UČENÍ SKRYTÉ STRUKTURY HUDBY: REPREZENTACE SKLADEB ZALOŽENÁ NA AUTOENKODÉRECH

Matouš Kovář
Faculty of Information Technology
Czech Technical University
Prague
kovarmat@fit.cvut.cz

ABSTRACT

Tato studie zkoumá učení sémanticky významných reprezentací hudebního signálu s využitím architektur neuronových sítí na datech z datasetu GTZAN. Úryvky jsou převedeny na Mel-spektrogramy s cílem prozkoumat, jak lze skryté struktury v hudbě mapovat do latentního prostoru. Porovnáváme tři odlišné přístupy: standardní konvoluční autoenkodér (CAE) optimalizovaný pro rekonstrukci signálu, model využívající Triplet Loss k vynucení shlukování na základě žánrů a hybridní konvoluční rekurentní neuronovou síť (CRNN) navrženou pro zachycení časového vývoje. Kvalita naučených embeddingů je vyhodnocena kvantitativně pomocí přesnosti klasifikace algoritmem K-nejbližších sousedů (KNN) a kvalitativně prostřednictvím vizualizace metodou UMAP.

Keywords CNN · CRNN · Embedding · Autoencoder · Machine learning · Audio processing · Spectrogram

1 Úvod

Zpracování obrazu pomocí konvolučních neuronových sítí je dnes ve strojovém učení standardní praxí. Zpracování zvuku je však o něco složitější, protože zvuk se mění v čase. Nejefektivnějším způsobem, jak se s tím vypořádat, je obvykle převod zvukových signálů na 2D **Mel-spektrogramy**—vizuální reprezentace zvuku—což nám umožňuje zpracovávat hudbu téměř stejně jako obrazy.

Tato práce se zaměřuje konkrétně na **učení autoenkodérů a embeddingů**. Naším cílem je natrénovat autoenkodér tak, aby objevil sémantickou strukturu hudby. Kompresí skladeb do embeddingů můžeme shlukovat podobné žánry nebo vytvořit základ pro doporučovací systém.

Navrhujeme a porovnáváme tři architektury: standardní konvoluční autoenkodér, verzi vylepšenou o **Triplet Loss** pro vynucení lepšího shlukování a model **CRNN**, který přidává rekurentní vrstvy pro lepší pochopení časového vývoje skladby.

2 Související práce

Komplexní přehled o metodách hlubokého učení v doméně zpracování audia podávají [1]. Ve své studii srovnávají různé vstupní reprezentace a uvádějí, že ačkoliv se v poslední době experimentuje se zpracováním surového signálu, log-mel spektrogramy zůstávají dominantní a výpočetně efektivní volbou. V kontextu architektur autoři upozorňují na limitace standardních konvolučních sítí, které jsou omezeny fixní velikostí receptivního pole, a nemusí tak zachytit dlouhodobé závislosti v hudbě. Jako efektivní řešení vyzdvihují hybridní modely typu CRNN, které využívají konvoluční vrstvy pro extrakci lokálních příznaků a následně rekurentní vrstvy pro integraci informací v čase.

V článku [2] se zabývají zpracováním lidského hlasu a hlavním předzpracovacím krokem je opět převod na spektrogram. Tento projekt používá pouze kombinaci konvolučních a plně propojených vrstev.

Článek [3] srovnává CNN (Convolutional neural network) a CRNN mezi sebou na datasetu Million songs dataset [4]. Modely, které jsou porovnávány co do schopnosti klasifikace jsou zobrazeny na snímku 1. CRNN využívají GRU jako RNN blok. Ve studii se ukazuje jako výkonnější model CRNN, i pro případy kde je konvoluční část modelů totožná.

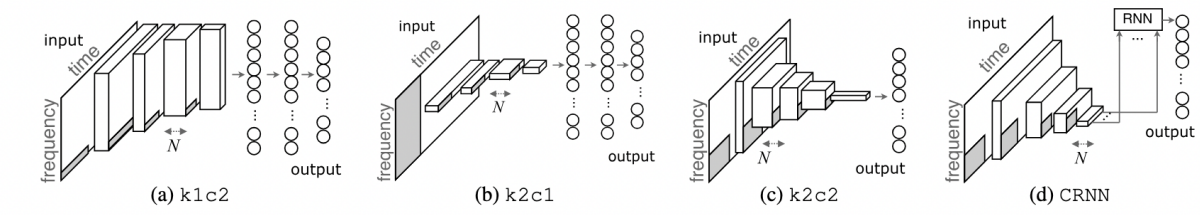

Figure 1: Porovnávané modely v [3].

[5] se zabývá Triplet loss ztrátovou funkcí a její aplikací na neolabelovaná data. Navrhuje některé způsoby jak získat pozitivní a negativní sample pro nějaký anchor například pomocí rozdělení písně, nebo zavedení šumu. Triplet loss je vyzdvihována jako funkce tlačící na umístění podobných písní do prostoru blízko sebe.

3 Metodika

V této sekci popisujeme postup zpracování dat a architektury navržených modelů.

3.1 Předzpracování dat

Vstupní data jsou zpracována do formy Mel-spektrogramů2. Pomocí STFT převádíme surový audio signál na frekvenční spektrum. Hodnoty jsou převedeny na decibely (logaritmická škála).

Spektrogramy jsou generovány s parametry vzorkovací frekvence $f_s = 22050$ Hz a počtem Mel filtrů $n_{mels} = 128$. Před vstupem do sítě jsou data normalizována metodou Min-Max do intervalu $[0, 1]$, což umožňuje použití aktivační funkce Sigmoid na výstupu sítě.

Výsledkem předzpracování je pro každý úryvek matice o fixních rozměrech 128×1292 bodů, kde svislá osa reprezentuje frekvenční pásma (Mel bands) a vodorovná osa časové rámce 2.

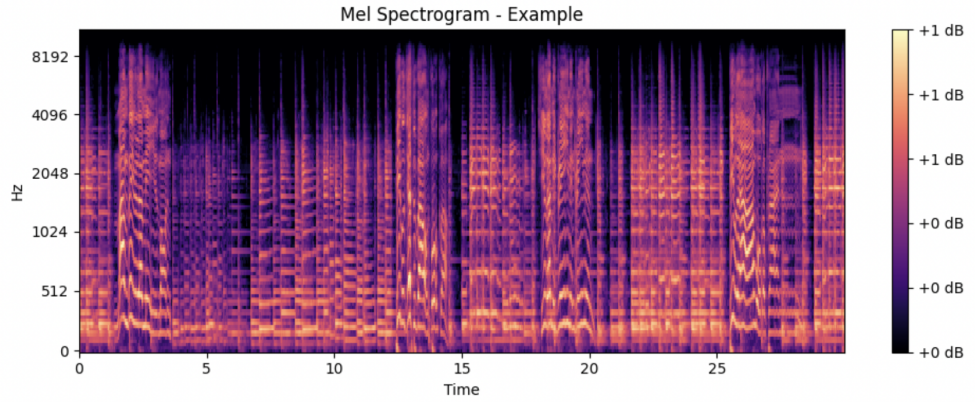

Figure 2: Ukázka Mel-spektrogramu jedné z písní.

3.2 Ztrátové funkce

Pro trénování využíváme rekonstrukční chybu (MSE), Triplet loss, nebo jejich váženou kombinaci.

Mean Squared Error (MSE): Zajišťuje, že embedding nese dostatek informací pro rekonstrukci skladby.

$$\mathcal{L}_{MSE}(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \tag{1}$$

Triplet Loss: Vynucuje shlukování podle žánrů. Minimalizuje vzdálenost mezi Anchor (a) a Positive (p) vzorkem a maximalizuje vzdálenost k Negative (n) vzorku.

$$\mathcal{L}_{Triplet}(a, p, n) = \max\{d(a, p) - d(a, n) + \alpha, 0\} \tag{2}$$

kde:

- a (**Anchor**): Referenční vzorek.
- p (**Positive**): Vzorek se **shodnou identitou** jako Anchor.
- n (**Negative**): Vzorek s **odlišnou identitou** od Anchoru.

Kombinovaná ztrátová funkce:

$$\mathcal{L}_{Total} = \mathcal{L}_{MSE} + \lambda \cdot \mathcal{L}_{Triplet} \tag{3}$$

Experimentálně byla váha nastavena na $\lambda = 0.01$. Po přeškálování to odpovídá, že MSE je o řád významnější než Triplet Loss. Ukázalo se, že samotná Triplet Loss vede k horší sémantice embeddingů, zatímco přidání MSE působí jako regularizace a nutí model soustředit se na celou podobu skladby.

3.3 Architektury modelů

Základem všech modelů je konvoluční neuronová síť využívající aktivace ReLU a volitelně Batch Normalization pro stabilizaci trénování.

3.3.1 Konvoluční Autoenkodér (Baseline)

Architektura typu Encoder-Decoder. Enkodér snižuje dimenzi vstupu do latentního vektoru, dekodér se snaží o rekonstrukci. Tento model slouží jako referenční bod. Cílem je zjistit jak dobré embeddingy — co do sémantického významu — jsme schopni získat bez jakékoliv další znalosti o datasetu.

3.3.2 Konvoluční Autoenkodér s MSE + Triplet Loss

Tento model vychází ze stejné architektury jako základní konvoluční autoenkodér (viz 3.3.1), ale zásadně se liší v trénovací strategii. Zatímco baseline model pracuje zcela bez učitele, tento přístup využívá dostupnou znalost o doméně – labely.

Cílem je vytvořit model, který nejen rekonstruuje vstup, ale hlavně strukturuje latentní prostor tak, aby respektoval sémantickou podobnost skladeb. Toho je dosaženo kombinací dvou ztrátových funkcí:

1. **MSE Loss:** Nutí enkodér zachovat dostatek informací pro rekonstrukci spektrogramu.
2. **Triplet Loss:** Nutí enkodér minimalizovat vzdálenost mezi skladbami stejného žánru a maximalizovat vzdálenost mezi odlišnými žánry.

Tento přístup představuje robustní alternativu. Významnou předností architektury je potenciál pro **semi-supervizované učení**: model dokáže efektivně těžit i z neanotovaných dat (optimalizací rekonstrukční chyby), zatímco penalizace pomocí Triplet Loss je aplikována selektivně pouze na vzorky se známou identitou.

3.3.3 CRNN Autoenkodér s MSE + Triplet Loss

Tato architektura kombinuje CNN pro extrakci příznaků a rekurentní blok (`GRU`) pro zachycení časových závislostí. RNN umožňuje modelu vnímat skladbu jako sekvenci a lépe pochopit kontext, což čistě konvoluční sítě postrádají.

3.4 Způsoby evaluace

Kvalita embeddingů není měřena hodnotou Loss funkce, ale jejich užitečností pro klasifikaci:

- **Kvantitativní (KNN):** Trénovací embeddingy jsou použity pro natrénování KNN klasifikátoru ($k = 10$). Měříme přesnost (Accuracy) na validační sadě.
- **Kvalitativní (UMAP):** Projekce do 2D prostoru pro vizuální kontrolu separace shluků.

4 Výsledky

Pro každý druh modelu je vytvořeno 5 různých konfigurací, které se liší v nastavení atributů sítě.

4.1 Kvantitativní evaluace

V experimentech se ukázalo, že autoenkodér, který se naučil nejlepší sémantické embeddingy písní je struktura používající CRNN s použitím Triplet loss a MSE. Struktura autoenkodéru je možná vidět na obrázku 3.

Jak si stály všechny modely, je možné vidět v tabulce 1

Table 1: Kvantitativní porovnání přesnosti klasifikace (KNN, $k = 10$) v latentním prostoru pro jednotlivé experimentální konfigurace.

Typ Modelu	Konfigurace	KNN Accuracy
1. Konvoluční Autoenkodér (Baseline - MSE)		
	Config 1 (Baseline)	**0.485**
	Config 2 (Větší jádro a latentní dimenze)	0.455
	Config 3 (Velký stride)	0.420
	Config 4 (Hluboká, malá latentní dimenze)	0.410
	Config 5 (Bez Batch norm)	0.440
2. Konvoluční AE s Triplet Loss (MSE + Triplet)		
	Config 1 (Baseline)	**0.885**
	Config 2 (Větší jádro a latentní dimenze)	0.840
	Config 3 (Velký stride)	0.850
	Config 4 (Hluboká, malá latentní dimenze)	0.655
	Config 5 (Bez Batch norm)	0.695
3. Hybridní CRNN Enkodér (MSE + Triplet)		
	Config 6 (CRNN Baseline)	0.810
	Config 7 (Větší kernel)	**0.915**
	Config 8 (Větší výstupní dimenze RNN)	0.865
	Config 9 (Baseline s větším jádrem)	0.875
	Config 10 (Malá latentní dimenze)	0.880

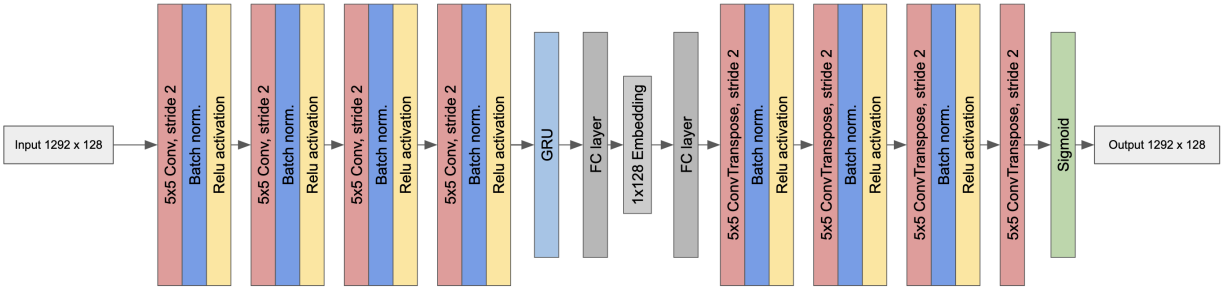

Figure 3: Architektura nejlepší CRNN.

4.2 Kvalitativní evaluace

Na následujících snímcích 4 5 6 jsou vizualizovány embeddingy naučené nejlepšími modely v jednotlivých kategoriích pomocí metody UMAP.

Z vizuální analýzy pomocí metody UMAP je patrné, že embedding naučený základním autoenkodérem (pouze MSE) nenese v projekci do dvou dimenzí žádnou zjevnou sémantickou informaci. Body reprezentující různé žánry jsou v prostoru chaoticky promíchány, což potvrzuje neschopnost modelu zachytit abstraktní hudební rysy bez učitele.

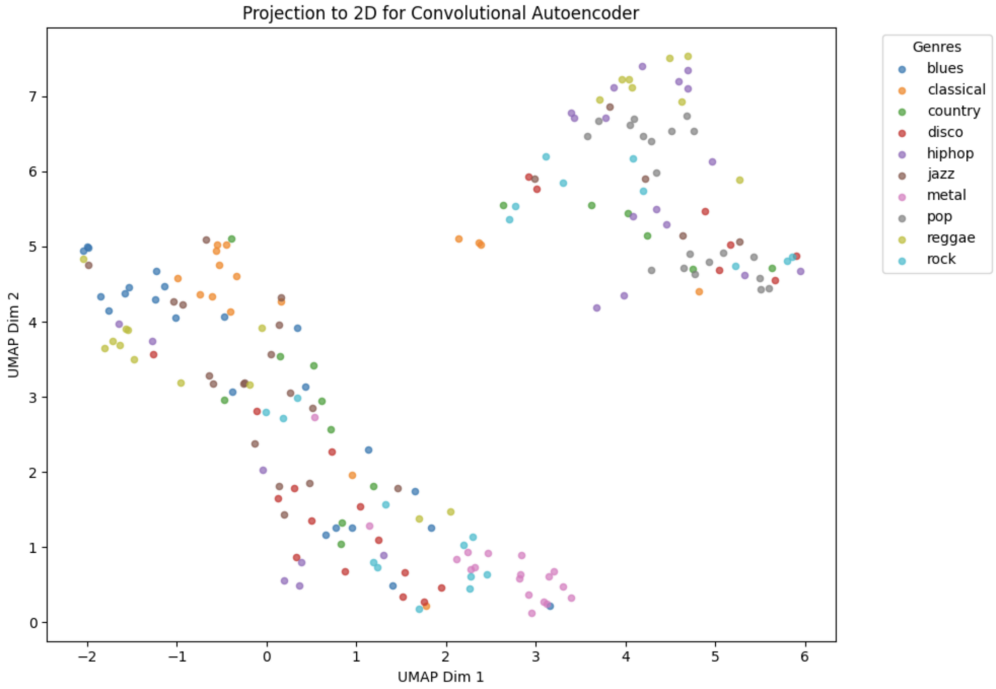

Figure 4: Embedding naučený se základním autoenkodérem.

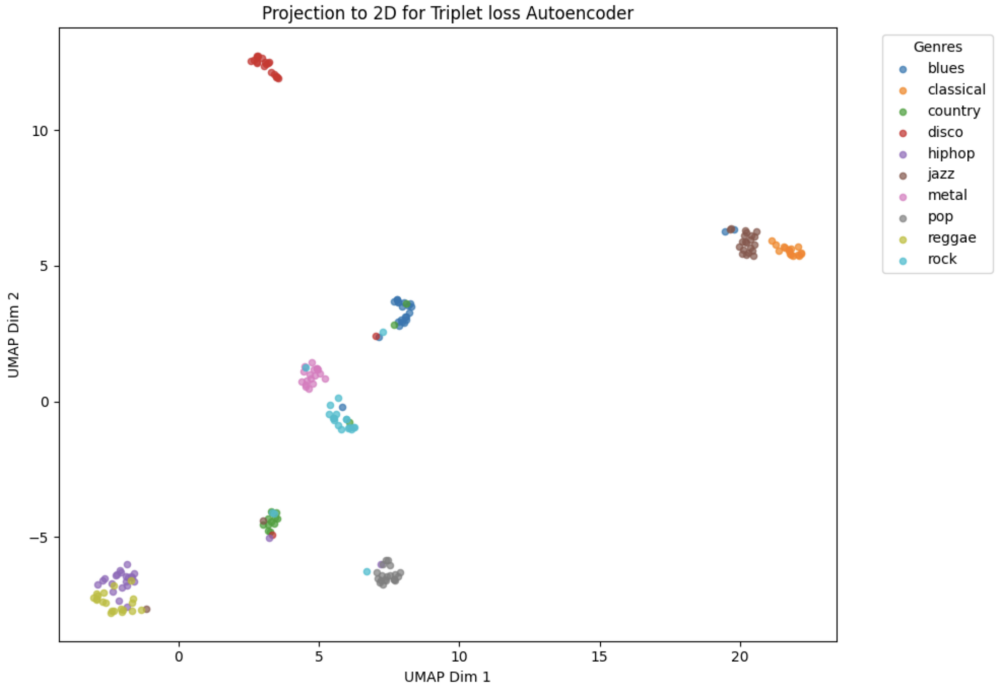

Figure 5: Embedding naučený se základním autoenkodérem se znalostí labelů.

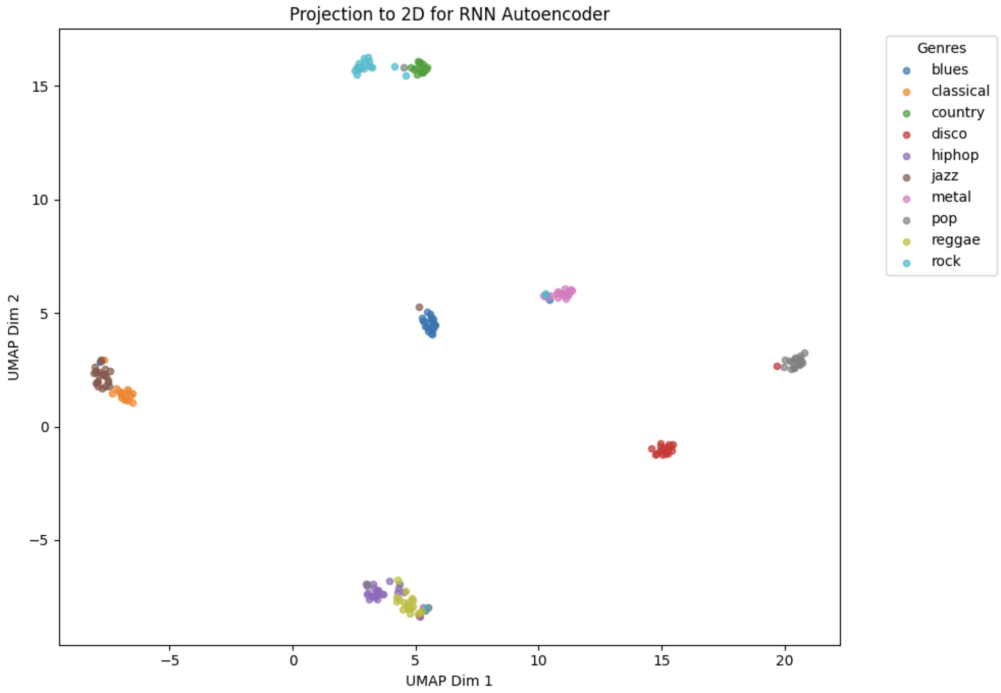

Figure 6: Embedding naučený s CRNN.

Naproti tomu modely využívající Triplet loss vykazují výrazně lepší strukturu latentního prostoru s jasně oddělenými shluky. Z kvalitativního hlediska dosahují modely s čistě konvoluční architekturou i modely s CRNN velmi podobných výsledků.

Důležitým pozorováním je zachování logických vztahů mezi žánry. Například akusticky a instrumentálně příbuzné kategorie, jako jsou *Classical* a *Jazz*, jsou v latentním prostoru projektovány do vzájemné blízkosti. To naznačuje, že model se naučil nejen rozlišovat kategorie, ale i chápat hlubší akustické souvislosti.

5 Diskuze

V této práci jsme zkoumali vliv různých architektur a trénovacích strategií na kvalitu naučených reprezentací hudebních skladeb.

5.1 Vliv architektury na kvalitu embeddingu

Naše experimenty jednoznačně ukázaly limity standardních konvolučních autoenkodérů (CAE) trénovaných pouze na rekonstrukční chybu. Přesnost klasifikace okolo 48 % na validační množině naznačuje, že vizuální podobnost spektrogramů neodpovídá sémantické podobnosti hudebních žánrů. Model se v tomto případě naučil efektivně komprimovat signál, avšak bez schopnosti extrahovat abstraktní hudební rysy.

Naproti tomu zavedení architektury CRNN s kombinovanou ztrátovou funkcí vedlo k nejlepším výsledkům (91,5 %) na validační množině. Tento výsledek potvrzuje předpoklad, že hudbu nelze vnímat pouze jako statickou texturu. Zatímco konvoluční vrstvy úspěšně extrahovaly timbrální vlastnosti (barvu zvuku), přidaná rekurentní vrstva (GRU) umožnila modelu integrovat informace v čase a zachytit tak rytmické struktury a dlouhodobý kontext, který je pro rozlišení žánrů klíčový.

Co se týče kvalitativního porovnání, tak si modely s použitím kombinované ztrátové funkce s a bez RNN vedou velmi podobně.

5.2 Význam kombinované ztrátové funkce

Kritickým poznatkem této práce je nutnost kombinace rekonstrukční a metrické ztrátové funkce. Nutí tak síť učit se i sémantický význam embeddingů.

Zavedení kombinované ztrátové funkce ($\mathcal{L}_{MSE} + \mathcal{L}_{Triplet}$) se ukázalo jako výhodné. Rekonstrukční složka (MSE) zde funguje jako regularizátor, který nutí enkodér zachovat v latentním prostoru dostatek informací o obsahu skladby.

5.3 Limitace a budoucí práce

Je nutné zmínit, že dosažené výsledky jsou ovlivněny velikostí a kvalitou datasetu GTZAN. S pouhými 1000 skladbami je hluboké učení náchylné k overfittingu.

Podle mého názoru by ale nejzajímavějším předmětem dalšího zkoumání byla úprava ztrátové funkce, v této práci se tvorbou příliš nezabýváme, používáme poznatky získané při prvotním učením. Nicméně využití jiných ztrátových funkcí a jiného váženého součtu by mohlo vést k lepším výsledkům.

V kontextu současných trendů se jako logický další krok jeví využití architektur založených na Audio Spectrogram Transformers (AST). V této práci nebyly zahrnuty primárně z důvodu jejich vysoké výpočetní náročnosti a potřeby výrazně větších trénovacích datasetů, než je GTZAN.

6 Závěr

V této práci jsme se zabývali učením sémantických reprezentací hudebních nahrávek. Cílem bylo navrhnout architekturu neuronové sítě, která dokáže transformovat Mel-spektrogramy do kompaktního latentního prostoru, ve kterém jsou hudební žánry přirozeně separovány.

Implementovali a porovnali jsme tři různé přístupy: standardní konvoluční autoenkodér, model založený na kombinaci Triplet Loss a MSE a hybridní CRNN model kombinující konvoluční a rekurentní vrstvy.

Na základě provedených experimentů na datasetu GTZAN jsme došli k následujícím závěrům:

1. Samotná rekonstrukce signálu (pomocí MSE) nevede k naučení sémanticky užitečných embeddingů. Základní autoenkodér dosáhl klasifikační přesnosti pouze 48,5 %.
2. Faktorem zlepšující úspěšnost je kombinace rekonstrukční a metrické ztrátové funkce.
3. Nejlepší architektura pro tento úkol je hybridní CRNN (přesnost 91,5 %). Prokázalo se, že zapojení rekurentních vrstev (GRU) je vhodné pro zachycení časového vývoje a rytmických struktur, které čistě konvoluční sítě ignorují.

Výsledná práce demonstruje, že i na relativně malém datasetu lze pomocí moderních technik hlubokého učení natrénovat model, který porozumí obsahu hudby a dokáže sloužit jako základ pro pokročilé doporučovací systémy.

Seznam použitých zkratek

CNN Convolutional Neural Network (Konvoluční neuronová síť)

RNN Recurrent Neural Network (Rekurentní neuronová síť)

CRNN Convolutional Recurrent Neural Network

GRU Gated Recurrent Unit

MSE Mean Squared Error (Střední kvadratická chyba)

KNN K-Nearest Neighbors (K-nejbližších sousedů)

STFT Short-Time Fourier Transform

UMAP Uniform Manifold Approximation and Projection

7 Obsah přiloženého archivu

Součástí odevzdané práce je elektronický archiv s následující adresářovou strukturou:

```
.
README.md           # Milestone
article.pdf         # Text této semestrální práce ve formátu PDF
mvi.ipynb           # Jupyter Notebook stažený z Google Colab
models.txt          # Odkaz na Google drive se složkou s modely
spectrograms.txt    # Odkaz na Google drive se složkou s spektrogramy
```

Poznámka: Původní dataset GTZAN není z důvodu své velikosti součástí přiloženého archivu. Pro ověření výsledků a inferenci lze využít přiložené předvypočítané spektrogramy ve složce `spectrograms/`.

References

[1] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schluter, Shuin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019.

[2] Harsh Sinha, Vinayak Awasthi, and Pawan K. Ajmera. Audio classification using braided convolutional neural networks. *IET Signal Processing*, 14(7):448–454, 2020.

[3] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[4] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[5] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P. W. Ellis, Shawn Hershey, Jiyang Liu, R. Channing Moore, and Rif A. Saurous. Unsupervised learning of semantic audio representations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 126–130. IEEE, 2018.