지도 attention 을 이용한 한국어 Transformer 음성 합성에 관한 연구

손병찬, 이준엽, 천성준, 최병진, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소

(bcson, jylee, sjcheon, bjchoi)@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on the Korean Transformer TTS systems with guided attention

Byung Chan Son, Joun Yeop Lee, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National Univ.

요 약

본 논문은 음성 합성 시스템 연구에서 입력 문자 열과 출력 오디오 열 간에 관계를 모델링하는 방식 중 attention 기법을 적용한 Transformer TTS 의 문제를 다루었다. 학습과정에서 배열 관계의 더딘 학습으로 인해 생성된 음성 샘플에 나타나는 발화의 생략, 반복 등의 문제를 해결하기 위해 지도 attention loss 를 적용하였으며 SNU 한국어 DB 에 대해 실험을 진행하였다. 그 결과 제안된 모델은 기존 모델보다 이른 시기에 배열 관계를 안정적으로 학습하는 모습을 보였다.

Ⅰ. 서 론

딥 러닝을 활용한 음성 합성 연구에서는 문자와 음성 신호 사이에 배열 관계(alignment)를 학습하는 것이 중요한 주제이다. 최근엔 입력과 출력의 배열 관계를 학습하는 attention 기법을 활용한 Transformer 음성 합성 시스템[1]이 개발되었다. 그러나 위 연구에서도 임의의 두 순차적인 데이터 간에 가능한 배열 관계가 다양하다는 특징 때문에 학습 시에 우리의 목적과는 다른 배열 관계가 만들어 질 수 있다는 문제를 여전히 가지고 있었다. 음성 합성 연구 분야에서는 이러한 문제를 해결하기 위해 지도 attention 기법을 사용한 학습 방법[2]이 제안되어 있었다. 이번 연구에서는 Transformer TTS 에 지도 attention 기법을 적용해 학습 초기 단계에서 배열 관계에 대한 학습을 개선하는 것을 보였다. 훈련 데이터는 SNU 한국어 음성 합성 DB 를 사용하였다.

Ⅱ. 본론

음성 합성 연구는 문자 입력이 주어졌을 때 그에 해당하는 오디오 파일을 생성하는 것이다. 최근에는 생성하는 오디오에 여러가지 스타일을 부여하는 연구가 진행되고 있다. 그 중에서도 음성 합성에 기본이 되는 연구 주제는 입력되는 문자 열과 목표 데이터로 주어지는 오디오 데이터 사이에 배열 관계를 학습하는 것이다. 이는 주어진 텍스트 입력을 몇 개 프레임을 갖는 오디오로 시간 축 위에서 어느 위치에 생성할 것인지에 대한 문제이다. 문자 열과 오디오 파일의 연관되는 부분을 찾아서 각 문자 입력들로부터 생성될 적당한 오디오 길이를 예측해주는 여러가지 방법들이 제안되고 있다.

그 중 하나의 방법으로 attention 기법[3]이 존재한다. 입력과 출력 사이에 존재하는 연관성을 학습할 수 있는 특징 덕분에 언어 간 번역, 음성 인식 분야에도 널리 사용되었다. Attention 기법을 통해 음성 합성 시스템을 구축한 Transformer TTS 는 self-attention 을 사용한 Encoder, Decoder 를 가지고 있다. 각각은 문자 열과 오디오 열 내부에 존재하는 연관성을 학습하게 된다. 그리고 이 두개의 모듈에서 출력된 값들 간에 다시 attention 기법을 적용하여서 문자 열과 오디오 열 사이에 존재하는 배열 관계를 학습하게 된다.

이처럼 attention 기법을 사용하여 음성 합성 시스템을 구축하면 문자 열과 오디오 열을 비교하여 어느 부분들이 서로 높은 연관성을 가지고 있는지 attention matrix 를 통해 확인할 수 있다.

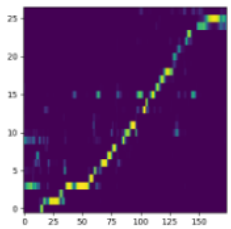

그림 1. 학습 마무리 단계 attention 그래프

예를 들어 학습이 마무리 단계인 경우 Encoder 와 Decoder 간에 생성된 attention 그래프를 그려보면 그림 1 과 같은 우 상향 대각선 모양이 나타난다. 가로축이 오디오 열이며 세로축이 입력된 문자 열을

나타낸다. 입력된 문자 열은 오디오가 존재하는 시간 축에서 순차적으로 발음되는 특성을 갖기 때문에 세로축과 가로축은 순차적인 연관성을 가지는 것이 자연스럽다. 그러한 이유에서 학습이 완료된 시점에 attention 그래프는 우 상향 대각선 모양을 갖게 된다.

그러나 attention 기법을 사용할 때 모델이 학습하는 과정에서 배열 관계를 제대로 학습하지 못하는 경우가 발생하기도 한다. 실제로 문자 열과 오디오 열은 시간 순으로 높은 연관성을 가지고 있고 학습이 완료되었을 때 우 상향 대각선 모양의 attention 그래프가 그려져야 한다는 것이 사실이다. 하지만 모델은 생성할 수 있는 가능한 배열 관계가 순차적이라는 가정을 하지 않은 채로 학습을 진행하기 때문에 원하는 그래프가 형성되기까지 오랜 시간이 걸리게 된다. 때문에 제대로 배열 관계를 학습했을 때 우 상향 대각선 그래프가 그려져야 한다는 가정을 loss 로 추가하는 방법[2]이 제안되었다. 그로 인해 attention 을 사용하여 음성 합성을 하는 연구들에서 보다 빠른 시간에 올바른 그래프가 형성되는데 도움을 주게 되었다. 때문에 이번 연구에서는 Transformer TTS 에 지도 attention loss 를 추가하여 기존 모델보다 빠른 시간에 안정적으로 배열 관계를 학습하는 모습을 관찰하였다.

이번 연구에서는 발화의 생략 등 문제를 해결하고 학습 과정에서 조금 더 빠른 수렴 속도를 얻기 위해 attention 그래프의 대각성분에는 0, 멀어질수록 1 에 가까운 값을 곱하는 지도 attention 목적함수를 추가하여 Transformer TTS 모델을 학습하였다. 목적 함수에 추가되는 수식은 다음과 같이 표현될 수 있다.

$$\mathcal{L}_{\text{att}}(A) = \mathbb{E}_{nt}[A_{nt}W_{nt}]$$

$$W_{nt} = 1 - \exp\{-(n/N - t/T)^2/2g^2\}$$

N 은 입력되는 문자 열의 길이이고 T 는 오디오 열의 길이이다. A_{nt}와 W_{nt}는 각각 attention 그래프에서 (n, t) 성분과 그것에 곱해지는 가중치 값들을 나타낸다. 식에 나타난 g 는 초 매개변수이며 이번 연구에서는 0.2 로 설정하였다.

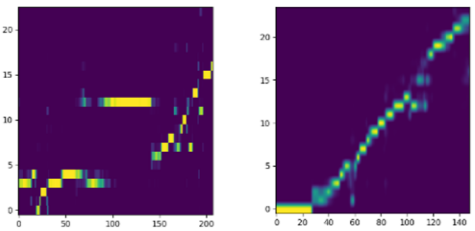

그림 2. 지도 attention 을 적용하기 전과 후 그래프

그림 2 는 학습 중 동일한 468k step 에서 기존 모델과 제안된 모델이 만들어내는 문자 열과 오디오 열 사이에 그려진 attention 그래프를 비교한 것이다. 왼쪽 그림은 기존 Transformer TTS 모델을 학습하는 중에 생성된 문자 열과 오디오 열 간에 배열 관계를 나타낸 그래프 이며 오른쪽 그림은 지도 attention loss 를 Transformer TTS 모델에 추가한 이번 연구에서 제안된 모델을 학습하며 생성된 문자 열과 오디오 열 간에 관계를 나타내는 그래프이다. 왼쪽 그림은 학습하는 과정에서 아직 그래프 상에 가로축 100~150 부근에 동떨어진 직선이 존재하고 있는 것을 볼 수 있다. 이런 경우에는 음성을 합성해 보았을 때 발화의 생략, 반복 등의 문제가 발생한다. 반면 오른쪽 그림은 동일한 학습 시점에 이미 우 상향 대각선 형태로 그래프가 그려지며 문자 열과 오디오 열 간에 순차적인 배열 관계가 학습된 것을 확인할 수 있다. 이렇게 문장의 처음부터 끝까지 연결이 생략되는 부분이 없이 우 상향 그래프가 나타난 것을 볼 수 있는 것은 문자 열과 오디오 열 사이에 시간 축에 존재하는 순차적인 연결관계를 모델이 강하게 학습하고 있는 것을 의미한다.

Ⅲ. 결론

이번 연구 에서는 기존 Transformer TTS 를 학습할 때 문자 열과 오디오 열 사이에 배열 관계에 대한 학습이 더디게 이루어지는 점을 개선하기 위해 지도 attention loss 를 목적 함수에 추가하여 실험을 진행하였다. 그 결과 학습 과정에서 같은 시점에 두 가지 모델을 비교하였을 때 제안된 방식이 배열 관계에 대해 더 안정적이고 빠르게 학습하는 모습을 볼 수 있었다. 기존 방식에서 나타나던 그래프 상에 늘어지는 모양이 사라졌고 우 상향 대각선에 가깝게 그래프가 그려지는 모습을 보였으며 기존 모델에 비해 학습 시에 이른 단계에서 순차적 배열 관계가 나타나는 것을 확인했다. 합성된 샘플을 들어보았을 때에도 발화가 생략되거나 늘어지는 부분이 고쳐진 걸 알 수 있었다.

ACKNOWLEDGMENT

이 연구는 방위 사업 청 및 국방과학 연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음.

참 고 문 헌

[1] N. Li, S. Liu, Y.Liu, S. Zhao, M.Liu, and M. T. Zhou, “Neural speech synthesis with transformer network,” in The AAAI Conference on Artificial Intelligence(AAAI), 2019.

[2] H. Tachibana, K.Uenovama, and S.Aihara, “Efficiently trainable text-to=speech system based on deep convolutional networks with guided attention,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 20118, pp. 4784-4788.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000- 6010.