Assessing and Improving the Quality of Image-Text Alignment in the Context of Visual Language Pretraining Models

MLBriefs Workshop - May 2023

François Role
PEReN and université Paris Cité

Plan

1. Introduction

2. Experimenting with Different Versions of the Bidirectional Contrastive Loss

3. A Simple yet Effective Solution to Improve VLP-based Image-Text Representations

Learning Aligned Image-Text Representations

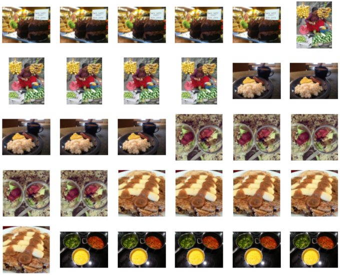
Filling the gap between different modalities is very useful in many situations :

- Multimodal information-retrieval
- Multimodal classification and clustering
- Automatic image captioning
- Scene understanding
- etc.

A Well-known Application : Multimodal Information Retrieval

Query : "Food"

Response :



Vision-language pretrained models (VLP models) are today the cornerstone for this type of application

Vision-Language Pretraining (VLP)

VLP models are the next logical step after pretrained CNN Models, and pretrained Transformer-based language models !

- Build models that jointly encode **vision and language**. Use these **pretrained** models to serve as a basis for multimodal downstream applications
- Learn from a large set of (image, text) pairs using the simple pre-training task of predicting which caption goes with which image
- It has been shown that so trained models can directly transfer to many interesting downstream tasks in a zero-shot manner

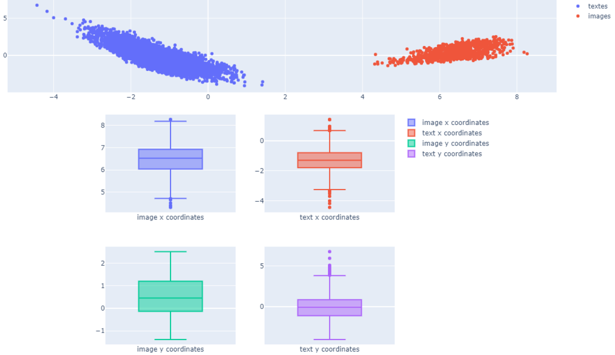
Aligning the Feature Spaces of Different Modalities

- We focus on models that align the feature spaces of different modalities (limiting ourselves to the image and text modalities) using late fusion such a CLIP [1] or ALIGN [2]
  - Please note that many other approaches are possible : first aligning the features from the different modalities and then fusing them using a transformer encoder, as proposed in ALBEF [3].
- Late fusion models mostly rely on cross-modal contrastive Learning technique

Cross-modal Contrastive Learning

- Use a set of input pairs (u, v) where u is an image and v represents a text which describes u.
- Learn encoders f_θ and g_ϕ that convert u and v into d-dimensional vectors $f_\theta(u)$ and $g_\phi(v)$ resp.
- For each "true" pair (u, v), try to maximize the "agreement" between $f_\theta(u)$ and $g_\phi(v)$
  - In the image-text case the training relies on a **bidirectional contrastive loss**

The Promise and the Reality



Why and how to Improve the Alignments ?

- Problem : a bad alignment may prevent direct use of the representations in many scenarios. For example, directly clustering the points shown in the previous figure would lead to closely related texts and images being put in different clusters
- In this presentation we report on solutions we designed to modify the out-of-the box, VLP-based representations so that they be better aligned :
  - by using a modified version of the standard, one-hot bidirectional contrastive loss
  - by using spectral methods

Plan

1. Introduction

2. Experimenting with Different Versions of the Bidirectional Contrastive Loss

3. A Simple yet Effective Solution to Improve VLP-based Image-Text Representations

Bidirectional Contrastive Loss : "standard" version

Targets

Use the one-hot ground-truth : negative pairs have a probability of 0 and the positive pair has a probability of 1

Definition

Let x_i and y_i be the normalized embeddings of the image and text in the i-th pair. For N pairs, we seek to minimize the following sum of two losses :

$$-\frac{1}{2N}\left(\sum_{i}^{N}\log\frac{\exp(x_i^T y_i)}{\sum_j^N \exp(x_i^T y_j)}+\sum_{i}^{N}\log\frac{\exp(x_i^T y_i)}{\sum_j^N \exp(y_i^T x_j)}\right)$$

Bidirectional Contrastive Loss : "soft" version

Targets

Compute the targets

$$P = (p_{ij}) = softmax((E_i E_i^T + E_w E_w^T)/2) \tag{1}$$

$$Q = (q_{ij}) = E_i E_w^T \tag{2}$$

where E_i and E_w are the image and text embedding matrices resp.

Cross-entropy loss

$$-\frac{1}{2N}\Big[\sum_i^N\Big(\sum_j^N p_{ij} \log softmax_r(q_{ij})\Big)+\sum_j^N\Big(\sum_i^N p_{ij} \log softmax_c(q_{ij})\Big)\Big]$$

where $softmax_r$ and $softmax_c$ stand for the softmax function computed along the rows or along the columns resp.

Tentative Conclusion

- We did not notice significant differences between the representations obtained using either of the two losses
- However, our experiments were of a very exploratory nature, so further research would be needed

Plan

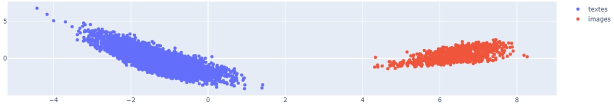
1. Introduction

2. Experimenting with Different Versions of the Bidirectional Contrastive Loss

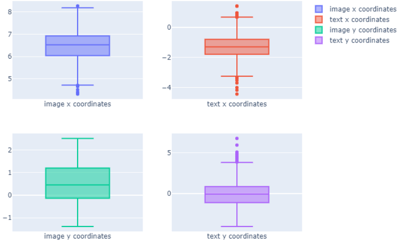
3. A Simple yet Effective Solution to Improve VLP-based Image-Text Representations

Recall the kind of representations that we get in the end

ACP



Boxplot views



Consider a Seemingly Unrelated Problem

Consider the easier case where we deal with n data points from the same modality

Assume a symmetric $(n \times n)$ weighted adjacency matrix W, where D is the diagonal degree matrix $d_i = \sum_{j=1}^{n} w_{ij}$.

Problem

Find a low-dimensional representation that is in agreement with the observed affinities. If we project the weighted graph onto a line it means minimizing $\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 = \mathrm{f}^T \mathrm{L} \mathrm{f}$ where $\mathbf{f}$ is an n-dimensional vector whose component f_i is the coordinate of the i-th vertex.

Solution

- Solution : the vector corresponding to the smallest non-null eigenvalue of the Laplacian matrix $L = D - W$.

More Generally

Problem

Embed the weighted graph into a k-dimensional space. Given the $k \times n$ matrix $F = [\mathrm{f}_1, \ldots, \mathrm{f}_\mathrm{n}]$, where vector f_i contains the coordinates of the i-th point, we have to minimize $\sum_{i,j=1}^{n} w_{ij} ||\mathrm{f}_i - \mathrm{f}_j||^2$ with the constraint $\mathrm{FDF}^\mathrm{T} = \mathrm{I}$.

Solution

- Solution : the matrix of the eigenvectors corresponding to the smallest eigenvalues of the generalized eigenvalue problem $\mathrm{Lf} = \lambda \mathrm{Df}$

Back to the Image-Text Case !

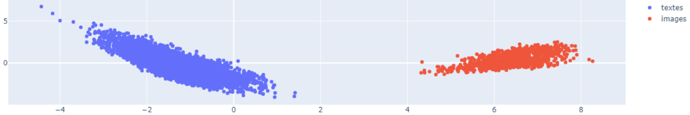
We can then solve our alignment problem in the following way :

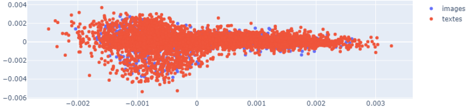
1. Given the image embedding matrix $\mathbf{U}$ and the text embedding matrix $\mathbf{V}$, compute $\mathbf{W} = \mathbf{U}\mathbf{V}^T$
2. Form the matrix $\mathbf{A} = \begin{bmatrix} 0 & \mathbf{W} \\ \mathbf{W}^T & 0 \end{bmatrix}$
3. Compute the matrix $\mathbf{F} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ where the u_i are the eigenvectors corresponding to the k smallest eigenvalues of $\mathbf{L}_{rw} = \mathbf{D}^{-1}\mathbf{L}$, with $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{D}$ being the degree matrix
4. Use the i-th row of $\mathbf{F}$ as the representation of the i-th point (be it an image or a text)

Experimenting with the COCO Dataset

- The first version of the COCO dataset "COCO 2014"
- Contains 164K images :
  - training set (83K images)
  - validation set (41K images)
  - test set (41K images)
- Each image is accompanied with 5 natural language descriptions

Before and After : it works !





Tentative Conclusion

- A mostly overlooked problem : the VLP embeddings are used out-of-the-box, without giving much thought to their nature
- However, better aligned data allow for better generalization and reuse on different tasks
- Paths for future research include :
    - Using additional models and datasets
    - Improving quantitative evaluation and comparison with other embedding methods
    - Comparing the native and post-processed embeddings in terms of cluster-friendliness

References I

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
Learning transferable visual models from natural language supervision.
In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

References II

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig.
Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[3] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi.
Align before fuse : Vision and language representation learning with momentum distillation, 2021.