BEAT TRACKING SYSTEM BASED ON A NEURAL NETWORK

Tomáš Suchánek
Master Degree Programme (2), FEEC BUT
E-mail: xsucha13@vutbr.cz

Supervised by: Matěj Ištvánek
E-mail: xistva02@stud.feec.vutbr.cz

Abstract: This thesis deals with systems for tempo and beat detection in music recordings, whose functionality is based on neural networks. The basic structure of such systems is briefly described and the emphasis is then placed on a comparison of recurrent and temporal convolutional networks, which have proven to be the most suitable for this task. The main outcome of this work is then proposal and comparison of modified temporal convolutional network with other state-of-the-art networks in a beat tracking system. The results suggest that simplification in existing architectures could benefit from faster training times, while it maintains or slightly improves the accuracy of a detection system.

Keywords: Beat tracking, machine learning, neural network, signal processing

1 ÚVOD

Detekce tempa a dob je jednou z podoblastí vědního oboru *Music Information Retrieval* (MIR), který si klade za cíl extrakci dále uplatnitelných informací z hudebních nahrávek. Ačkoli lze tuto extrakci s velkou spolehlivostí provádět manuálně, je v kontextu posledních let a množství dat vytvářen tlak na co nejrychlejší a nejpřesnější automatizaci těchto postupů. S navyšující se dostupností výpočetní techniky se pak dostávají do popředí neuronové sítě, s jejichž možnostmi lze počítačovým systémům mj. vtisknout právě schopnost identifikace rytmu a hudebních dob. Získané informace se pak uplatňují při synchronizaci moderních technologií na specifické události v hudbě, při doporučování obsahu u streamovacích služeb, nebo v neposlední řadě také při porovnávání různých interpretací totožných skladeb. Tato práce si stanovuje za cíl výzkum dosavadních přístupů a zaměřuje se na efektivitu různých neuronových sítí v těchto systémech. Vlastní systém s různými architekturami sítí je také výsledkem a předmětem závěrečné diskuze této práce.

2 STRUKTURA DETEKČNÍCH SYSTÉMŮ

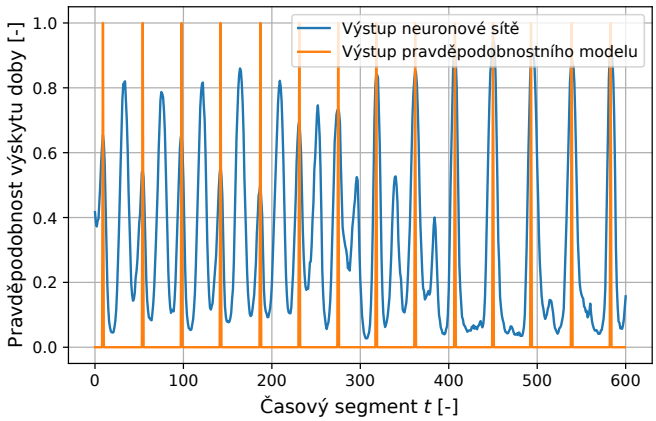
Proces detekce dob ve své podstatě vychází z myšlenky, že skladba ve frekvenčním spektru obsahuje opakující se vzorec zvukových událostí, na jejichž základě lze tempo a doby odvodit. Cílem systému je tedy nalezení těchto vzorců a odvození nejpravděpodobnější sekvence hudebních dob, které takové vzorce vysvětlují. Zvukový signál je pak za tímto účelem zpracováván zpravidla ve třech bodech.

Nejprve je provedena jeho transformace do časově-frekvenční reprezentace, nejčastěji mel spektrogramu, který v porovnání s běžným spektrogramem blíže simuluje nelinearitu a maskování lidského slyšení. Za účelem zisku robustnější informace o vývoji spektra v čase jsou pak mel spektrogramy s různými délkami segmentů kombinovány do jediné časově-frekvenční reprezentace, která slouží jako vstup do další části systému.

V té již figurují neuronové sítě, jejichž úkolem je na základě předem daných časových anotací dob každé skladby vyvodit vztah mezi těmito anotacemi a událostmi ve zvukovém spektru, přičemž jejich následným výstupem je vyjádření pravděpodobnosti, s jakou se v každém časovém segmentu vstupní

reprezentace vyskytuje doba. Ačkoli je zřejmé, že z výsledných dat lze pomocí algoritmů na hledání lokálních maxim sestavit žádaný výstup, může takový proces vzhledem k určité nejistotě neuronové sítě u různých skladeb vykazovat horší přesnost v důsledku falešných detekcí.

Tento problém aktuálně nejúspěšněji eliminuje využití pravděpodobnostních modelů [1], s jejichž vhodným přizpůsobením lze snadněji identifikovat i změny v tempu nebo metrické struktuře skladby. Tato práce, podobně jako mnohé další, využívá algoritmus založený na skrytých Markovových modelech. Ten v předem nastavených podmínkách, jako je například maximální tempové zrychlení mezi sousedními dobami, vyvozuje nejpravděpodobnější sekvenci dob, která těmto podmínkám a vstupním datům daných neuronovou sítí vyhovuje. Obrázek 1 ukazuje příklad výstupu sítě a odvození konečných pozic dob pravděpodobnostním modelem.



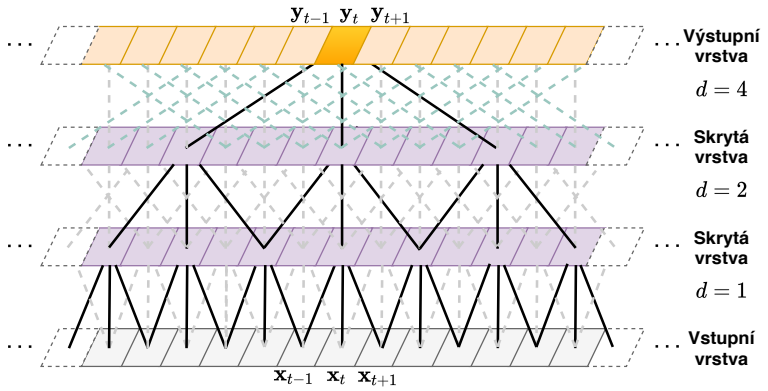
Obrázek 1: Výstup neuronové sítě a pravděpodobnostního modelu.

2.1 Využívané neuronové sítě

Požadavky na neuronové sítě pro efektivní detekci lze shrnout dvěma body. Prvním z nich je možnost zpracování datové posloupnosti libovolné délky, druhým je možnost vytváření interních souvislostí mezi časově proměnnými událostmi. Pro tyto účely se již v jiných pracích [2, 3] prokázaly jako nejefektivnější obousměrné rekurentní sítě s buňkami LSTM a temporální konvoluční sítě TCN. První jmenované si prostřednictvím zpětných vazeb vytvářejí paměť, která je pro detekci dob zásadní, přičemž struktura LSTM buněk tuto paměť dále navyšuje. TCN sítě pak paměť nemají, nicméně jejich architektura disponuje konvolučními filtry, které se s každou další vrstvou sítě rozšiřují. Hodnota výstupní vrstvy pro určitý časový segment vstupní reprezentace pak v důsledku závisí i na okolních vstupních datech, jejichž pomyslné časové rozpětí je definováno právě hloubkou sítě a počáteční šířkou konvolučního filtru. Princip této sítě je blíže vyjádřen na obrázku 2. Tyto sítě se také zásadně odlišují procesem učení – zatímco u rekurentních sítí je chyba na výstupu zpětně šířena přes všechny časové kroky vstupní reprezentace, u sítě konvolučního typu je chyba šířena pouze přes pevně daný počet vrstev a jejich parametrů, z čehož plyne markantní zrychlení učícího se procesu především u delších nahrávek.

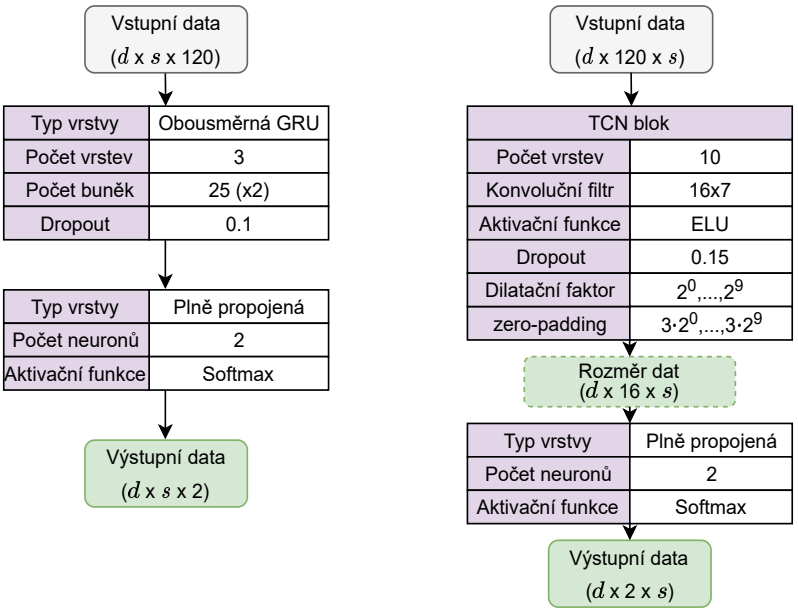
2.2 Návrh systému

V rámci práce byl vytvořen systém, ve kterém byly pro další srovnání testovány čtyři architektury neuronových sítí. Algoritmus jako takový sdílí pro každou síť stejnou fázi předzpracování vstupních dat a také konečné odvození sekvence dob, které vychází z prací [1, 3] a je dostupné v knihovně



Obrázek 2: Struktura TCN sítě. Černé čáry vyznačují závislosti prvků mezi sousedními vrstvami. Na příkladu je použit konvoluční filtr o šířce $k = 3$ s dilatačními faktory $d = 1, 2, 4$, čímž se rozumí roztažení filtru v každé z dalších vrstev.

madmom pro jazyk Python. Vstupní zvuková data jsou převedena do tří mel spektrogramů s délkami segmentů $N = 1024, 2048, 4096$, fixním posunem analyzačního okna $h = 441$ a $M = 20$ mel frekvenčními pásmy na jeden segment. Každý mel spektrogram je dále zvláště filtrovaný mediánovým filtrem s posuvným oknem o délce $N_{med} = N/100$, přičemž výsledná reprezentace se dále odečítá od té původní. Jednotlivé spektrogramy jsou poté spojeny do jediné reprezentace, která tímto postupem získává rozměr $M = 120$ mel frekvenčních pásem s časovým rozlišením 100 fps při $f_{vz} = 44,1$ kHz. Mezi testované sítě pak patří dvě replikované architektury z prací [2, 3] (dále CNNTCN a BLSTM) a dvě vlastní modifikované architektury (dále BGRU a STCN), jejichž strukturu popisuje obrázek 3. Rozdíl mezi BGRU a BLSTM tkví pouze v typu buněk, CNNTCN od STCN pak odlišuje kromě jiného nastavení vnitřních parametrů také předcházející 2D konvoluční blok, který slouží k dalšímu předzpracování vstupních dat.



Obrázek 3: Architektura sítí BGRU (vlevo) a STCN (vpravo). Počet vzorů jedné vstupní dávky označuje d, jejich délku pak značí s. Výstupní data jsou dále vstupem pravděpodobnostního modelu.

3 VÝSLEDKY

Všechny sítě byly trénovány na datasetu čítajícím 220 skladeb o délce 30 s pomocí 5složkové křížové validace, přičemž celý proces byl třikrát opakován, získané výsledky byly zprůměrovány a jsou vypsány v tabulce 1. Pro vytvoření základního náhledu na výkonnost systému byla použita nejběžnější metrika F-score, přičemž hodnota F-score = 1 vyjadřuje absolutní shodu detekce s příslušnou anotací skladby. Z výsledků vyplývá, že navržená síť STCN dosahuje srovnatelného či mírně lepšího skóre než další testované sítě a to při kratší době zpracování jedné iterace vzorových dat při učení, což je výhodné při učení sítě na velkých datových sadách. Z jiných prací se stejným zaměřením však vyplývá, že různé sítě mohou vykazovat různé skóre na např. žánrově odlišných datasetech a pro potvrzení nebo naopak vyvrácení dosažených výsledků je zapotřebí testování zopakovat na větším množství dat. Je také vhodné podotknout, že časy učení byly prozatím dosaženy na jednotce CPU a v případě využití grafické karty budou časy nadále menší a mohou se do jisté míry poměrově lišit.

Tabulka 1: Srovnání přesnosti detekčního systému s různými neuronovými sítěmi.

Síť	F-score	Doba učení
BLSTM	0,664	10,91 s/it
BGRU	0,653	10,86 s/it
CNNTCN	0,681	4,13 s/it
STCN	0,685	1,36 s/it

4 ZÁVĚR

Práce shrnuje zásadní poznatky o systémech na detekci dob založených na využití neuronové sítě a zprostředkovává vhled do přesnosti systému s různými síťovými architekturami. Praktickým výstupem práce je funkční algoritmus, pomocí kterého pak byly jednotlivé sítě testovány a vyhodnoceny, přičemž jedna z navržených sítí vykazuje konkurenceschopnou přesnost v porovnání se zavedenými state-of-the-art architekturami při rychlejším čase jejího učení. V další fázi práce bude zkoumán vliv úprav předzpracování dat na přesnost detekce a sítě budou podrobeny rozsáhlejšímu testování na robustnější datové sadě.

REFERENCE

[1] KREBS, Florian, Sebastian BOCK a Gerhard WIDMER. *Rhythmic pattern modeling for beat and downbeat tracking in musical audio* [online]. 2013 [cit. 2020-11-11]. Dostupné z: `http://phenicx.upf.edu/system/files/publications/Krebs_ISMIR_2013.pdf`

[2] DAVIES, Matthew a Sebastian BOCK. *Temporal convolutional networks for musical audio beat tracking*. European Signal Processing Conference (EUSIPCO) [online]. 2019 [cit. 2020-12-05]. Dostupné z: `http://telecom.inesctec.pt/~mdavies/pdfs/DaviesBoeck19-eusipco.pdf`

[3] BOCK, Sebastian, Florian KREBS a Gerhard WIDMER. *Joint beat and downbeat tracking with recurrent neural networks*. 17th International Society for Music Information Retrieval Conference [online]. 2016 [cit. 2021-03-11]. Dostupné z: `http://www.cp.jku.at/research/papers/Boeck_etal_ISMIR_2016.pdf`