FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

WaveNet - nová metoda syntézy řeči

Jakub Vít[1]

1 Úvod

Hluboké neuronové sítě zažívají ohromný rozmach. Přestože je koncept znám už více než padesát let, jejich potenciál se odkrývá díky stále se zvyšujícímu výpočetnímu výkonu až nyní. Neuplyne týden, kdy by nebyla představena neuronová síť, která zvládla vyřešit problém, jež se až doteď zdál neřešitelný a to hlavně v oblastech umělé inteligence. Jejich použití se stále rozšiřuje do více a více stávajících problémů, kde nahrazují dosavadní state of the art algoritmy. Ani syntéza řeči nezůstala ušetřena.

2 WaveNet

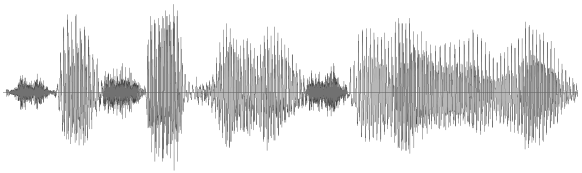
Po mnoho let existovaly dva ustálené přístupy v syntéze řeči. Asi nejpoužívanější je metoda konkatenační, která je založená na spojování úseků existující řeči. Protipólem jsou parametrické metody, které se snaží modelovat řečové parametry, z nichž se pak pomocí filtru a buzení (tzv. vokodéru) generuje lidská řeč. Pro generování parametrů se dlouhou dobu používaly skryté markovské modely. Ty byly postupně nahrazeny rekurentními neuronovými sítěmi. Slabinou parametrických metod ale stále zůstává použití vokodéru.

Skutečná revoluce přišla až koncem minulého roku, kdy společnost DeepMind vlastněná Googlem představila v článku síť WaveNet. WaveNet je neuronová síť, která dokáže generovat lidskou řeč s mnohem vyšší kvalitou než parameterické metody a bez artefaktů, které vznikají při řetězení signálu v konkatenační metodě. Asi nejzajímavější je však způsob, jak funguje. Tato síť totiž generuje řečový signál vzorek po vzorku, což pro audio bývá 16 tisíc vzorků za sekundu. Tento postup je obdivuhodný, neboť řečový signál je velmi komplexní a složitý (obr. 1). Takový přístup se dříve zdál být nemožný. Vstupem sítě je její předchozí výstup, takže je tzv. autoregresivní. Součástí vstupu jsou také podmiňující lingvistické příznaky, které docílí, že vygenerovaná řeč obsahuje zadaný text.

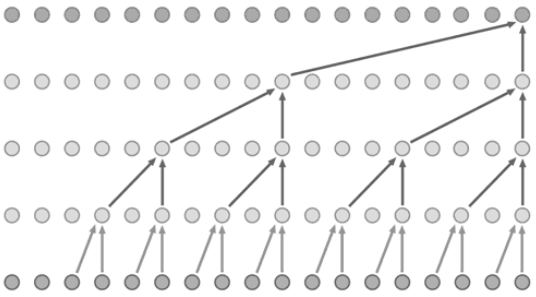
3 Architektura

WaveNet je hluboká konvoluční neuronová síť. Pro zvýšení receptivního pole se používá tzv. dilatovaná konvoluce (obr. 2), kde s každou další vrstvou roste dilatace o dvojnásobek. Pro urychlení trénování jsou použity tzv. skip connections (obr. 3), které byly představeny rovněž nedávno. Síť negeneruje konkrétní hodnotu následujícího vzorku, ale generuje histogram rozložení. Z tohoto rozložení se poté hodnota následujícího vzorku určí náhodným výběrem. Tato síť tak vlastně neprovádí regresi ale klasifikaci (obr. 4). Vzorky audia se standardně ukládají pomocí 16 bitového kódování. Jelikož klasifikace do tolika tříd by nebyla možná, používá se 8-bitové μ-law kódování.

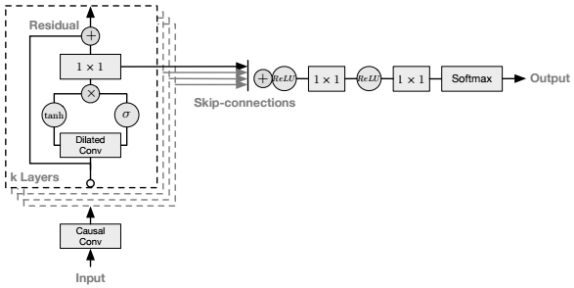
[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Umělá inteligence, e-mail: jvit@students.zcu.cz



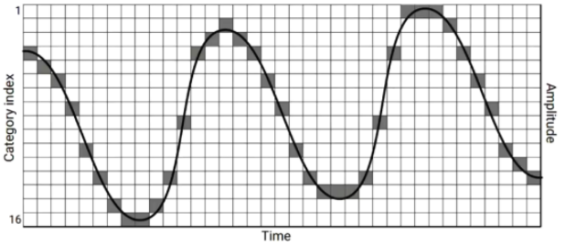
Obrázek 1: Vygenerovaný audio signál



Obrázek 2: Dilatovaná konvoluce



Obrázek 3: Architektura sítě



Obrázek 4: Kategorizace vstupu

4 Realizace

Pro implementaci jsem zvolil framework TensorFlow, což je open-source knihovna pro numerické výpočty od společnosti Google. Dnes je to asi nejpopulárnější framework pro neuronové sítě. Pro kvalitní syntézu je nutné trénovat několik miliónů velmi citlivých parametrů a to pro 20 hodin záznamu. K tomu je nutný velký výpočetní výkon. Trénování jsem proto prováděl na gridové infrastruktuře MetaCentrum. Samotné trénování běží až na 100 strojích po několik dnů nebo týdnů. Výpočet lze urychlit použitím strojů s GPU.

5 Závěr

Implementace sítě a zprovoznění distribuovaného trénování byly velmi nelehké úkoly, ale výsledek rozhodně stál za vynaložené úsilí. Díky tomu disponujeme na naší katedře novou metodou, která je teprve půl roku stará a představuje aktuální state of the art v úloze syntézy řeči V dalších měsících bude nutné provést mnoho experimentů pro vyladění a nastavení sítě tak, aby se dala použít i v produkčním prostředí. Jednou z výzev je například syntéza v reálném čase. Ta je velmi obtížná, neboť vzorkovací frekvence je vysoká, a vzhledem k architektuře sítě nelze proces generování paralelizovat. Nedávné články však dokazují, že to není nemožné.

Poděkování

Děkuji MetaCentru za výpočetní zdroje poskytnuté v rámci programu „Projects of Large Research, Development, and Innovations Infrastructures“(CESNET LM2015042).

Literatura

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. (2016) *WaveNet: A Generative Model for Raw Audio.* Available from: `https://arxiv.org/abs/1609.03499`