# A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations

ORAL PRESENTATION ACL-2021

**Pierre Colombo**[†⋆], Pablo Piantanida[*], Chloé Clavel[⋆]
[⋆]Télécom ParisTech, Université Paris Saclay
[†] IBM GBS France
[*] L2S, CentraleSupelec CNRS Universite Paris-Saclay

# Importance of Disentangled Representations

- Audio processing [11]
- Video processing [10]
- Visual reasoning [15]
- Robust and fair classification [1]
- Few-shot learning [12]
- Style transfer [8]
- Conditional generation [5, 3]
- ...

# Learning to Disentangle representations

## Problem Definition

**Main goal :** To learn a model $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^d$ that retains as much as possible information of the original content from the input sentence $X \in \mathcal{X}$ but as little as possible about the sensitive attribute $Y \in \mathcal{Y}$.

**Assumption :** $Y \in \mathcal{Y}$ is a discrete attribute or concept.

# Contributions

- A novel variational-based estimator of the Mutual Information (MI)

- Applications and numerical results :
  - Fair textual classification
  - Text style transfer

## Learning Objective

**Mutual Information :** Given two r.v $Z$ and $Y$, the MI is defined by

$$I(Z;Y) = \mathbb{E}_{ZY}\left[\log \frac{p_{ZY}(Z,Y)}{p_Z(Z)p_Y(Y)}\right] = H(Y) - H(Y|Z),$$

where $p_{ZY}$ is the joint pdf and $p_Z$ and $p_Y$ are the marginal pdfs.

## Learning Objective

**Mutual Information :** Given two r.v $Z$ and $Y$, the MI is defined by

$$I(Z;Y) = \mathbb{E}_{ZY}\left[\log \frac{p_{ZY}(Z,Y)}{p_Z(Z)p_Y(Y)}\right] = H(Y) - H(Y|Z),$$

where $p_{ZY}$ is the joint pdf and $p_Z$ and $p_Y$ are the marginal pdfs.

Computing the MI is a long standing challenge [2, 9, 14].

## Learning Objective

**General Loss to Minimize :**

$$\mathcal{L}(f_{\theta_e}) \equiv \underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X); Y)}_{\text{disentangled}},$$

$\mathcal{L}_{down.}$ is the task loss, $f_{\theta_e}$ is the encoding function.

## A novel Variational-Based Estimator of MI

$(Z, Y) \sim p_{ZY}$, $q_{\hat{Y}|Z}$ be a conditional variational distribution

# A novel Variational-Based Estimator of MI

$(Z, Y) \sim p_{ZY}$, $q_{\hat{Y}|Z}$ be a conditional variational distribution

$$I(Z;Y) = H(Y) - H(Y|Z)$$

## A novel Variational-Based Estimator of MI

$(Z, Y) \sim p_{ZY}$, $q_{\hat{Y}|Z}$ be a conditional variational distribution

$$I(Z; Y) = H(Y) - H(Y|Z)$$

**Upper Bound on H(Y) :**

$$\begin{aligned} H(Y) &\leq \mathbb{E}_Y\left[-\log q_Y(Y)\right] \\ &= \mathbb{E}_Y\left[-\log \int q_{\widehat{Y}|Z}(Y|z)p_Z(z)dz\right] \end{aligned}$$

# A novel Variational-Based Estimator of MI

**Lower Bound on $H(Y|Z)$ :**

$$H(Y|Z) = \mathbb{E}_{YZ}\left[-\log q_{\widehat{Y}|Z}(Y|Z)\right] - \mathrm{KL}(p_{YZ}\|p_Z \cdot q_{\widehat{Y}|Z})$$

# A novel Variational-Based Estimator of MI

**Lower Bound on $H(Y|Z)$ :**

$$H(Y|Z) = \mathbb{E}_{YZ}\left[-\log q_{\widehat{Y}|Z}(Y|Z)\right] - \mathrm{KL}(p_{YZ}\|p_Z \cdot q_{\widehat{Y}|Z})$$

Let be $D_\alpha(\cdot\|\cdot)$ the Renyi divergence with $\alpha > 1$ :

$$H(Y|Z) \leq \mathbb{E}_{YZ}\left[-\log q_{\widehat{Y}|Z}(Y|Z)\right] - D_\alpha(p_{YZ}\|p_Z \cdot q_{\widehat{Y}|Z}).$$

# A novel Variational-Based Estimator of MI

## Variational upper bound on MI

$$I(Z;Y) \leq \mathbb{E}_Y\left[-\log \int_{R^d} Q_{\widehat{Y}|Z}(Y|z)P_Z(dz)\right] + \mathbb{E}_{YZ}\left[\log Q_{\widehat{Y}|Z}(Y|Z)\right] + D_\alpha\left(P_{ZY}\|P_Z Q_{\widehat{Y}|Z}\right),$$

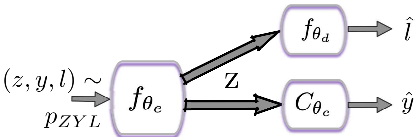
where

$$D_\alpha\left(P_{ZY}\|P_Z Q_{\widehat{Y}|Z}\right) = \frac{1}{\alpha-1}\log \mathbb{E}_{ZY}[R^{\alpha-1}(Z,Y)]$$

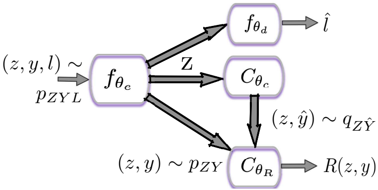
denotes the Renyi divergence and $R(z,y) = \frac{P_{Y|Z}(y|z)}{Q_{\widehat{Y}|Z}(y|z)}$.

# Comparison with Previous Works

**Adversarial Losses :** [16, 1, 6]




**Our model :**




$$\underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{CE(\hat{Y}, Y)}_{\text{adv}}$$

$$\underbrace{\mathcal{L}_{down.}(f_{\theta_e})}_{\text{downstream task}} + \lambda \cdot \underbrace{I(f_{\theta_e}(X); Y)}_{\text{disentangled}}$$

## Related Work

**Adversarial training loss** : $CE(\hat{Y}, Y)$ is a lower bound (up to a constant) of the MI

# Related Work

**Adversarial training loss** : $CE(\hat{Y}, Y)$ is a lower bound (up to a constant) of the MI

### Limitation of Adversarial Losses

- Disentanglement is not perfect [6]
- Adversarial Losses Fail for $|\mathcal{Y}| > 2$

## Related Work

**vCLUB** [7, 4]

$$I_{\text{vCLUB}}(Y;Z) = \mathbb{E}_{YZ}[\log p_{Y|Z}(Y|Z)] - \mathbb{E}_Y\mathbb{E}_Z[\log p_{Y|Z}(Y|Z)]$$

**Limitation of vCLUB**

- No fine-grained control of the degree (or force) of the disentanglement [7].

# Application to Fair Classification

**DIAL corpus** : The main task consists in predicting a binary sentiment (positive/negative). The considered protected attribute is the race.

# Application to Fair Classification

**DIAL corpus** : The main task consists in predicting a binary sentiment (positive/negative). The considered protected attribute is the race.

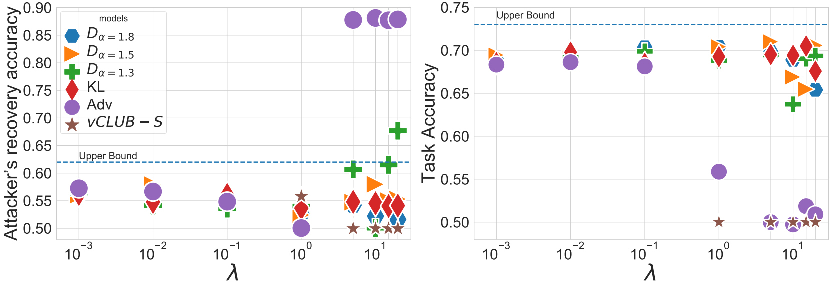



FIGURE – Numerical results on fair classification. Trade-offs between target task and attacker accuracy are reported for sentiment task.

# Application to Textual Style Transfer

**Yelp corpus** : Review from Yelp. The task consists in transferring a **binary label (left) or multiple category (right)**. [17, 13]

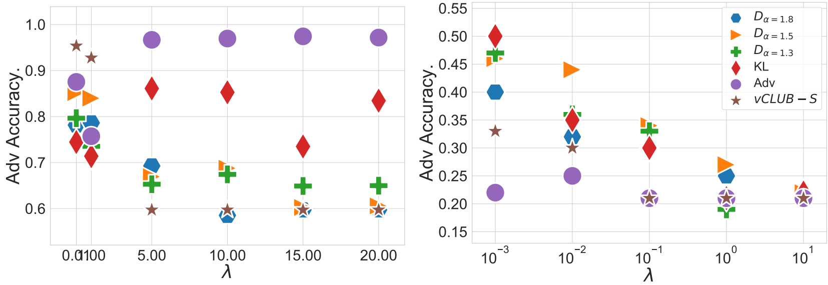
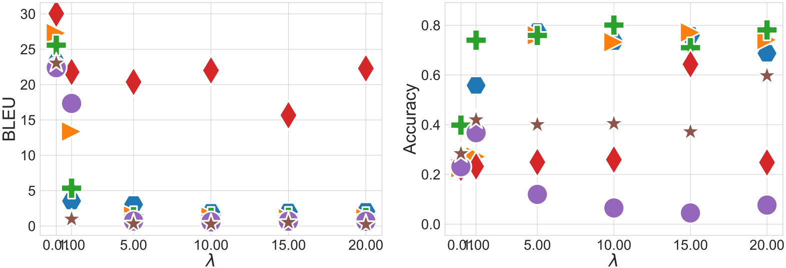



FIGURE – Disentanglement of the representations learnt by the encoder $f_{\theta_e}$ when the model is trained on a **binary (left) and multi-label (right)** sentence generation task.

# Binary Style Transfer




(a) (b)

FIGURE – Numerical experiments on binary style transfer. Quality of generated sentences are evaluated using BLEU (3a) ; style transfer accuracy (3a).

# Binary Style Transfer

| | Input | **It's freshly made, very soft and flavorful.** |
|---|---|---|
| 0.1 | Adv | it's crispy and too nice and very flavor. |
| | KL | it's a huge, crispy and flavorful. |
| | $D_{\alpha=1.3}$ | it's hard, and the flavor was flavorless. |
| | $D_{\alpha=1.5}$ | it's very dry and not very flavorful either. |
| | $D_{\alpha=1.8}$ | it's a good place for lunch or dinner. |
| | Input | it's freshly made, very soft and flavorful. |
| 1 | Adv | it's not crispy and not very flavorful flavor. |
| | KL | it's very fresh, and very flavorful and flavor. |
| | $D_{\alpha=1.3}$ | it's not good, but the prices are good. |
| | $D_{\alpha=1.5}$ | it's not very good, and the service was terrible. |
| | $D_{\alpha=1.8}$ | it was a very disappointing experience and the food was awful. |
| | Input | it's freshly made, very soft and flavorful. |
| 10 | Adv | i hate this place. |
| | KL | it's a little warm and very flavorful flavor. |
| | $D_{\alpha=1.3}$ | it was a little overpriced and not very good. |
| | $D_{\alpha=1.5}$ | it's a shame, and the service is horrible. |
| | $D_{\alpha=1.8}$ | it's not worth the $ NUM. |

TABLE – Sequences generated on the binary sentiment transfer task.

# Concluding Remarks and Perspectives

**Summary of our contributions**

- A New estimator of the MI based on a variational upper bound.
- New method capable of learning disentangled textual representation.
- Our method provides better tradeoffs for Fair Classification tasks.
- There is no free-lunch for sentence generation tasks : transferring style is easier with disentangled representations, but removes important information about the content.

## References I


Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard.
Adversarial removal of demographic attributes revisited.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm.
Mine : mutual information neural estimation.
*arXiv preprint arXiv :1801.04062*, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner.
Understanding disentangling in $\beta$-vae.
*arXiv preprint arXiv :1804.03599*, 2018.

# References II


Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin.
Club : A contrastive log-ratio upper bound of mutual information.
In *International Conference on Machine Learning*, pages 1779–1788. PMLR, 2020.

Emily L Denton et al.
Unsupervised learning of disentangled representations from video.
In *Advances in neural information processing systems*, pages 4414–4423, 2017.

Yanai Elazar and Yoav Goldberg.
Adversarial removal of demographic attributes from text data.
*arXiv preprint arXiv :1808.06640*, 2018.

Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel.
Learning anonymized representations with adversarial neural networks, 2018.

## References III

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan.
Style transfer in text : Exploration and evaluation.
*arXiv preprint arXiv :1711.06861*, 2017.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio.
Learning deep representations by mutual information estimation and maximization.
*arXiv preprint arXiv :1808.06670*, 2018.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles.
Learning to decompose and disentangle representations for video prediction.
In *Advances in Neural Information Processing Systems*, pages 517–526, 2018.

## References IV


Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang.
Learning disentangled representations for timber and pitch in music audio.
*arXiv preprint arXiv :1811.03271*, 2018.

Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai.
Generalized zero-shot learning via synthesized examples.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4281–4289, 2018.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han.
Mining quality phrases from massive text corpora.
In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM, 2015.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals.
Representation learning with contrastive predictive coding.
*arXiv preprint arXiv :1807.03748*, 2018.

# References V


Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem.
Are disentangled representations helpful for abstract visual reasoning ?
In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig.
Controllable invariance through adversarial feature learning.
In *Advances in Neural Information Processing Systems*, pages 585–596, 2017.

Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han.
Personalized entity recommendation : A heterogeneous information network approach.
In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 283–292. ACM, 2014.