УДК 519.65

A. Ivaniuk

DOI: 10.18523/2617-70805202253-56

SPEECH AUDIO MODELING BY MEANS OF CAUSAL MOVING AVERAGE EQUIPPED GATED ATTENTION

In the paper we compare different attention mechanisms on the task of audio generation using unsupervised approaches following previous work in language modeling. It is important problem, as far as speech synthesis technology could be used to convert textual information into acoustic waveform signals. These representations can be conveniently integrated into mobile devices and used in such applications as voice messengers or email apps. Sometimes it is difficult to understand and read important messages when being abroad. The lack of appropriate computer systems or some security problems may arise. With this technology, e-mail messages can be listened quickly and efficiently on smartphones, boosting productivity. Apart from that, it is used to assist visually impaired people, so that, for instance, the screen content can be automatically read aloud to a blind user. Nowadays, home appliances, like slow cookers can use this system too for reading culinary recipes, automobiles for voice navigation to the destination spot, or language learners for pronunciation teaching. Speech generation is the opposite problem of automatic speech recognition (ASR) and is researched since the second half of the eighteen's century. Also, this technology also helps vocally handicapped people find a way to communicate with others who do not understand sign language. However, there is a problem, related to the fact that the audio sampling rate is very high, thus leading to very long sequences which are computationally difficult to model. Second challenge is that speech signals with the same semantic meaning can be represented by a lot of signals with significant variability, which is caused by channel environment, pronunciation or speaker timbre characteristics. To overcome these problems, we train an autoencoder model to discretize continuous audio signal into a finite set of discriminative audio tokens which have a lower sampling rate. Subsequently, autoregressive models, which are not conditioned on text, are trained on this representation space to predict the next token, based on previous sequence elements. Hence, this modeling approach resembles causal language modeling. In our study, we show that unlike in the original MEGA work, traditional attention outperforms moving average equipped gated attention, which shows that EMA gated attention is not stable yet and requires careful hyper-parameter optimization.

Keywords: audio modeling, artificial neural networks, attention mechanism.

Introduction

Audio signals consists of several abstraction layers. For example, speech audio can be analyzed at a very fine-grained acoustic or text level but also in terms of speaking style, syntax, grammar, or semantics. Music and singing also have a long-term structure, while being composed of complex non-periodic acoustic signals. In the case of audio synthesis and generation, these multiple abstraction layers interact in such a way that getting high audio quality while demonstrating good consistency level remains a challenge, in particular in unsupervised training scenarios. Latest audio generation models have reached nearly genuine signal quality by using methods such as auto-regressive waveform modeling, adversarial training, flow[1] or diffusion models[2].

During the recent years audio generation quality significantly developed, mainly attributed to the introduction of cost functions that outperforms basic audio time-domain regression. In particular, WaveNet [3] introduced an autoregressive generation approach to audio generation, with quality that was significantly better than traditional concatenative and parametric methods at the cost of slow inference. While WaveNet was a good baseling for more computationally efficient models such as WaveRNN [4] or parallel WaveNet [5], a significant paradigm shift happened with the introduction of adversarial audio synthesis [6; 7], which enabled high fidelity generation without any autoregressive component. Moreover, combining such high-quality generation systems with differentiable vector quantization [8; 9], made possible to train jointly neural audio codecs by compressing activations in a bottleneck layer. In this work, it was used tokens produced by a VQ-VAE neural codec [8], not as intermediate features for signal reconstruction, but rather as ground truth for a sequence modeling task operating at a lower frame rate, which can be reverted back to audio spectrogram at the original frame rate.

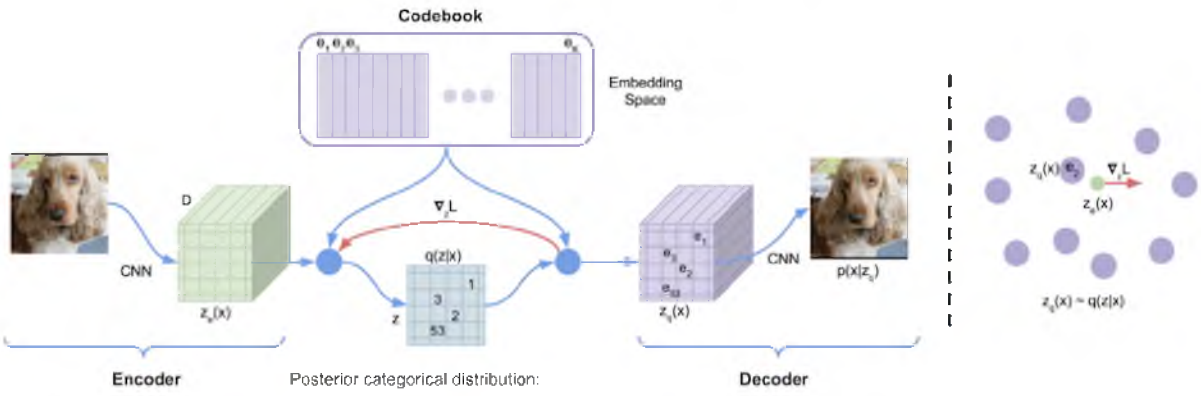

Figure 1. VQ-Vae architecture

Theoretical Background

Discrete audio representation. It is a common technique to discretize image or audio signal using an autoencoder with vector quantization as in VQ-VAE [8]. The main idea is to map the encodur output vectors to the nearest codebook vector e. Subsequently, mapped codebook vectors are passed to the decoder. Training objective is the following:

$$L = \log p(x|z_q(x)) + \beta\|z_e(x) - sg[e]\|_2^2$$

where sg is the stopgradient operator which is an identity at forward pass time and has zero gradient, thus effectively constraining its parameter to be a constant variable. The decoder optimises the first loss term only, the encoder optimises the first and the second loss terms. Autoencoder architecture can be seen in the Figure 1.

Self Attention mechanism. Traditional self-attention mechanism is the following function:

$$Y = \text{Attention}(X) = f(\frac{QK^T}{\sqrt{d_k}})V, \qquad (1)$$

where $X = (x_1, ..., x_n)$ is the input sequence with length n, Attention : $\mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ is the self-attention function and d_k is the input dimensionality. It is also assumed that input and outputs sequences have the same length.

$$Q = XW_q + b_q,$$
$$K = XW_k + b_k,$$
$$V = XW_v + b_v$$

are the sequences of queries, keys and values, with learnable parameters W_q, W_k, $W_v \in \mathbb{R}^{d\times d}$, and b_q, b_k, $b_v \in \mathbb{R}^d$. $f(\cdot)$ is an activation function, e.g. the softmax function.

The matrix $A = f(\frac{QK^T}{d_k}) \in \mathbb{R}^{n\times n}$ is called the *attention matrix*, as it specifies the weight of the dependency strength between every pair of tokens in X. Since it models pairwise dependency weights, the matrix A in principle delivers a flexible and powerful mechanism to learn long-distance dependencies with minimal inductive biases. However, it is in practice a complex task to detect all the relationship patterns in A directly from data, especially when working with long sequences. Also, computing A with h attention heads takes $O(hn^2)$ space and time, and the quadratic dependency on sequence length becomes a significant bottleneck.

Moving Average Equipped Gated Attention. The gated attention mechanism in Mega[10] uses Gated Recurrent Unit and Gated Attention Unit (GAU)[11] as a backbone. Firstly, shared representation is computed

$$X' = \text{EMA}(X) = \alpha \odot \mathbf{x}_t + (1-\alpha) \odot \mathbf{y}_{t-1} \qquad (2)$$
$$Z = \phi_{\text{silu}}(X'W_z + b_z) \qquad (3)$$

where X' is the contextual input and Z is the shared context with z dimensions, with projection matrix $W_z \in \mathbb{R}^{d\times z}$ and bias term $b_z \in \mathbb{R}^z$.

Similar to GAU, the query and key representations are computed by using element-wise multipliers and offsets to Z, and the value sequence is from the original X:

$$Q = \kappa_q \odot Z + \mu_q \qquad \in \mathbb{R}^{n\times z} \qquad (4)$$
$$K = \kappa_k \odot Z + \mu_k \qquad \in \mathbb{R}^{n\times z} \qquad (5)$$
$$V = \phi_{\text{silu}}(XW_v + b_v) \qquad \in \mathbb{R}^{n\times v} \qquad (6)$$

where κ_q, μ_q, κ_k, $\mu_k \in \mathbb{R}^z$ are the learnable scalars and offsets of queries and keys, respectively. v is the expanded intermediate dimension for the value sequence. The output of attention is computed as follows:

$$O = f\left(\frac{QK^T}{\tau(X)} + b_{\text{rel}}\right)V \qquad \in \mathbb{R}^{n\times v}. \qquad (7)$$

where $\tau(X)$ is a scaling factor which was set to d_k.

In the expression the term $b_{\text{rel}} \in \mathbb{R}^{n \times n}$ is the relative positional bias.

Subsequently, MEGA introduces the reset γ and update φ gates, and computes the candidate activation output $\hat{H}$:

$$\begin{aligned} \gamma &= \phi_{\text{silu}}(X'W_\gamma + b_\gamma) && \in \mathbb{R}^{n \times v} \\ \varphi &= \phi_{\text{sigmoid}}(X'W_\varphi + b_\varphi) && \in \mathbb{R}^{n \times d} \\ \hat{H} &= \phi_{\text{silu}}(X'W_h + (\gamma \odot O)U_h + b_h) && \in \mathbb{R}^{n \times d} \end{aligned}$$

The final output Y is computed with the update gate φ:

$$Y = \varphi \odot \hat{H} + (1 - \varphi) \odot X \qquad \in \mathbb{R}^{n \times d} \qquad (8)$$

Numerical Experiment

Firstly, VQ-VAE model was pretrained on LJ-Speech dataset in order to obtain discrete audio representation. The input are mel spectrograms. This yields latent space 4x smaller that the original spectrogram. It consists of one dimensional sequence of discrete points which is 512-dimensional vectors. There are 8192 discrete codebook vectors in total.

Subsequently, two autoregressive models were trained on the latent representations. The first model is the traditional transformer decoder with causal self attention, resembling GPT [12]. The second model is transformer with EMA gated attention. Both model were trained to maximize the following objective:

$$L = \sum_{x,y} \log P(y|x_1, ..., x_m).$$

Model sizes are the same, which is approximately 23.5 million parameters, as well as other hyper-parameters. The loss curves are show on the Figure 2.

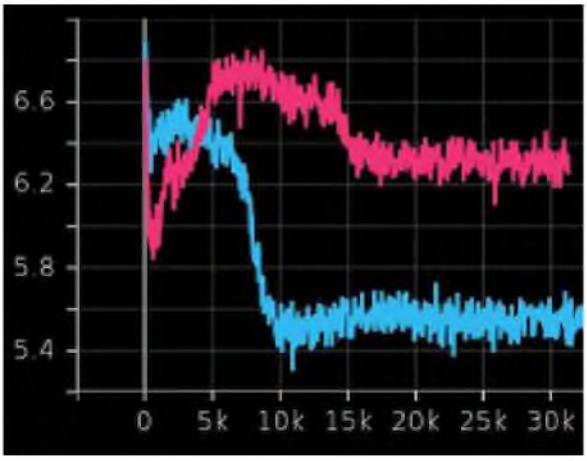

Figure 2. Auto-regressive train losses. Pink curve: EMA Gated Attention, blue curve: traditional transformer

As can be seen, transformer model with traditional attention mechanism performs better, however EMA Gated transformer converges much faster.

Conclusion

In this work, it was conducted experiment to compare different attention mechanisms on the discrete speech representation. It can be concluded that traditional self-attention performs better, although the model based on EMA gated attention converges much faster. It shows, that EMA gated attention mechanism is not robust and stable yet, and probably requires careful hyper-parameter tuning.

References

1. W. Ping, "Waveflow: A compact flow-based model for raw audio" (2019), https://arxiv.org/abs/1912.01219.
2. Z. Kong, "Diffwave: A versatile diffusion model for audio synthesis" (2020), https://arxiv.org/abs/2009.09761.
3. A. Oord, "Wavenet: A generative model for raw audio" (2016), https://arxiv.org/abs/1609.03499.
4. N. Kalchbrenner, "Efficient neural audio synthesis" (2018), https://arxiv.org/abs/1802.08435.
5. A. Oord, "Parallel wavenet: Fast high-fidelity speech synthesis". (2017), https://arxiv.org/abs/1711.10433.
6. C. Donahue, "Adversarial audio synthesis" (2018), https://arxiv.org/abs/1802.04208.
7. Jesse Engel, Kumar Krishna Agrawal, Shuo Chen [et al.], *Gansynth: Adversarial neural audio synthesis* ([S. l. : s. n.], 2019), https://openreview.net/pdf?id=H1xQVn09FX.
8. A. Oord, "Neural discrete representation learning" (2017), https://arxiv.org/abs/1711.00937.
9. N. Zeghidour, "Soundstream: An end-to-end neural audio codec" (2021), https://arxiv.org/abs/2107.03312.
10. X. Ma, "Mega: Moving average equipped gated attention" (2022), https://arxiv.org/abs/2209.10655.
11. J. Gaan Zhang, "Gated attention networks for learning on large and spatiotemporal graphs" (2018), https://arxiv.org/abs/1803.07294.
12. A. Radford, K. Narasimhan, *Improving language understanding by generative pre-training* ([S. l.: s. n.], 2018).

Іванюк А. О.

МОВНЕ МОДЕЛЮВАННЯ АУДІО З ДОПОМОГОЮ МЕХАНІЗМУ УВАГИ З РУХОМИМ СЕРЕДНІМ

У цій роботі ми порівнюємо різні механізми уваги на прикладі задачі генерації аудіо, використовуючи підходи «навчання без вчителя», беручи за основу попередні дослідження в моделюванні мови. Це важлива проблема, оскільки технологію синтезу мови можна використовувати для конвертації текстової інформації в звукові сигнали. Таке представлення можна зручно інтегрувати в мобільні пристрої та використовувати в таких програмах, як голосові месенджери або програми електронної пошти. Іноді важко зрозуміти та прочитати важливі повідомлення, перебуваючи за кордоном. Таким чином, може виникнути нестача відповідних комп'ютерних систем або проблеми з безпекою. Завдяки цій технології повідомлення електронної пошти можна швидко й ефективно прослуховувати на смартфонах, підвищуючи продуктивність. Крім того, вона може використовуватись для допомоги людям із вадами зору, щоб, наприклад, вміст екрана міг автоматично читатися вголос для незрячого користувача. Сьогодні побутова техніка, як-от мультиварки, також може використовувати цю систему для читання кулінарних рецептів, автомобілі для голосової навігації до місця призначення, або особи які вивчають мову, — для навчання вимови. Генерація мови є протилежною проблемою автоматичного розпізнавання мови (ASR) і досліджується з другої половини XVIII століття. Крім того, ця технологія також допомагає людям із вадами голосу знайти спосіб спілкування з іншими, хто не розуміє мови жестів. Однак існує проблема, пов'язана з тим, що частота дискретизації звуку є дуже високою, що призводить до дуже довгих послідовностей, які обчислювально важко змоделювати. Друга проблема полягає в тому, що мовні сигнали з однаковим семантичним значенням можуть бути представлені великою кількістю сигналів зі значною мінливістю, яка спричинена каналом передавання даних, вимовою або характеристиками тембру мовця. Щоб подолати ці проблеми, ми навчаємо модель автоенкодера, щоб дискретизувати безперервний аудіосигнал у скінченний набір дискримінативних аудіотокенів, які мають нижчу частоту дискретизації. Після цього, авторегресивні моделі, які не залежать від тексту, навчаються на цих репрезентаціях, щоб передбачити наступний токен на основі попередніх елементів послідовності. Отже, цей підхід до моделювання нагадує авторегресивне моделювання мови. У нашому дослідженні ми показуємо, що, на відміну від оригінальної роботи MEGA, традиційний механізм перевершує механізм з рухомим середнім, що показує, що останній ще не є стабільним та потребує ретельної оптимізації гіперпараметрів.

Ключові слова: аудіомоделювання, штучні нейронні мережі, механізм уваги.

Матеріал надійшов 28.06.2022