Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data

Report

Stefan Wezel
stefan.wezel@student.uni-tuebingen.de
4080589
ML4S

Abstract

Information in sequential data is often distributed over multiple time scales. While if viewed as a single signal, such data might appear noisy. However, patterns can emerge if temporal scales are viewed separately from one another. Hsu et al. [1] leverage this intrinsic structure to learn disentangled representations from sequential data in an unsupervised manner with a proposed factorized hierarchical variational autoencoder (FHVAE). They aim to factorize sequence level and segment level attributes into distinct latent subspaces. Architectural and sequence dependent priors create an inductive bias to encourage the proposed factorization. Here, we put their work into a formal context, explore the proposed methodology, and reflect critically on their work.

Introduction

Intuitively, disentangled representations are reflective of the underlying generating factors of observed data in thus they are encoded as separate latent subspaces. This notion is already present in classical factor analysis work, where it is referred to as independent component analysis (ICA) [2]. However, many problems cannot be solved in linear fashion. The vast success of deep neural networks (DNN) can be largely attributed to the fact that they are very powerful non-linear function approximators. Thus, making them an promising method to solve the long standing problem of non-linear ICA.
Different methods have been proposed to learn such disentangled representations [3, 4, 5] with varying success. Many of these works focus on image data. However, it has been shown by Locatello et al. [6] that disentangled representations cannot by learned without introducing any kind of supervision or inductive biases. Sequential data, while having been explored less, despite offers inherent structure that can be exploited to construct inductive biases as has been proposed by Hsu et al. [1].
Besides technical challenges, this strain of research suffers from the lack of formally defined and agreed upon foundations. The very term of disentangled representations for example is often understood differently in between works. In the following section, we will use the definition, proposed by Higgins et al. [7] to put the work by Hsu et al. [1] into formal context.

Viewing the FHVAE though a Formal Lense

Works on disentanglement often lack a proper formal framework [7]. While claiming to achieve disentanglement, Hsu et al. [1] do not provide any formal foundation for their claim. With no formal tools at hand, we cannot formally discuss whether disentanglement was achieved. Thus, we use

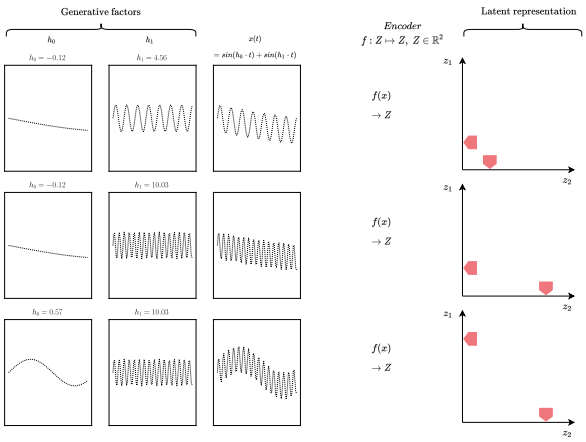

Figure 1: Changing values of generating factors are reflected in the corresponding latent variables.

the following sections to respectively give a theoretical background that defines disentanglement in group theoretic terms, and describe the problem, the FHVAE was designed to solve, in these terms.

The Tools of Group Theory

A group is described by a tuple of an operation $\circ$ and a non-empty set G. The set has to be closed under the operation, it must contain an identity element, the operation must be associative, and for every element in G, there must be an inverse element [8]. Multiple groups G_i can be combined via a direct product $G = G_1 \times ... \times G_n$. The result of this direct product is itself a group.
A group that is of particular interest for the field of disentangled representation learning is the symmetry group [8]. A symmetry group consists of a set of transformations that leave a given object X invariant and the operation is the composition of such transformations. A prominent example of a symmetry group is SE_3. This symmetry group can be visualized as a set set of vertices X that form an equilateral triangle. Permutations of this set would result in rotations or flips of the triangle. These permutations would be the symmetry transformations of our symmetry group. The operation would be composing multiple permutations.
Another important concept is the group action. It is the result of applying a symmetry transformation to an object. In our triangle example, a group action would be the permuted set of vertices.
If a group action on X is the result of a subset G_i of symmetries $G = G_1 \times ... \times G_n$ and only affects a subset X_i of X but leaves all other $X_{j \neq i}$ unchanged, we say it is a disentangled group action. Now, such disentangled group actions are of particular interest for disentangled representation learning, and are generally what we want to model. Note, that if we observe such disentangled group actions, we can infer that G can be decomposed into a direct product of symmetry groups G_i without knowing the underlying processes.
To model such a process, we need to find some symmetry preserving mapping $f : X \mapsto Z$. It should not matter, whether X is mapped to Z followed by applying a symmetry G or G is applied to X and then mapped to Z. The resulting space Z should be the same. This is visualized below.

$$\begin{array}{ccc} X & \xrightarrow{G} & X \\ f \downarrow & & f \downarrow \\ Z & \xrightarrow{G} & Z \end{array}$$

Such a mapping is called an equivariant map and its result is a disentangled representation. The space Z (which we will refer to as latent space in the following) naturally decomposes into a direct product of independent subspaces $Z_1 \times ... \times Z_n$. Moreover, each subspace Z_i is only affected by symmetry G_i on X (as only X_i is changed) and remains invariant to all other $G_{j \neq i}$.
Note that it is only disentangled with respect to a certain decomposition. This is important, as it

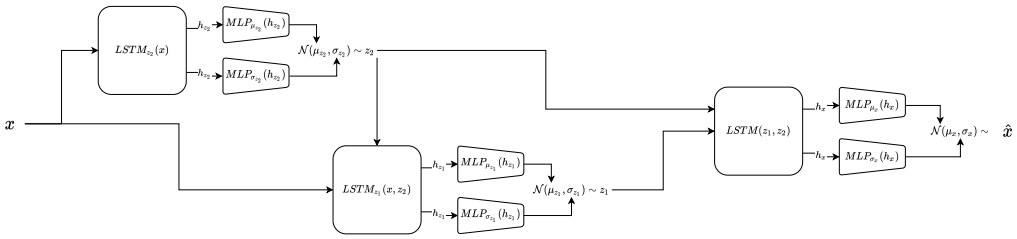

Figure 2: Architecture of the proposed FHVAE.

means we can only discuss whether disentanglement was achieved, if the decomposition is sufficiently clear. Not all decompositions make sense or are possible to model.
As these concepts are rather abstract, we will use the next section to frame the real-world setting of Hsu et al. [1] using group theoretic terms to build further intuition and build a formal foundation to discuss their work.

Symmetries in Sequential Data

Hsu et al. [1] argue that certain sequential data can be factorized into attributes of different temporal scale. For example, voice recordings can be decomposed into sequence attributes and segment attributes. In this context, sequence attributes are features that remain unchanged over multiple sequences if spoken by the same speaker. Segment attributes on the other hand, vary in(-between?) sequences and are independent from the speaker. They are determined by variables such as linguistic content.
In group theoretic terms, we could state this as the group G acting on audio recordings X is a direct product of $G_{sequence} \times G_{segment}$, because we observe the resulting disentangled group actions of this decomposition. Hsu et al. [1] now want to find a representation that is disentangled with respect to this decomposition. They need to find an equivariant map $f : X \mapsto Z$, where $Z = (z_1, z_2)$ is a latent space, so that i.e. z_2 is only affected by $G_{sequence}$ and z_1 only is affected by $G_{segment}$. Then, the proposed decomposition of symmetries is reflected in Z.
For this specific setting, finding such an equivariant map, would allow to separate speaker information from content information. This in turn would enable us to reconstruct given content information using another speaker's voice information. Hsu et al. [1] propose this as qualitative assessment of their disentanglement. Quantitative evaluation of disentanglement are notoriously challenging and various metrics have been proposed by an active field of research [6, 3].

FHVAE - Constructing an Equivariant Map

To find an equivariant map for the sequential data setting, Hsu et al. [1] porpose the FHVAE (see Figure).

Results

Discussion and Future Work

Lack of formal context. Lack of evidence for disentaglement. Further exploit the available data using cross reconstruction.

Conclusion

References

[1] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *arXiv preprint arXiv:1709.07902*, 2017.

[2] Pierre Comon. Independent component analysis, 1992.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[5] Tejas D Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B Tenenbaum. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167*, 2015.

[6] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[7] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[8] Georgia Benkart. Abstract algebra, by in herstein. *The American Mathematical Monthly*, 94(8): 804–806, 1987.