Généralisation de domaine par translation dans l'espace CLIP

Louis Hémadou, Héléna Vorobieva, Ewa Kijak, Frédéric Jurie

Généralisation de domaine par translation dans l'espace CLIP

L. Hémadou[1,2,3] H.Vorobieva [1] E. Kijak[2] F. Jurie[3]

[1] Safran Tech, Digital Sciences & Technologies Department
[2] Université de Rennes 1, INRIA, CNRS
[3] Université de Caen Normandie, ENSICAEN, CNRS
email : louis.hemadou@safrangroup.com

Résumé

Ce travail aborde la problématique de la généralisation de domaine en apprentissage profond. L'objectif est de transférer un classifieur d'images entraîné sur un domaine source vers un domaine cible pour lequel seules des descriptions textuelles (et non des images) sont disponibles. Pour cela, nous utilisons CLIP, un modèle de représentation d'images et de textes dans un espace latent commun. Nous proposons une méthode originale qui consiste à transférer les images du domaine cible vers le domaine source en exploitant la différence entre les descriptions textuelles des domaines cible et source dans l'espace latent de CLIP. Il s'agit d'un travail préliminaire et nous présentons ici les grandes lignes de notre approche.

Mots Clef

Généralisation de domaine, représentation textuelle, CLIP

Abstract

This work addresses the problem of domain generalization in deep learning. The objective is to transfer an image classifier trained on a source domain to a target domain for which only textual descriptions (and not images) are available. To do this, we use CLIP, a model for representing images and texts in a common latent space. We propose an original method that consists of transferring the images from the target domain to the source domain by exploiting the difference between the textual descriptions of the target and source domains in CLIP's latent space. This is a preliminary work and we present the main outline of our approach in this document.

Keywords

Domain generalization, text representation, CLIP

1 Introduction et travaux liés

Les algorithmes d'apprentissage machine, et en particulier ceux d'apprentissage profond, font l'hypothèse que la distribution des données d'entraînement est identique à la distribution des données que devra traiter le modèle entraîné (données de validation). Dans beaucoup d'applications, il existe un écart entre ces distributions, et, sans attention particulière, les performances d'un algorithme d'apprentissage peuvent être impactées négativement par cet écart. Nous parlerons de domaine source pour se référer au domaine des données d'entraînement et de domaine cible pour celui des données de validation. Dans certains cas, la différence de domaine vient du fait qu'il n'existe aucune donnée réelle dans le domaine cible et que les seules données d'entraînement proviennent d'autres domaines. La généralisation de domaine vise à rendre un algorithme robuste face à un changement de distribution. Par exemple, nous pouvons vouloir utiliser des images de synthèse pour pouvoir reconnaître des objets réels pour lesquels nous n'avons pas d'images. Le lecteur peut, par exemple, se référer à [6] pour trouver une synthèse des travaux récents sur ce sujet.

Lorsqu'aucune donnée du domaine cible n'est disponible, il est intéressant de pouvoir transférer les données entre domaines en utilisant pour cela leur représentation textuelle. Un domaine pourrait par exemple se décrire par des phrases du type : "Des voitures dans des images de synthèse".

La transposition de représentation du texte vers l'image et de l'image vers le texte a été rendue possible récemment par l'introduction de représentation alignant les deux modalités. Par exemple, CLIP (Contrastive Language-Image Pre-Training) [5] est un modèle constitué d'un encodeur d'image et d'un encodeur textuel, entraîné avec un grand nombre de paires (image, texte). Son but est de rapprocher la représentation des éléments de paires (image, texte) dans un espace latent commun aux deux encodeurs, tout en éloignant les images et les textes ne provenant pas de la même paire. Comme CLIP a été entraîné avec des images et des descriptions textuelles nombreuses et dans des styles/domaines variés, nous pensons qu'il peut avoir de grandes facultés de généralisation.

Quelques articles récents s'intéressent au potentiel de CLIP pour la généralisation de domaine. Les travaux de Niu *et al*. [4] proposent de générer des descriptions textuelles d'une même classe dans des styles différents ("a photo of a dog", "a painting of a dog", ...) puis de les encoder avec l'encodeur de texte de CLIP. Les représentations ainsi obtenues sont agrégées afin d'obtenir une représentation invariante au domaine d'une classe.

L'espace latent de CLIP est aussi exploité pour faire du transfert de style. Kwon *et al*. [2] proposent de rapprocher une image d'une description textuelle du style désiré dans l'espace latent de CLIP, tout en conservant le contenu de manière analogue au transfert de style original. Enfin, Gal *et al*. [1] transfèrent un générateur d'images entraîné sur un domaine source (description textuelle t_{src}) vers un domaine cible (description textuelle t_{tgt}). Pour cela, les auteurs contraignent le générateur à générer des images ne variant que selon l'axe $t_{src} \rightarrow t_{tgt}$ dans l'espace latent de CLIP.

2 Idée proposée et validation expérimentale

Nous nous intéressons à une tâche de classification d'images pour laquelle les données d'entraînement sont dans un domaine particulier alors que les données de validation sont dans d'autres domaines. Nous nous proposons de transférer les images du domaine cible vers le domaine source en exploitant la différence entre les descriptions textuelles des domaines cible et source dans l'espace latent de CLIP. Notons E_I et E_T les encodeurs d'image et de texte de CLIP.

Le classifieur que nous utilisons est un réseau de neurones composé d'un encodeur d'image E_I fixé pendant l'entraînement suivi d'un régresseur logistique C. L'entraînement se fait avec des images du domaine source pour lequel une description textuelle est disponible, notée t_{src}, par exemple "une photo", "un cartoon"... Lors de l'inférence sur une image x provenant d'un domaine cible pour lequel on ne connaît que la description textuelle t_x, on applique la translation $E_I(x) + \Delta$ dans l'espace latent de CLIP, avec $\Delta = E_T(t_{src}) - E_T(t_x)$. Il est alors possible de prédire la classe de x avec $C(E_I(x) + \Delta)$. Nous mesurons la capacité d'une description textuelle $t_{\mathcal{D}}$ à représenter un domaine $\mathcal{D}$ en calculant la similarité moyenne entre $t_{\mathcal{D}}$ et les images de $\mathcal{D}$ dans l'espace latent de CLIP :

$$Sim(t_{\mathcal{D}}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{E_t(t_{\mathcal{D}}) \cdot E_i(x)}{\|E_t(t_{\mathcal{D}})\|_2 \|E_i(x)\|_2}$$

Pour évaluer cette technique de généralisation de domaine, nous avons utilisé la base de données PACS [3]. Elle comprend 4 domaines ('photo', 'art painting', 'cartoon', 'sketch') et 7 classes. La Table 1 donne la matrice de similarité moyenne entre les descriptions textuelles des domaines et les images des domaines considérés. Les similarités sont conformes à nos attentes. Pour ce qui concerne les expériences de classification avec changement de domaine, le classifieur est entraîné sur les images d'un des domaines, les 3 autres étant utilisés comme domaines cibles. Nous ne présentons pas ici les résultats faute de place.

description textuelle / domaine	photo	art painting	cartoon	sketch
"a photo"	<u>0.226</u>	0.190	0.175	0.183
"an art painting"	0.220	<u>0.249</u>	0.204	0.216
"a cartoon"	0.215	0.199	<u>0.228</u>	0.204
"a black and white drawing"	0.226	0.226	0.247	<u>0.265</u>

TABLE 1 – Similarité moyenne entre les images d'un domaine et une description textuelle.

3 Discussions

Des résultats préliminaires que nous avons obtenus en classification ne semblent pas apporter de gains de performance, comparé à une utilisation directe du modèle sur les données de validation. Cela soulève deux questions que nous discutons ci-dessous.

(1) Comment représenter les domaines par des descriptions textuelles ? Il semble raisonnable de supposer que les descriptions textuelles des images d'entraînement de CLIP ressemblant à un "cartoon", un "art painting" ou un "sketch" incluaient l'information de style de l'image. De plus, il est probable que les images d'entraînement photoréalistes n'incluaient pas dans leur description textuelle le style de l'image, car le style photoréaliste est le style "par défaut". Ainsi, CLIP comprend mieux ce qu'est "an art painting" plutôt que "a photo". C'est pourquoi le domaine source est moins bien représenté par sa description textuelle. On peut imaginer d'autres manières de formuler des descriptions textuelles. Par exemple, en s'inspirant de [4], pour un domaine donné (par exemple le domaine "photo"), générer N descriptions textuelles spécifiques à une classe factice : "a photo of a goose", "a photo of a goldfish", etc., puis, moyenner les encodages de ces descriptions textuelles pour obtenir une représentation du domaine invariant à la classe.

(2) La translation est-elle la bonne opération pour aligner les domaines ? En supposant qu'il soit possible de construire des représentations textuelles fidèles au domaine et invariantes à la classe, il n'est pas garanti que la translation soit une opération adaptée à l'espace latent de CLIP. En effet, CLIP n'a pas été entraîné pour que des translations aient du sens dans son espace latent. Nous comptons étudier, dans le futur, d'autres opérateurs de transfert de domaines.

Références

[1] Rinon Gal, Or Patashnik, Haggai Maron, and et.al. Stylegannada : Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4), 2022.

[2] Gihyun Kwon and Jong Chul Ye. CLIPstyler : Image Style Transfer with a Single Text Condition. In *CVPR*. arXiv, 2022.

[3] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *ICCV*. arXiv, October 2017.

[4] Hongjing Niu, Hanting Li, Feng Zhao, and Bin Li. Domain-Unified Prompt Representations for Source-Free Domain Generalization, 2022.

[5] Alec Radford, Jong Wook Kim, and et. al. Hallacy. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[6] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization in Vision : A Survey. *arXiv :2103.02503 [cs]*, July 2021.