoutu. Převzato z [16].

Pravděpodobnost, že v masce na danou pozici bude vybrána hodnota 1, je dána předem zvoleným parametrem. Většinou se volí pravděpodobnost 0.8, že bude při trénování využita vstupní jednotka, a pravděpodobnost 0.5, že bude využita skrytá jednotka.

Maskováním se do učícího procesu vnáší jistá forma šumu, která má za následek adaptivní poškození informačního obsahu na vstupu sítě a na vstupu skrytých jednotek. Tím je zamezeno tomu, aby se určitá jednotka zaměřila na rozpoznávání jednoho signifikantního příznaku, zatímco jiná jednotka by byla téměř nevyužita. Vyloučením náhodných jednotek z učícího se procesu jsou tak jednotky donuceny "rozdělit" si informace mezi sebou.

Dropout opět může být využit spolu s ostatními metodami regularizace [2, 3, 16].

5 CTC

Tato práce se věnuje úloze klasifikace fonémů, jejímž cílem je pro danou zvukovou nahrávku získat odpovídající sekvenci fonému. Zvuková nahrávka je rozdělena na malé segmenty a těmto segmentům jsou pak přiřazeny fonémy, které se v jednotlivých segmentech vyskytují (více v [1]). K tomuto označování segmentů se většinou využívají předtrénované modely, které ovšem vyžadují jisté předpoklady, mohou být náchylné vůči šumu a výsledné označování může být nepřesné, jelikož ne vždy lze přesně určit hranici mezi jednotlivými fonémy.

Běžně využívané cenové funkce jsou ovšem definovány pro jednotlivé prvky trénovací sekvence (segmenty) a umožňují pouze klasifikaci nezávislých fonémů. To znamená, že takto definované cenové funkce je možné využít pouze pro nezávislou predikci fonémů v jednotlivých segmentech, kdy výstupní sekvence má stejnou délku jako vstupní sekvence. K získání výsledné sekvence fonémů je třeba výstupy sítě dále zpracovat pomocí dekodéru (např. pro vstupní sekvenci o dvanácti segmentech: $aahahhoohooj \rightarrow ahoj$).

V této kapitole si uvedeme metodu CTC (z anglického "connectionist temporal classification"), díky které lze využít rekurentní neuronové sítě přímo k predikci výsledné sekvence fonémů bez potřeby dalšího zpracování výstupu sítě [1, 17, 18].

5.1 Formální definice úlohy

Mějme trénovací množinu S z pevně daného rozložení $\mathcal{D}_{\mathcal{X} \times \mathcal{Z}}$. Prostor vstupních hodnot $\mathcal{X} = (\mathcal{R})^*$ je množina všech sekvencí tvořených m dimenzionálními reálnými vektory. Prostor výstupních hodnot $\mathcal{Z} = L^*$ je množina všech sekvencí nad konečnou abecedou L značek (fonémů). Prvky L^* budeme nazývat sekvencí značek (sekvence fonému) či značkováním. Každé pozorování v S je tvořeno párem (\mathbf{x}, \mathbf{z}) , kde cílová sekvence $\mathbf{z} = (z_1, z_2, \dots, z_U)$ je nejvýše tak dlouhá, jako vstupní sekvence $\mathbf{x} = (x_1, x_2, \dots, x_T)$, tedy $U \leq T$. Vzhledem k tomu, že vstupní a cílové sekvence nejsou obecně stejně dlouhé, není možné je předem zarovnat, tj. přiřadit prvky cílové sekvence k vstupním prvkům sekvence.

Cílem je natrénovat klasifikátor $h : \mathcal{X} \mapsto \mathcal{Z}$ za využití S , který minimalizuje vhodně zvolenou metriku pro danou úlohu (např. minimalizace počtu transkripčních chyb) [17, 18].

5.2 Reprezentace výstupu

Aby bylo možné využít RNN spolu s CTC, je třeba vhodně zvolit reprezentaci výstupu sítě. Využijeme-li aktivační funkci softmax pro výstupní jednotky, můžeme transformovat výstupy sítě do tvaru podmíněné hustoty pravděpodobnosti nad danou abecedou značek a využít síť jako klasifikátor, kdy vstupní sekvence klasifikujeme výběrem nejvíce pravděpodobného značkování. Výstupní vrstva zároveň musí obsahovat o jednu jednotku více, než je počet značek v L . Aktivace prvních $|L|$ jednotek je interpretována jako pravděpodobnost pozorování odpovídajících značek a aktivace $|L| + 1$ jednotky je interpretována jako pravděpodobnost pozorování prázdné značky. Prázdná značka umožňuje zobrazit více prvků vstupní sekvence na jednu značku či ignorovat prvky sekvence, které odpovídají tichu, a z výsledné sekvence je odstraněna. Celkovou pravděpodobnost libovolného značkování pak můžeme vyčíslit sečtením pravděpodobností jeho různých zrovnání.

Mějme RNN s m vstupními jednotkami, n výstupními jednotkami a váhovým vektorem \mathbf{w} , která slouží jako zobrazení $\mathcal{N}_{\mathbf{w}} : (\mathbb{R}^m)^T \mapsto (\mathbb{R}^n)^T$. Dále mějme výstupní sekvenci sítě $\mathbf{y} = \mathcal{N}_{\mathbf{w}}(\mathbf{x})$, jejíž prvky $\mathbf{y}_k^{(t)}$ odpovídají aktivaci výstupní jednotky k v čase t a lze je interpretovat jako pravděpodobnost pozorování značky k v čase t . Hustota pravděpodobnosti značkování nad množinou L'^T sekvencí délky T nad abecedou $L' = L \cup \{\text{prázdná značka}\}$, je pak

$$p(\boldsymbol{\pi} \mid \mathbf{x}) = \prod_{t=1}^T \mathbf{y}_{\pi_t}^{(t)}, \forall \boldsymbol{\pi} \in L'^T, \quad (5.1)$$

kde $\boldsymbol{\pi}$ budeme nazývat cestami a odpovídají prvkům L'^T .

Dále definujme zobrazení $\mathcal{B} : L'^T \mapsto L^{\leq T}$, kde $L^{\leq T}$ je množina všech možných značkování, tj. množina výstupních sekvencí o délce menší či rovné T nad původní abecedou značek L . Toto zobrazení získáme odstraněním opakujících se značek a prázdných značek ze získaných cest (např. $\mathcal{B}(aa - ab -) = \mathcal{B}(a - a - bb) = aab$) a využijeme ho k získání podmíněné pravděpodobnosti daného značkování $\mathbf{l} \in L^{\leq T}$ [17, 18]

$$p(\mathbf{l} \mid \mathbf{x}) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{l})} p(\boldsymbol{\pi} \mid \mathbf{x}), \quad (5.2)$$

5.3 Dekódování

Výstupem klasifikátoru by mělo být nejvíce pravděpodobné značkování pro vstupní sekvenci

$$h(\mathbf{x}) = \operatorname{argmax}_{\mathbf{l} \in L^{\leq T}} p(\mathbf{l} \mid \mathbf{x}). \quad (5.3)$$

Proces hledání značkování nazýváme dekódováním a pro metodu CTC lze využít dvě aproximativní metody.

První metoda, dekódování nejlepší cesty, je založena na předpokladu, že cesta s nejvyšší pravděpodobností odpovídá nejvíce pravděpodobnému značkování

$$h(\mathbf{x}) \approx \mathcal{B}(\boldsymbol{\pi}^*), \quad (5.4)$$

$$\boldsymbol{\pi}^* = \operatorname{argmax}_{\boldsymbol{\pi} \in N^t} p(\boldsymbol{\pi} \mid \mathbf{x}). \quad (5.5)$$

Dekódování nejlepší cesty lze snadno najít zřetězením výstupu s nejvyšší aktivační hodnotou v každém časovém kroku. Není ovšem zaručeno, že tato metoda nalezne nejlepší možné značkování.

Druhá metoda využívá slučování cest, které procházejí stejnou sekvencí značek, a zaručuje nalezení nejlepšího značkování. Je však výpočetně náročná a aby tato metoda byla upočitatelná, je třeba využít další heuristiky (např. BEAM prořezávání, které umožní kompromis mezi rychlostí výpočtu a přesností).

Obě tyto metody mohou být zároveň obohaceny o jazykový model

$$h(\mathbf{x}) = \operatorname{argmax}_{\mathbf{l} \in L^{\leq T}} p(\mathbf{l} \mid \mathbf{x}) \cdot p(\mathbf{l})^\gamma \cdot g(\mathbf{l})^\zeta, \quad (5.6)$$

kde $p(\mathbf{l})$ je pravděpodobnost značkování daná jazykovým modelem, $g(\mathbf{l})$ je penalizace za vložení slova a parametry γ a ζ jsou předem dané váhy [17, 18].

5.4 Dopředný a zpětný algoritmus

Pro získání výsledného značkování je třeba spočítat podmíněnou pravděpodobnost $p(\mathbf{l} \mid \mathbf{x})$ jednotlivých značkování. To může být značně problematické, jelikož je potřeba spočítat sumu přes všechny cesty odpovídající danému značkování. Využijeme k tomu algoritmus dynamického programování, který vychází z předpokladu, že suma všech cest odpovídajících značkování může být rozložena na iterativní sumu přes cesty, které v daném

časovém kroku dosáhly stejné značky. Tyto iterace pak mohou být snadno spočteny pomocí rekurzivních dopředných a zpětných proměnných.

Mějme sekvenci \mathbf{q} délky r , kde $\mathbf{q}_{1:p}$ a $\mathbf{q}_{r-p:r}$ značí prvních a posledních p značek. Potom pro značkování \mathbf{l} zavedeme dopřednou proměnnou $\alpha_t(s)$, která odpovídá celkové pravděpodobnosti $\mathbf{l}_{1:s}$ v čase t

$$\alpha_t(s) = \sum_{\substack{\pi \in N^T: \\ \mathcal{B}(\pi_{1:t}) = \mathbf{l}_{1:s}}} \prod_{t'=1}^t \mathbf{y}_{\pi_{t'}}^{(t)'}, \quad (5.7)$$

kde $\alpha_t(s)$ může být rekurzivně spočtena z $\alpha_{t-1}(s)$ a $\alpha_{t-1}(s-1)$. Aby bylo možné využít prázdné značky ve výstupní cestě, je třeba upravit značkování \mathbf{l} na \mathbf{l}' vložením prázdné značky na začátek a konec sekvence a mezi každé dvě značky. Délka \mathbf{l}' je $2|\mathbf{l}| + 1$. Při výpočtu pravděpodobnosti značkování \mathbf{l}' umožníme přechody mezi prázdnými a neprázdnými značkami a také mezi každým párem unikátních neprázdných značek. Zároveň musí všechny sekvence začínat buď prázdnou značkou b , nebo prvním značkou v \mathbf{l} . Těmito omezeními jsou definovány inicializační hodnoty algoritmu

$$\alpha_1(1) = \mathbf{y}_b^{(1)}, \quad (5.8)$$

$$\alpha_1(2) = \mathbf{y}_{l_1}^{(1)}, \quad (5.9)$$

$$\alpha_1(s) = 0, \forall s > 2 \quad (5.10)$$

a rekurze

$$\alpha_t(s) = \begin{cases} \bar{\alpha}_t(s) \mathbf{y}_{l'_s}^{(t)} & \text{pokud } \mathbf{l}'_s = b \text{ nebo } \mathbf{l}'_{s-2} = \mathbf{l}'_s, \\ (\bar{\alpha}_t(s) + \alpha_{t-1}(s-2)) \mathbf{y}_{l'_s}^{(t)} & \text{jinak} \end{cases}, \quad (5.11)$$

kde

$$\bar{\alpha}_t(s) = \alpha_{t-1}(s) + \alpha_{t-1}(s-1). \quad (5.12)$$

Pravděpodobnost \mathbf{l} je pak dána sumou výsledných pravděpodobností \mathbf{l}' s prázdnou značkou a bez prázdné značky v čase T

$$p(\mathbf{l} \mid \mathbf{x}) = \alpha_T(|\mathbf{l}'|) + \alpha_T(|\mathbf{l}'| - 1). \quad (5.13)$$

Obdobně zadefinuujeme zpětnou proměnnou $\beta_t(s)$ jako celkovou pravděpodobnost $\mathbf{l}_{s:|\mathbf{l}|}$ v čase t

$$\beta_t(s) = \sum_{\substack{\boldsymbol{\pi} \in N^T: \\ \mathcal{B}(\boldsymbol{\pi}_{t:T}) = \mathbf{l}_{s:|\mathbf{l}|}}} \prod_{t'=t}^T \mathbf{y}_{\boldsymbol{\pi}_{t'}}^{(t)'}, \quad (5.14)$$

$$\beta_T(|\mathbf{l}'|) = \mathbf{y}_b^{(T)}, \quad (5.15)$$

$$\beta_T(|\mathbf{l}'| - 1) = \mathbf{y}_{l'_i}^{(T)}, \quad (5.16)$$

$$\beta_T(s) = 0, \forall s < |\mathbf{l}'| - 1, \quad (5.17)$$

$$\beta_t(s) = \begin{cases} \bar{\beta}_t(s) \mathbf{y}_{l'_s}^{(t)} & \text{pokud } \mathbf{l}'_s = b \text{ nebo } \mathbf{l}'_{s+2} = \mathbf{l}'_s \\ (\bar{\beta}_t(s) + \beta_{t+1}(s+2)) \mathbf{y}_{l'_s}^{(t)} & \text{jinak} \end{cases}, \quad (5.18)$$

$$\bar{\beta}_t(s) = \beta_{t+1}(s) + \beta_{t+1}(s+1). \quad (5.19)$$

Aby byla při výpočtech zajištěna numerická stabilita, je třeba dopředné a zpětné proměnné přeškálovat

$$\hat{\alpha}_t(s) = \frac{\alpha_t(s)}{\sum_s \alpha_t(s)}, \quad (5.20)$$

$$\hat{\beta}_t(s) = \frac{\beta_t(s)}{\sum_s \beta_t(s)}. \quad (5.21)$$

Dosazením přeškálovaných proměnných do vzorců a jejich úpravou (podrobněji v [18]) můžeme síť natrénovat pomocí metod založených na maximální věrohodnosti. Zpětné šíření gradientu skrz výstupní vrstvu s aktivační funkcí softmax je definováno následovně

$$\frac{\partial O^{ML}(\{\mathbf{x}, \mathbf{z}\}, \mathcal{N}_w)}{\partial \mathbf{u}_k^{(t)}} = \mathbf{y}_k^{(t)} - \frac{1}{\mathbf{y}_k^{(t)} Z_{(t)}} \sum_{s \in \text{lab}(\mathbf{z}, k)} \hat{\alpha}_t(s) \hat{\beta}_t(s), \quad (5.22)$$

kde $\mathbf{u}_k^{(t)}$ jsou nenormalizované pravděpodobnosti na výstupu sítě, $\text{lab}(\mathbf{z}, k) = \text{lab}(\mathbf{l}, k) = \{s : \mathbf{l}'_s = k\}$ je množina pozic, na kterých se ve značkování \mathbf{l} vyskytuje značka k , a kde [17, 18]

$$Z_t = \sum_{s=1}^{|\mathbf{l}'|} \frac{\hat{\alpha}_t(s) \hat{\beta}_t(s)}{\mathbf{y}_{l'_s}^{(t)}}. \quad (5.23)$$

6 Klasifikace fonémů

V této práci bylo porovnáno šest architektur neuronových sítí pro úlohu klasifikace fonémů. Experimenty byly provedeny nad dvěma datovými sady pro čtyři různé parametrizace příznaků. K implementaci byl využit programovací jazyk Python 3.4 spolu s knihovnami pro neuronové sítě Tensorflow a Keras (kód dostupný na [19]). Sítě byly trénovány na grafické kartě Nvidia Tesla K20m s pamětí 5GB a knihovnami CUDA 8.0 a CuDNN 6.0.

6.1 Předzpracování dat

Architektury byly testovány na dvou datových sadách, ŠkodaAuto a SpeechDat-E, které obsahují české nahrávky se vzorkovací frekvencí 8 kHz. Obě tyto sady byly dále rozděleny na sadu trénovací (75%), validační (12.5%) a testovací (12.5%) tak, aby každý řečník byl v právě jedné sadě. Základní informace o datových sadách jsou uvedeny v tabulce 1.

	ŠkodaAuto	SpeechDat-E
počet řečníků	47	924
počet nahrávek	14523	39560
počet řečníků v trénovací sadě	35	693
počet řečníků ve validační sadě	6	115
počet řečníků v testovací sadě	6	116

Tabulka 1: Popis datových sad a jejich rozdělení.

Pro obě datové sady pak byly spočteny čtyři parametrizace příznaků [1], na které se dále budeme odkazovat tučně uvedenými zkratkami:

- **LFE** - logaritmované energie banky filtrů, pro výpočet využito 40 filtrů,
- **LFE $\Delta\Delta$** - logaritmované energie banky filtrů, pro výpočet využito 40 filtrů, vypočteny delta a delta-delta koeficienty,
- **MFCC** - mel-frekvenční keprstrální koeficienty, pro výpočet využito 26 filtrů, spočteno 13 koeficientů,

- **MFCC $\Delta\Delta$** - mel-frekvenční keprální koeficienty, pro výpočet využito 26 filtrů, spočteno 13 koeficientů, vypočteny delta a delta-delta koeficienty.

Všechny parametrizace byly vygenerovány s využitím Hammingova okénka o délce 25ms s posunem 10ms (jednotlivé segmenty se překrývají) a 512 bodové FFT.

Takto spočtené příznaky pak byly po sloupcích normalizovány tak, aby měly nulovou střední hodnotu a jednotkový rozptyl. K tomu byla využita Z-score normalizace, která přetransformuje vstupní data \mathbf{x} na data \mathbf{z} dle vzorce

$$\mathbf{z} = \frac{\mathbf{x} - \mu_x}{\sigma_x}, \quad (6.1)$$

kde μ_x je střední hodnota \mathbf{x} a σ_x je rozptyl \mathbf{x} .

6.2 Architektury a trénování sítí

Celkem bylo navrženo šest architektur sítí, na které se dále budeme odkazovat tučně uvedenými názvy:

- **dopředná**

typ vrstvy	počet neuronů	aktivační funkce
plně propojená	1024	ReLU
dropout (0.5)	-	-
plně propojená	512	ReLU
dropout (0.5)	-	-
plně propojená	256	ReLU
dropout (0.5)	-	-
plně propojená	40	softmax

- **LSTM**

typ vrstvy	počet neuronů/buněk	aktivační funkce
LSTM	{100, 150, 200, 250}	tanh
LSTM	{100, 150, 200, 250}	tanh
plně propojená	40	softmax

- **GRU**

typ vrstvy	počet neuronů/buněk	aktivační funkce
GRU	{100, 150, 200, 250}	tanh
GRU	{100, 150, 200, 250}	tanh
plně propojená	40	softmax

- **obousměrná CTC LSTM**

typ vrstvy	počet neuronů/buněk	aktivační funkce
obousměrná LSTM	{100, 150, 200, 250}	tanh
obousměrná LSTM	{100, 150, 200, 250}	tanh
plně propojená	41	softmax

- **dávková CTC LSTM**

typ vrstvy	počet neuronů/buněk	aktivační funkce
LSTM	{100, 150, 200, 250}	tanh
LSTM	{100, 150, 200, 250}	tanh
plně propojená	41	softmax

- **obousměrná dávková CTC LSTM**

typ vrstvy	počet neuronů/buněk	aktivační funkce
obousměrná LSTM	{100, 150, 200, 250}	tanh
obousměrná LSTM	{100, 150, 200, 250}	tanh
plně propojená	41	softmax

Dopředná síť byla trénována pomocí stochastického gradientního sestupu s konstantou učení $\epsilon = 0.01$ a Nesterovovým momentem $\alpha = 0.9$. Pokud během 5 trénovacích epoch nedošlo ke snížení ztráty na validační sadě, konstanta učení byla snížena na polovinu. Všechny rekurentní neuronové sítě pak využívaly optimalizační algoritmus ADAM s doporučenými hodnotami [15]. Všechny sítě byly trénovány po dávkách o velikosti 64 pozorování.

Obousměrná CTC LSTM využívá k predikci celou trénovací sekvenci. Celá sekvence ovšem v případě predikce v reálném čase není dostupná, a proto byly také testovány dávkové CTC sítě, které stejně jako dopředné či LSTM a GRU sítě využívají k predikci pouze kratší sekvence o předem definované délce .

Dopředná síť, síť LSTM a síť GRU využívají k dekodování výsledné sekvence Viterbiho algoritmus se zerogramovým jazykovým modelem. Viterbiho dekodér byl implementován pomocí algoritmu cestování žetonů (anglicky token passing), kde byly využity jednostavové skryté Markovovy modely reprezentující jednotlivé fonémy. CTC sítě využívají dekodování nejlepší cesty.

Jako regularizační metody byly využity zanesení šumu do trénovacích dat a předčasné ukončení o toleranci 10 epoch. Gaussovský šum o směrodatné odchylce 0.6 [20] byl nově vygenerován pro každé pozorování před vstupem do sítě v každé epoše. Metoda dropout mohla být využita pouze pro dopřednou síť, jelikož pro rekurentní sítě byly využity optimalizované vrstvy pro knihovnu CuDNN, které tuto metodu zatím nepodporují. Nicméně se ukázalo, že regularizace metodou zanesení šumů je pro tuto úlohu dostačující a dropout, jehož rekurentní implementace bývá výpočetně velice pomalá, není třeba.

7 Experimenty a vyhodnocení

Nejprve byly provedeny experimenty nad datovou sadou ŠkodaAuto a její parametrizací LFE. Nejlepší nastavení jednotlivých architektur pak bylo testováno nad zbylými parametrizacemi jak na datové sadě ŠkodaAuto, tak na datové sadě SpeechDat-E. Byly uvažovány tři metriky (všechny na testovací sadě), na které se dále budeme odkazovat tučně uvedenými názvy:

- **frameAcc** [%] - přesnost klasifikace modelu vyhodnocována po jednotlivých segmentech nahrávek

$$\text{frameAcc} = \frac{C_s}{N_s} \cdot 100, \quad (7.1)$$

kde C_s je počet správně klasifikovaných pozorování (segmentů) a N_s je počet všech pozorování,

- **phoneCorr** [%] - procento správně klasifikovaných fonémů po dekodování

$$\text{phoneCorr} = \frac{C_f}{N_f} \cdot 100, \quad (7.2)$$

kde C_f je počet správně klasifikovaných fonémů a N_f je počet všech fonémů,

- **phoneAcc** [%] - přesnost modelu po dekodování

$$\text{phoneAcc} = \frac{C_f - I}{N_f} \cdot 100, \quad (7.3)$$

kde C_f je počet správně klasifikovaných fonémů, I je počet navíc vložených znaků (fonémů) a N_f je počet všech fonémů.

Pro síť využívající metodu CTC nebyla sledována přesnost modelu (frameAcc), ale pouze přesnost dekodování (phoneAcc). Během trénování modelů pak docházelo ke třem jevům

1. předčasné ukončení z důvodu stagnace učícího se procesu, tj. chyba na validační sadě se po dobu 10 epoch téměř nezměnila (*),
2. předčasné ukončení z důvodu overfittingu, tj. chyba na validační sadě začala růst a trénování bylo ukončeno (**),
3. síť se nepodařilo natrénovat (***)

Zároveň je třeba rozlišit tři důvody, proč se síť nepodařilo natrénovat:

1. síť se nepodařilo natrénovat kvůli vysoké výpočetní a časové náročnosti (např. pro datovou sadu SpeechDat-E s parametrizací LFE $\Delta\Delta$ trvala jedna epocha trénování dopředné sítě přes 45 hodin),
2. síť se nepodařilo natrénovat kvůli jevům jako je exploze či vymizení gradientu (např. síť GRU pro datovou sadu ŠkodaAuto s parametrizací LFE),
3. síť se nepodařilo natrénovat kvůli neznámé chybě v knihovně Keras, kdy během jedné epochy dojde k výraznému poklesu přesnosti - pravděpodobně se jedná o numerickou nestabilitu (např. síť LSTM a GRU pro datovou sadu SpeechDat-E s parametrizacemi LFE).

7.1 Experimenty nad datovou sadou ŠkodaAuto

První testovanou architekturou byla dopředná síť, která na rozdíl od rekurentních sítí nemůže implicitně využívat kontextu v datech. Proto bylo k aktuálně klasifikovanému segmentu přidáno n předešlých segmentů zleva (levý kontext - minulost) a m následujících segmentů zprava (pravý kontext - budoucnost). Z tabulky 2 je patrné, že čím větší kontext má model k dispozici, tím je model a dekodování přesnější. Ačkoliv nejlepších výsledků dosahuje model s levým kontextem $n = 20$ a pravým kontextem $m = 10$, k dalším experimentům byl využit model s levým kontextem $n = 10$ a pravým kontextem $m = 5$, který byl vybrán na základě několika kritérií:

- využívá kontextu o srovnatelné délce (15) s délkou sekvence zvolenou pro vybrané rekurentní síť (20),
- dosahuje relativně vysoké přesnosti,
- využívá nižší počet parametrů (rychlejší trénování a "spravedlivější" srovnání s vybranými rekurentními sítěmi, které využívají 100000 - 500000 parametrů).

Jelikož cílem práce je porovnat vybrané architektury a ne optimalizace topologie a natrénování jedné sítě za účelem co nejvyšší přesnosti, trénování dopředných sítí bylo omezeno na 150 epoch. Lze předpokládat, že s vyšším počtem epoch by bylo možné dosáhnout ještě lepších výsledků.

levý kontext	pravý kontext	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
0	0	60.54	64.79	53.17	150	710604
10	5	81.27	87.26	79.99	150	1327404
20	5	82.07	87.46	81.19	150	1738604
20	10	83.34	87.13	82.09	150	1994204

Tabulka 2: Výsledky experimentů pro dopřednou síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

Jak pro síť LSTM, tak pro síť GRU, bylo testováno několik různých kombinací počtu buněk v jednotlivých vrstvách a délek vstupní sekvence. U všech sítí s délkou vstupní sekvence 20 a 30 segmentů, kromě sítě s délkou vstupní sekvence 20 a 100 buňkami v jednotlivých vrstvách, docházelo k silnému overfittingu již od třetí epochy a sítě byly vyhodnoceny jako nevhodné kvůli nízké schopnosti generalizace (viz. tabulka 3). Pro další experimenty byla tedy zvolena síť, která obsahuje 100 buněk v každé vrstvě a přijímá vstupní sekvenci o délce 20 segmentů. Tím byl zároveň splněn požadavek na malou síť, která by mohla být využita pro predikci v reálném čase i na mobilních zařízeních.

I experimenty se sítěmi GRU potvrdily, že výše zmíněné parametry sítí, které trpěly overfittingem, nejsou pro tuto úlohu a datovou sadu vhodné. Síť GRU s těmito parametry se nepodařilo natrénovat z důvodu saturace gradientu (viz. tabulka 4). To zároveň ukazuje na jednu z předností LSTM vrstev a to, že u nich nemůže dojít k saturaci gradientu při zpracování dlouhých sekvencí. Stejně jako pro síť LSTM byla zvolena síť GRU se 100 buňkami v každé vrstvě, která přijímá vstupní sekvenci o délce 20 segmentů. Díky sjednocení komponent v GRU buňce zmíněného v teoretické části si můžeme všimnout, že sítě GRU mají výrazně nižší počet parametrů oproti odpovídajícím LSTM sítím.

počet buněk	délka sekvence	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
100	10	70.37*	79.35	67.73	31	141640
150	10	72.12*	81.83	69.54	30	302440
200	10	72.80*	82.89	70.39	25	523240
250	10	73.49*	83.35	69.79	31	804040
100	20	76.03*	84.37	76.45	58	141640
150	20	77.73**	85.11	76.31	48	302440
200	20	78.51**	86.42	77.66	27	523240
250	20	78.63**	86.72	77.76	19	804040
100	30	79.59**	78.45	73.12	39	141640
150	30	81.57**	79.95	74.76	24	302440
200	30	82.02**	80.93	75.23	25	523240
250	30	82.70**	80.93	75.24	26	804040

Tabulka 3: Výsledky experimentů pro LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

počet buněk	délka sekvence	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
100	10	69.22	77.99	66.81	41	107240
150	10	70.32*	79.91	68.30	34	228340
200	10	70.86*	80.75	68.70	39	394440
250	10	71.14*	81.05	68.46	56	605540
100	20	73.63*	80.40	72.84	30	107240
150	20	56.13***	62.28	55.41	25	228340
200	20	50.45**** ²	55.69	49.19	16	394440
250	20	52.18**** ²	57.89	51.03	12	605540
100	30	53.80**** ²	53.23	47.22	43	107240
150	30	56.00**** ²	56.92	49.57	18	228340
200	30	48.02**** ²	45.62	40.04	15	394440
250	30	49.97**** ²	48.87	43.43	12	605540

Tabulka 4: Výsledky experimentů pro GRU síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

Dále byly provedeny experimenty s LSTM sítěmi, které využívají metodu CTC. Obousměrná

CTC LSTM síť, která při klasifikaci využívá celou délku sekvence, dosahovala velmi vysoké přesnosti jak na trénovací, tak na validační a testovací sadě (viz. tabulka 5). Z obdobných důvodů jako u dopředné sítě nebyla pro další experimenty zvolena síť dosahující nejvyšší přesnosti, ale síť se 100 buňkami v každé dopředné i zpětné LSTM vrstvě.

počet buněk	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
100	94.03	92.97**	124	363441
150	94.07	92.65**	68	785141
200	94.32	93.36**	60	1366841
250	95.18	94.43**	67	2108541

Tabulka 5: Výsledky experimentů pro obousměrnou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

Síť, která k zahájení predikce vyžaduje celou sekvenci, je pro účely predikce v reálném čase nevhodná. Proto byly provedeny experimenty se sítěmi využívající metodu CTC, které k predikci využívají pouze kratší sekvence (úseky z celé sekvence). Byly uvažovány dvě varianty - síť s obyčejnými LSTM vrstvami a síť s obousměrnými LSTM vrstvami. Zde již byly provedeny experimenty pouze pro dvě nejlepší nastavení sítě LSTM, a to buď s původní trénovací datovou sadou nebo s rozšířenou datovou sadou. Rozdílem mezi těmito trénovacími sadami je, že jednotlivé trénovací sekvence v původní datové sadě na sebe navazují a jejich zřetězením lze získat celou sekvenci, zatímco v rozšířené trénovací sadě dochází k 50% přesahu mezi jednotlivými trénovacími sekvencemi, tj. pro trénovací sekvence o délce m segmentů je posledních $\frac{m}{2}$ segmentů trénovací sekvence stejných, jako prvních $\frac{m}{2}$ segmentů následující trénovací sekvence). Tímto rozšířením získá síť větší množství trénovacích dat a je schopna přesnější klasifikace.

Z tabulek 6 a 7 je patrné, že přesnost obousměrné sítě vždy převyšuje přesnost sítě obyčejné, ovšem za cenu dvojnásobného počtu trénovatelných parametrů.

Pro obě architektury bylo zvoleno 100 buněk v každé vrstvě (pro obousměrnou síť 100 buněk v každé dopředné i zpětné vrstvě), délka vstupní sekvence 20 a trénování pomocí rozšířené trénovací sady.

počet buněk	délka sekvence	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
100	20 (přesah 10)	81.31	63.04*	73	141741
100	20	78.61	61.55*	63	141741
200	10 (přesah 5)	85.8	51.13*	72	523441
200	10	83.76	45.61*	54	523441

Tabulka 6: Výsledky experimentů pro dávkovou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

počet buněk	délka sekvence	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
100	20 (přesah 10)	87.55	72.31*	73	363441
100	20	86.42	67.93**	65	363441
200	10 (přesah 5)	87.34	53.22*	67	1366841
200	10	85.83	49.39**	42	1366841

Tabulka 7: Výsledky experimentů pro obousměrnou dávkovou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

Vybraná nastavení jednotlivých architektur uvedených výše, která byla využita pro další experimenty, jsou shrnuty v tabulce 8.

architektura	levý kontext	pravý kontext
dopředná	10	5
architektura	počet buněk	délka sekvence
LSTM	100	20
GRU	100	20
obousměrná CTC LSTM	100	-
dávková CTC LSTM	100	20 (přesah 10)
obousměrná dávková CTC LSTM	100	20 (přesah 10)

Tabulka 8: Vybraná nastavení jednotlivých architektur získaných experimenty.

Modely se zvoleným nastavením pak byly natrénovány pro všechny parametrizace nad datovou sadou ŠkodaAuto. Na výsledcích pro jednotlivé parametrizace (viz. tabulky

9 až 12) je možné pozorovat jasný trend, co se přesnosti jednotlivých architektur týče. Nejvyšší přesnosti dosáhla obousměrná CTC LSTM síť a to přes 90% pro všechny parametrizace. Druhé nejvyšší přesnosti pak stabilně dosahovala dopředná síť, která se pohybovala okolo přesnosti 80%. Ukázalo se, že síť LSTM ve většině případech dosahuje lepších výsledků než síť GRU a po dalším odlazení architektury by mohla konkurovat dopředné síti. Obousměrná dávková CTC LSTM síť pak dle předpokladů dosahuje vyšší přesnosti než obyčejná dávková CTC LSTM síť, jelikož využívá kontext i z takto krátkých sekvencí. S rostoucí délkou sekvence pak lze očekávat vyšší přesnost blížící se k přesnosti obousměrné CTC LSTM sítě - délku sekvence je tedy vhodné volit na základě požadované přesnosti rozpoznání fonémů a na základě omezení jako je například doba odezvy (kompromis mezi delší nebo kratší délkou sekvence).

Pro velké množství trénovaných modelů byly experimenty provedeny vždy jednou, tudíž není známá střední hodnota a standardní odchylka přesností jednotlivých architektur nad danými parametrizacemi. Nicméně na základě výsledků jak na datové sadě ŠkodaAuto, tak datové sadě SpeechDat-E, lze předpokládat, že výše zmíněné závěry jsou reprezentativní a vyšší počet experimentů by je pouze potvrdil.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	81.27	87.26	79.99	150	1327404
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	76.03*	84.37	76.45	58	141640
GRU	73.63*	80.40	72.84	30	107240
obousměrná CTC LSTM	-**	94.03	92.97	124	363441
dávková CTC LSTM	-*	81.31	63.04	73	141741
obousměrná dávková CTC LSTM	-*	87.55	72.31	73	363441

Tabulka 9: Výsledky experimentů pro parametrizaci LFE nad datovou sadou ŠkodaAuto.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	80.15	87.10	80.15	150	2643244
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	77.13**	85.39	78.17	59	173640
GRU	72.53*	81.23	73.49	22	131240
obousměrná CTC LSTM	**	94.22	93.34	120	427441
dávková CTC LSTM	*	86.05	68.58	98	173741
obousměrná dávková CTC LSTM	**	89.97	72.74	83	427441

Tabulka 10: Výsledky experimentů pro parametrizaci LFE $\Delta\Delta$ nad datovou sadou ŠkodaAuto.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	81.16	87.06	79.63	150	883308
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	71.53**	81.01	67.10	32	501640
GRU	73.70*	80.56	73.12	35	99140
obousměrná CTC LSTM	**	94.20	93.03	116	341841
dávková CTC LSTM	*	80.34	65.28	69	130941
obousměrná dávková CTC LSTM	*	87.54	72.19	68	341841

Tabulka 11: Výsledky experimentů pro parametrizaci MFCC nad datovou sadou ŠkodaAuto.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	82.35	87.76	81.06	150	1310956
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	78.82**	86.58	79.56	44	141240
GRU	77.03**	84.61	77.65	71	106940
obousměrná CTC LSTM	_**	94.66	93.70	12	5 362641
dávková CTC LSTM	_*	87.02	70.80	73	141341
obousměrná dávková CTC LSTM	_**	90.93	75.19	45	362641

Tabulka 12: Výsledky experimentů pro parametrizaci MFCC $\Delta\Delta$ nad datovou sadou ŠkodaAuto.

7.2 Experimenty nad datovou sadou SpeechDat-E

Ačkoliv jsou výsledné přesnosti nad parametrizacemi datové sady ŠkodaAuto poměrně vysoké, jedná se o malou datovou sadu, kterou se vhodně zvolené architektury mohou poměrně snadno naučit. Proto byly provedeny stejné experimenty i nad parametrizacemi komplexnější datové sady SpeechDat-E. Z tabulek 13 až 16 je na první pohled patrné, že přesnost všech sítí je nižší, a to o 5 až 15%. Výše zmíněné závěry jsou ovšem stále platné i pro tuto datovou sadu.

Pro parametrizaci LFE se bohužel nepodařilo natrénovat síť LSTM a GRU kvůli neznámé a dosud neopravené chybě v knihovně Keras. Trénování těchto modelů bylo spuštěno vícekrát, nicméně vždy došlo k chybě dříve, než byl model natrénován do takové míry, aby bylo možné reprezentativně vyhodnotit jeho přesnost. Pro parametrizaci LFE a LFE + $\Delta\Delta$ se nepodařilo natrénovat síť LSTM, GRU a dopřednou síť kvůli vysoké časové náročnosti (až dva dny na jednu epochu).

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	72.5	71.68	68.65	150	1329717
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	_***3	-	-	-	142549
GRU	_***3	-	-	-	108149
obousměrná CTC LSTM	_**	83.10	81.22	104	365250
dávková CTC LSTM	_*	69.40	58.32	50	142650
obousměrná dávková CTC LSTM	_*	73.12	62.42	40	365250

Tabulka 13: Výsledky experimentů pro parametrizaci LFE nad datovou sadou SpeechDat-E.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	_***1	-	-	-	2645557
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	_***1	-	-	-	174549
GRU	_***1	-	-	-	132149
obousměrná CTC LSTM	_**	83.79	82.11	112	429250
dávková CTC LSTM	_*	73.28	60.87	61	174650
obousměrná dávková CTC LSTM	_*	77.85	66.81	28	429250

Tabulka 14: Výsledky experimentů pro parametrizaci LFE $\Delta\Delta$ nad datovou sadou SpeechDat-E.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	72.45	71.99	68.92	150	885621
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	65.41*	66.64	62.87	41	131749
GRU	63.37*	63.31	59.86	23	100049
obousměrná CTC LSTM	-**	81.73	80.42	96	343650
dávková CTC LSTM	-*	70.70	60.05	86	131850
obousměrná dávková CTC LSTM	-*	73.66	63.42	25	343650

Tabulka 15: Výsledky experimentů pro parametrizaci MFCC nad datovou sadou SpeechDat-E.

architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
dopředná	73.13	72.64	69.88	150	1313269
architektura	frameAcc [%]	phoneCorr [%]	phoneAcc [%]	počet epoch	počet parametrů
LSTM	69.94*	72.18	68.80	73	142149
GRU	67.34	70.23	65.53	42	107489
obousměrná CTC LSTM	-*	84.82	83.23	71	364450
dávková CTC LSTM	-*	75.07	63.81	91	142250
obousměrná dávková CTC LSTM	-*	78.21	67.87	48	364450

Tabulka 16: Výsledky experimentů pro parametrizaci MFCC $\Delta\Delta$ nad datovou sadou SpeechDat-E.

7.3 Porovnání výsledku nad ŠkodaAuto a SpeechDat-E

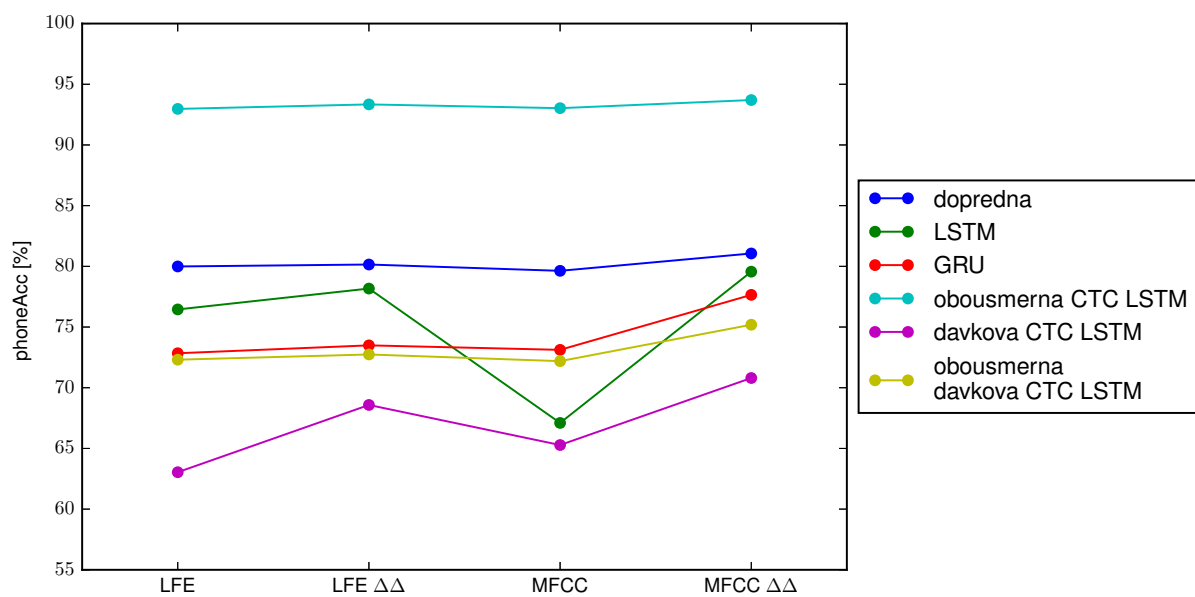
Pro snadnější srovnání jednotlivých architektur nad všemi parametrizacemi obou datových sad byly přesnosti architektur a jejich přesnosti dekódování shrnuty do tabulek 17 a 18.

architektura	frameAcc [%]				phoneAcc [%]			
	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$
dopředná	81.27	80.15	81.16	82.35	79.99	80.15	79.63	81.06
LSTM	76.03	77.13	71.53	78.82	76.45	78.17	67.10	79.56
GRU	73.63	72.53	73.70	77.03	72.84	73.49	73.12	77.65
obousměrná CTC LSTM	-	-	-	-	92.97	93.34	93.03	93.70
dávková CTC LSTM	-	-	-	-	63.04	68.58	65.28	70.80
obousměrná dávková CTC LSTM	-	-	-	-	72.31	72.74	72.19	75.19

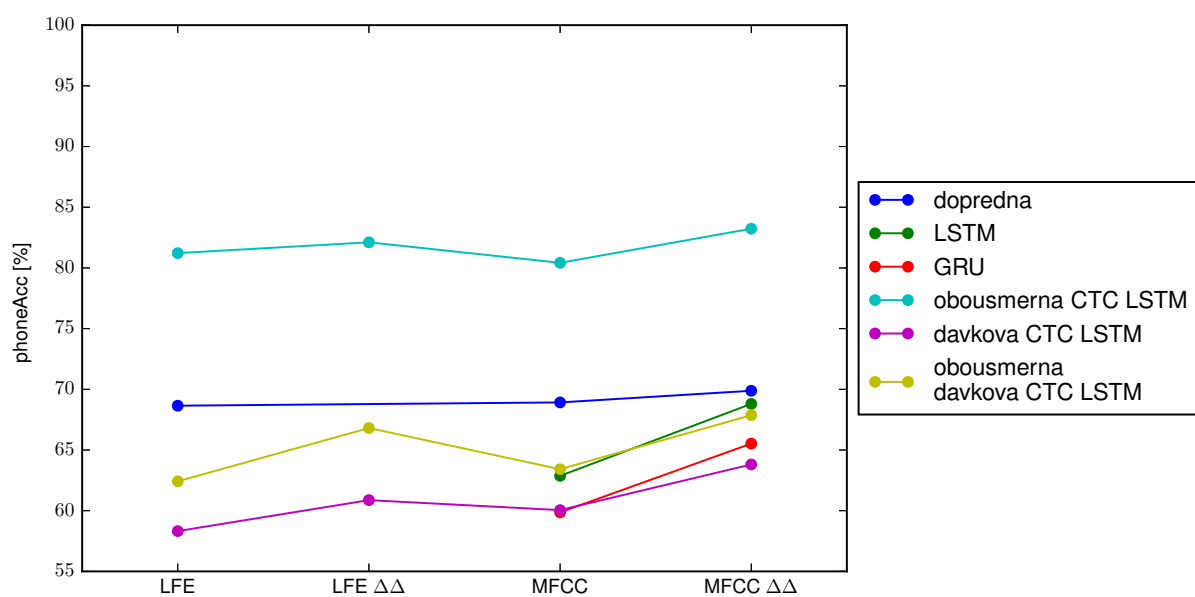
Tabulka 17: Výsledky experimentů pro datovou sadu ŠkodaAuto.

architektura	frameAcc [%]				phoneAcc [%]			
	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$
dopředná	72.50	-	72.45	73.13	68.65	-	68.92	69.88
LSTM	-	-	65.41	69.94	-	-	62.87	68.80
GRU	-	-	63.37	67.34	-	-	59.86	65.53
obousměrná CTC LSTM	-	-	-	-	81.22	82.11	80.42	83.23
dávková CTC LSTM	-	-	-	-	58.32	60.87	60.05	63.81
obousměrná dávková CTC LSTM	-	-	-	-	62.42	66.81	63.42	67.87

Tabulka 18: Výsledky experimentů pro datovou sadu SpeechDat-E.



Obrázek 8: Srovnání přesností dekódování jednotlivých architektur pro parametrizace nad datovou sadou ŠkodaAuto.



Obrázek 9: Srovnání přesností dekódování jednotlivých architektur pro parametrizace nad datovou sadou SpeechDat-E.

8 Závěr

Cílem této práce bylo porovnat různé architektury neuronových sítí pro úlohu klasifikace fonémů. V první části práce byly představeny dopředné a rekurentní neuronové sítě a běžně využívané optimalizační algoritmy. Zároveň byla odvozena metoda CTC, která umožňuje přímý přepis fonémů do výsledné podoby bez potřeby dalšího zpracování výstupů sítě. Ve druhé části pak byly představeny navržené architektury a byla porovnána jejich přesnost na čtyřech různých parametrizacích dvou datových sad. Při návrhu architektur bylo zároveň zohledněno jejich případné použití pro rozpoznávání v reálném čase, a proto byly voleny především sítě o dvou vrstvách s nízkým počtem parametrů.

architektura	ŠkodaAuto				SpeechDat-E			
	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$	LFE	LFE $\Delta\Delta$	MFCC	MFCC $\Delta\Delta$
dopředná	79.99	80.15	79.63	81.06	68.65	-	68.92	69.88
LSTM	76.45	78.17	67.10	79.56	-	-	62.87	68.80
GRU	72.84	73.49	73.12	77.65	-	-	59.86	65.53
obousměrná CTC LSTM	92.97	93.34	93.03	93.70	81.22	82.11	80.42	83.23
dávková CTC LSTM	63.04	68.58	65.28	70.80	58.32	60.87	60.05	63.81
obousměrná dávková CTC LSTM	72.31	72.74	72.19	75.19	62.42	66.81	63.42	67.87

Tabulka 19: Porovnání přesnosti dekodování (phoneAcc [%]) mezi datovou sadou ŠkodaAuto a SpeechDat-E.

Porovnáním přesností rozpoznání se ukázalo (shrnutí v tabulce 19), že nejpřesnější architekturou je síť využívající obousměrné LSTM vrstvy a metodu CTC, která dosáhla přesnosti přes 90% na jednodušší datové sadě ŠkodaAuto a přes 80% na komplexnější datové sadě SpeechDat-E. Tato síť ovšem k rozpoznání využívala informaci z celé nahrávky a pro využití v reálném čase není vhodná. Poměrně dobrých výsledků ovšem také dosahovala dopředná síť s pyramidovou strukturou a rekurentní síť obsahující dvě LSTM vrstvy. Obousměrná verze této rekurentní sítě, která využívala metodu CTC dosahovala obdobných výsledků a pro rozpoznávání v reálném čase má poměrně velký potenciál omezený pouze výkonem zařízení.

Pro další zlepšení dosažených výsledků by bylo vhodné zaměřit se na síť využívající LSTM vrstvy a na její obousměrnou obdobu využívající metodu CTC a optimalizovat

topologii těchto architektur. Na základě podkladů daných touto prací a výzkumem Alexe Gravesa [20], který již řešil vliv počtu vrstev na přesnost rozpoznání fonémů pro síť využívající celou sekvenci, se jako další postup jeví otestování vlivu délky vstupní sekvence spolu s hloubkou sítě a šířkou jednotlivých vrstev na přesnost rozpoznání. Zároveň je třeba najít jistý kompromis mezi velikostí sítě a délkou vstupní sekvence tak, aby byla dosažena dostatečná přesnost rozpoznání a zároveň byla zaručena odezva sítě v reálném čase.

Seznam obrázků

1	Schéma rekurentní neuronové sítě. Převzato z [9].	9
2	Ilustrace rozvinutí rekurentní neuronové sítě. Převzato z [10].	9
3	Schéma obousměrné rekurentní neuronové sítě. Převzato z [9].	12
4	Schéma LSTM buňky. Převzato z [10].	14
5	Schéma GRU buňky. Převzato z [10].	15
6	Příklad overfittingu a underfittingu.	20
7	Vlevo - před aplikací dropoutu, vpravo - po aplikaci dropoutu. Převzato z [16].	22
8	Srovnání přesností dekodování jednotlivých architektur pro parametrizace nad datovou sadou ŠkodaAuto.	45
9	Srovnání přesností dekodování jednotlivých architektur pro parametrizace nad datovou sadou SpeechDat-E.	45

Seznam tabulek

1	Popis datových sad a jejich rozdělení.	29
2	Výsledky experimentů pro dopřednou síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	35
3	Výsledky experimentů pro LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	36
4	Výsledky experimentů pro GRU síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	36
5	Výsledky experimentů pro obousměrnou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	37
6	Výsledky experimentů pro dávkovou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	38
7	Výsledky experimentů pro obousměrnou dávkovou CTC LSTM síť pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	38
8	Vybraná nastavení jednotlivých architektur získaných experimenty.	38
9	Výsledky experimentů pro parametrizaci LFE nad datovou sadou ŠkodaAuto.	39
10	Výsledky experimentů pro parametrizaci LFE $\Delta\Delta$ nad datovou sadou ŠkodaAuto.	40

11	Výsledky experimentů pro parametrizaci MFCC nad datovou sadou ŠkodaAuto.	40
12	Výsledky experimentů pro parametrizaci MFCC $\Delta\Delta$ nad datovou sadou ŠkodaAuto.	41
13	Výsledky experimentů pro parametrizaci LFE nad datovou sadou SpeechDat-E.	42
14	Výsledky experimentů pro parametrizaci LFE $\Delta\Delta$ nad datovou sadou SpeechDat-E.	42
15	Výsledky experimentů pro parametrizaci MFCC nad datovou sadou SpeechDat-E.	43
16	Výsledky experimentů pro parametrizaci MFCC $\Delta\Delta$ nad datovou sadou SpeechDat-E.	43
17	Výsledky experimentů pro datovou sadu ŠkodaAuto.	44
18	Výsledky experimentů pro datovou sadu SpeechDat-E.	44
19	Porovnání přesnosti dekódování (phoneAcc [%]) mezi datovou sadou ŠkodaAuto a SpeechDat-E.	46

Reference

- [1] M. Majer. Detekce klíčových frází. Bakalářská práce, Západočeská univerzita v Plzni, 2016.
- [2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [3] A. Karpathy. Cs231n convolutional neural networks for visual recognition, 2016. [Online], Dostupné z: <http://cs231n.github.io>.
- [4] Ch. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [5] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.
- [6] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 22(10):1533–1545, 2014.
- [7] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*. Springer, 2015. ISBN: 978-1-461-47137-0.
- [8] T. Parr and J. Howard. The matrix calculus you need for deep learning. abs/1802.01528, 2018.
- [9] C. Olah. Neural networks, types, and functional programming, 2015. [Online], Dostupné z: <http://colah.github.io/posts/2015-09-NN-Types-FP>.
- [10] C. Olah. Understanding lstm networks, 2015. [Online], Dostupné z: <http://colah.github.io/posts/2015-08-Understanding-LSTMs>.
- [11] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997.
- [12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [13] S. Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

- [14] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. *CoRR*, abs/1212.0901, 2012.
- [15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [17] A. Hannun. Sequence modeling with ctc. *Distill*, 2017. <https://distill.pub/2017/ctc>.
- [18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, pages 369–376, New York, NY, USA, 2006. ACM.
- [19] M. Majer. Diplomová práce - GitHub repozitář, 2018. https://github.com/MajerMartin/phoneme_classification.
- [20] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013.