Joint Wasserstein Autoencoders for Aligning Multimodal Embeddings

Shweta Mahajan Teresa Botschen Iryna Gureyvch Stefan Roth

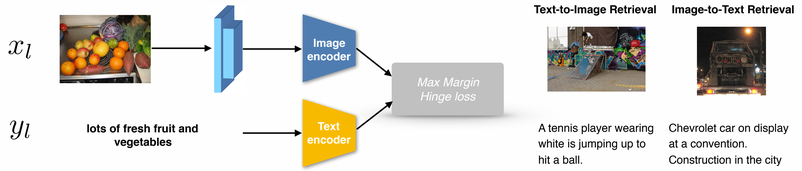
Department of Computer Science, TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction

- **Goal:** Alignment of visual and textual distributions for image text matching tasks
- **Supervised approaches** leverage information *only* from paired data
- **Drawbacks:** limited cross dataset generalization for real world scenarios

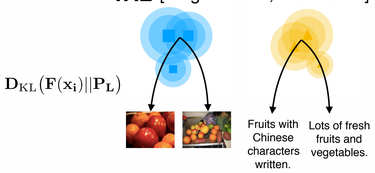
Previous work



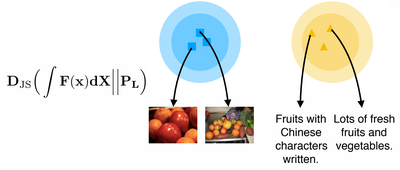
Overfits to particularities of the dataset

Wasserstein Autoencoder Backbone

VAE [Kingma et al., ICLR 2014]



WAE [Tolstikhin et al., ICLR 2018]



Semantic Continuity in Latent Spaces

Embedding Network

jWAE

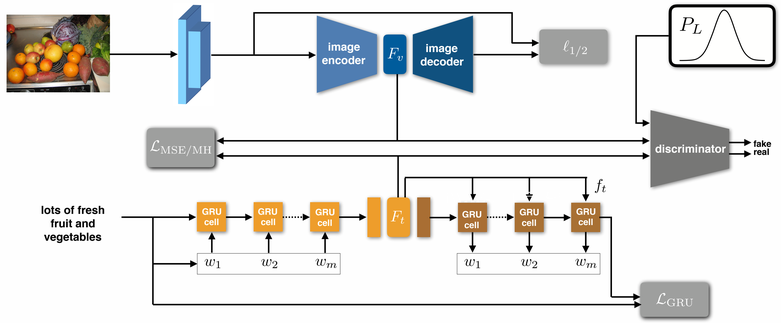
Similar images are close to each other
Similar sentences are close to each other
Similar images *and* sentences are close to each other

[Embedding Network] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text match-ing tasks. TPAMI, 41(2), 2019
[VSE++] F. Faghri, D. Fleet, R. Kiros and S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In BMVC, 2018
[SCAN] K. Lee, Xi Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In ECCV, 2018
[CITE] B. Plummer, P Kordas, M. Kiapour, S Zheng, R Piramuthu and S Lazebnik. Conditional image-text embedding networks. In ECCV, 2018

AIPHES

This work has been supported by the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources (AIPHES)* under grant No. GRK 1994/1.

jWAE Network Architecture



Loss Function

$$\mathcal{L}_{\mathrm{jWAE}} = \lambda_1 \sum_{i=1}^{N_X} \left\| x_i - g_v\left(f_v(x_i)\right) \right\| - \lambda_2 \sum_{j=1}^{N_Y} \sum_{m=0}^{M-1} \log p_{g_t}\left(w_m^j \middle| w_{0:m-1}^j, f_t(y_j); g_t\right)$$

Reconstruction Errors

$$+ \lambda_3 \mathbf{D}_{\mathrm{JS}}\left(F_v || P_L\right) + \lambda_4 \mathbf{D}_{\mathrm{JS}}\left(F_t || P_L\right)$$

Jensen Shannon Divergence

- Regularization parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$
- Encoders f_v, f_t
- Decoders g_v, g_t
- Encoded distributions F_v, F_t
- Decoded distributions G_v, G_t
- Output probability of word p_{g_t}
- Shared Gaussian prior P_L

$$\mathcal{L}_{\text{jWAE-MH}} = \mathcal{L}_{\mathrm{jWAE}} + \mathcal{L}_{\mathrm{MH}}$$

Image latent space without supervision

Embedding Network

jWAE

jWAE achieves semantic continuity in the coordinated image space in absence of within-domain supervision

Experiments

Cross-modal Retrieval on COCO dataset

Method	Image-to-text R@1	Text-to-image R@1
Emb. Network	50.1	39.6
VSE++	64.6	52.0
SCAN	70.9	56.4
jWAE-MH	**66.6**	**53.1**
jWAE-MSE	**50.3**	**25.2**
SCAN+jWAE-MH	**72.0**	**57.1**

Improves accuracy over state of the art

Cross-modal Generalization across Datasets

Method	Image-to-text R@5	Text-to-image R@5
Emb. Network	33.7	8.4
SCAN	73.7	61.3
jWAE-MH	**51.1**	**37.0**
SCAN+jWAE	**80.0**	**66.7**

Train - COCO Test- Flickr30k

Better generalization across datasets

Examples of images or texts retrieved on Flickr30k based on embeddings trained on COCO

Text	Supervised Approach	jWAE (Ours)
Two tan dogs play on the grass near the wall.		
A tennis player wearing white is jumping up to hit a ball.		

Image	Supervised Approach	jWAE (Ours)
	The dog is running around the cow.	A lioness is chasing a black bison across a grassy plain.

jWAE improves the generalization across datasets owing to the semantic continuity from the Gaussian regularization

Phrase Localization on Flickr30k Entities dataset

Method	R@1
Emb. Network	51.0
CITE	59.2
jWAE-MSE	**52.5**
CITE+jWAE	**60.4**

jWAE can be used across cross modal tasks

A young man in a t-shirt is speaking at a podium while another young man stands by.