Natural Language Processing

Disentangled Representation Learning

Pranjul Mishra
Saurabh Singh
Nazira Tukeyeva

Research Paper

Title of the paper: Disentangled Representation Learning
Date of Publication: 01 July 2024
Authors: Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, Wenwu Zhu
Description: Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
Published in: IEEE Transactions on Pattern Analysis and Machine Intelligence

Link

Objective

1. Present two well-recognized definitions:
    a. Intuitive Definition (semantic independence).
    b. Group Theory Definition (mathematical).
2. Categorize DRL approaches based on model type, representation structure, supervision signal, and independence assumption.
3. Analyze principles to design different DRL models.
4. Applications of DRL
5. Challenges & Potential Improvements

Introduction to DRL

What is Disentangled Representation Learning (DRL)?

- DRL focuses on separating underlying factors in data into independent, semantically meaningful representations.
- Inspired by how humans understand the world (e.g., separating object shape, size, and color).

DRL Importance:

1. Explainability: Makes models more interpretable.
2. Generalization: Separates relevant factors for improved downstream task performance.
3. Controllability: Enables controllable generation (e.g., changing size without altering color).

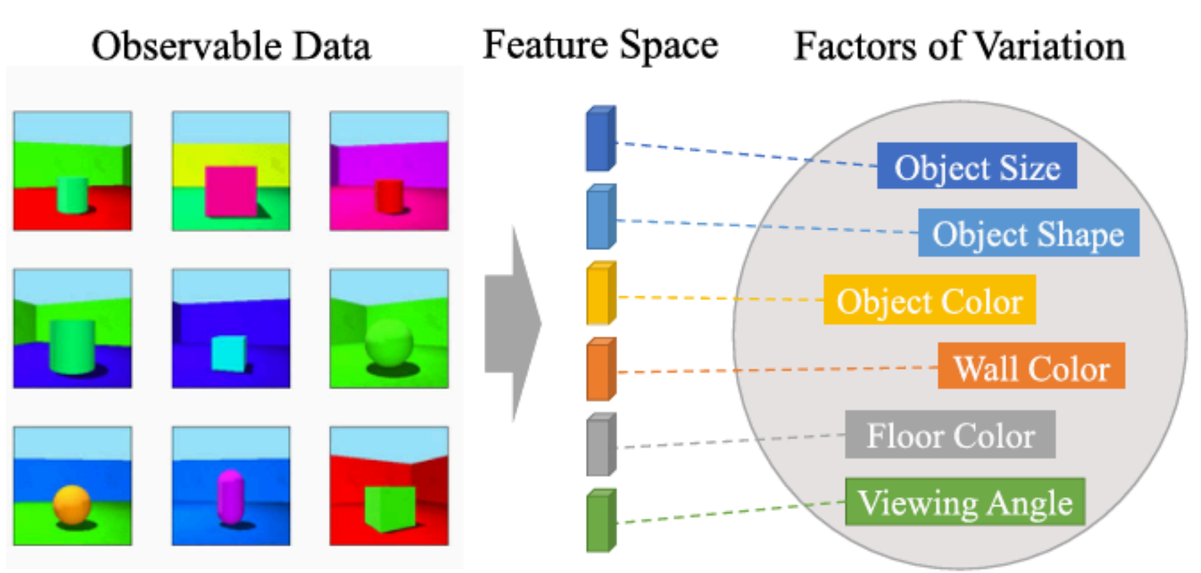

Fig. 1. Six factors of variation in Shape3D dataset

Definitions

Intuitive Definition (By Bengio)

- Each latent variable corresponds to one specific factor of variation in the data.
- Example: If one latent variable represents ‘size’, then changing that variable would only affect the size of the objects, keeping others constant.

Group Theory Definition

- A formal, mathematical approach.
- Based on mathematical symmetry, latent variables are organized into groups where each group represents specific generative factors, ensuring hierarchical decomposition of information.

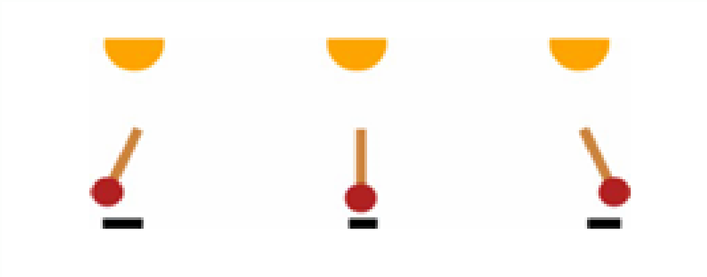

Fig. 2. Causal factors such as light source position and pendulum angle affecting shadow

Taxonomy of the DRL Approaches

Approaches

DRL approaches categorized based on:

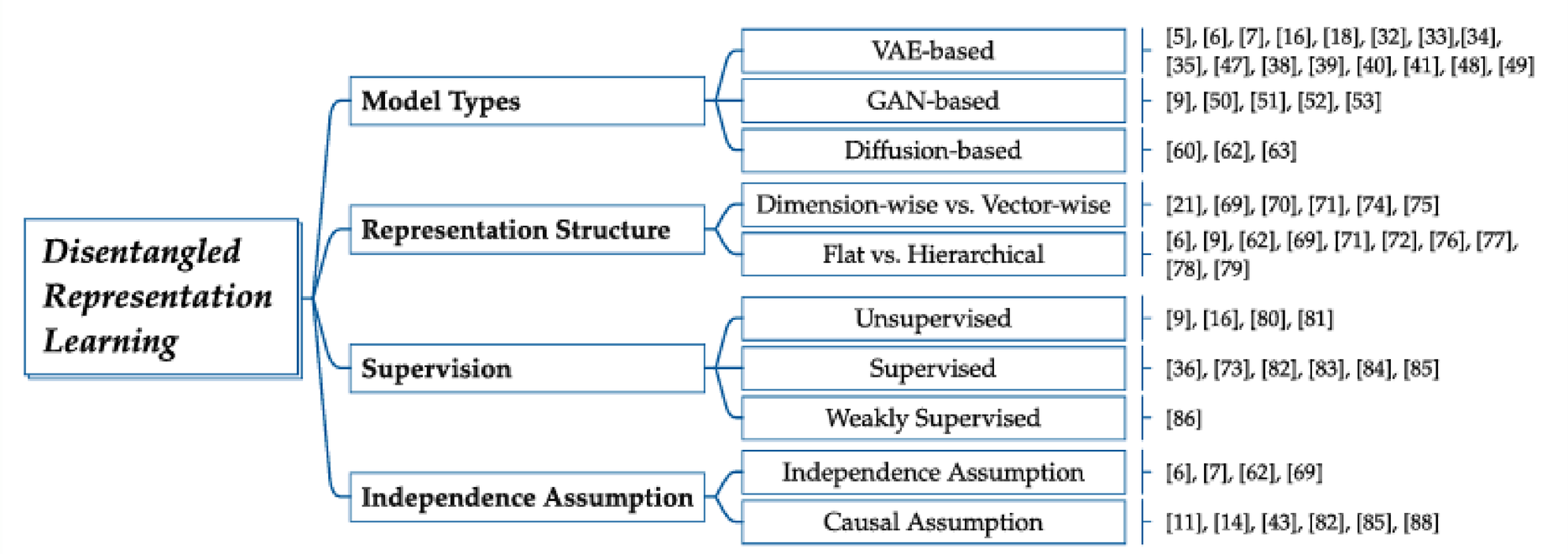

Fig. 3. A categorization of DRL approaches.

VAE-Based Methods

Variational auto-encoder (VAE) is a variant of the auto encoder, which adopts the idea of variational inference.

The key idea of VAE:

- Model the data distribution in a way that maximizes the likelihood of the observed data while using variational inference
- Variational inference is a technique that allows us to approximate complex probability distributions, making the model efficient and effective.
- Implemented first on Frey Faces and MINIST datasets

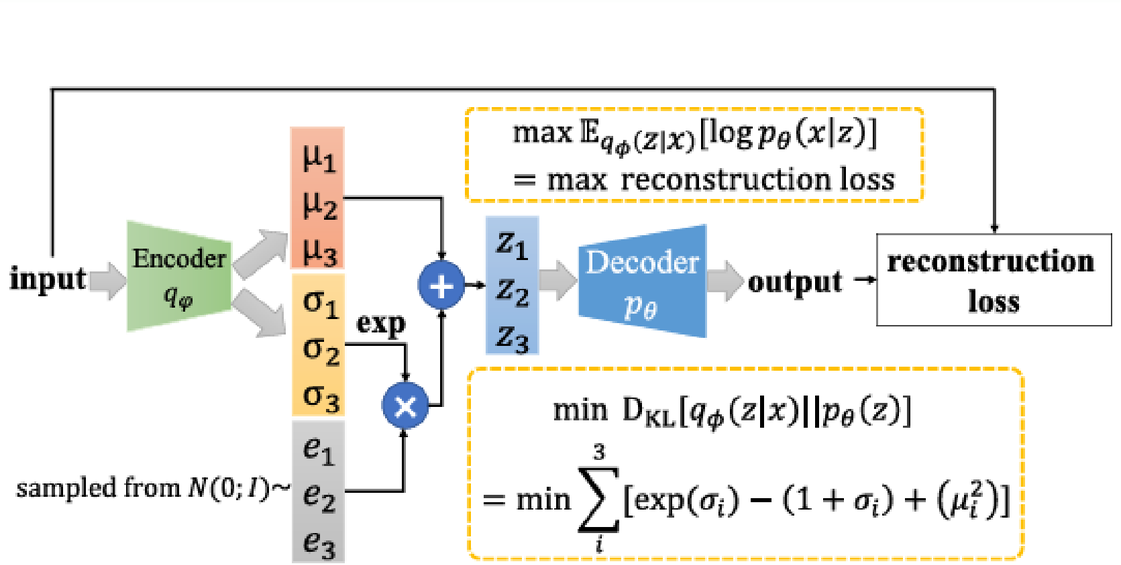

FIg. 4. The general framework of variational auto-encoder (VAE)

THE SUMMARY OF VAE BASED APPROACHES

Method	Regularizer	Description
β-VAE	$-\beta D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)$	β controls the trade-off between reconstruction fidelity and the quality of disentanglement in latent representations.
Understanding disentangling in β-vae	$-\gamma\left\|D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(z)\right)-C\right\|$	The quality of disentanglement can be improved as much as possible without losing too much information from original data by linearly increasing C during training.
DIP-VAE	$-\lambda D_{KL}\left(q_\phi(\mathbf{z})\|p(\mathbf{z})\right)$	Enhance disentanglement by minimizing the distance between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$. In practice, we can match the moments between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$.
FactorVAE	$-\gamma D_{KL}\left(q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\right)$	Directly impose independence constraint on $q_\phi(\mathbf{z})$ in the form of total correlation.
β-TCVAE	$-\alpha I_q(\mathbf{z};\mathbf{x})-\beta D_{KL}\left(q(\mathbf{z})\|\prod_j q(z_j)\right)-\gamma\sum_j D_{KL}\left(q(z_j)\|p(z_j)\right)$	Decompose $D_{KL}\left(q(\mathbf{z}\|\mathbf{x})\|\|p(\mathbf{z})\right)$ into three terms: i) mutual information, ii) total correlation, iii) dimension-wise KL divergence and then penalize them respectively.
JointVAE	$-\gamma\left\|D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)-C_z\right\|-\gamma\left\|D_{KL}\left(q_\phi(\mathbf{c}\|\mathbf{x})\|p(\mathbf{c})\right)-C_c\right\|$	Separate latent variables into continuous $\mathbf{z}$ and discrete $\mathbf{c}$, then modify the objective function of β-VAE to capture discrete generative factors.
RF-VAE	$-\sum_{j=1}^{d}\lambda(r_j)D_{KL}\left(q(z_j\|\mathbf{x})\|p(z_j)\right)-\gamma D_{KL}\left(q(\mathbf{r}\circ\mathbf{z})\|\prod_{j=1}^{d}q(r_j\circ z_j)\right)-\eta\|\mathbf{r}\|_1$	Introduce relevance indicator variables $\mathbf{r}$ by only focusing on relevant part when computing the total correlation, penalize $D_{KL}\left(q(z_j\|\mathbf{x})\|\|p(z_j)\right)$ less for relevant dimensions and more for nuisance (noisy) dimensions.

GAN-Based Methods

GANs consist of a generator that creates data and a discriminator that evaluates whether the data is real or fake. While, the InfoGAN method extends GANs by adding a mutual information regularizer.

The goal of GAN-based DRL methods:

- ensure that latent variables control interpretable and independent factors of variation

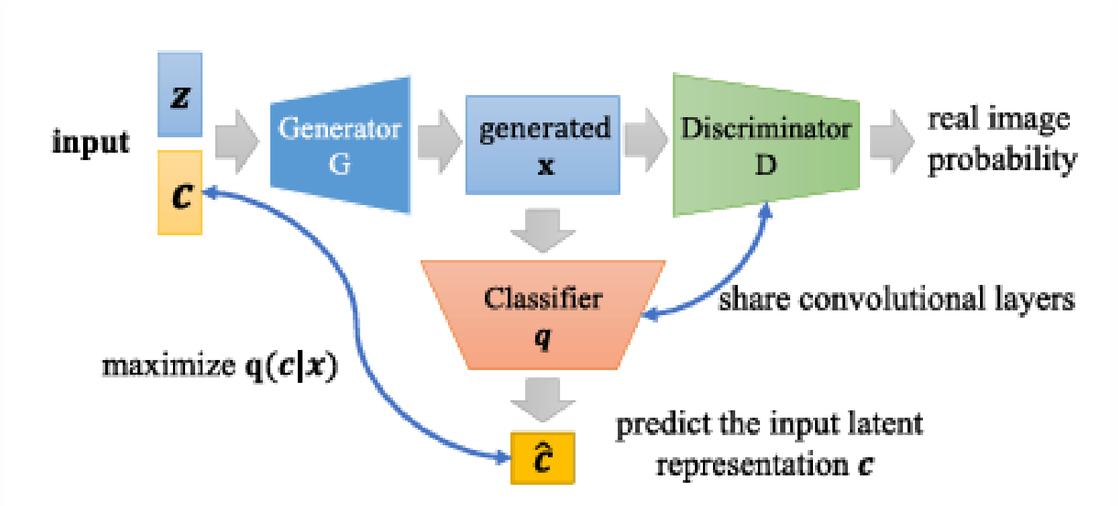

Fig. 5. The overall framework of InfoGAN

Diffusion-Based Methods

Generative models that transform data from noise to meaningful data through a process of gradual noise addition and removal. While the DisDiff method applies diffusion to disentangle these factors in the latent space.

The goal of using diffusion models in DRL:

- separate generative factors in a robust and interpretable manner.
  - DisDiff disentangles these factors by learning separate latent variables for each factor, resulting in high-quality data generation with control over specific attributes.

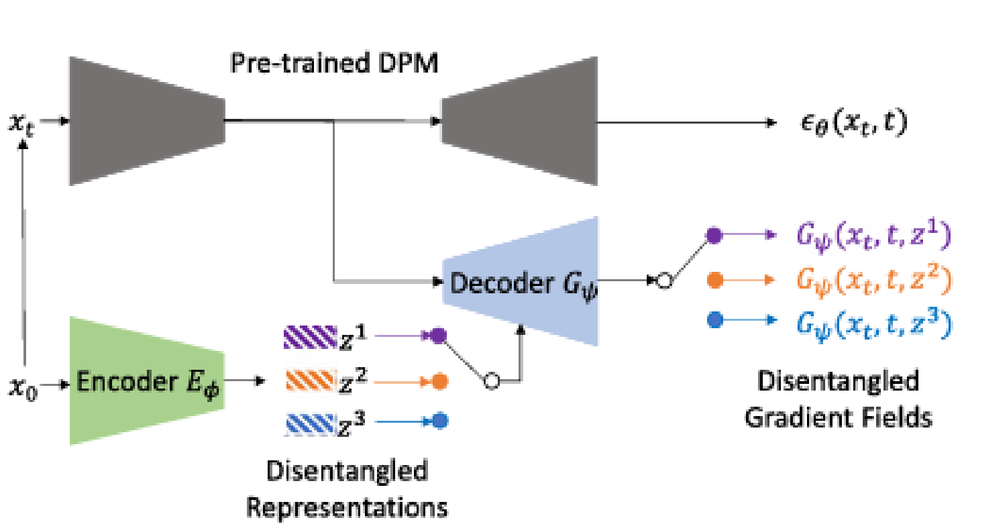

Fig. 6. The network structure of DisDiff

DRL Design

1. **Design of Representation Structure:**

- Dimension-Wise DRL:
  - Each latent variable represents a fine-grained factor (e.g., shape, size).
  - Best for simple datasets with clear factors.
- Vector-Wise DRL:
  - Groups of latent variables represent coarse-grained factors (e.g., identity, pose).
  - Suitable for real-world tasks requiring more complex information.

- Flat DRL:
  - All latent variables are treated equally without abstraction.
  - Example: DRL models for datasets like CelebA (attributes like smile, rotation).
- Hierarchical DRL:
  - Latent variables are organized into layers of abstraction (e.g., gender → smile).
  - Useful for complex generative processes.

2. Design of Loss Function

- Purpose of the Loss Function in DRL:
  - Ensures the model learns disentangled latent variables while balancing other objectives like reconstruction quality or task performance.

- Key Components:
  - **KL Divergence Regularizer:** Forces independence among latent variables by penalizing their overlap.
  - **Total Correlation (TC):** Measures statistical dependence among variables, encouraging each to capture unique information.
  - **Reconstruction Loss:** Ensures the latent variables can be used to accurately reconstruct the original data.

Representation Structure

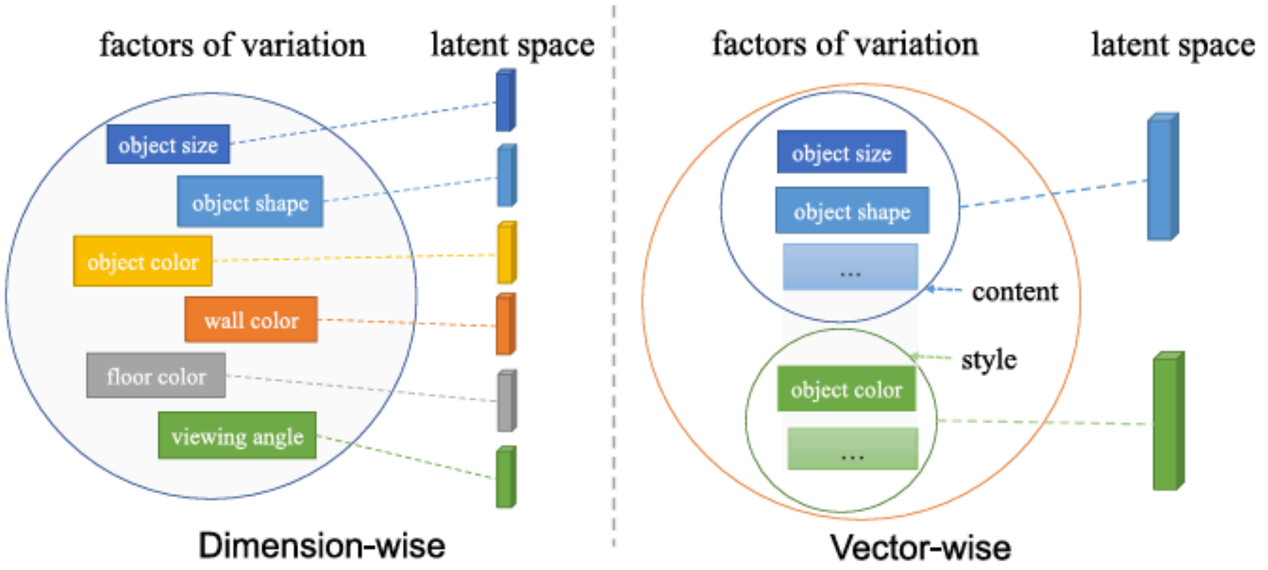

Fig. 7. Dimension-wise VS vector wise DRL

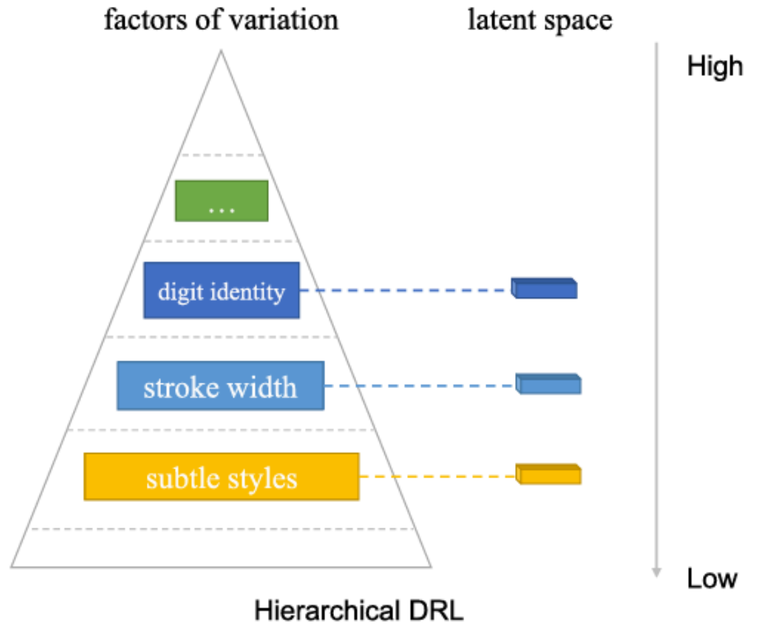

Fig. 8. Hierarchical DRL

Applications

01 Image Generation

02 Video

03 Natural Language Processing
- Text representation
- Style transfer

04 Recommender Systems

05 Graph Representation Learning

Limitations & Future Improvements

Limitations:

1. Unsupervised DRL Challenges:
   - Pure unsupervised disentanglement is impossible without inductive biases on methods or datasets.
2. Evaluation Metrics:
   - Current metrics for measuring disentanglement are task-specific and may not generalize.
3. Scalability Issues:
   - Many existing DRL methods struggle to handle complex, real-world datasets.

Future Improvements:

1. Scalable and Robust Models:
   - Design architectures that are resilient to noise and can scale to diverse datasets.
2. Better Evaluation Metrics:
   - Establish standardized, general-purpose metrics to evaluate disentanglement.

References

X. Wang, H. Chen, S. Tang, Z. Wu and W. Zhu, "Disentangled Representation Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 9677-9696, Dec. 2024, doi: 10.1109/TPAMI.2024.3420937.

Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013.Add a little bit of body text

R.T.Chenetal.,“IsolatingsourcesofdisentanglementinVAEs,”inProc. 32nd Int. Conf. Neural Inf. Process. Syst., 2019, pp. 2615–2625.

QUIZ

Collab Notebook