CONDITIONED-U-NET: INTRODUCING A CONTROL MECHANISM IN THE U-NET FOR MULTIPLE SOURCE SEPARATIONS.

Meseguer-Brocal, Gabriel, and Geoffroy Peeters. Proceedings of the 20th International Society for Music Information Retrieval Conference. 2019.

paper link (arxiv)

Abstract (1)

- Data-driven models for audio source separation such as U-Net or Wave-U-Net are usually models dedicated to and specifically trained for a **single task**, e.g. a particular instrument isolation.
- Training them for various tasks at once commonly results in worse performances than training them for a single specialized task.

Abstract (2)

- In this work, we introduce the Conditioned-U-Net (C-U-Net) which adds a control mechanism to the standard U-Net.
    - The control mechanism allows us to train a unique and generic U-Net to perform the separation of various instruments.
    - The C-U-Net decides the instrument to isolate according to a one-hot-encoding input vector.
    - The input vector is embedded to obtain the parameters that control Feature-wise Linear Modulation (FiLM) layers.
        - FiLM layers modify the U-Net feature maps in order to separate the desired instrument via affine transformations.
    - The C-U-Net performs different instrument separations, all with a single model achieving the same performances as the dedicated ones at a lower cost.

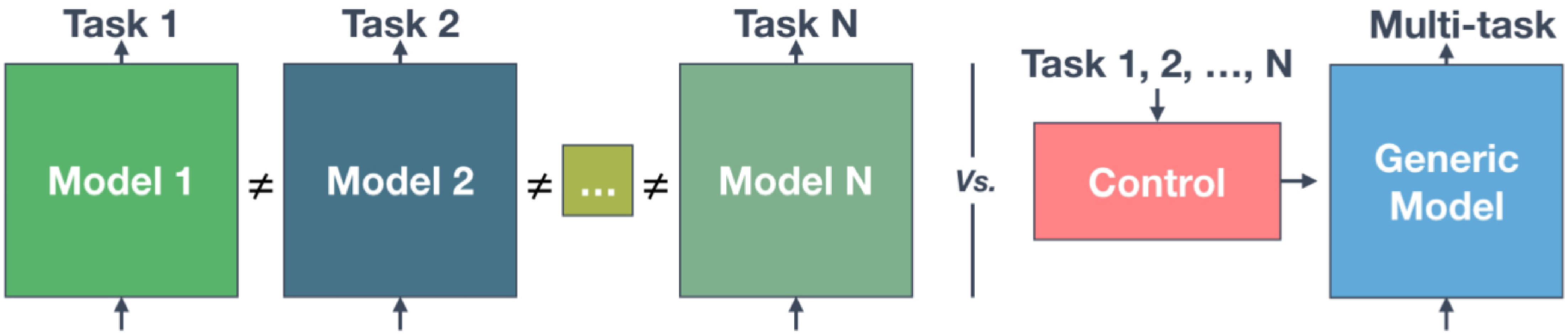
Related Works: Source Separation (1-to-1 or 1-to-many)

- The usual approach is to build dedicated models for each task to isolate [1, 20]
- Since isolating an instrument requires a specific system, we can easily run into problems such as scaling issues (100 instruments = 100 systems)
- Besides, these models do not use the commonalities between instruments. If we modify them to do various tasks at once i.e., adding fix numbers of output masks in last layers, they reduce their performance
- Multi-instruments version performs worse than the dedicated one (for vocal isolation) and has to be retrained to different source combinations
    - Reality Check: Liu, Jen-Yu, and Yi-Hsuan Yang. "Dilated convolution with dilated GRU for music source separation." Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 2019.

        > It then outputs the separation predictions of all the sources at once.

Related Works: Conditioning Learning

- a solution to problems that need the integration of multiple resources of information
- Conditioning learning divides problems into two elements
    i. a generic system
    ii. a control mechanism that governs it according to external data
- Vision -> Source Separation
    - This paradigm can be integrated into source separation creating a generic model that adapts to isolate a particular instrument via a control mechanism


Task 1
Task 2
Task N
Multi-task
Task 1, 2, ..., N
Model 1
≠
Model 2
≠
...
≠
Model N
Vs.
Control
Generic Model

Related Works: Conditioning in Audio

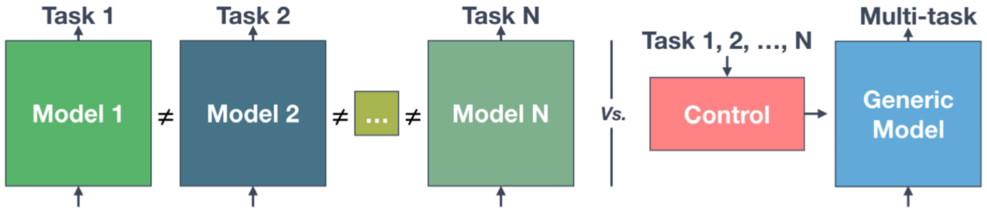
- Actively studed in speech generation
  - WaveNet: speaker identity => speaker-conditioned speech genearation
- not directly related to conditioning mechanism
  - speech recognition
  - music generation
  - piano transcription

VS

- The closest work to ours is [10].
    - Kameoka, Hirokazu, et al. "Semi-blind source separation with multichannel variational autoencoder." arXiv preprint arXiv:1808.00892 (2018).
        - did not use MUSDB18
    - multi-channel audio as input to a Variational AutoEncoder (VAE) to separate 4 different speakers
    - The VAE is conditioned on the ID of the speaker to be separated.
    - The proposed method outperforms its baseline
        - not found in the experiment section

Scope of this work

- propose a **standard U-Net system** not specialized in a specific task but rather in finding a set of generic source separation filters, that we **control differently** for isolating a particular instrument



- Our system takes as input
  - the spectrogram of the mixed audio signal
  - the control vector (ont-hot encoded)
- Our system gives as output
  - the separated instrument defined by the control vector

Contribution

1. the Conditioned-U-Net (C-U-Net), a joint model that changes its behavior depending on external data and performs for any task **as good as a dedicated model** trained for it.
    - C-U-Net has a fixed number of parameters no matter the number of output sources.
2. The C-U-Net proves that **conditioning learning** (via Feature-wise Linear Modulation (FiLM) layers) is an efficient way of inserting external information to MIR problems.
3. A new FiLM layer (FiLM-simple) that works as good as the original one but with a lower cost (fewer parameters)

Method and Evaluation

http://localhost:8888/notebooks/seminar/2020-09-01-wschoi-Conditioned_Unet/Conditioned-U-Net.ipynb