T-아카데미

딥러닝 기반 음성 합성

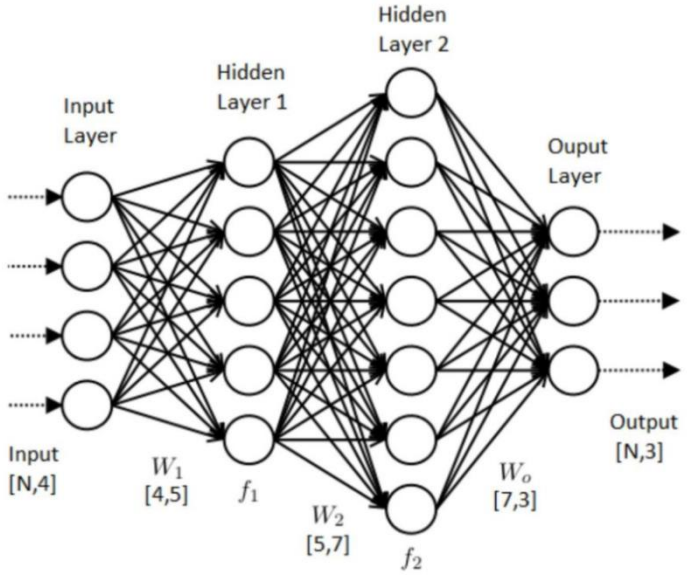
허승연

04. Deep Learning based Speech Synthesis

내용 추가 조사

- ANN(Artificial Neural Network)
  - 사람의 신경망 원리와 구조를 모방하여 만든 기계 학습 알고리즘

    : 신경망 = 입력층(Input) + 출력층(Output) + 입력층과 출력층 사이의 레이어들, 은닉층(Hidden)
  - 신경망 모델을 잘 구성하여 원하는 Output 값을 잘 예측하는 것이 목적
  - 은닉층은 활성화 함수를 사용하여 최적의 weight와 bias를 찾아내는 역할을 함



04. Deep Learning based Speech Synthesis

내용 추가 조사

- ANN의 단점
  - 학습 과정에서 파라미터의 최적값 찾기 어려움

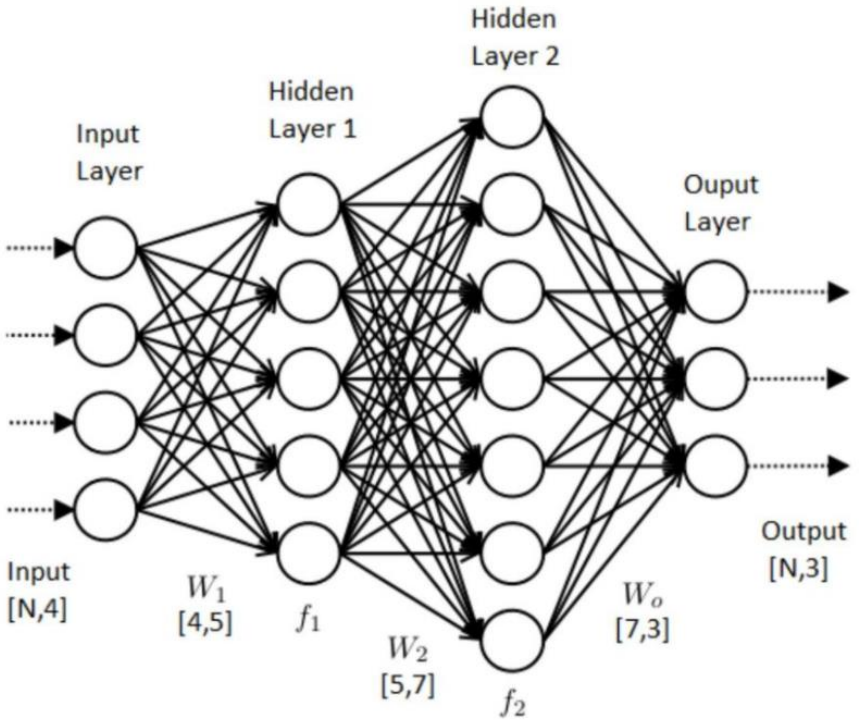
    Ex) 활성화 함수 사용시 기울기 값에 의해 weight가 결정되었는데 이 값이 뒤로 갈수록 0에 수렴하는 오류

    부분적인 에러를 최저 에러로 인식하여 더 이상 학습 X
  - Overfitting(과적합)에 따른 문제

    → 사전 훈련을 통해 방지
  - 학습 속도가 너무 느림

    : 은닉층의 수에 따라 연산량 기하급수적으로 증가

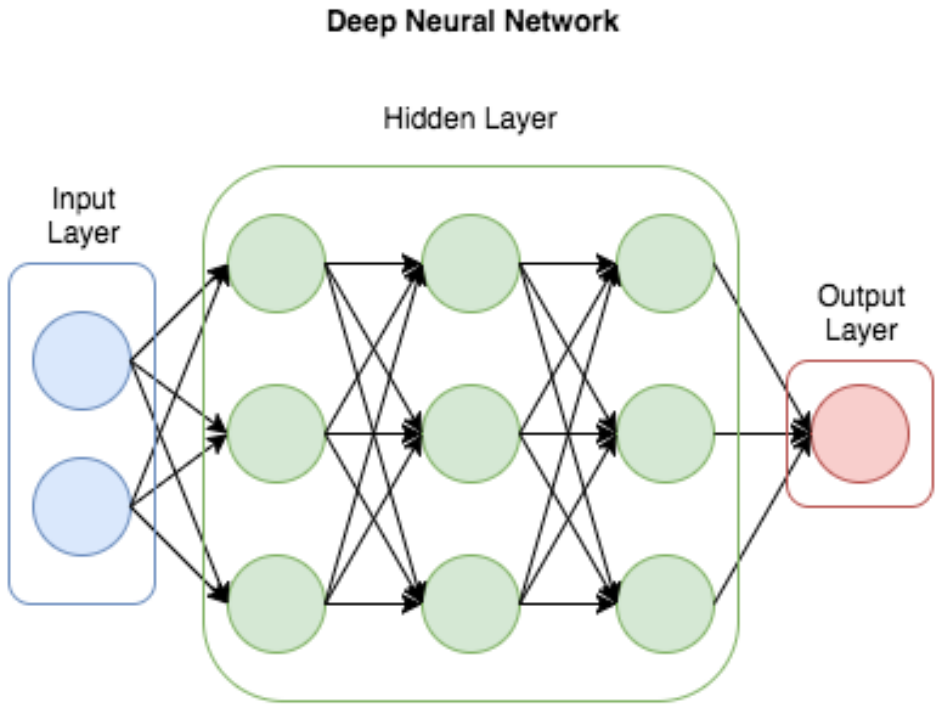
    → 하드웨어의 발전으로 감당 가능



04. Deep Learning based Speech Synthesis

내용 추가 조사

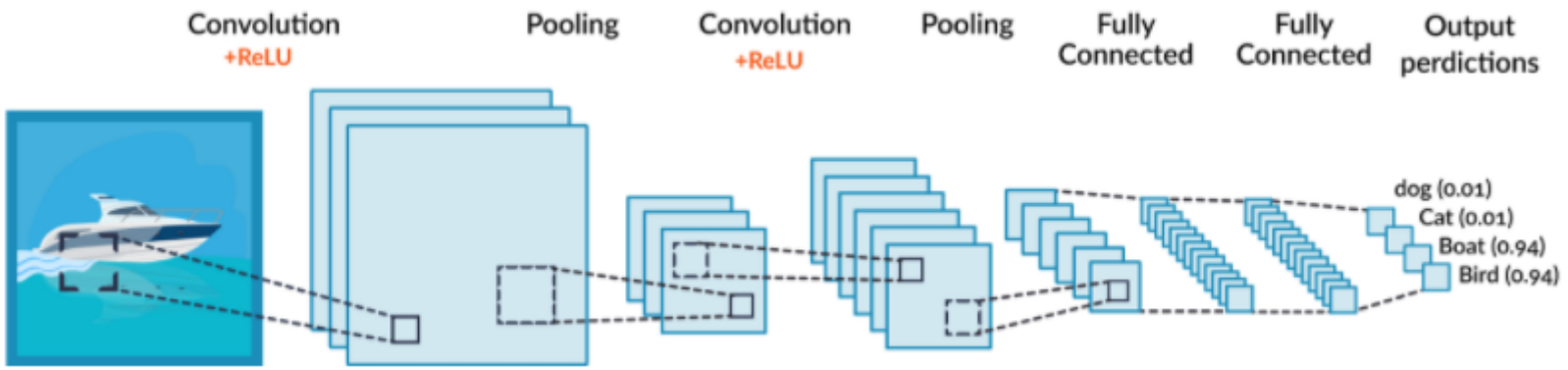
- DNN(Deep Neural Network)
  - 은닉층을 2개 이상 지닌 학습 방법 → ANN의 단점이 해결되면서 등장
  - 스스로 분류 레이블을 만들어 내고 데이터를 구분짓는 과정을 반복하여 최적의 구번선(implementation) 도출
  - DNN의 응용 → CNN, RNN 등



04. Deep Learning based Speech Synthesis

내용 추가 조사

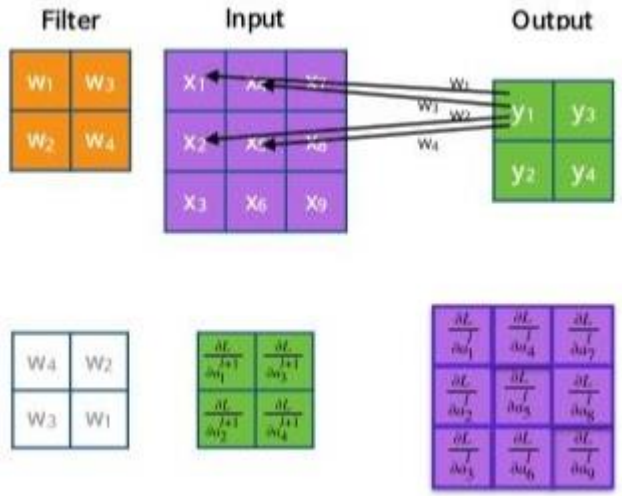
- CNN (Convolution Neural Network)
  - 데이터의 특징을 추출하여 데이터의 특징들의 패턴을 파악하는 구조
  - Convolution Layer + Pooling Layer를 복합적으로 구성하여 알고리즘 만듬



04. Deep Learning based Speech Synthesis

내용 추가 조사

- Convolution 과정
  - 데이터에 각 성분의 인접 성분들을 조사해 특징을 파악하고 파악한 특징을 한 장(Layer)으로 도출
  - 압축 과정으로 파라미터의 개수를 효과적으로 줄여주는 역할

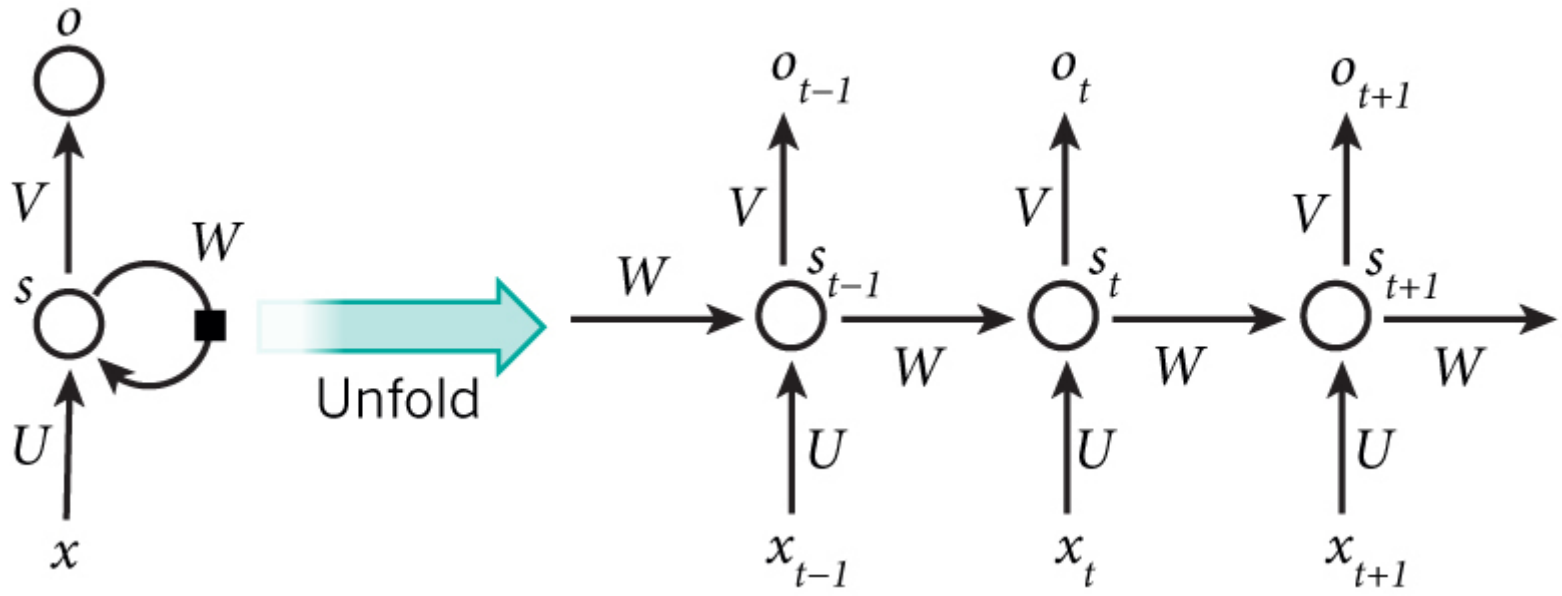


- Pooling 과정
  - Convolution 과정을 거친 Layer의 사이즈를 줄여줌
  - 데이터의 사이즈를 줄이고, 노이즈를 상쇄시키고, 미세한 부분에서 일괄적인 특징 제공

04. Deep Learning based Speech Synthesis

내용 추가 조사

- RNN (Recurrent Neural Network)
  - 반복적이고 순차적인 데이터 학습에 특화된 인공신경망의 한 종류
  - 내부에 순환 구조 有 → 과거의 학습을 weight을 통해 현재 학습에 반영



04. 프로젝트

깃허브 및 논문

- **Real Time Voice Cloning**

  CorentinJ/Real-Time-Voice-Cloning: Clone a voice in 5 seconds to generate arbitrary speech in real-time (github.com)

- **[논문] 세밀한 감정 음성 합성 시스템의 속도와 합성음의 음질 개선 연구**

  : 세밀한 감정 표현을 위해 감정 클러스터의 특성을 고려한 SA-I2I (Spread Aware Inter-to-Intra distance ratio)를 제안 했으나 *학습 속도가 느리고* 감정의 *세기가 강해질수록 음질 저하*

  - 느린 학습 속도 → WaveRNN을 통해 약 8배 빠르게 합성
  - 음질 저하 → 데이터 전처리 과정 (각 감정의 데이터 셋의 신호 크기 조절) 거쳐

  → 음성 합성 시스템에 전역 스타일 토큰을 추가한 GST-Tacotron을 기준 모델로 사용하고 보코더로 WaveRNN을 사용함.