Intro: Tacotron2 (2018)이 무엇인가?

1. End to end 모델이 뭔지 간략하게 말해주자.

<end to end deep learning>

장점

1. Let the data speak: 사람의 선입견에 영향을 받지 않고, 데이터 그 자체를 학습 (순수한 머신러닝 학습법)
2. Less hand-designing of components needed

단점

1. May need large amount of data
2. Excludes potentially useful hand-designed components (인간의 손으로 만들어진 중간 영역들이 배제가 된다 🡪 데이터가 적을 경우, 알고리즘에 수동으로 특정 지식을 넣을 수도 있는데 그러한 가능성이 원천 배제 된다)

+ 학습 알고리즘에는 두 가지 원천이 존재: 데이터 + 인간의 개입(중간 요소와 같은 것들)

<네이버 자료>

* end-to-end 딥러닝은 자료처리 시스템 / 학습시스템에서 여러 단계의 필요한 처리과정을 한번에 처리합니다. 즉, 데이터만 입력하고 원하는 목적을 학습시키는 것입니다.
* 이는 기존의 처리 파이프라인 중 일부를 대체할 수 도 있습니다. 다만, end-to-end 딥러닝을 사용하기 위해서는 엄청나게 많은 양의 데이터가 필요합니다.
* 하지만 end-to-end 딥러닝이 만능은 아닙니다. 단계를 나눠서 학습 시키는 것이 효율적일 때도 있습니다. 효율적인 이유는 아래와 같습니다.
  + 복잡한 문제를 분리하여 각각의 간단한 문제로 바꿉니다.
  + 데이터의 정보가 각각의 작업에 더 적합되게 사용됩니다.
* 따라서, 순수 end-to-end 딥러닝 보다는 문제를 쪼개서 해결하는 것이 좋습니다.

1. Tactoron end to end model

<end to. End model이 아니라면>

* 방대한 Domain 전문 지식, 여기서는 음성에 대한 전문 지식들이 필요하다.
* 디자인에 어려움이 있을 수 있다.
* 트레이닝이 파이프라인 별로 따로 되면서 에러가 누적될 수 있고 복잡하다.

결과적으로 End-to-End로 학습하는 경우의 장점은 다음과 같이 정리할 수 있다.

* <text, audio> 페어로만 가지고 학습이 가능하다.
* Feature Engineering이 간단하다. 즉, 어떤 Feature를 뽑아서 넘겨주고 해야할지에 대한 디자인이 간단하다.
* 발화자, 언어, 감정 등의 Feature 등을 손쉽게 조절이 가능하다.
* 새로운 데이터에 더 Adaptable하다.
* 노이즈에 더 강하다.

1. 기존 모델과의 비교
2. 2016년 Wavenet

매우 강력한 오디오 생성 모델이다. 성능이 매우 좋아서 Tacotron의 Vocoder로 사용된다.

샘플 수준의 Autoregressive Model이라는 이유로 너무 느리다는 단점이 있다.

TTS로 바로 활용할 수 없으며 TTS-Frontend로부터 Linguistic Feature를 입력으로 넣어줘야 하는 단점이 있다.

1. DeepVoice (2017):

* 각각의 TTS 파이프라인을 뉴럴넷 모델로 대체하였다.
* 학습이 End-to-End로 되지 않는다는 단점이 있다.

1. Encoder – Decoder 모델

이렇게 생성된 문장 임베딩은 다시 RNN 디코더의 첫번째 타임 스텝의 입력으로 들어가게 되고 인코더의 마지막 타임 스텝의 RNN Hidden State Vector는 RNN 디코더의 첫번째 타임 스텝의 Hidden State Vector로 들어가게 된다. 이렇게 입력 문장 임베딩 및 Hidden State Vector를 받은 디코더는 순차적으로 문장을 생성하여 결국 번역된 문장을 완성하게 된다.

1. Attention model

기존 encoder-decoder 모델에 Attention mechanism을 추가하였다.

문장을 구성하는 모든 단어들을 각각 임베딩 하여 디코더가 특정 타임 스텝에서 어떤 단어에 더 집중하여 현재 타임 스텝의 출력을 생성할지에 Attention mechanism.

Tacotron은 Bahdanau attention을 사용한다.

1. Tactoron 모델 구조

Tactoron + wave net – griffin-Lim = Tacotron 2

TacoTron2는

1. Network: (LSTM & CNN layers)

Convert character sequences to mel spectorogram

1. Wave net : mel spectrogram to speech로 바꿔준다!

Why mel spectrogram?

* Faster, Parallel training of Network & WaveNet parts.
* Emphasize low frequency signals, allowing better intelligibility when converting to speech

Spectrogram Prediction Network

1. Encoder: Convert character sequence to internal feature representation
2. Decoder: Convert feature representation to frames of mel spectro gram.

Question 1 taco tron의 이름?

Question 2 taco tron1과 tacotron2의 차이점

Question3 mel spectorogram이 도대체 무엇인가?