« You Only Need Adversarial Supervision For Semantic Image Synthesis »

Claire BIZON MONROC, Idles MAMOU, Amine DJEGHRI
SORBONNE UNIVERSITE

Introduction

La Semantic Image Synthesis se définit comme l'art de générer une image à partir d'une carte sémantique (ex une image segmentée). Dans ce domaine, les approches les plus performantes sont celles basées sur les GANs (ex Pix2Pix, SPADE…). Cependant, le calcul du coût adversaire sur les images en sortie du générateur seulement ne produit pas de résultats qualitatifs dans ce domaine. Une solution communément adoptée est d'ajouter le coût calculé sur des représentations intermédiaires obtenues par des modèles pré-entraînés (ex VGG). Mais elle introduit une dépendance vis à vis de ce modèle et du Dataset sur lequel il a été entraîné !

L'objectif de ce papier est de **s'affranchir des modèles pré-entraînés en se basant seulement sur le coût adversarial** tout en garantissant une bonne qualité de génération.

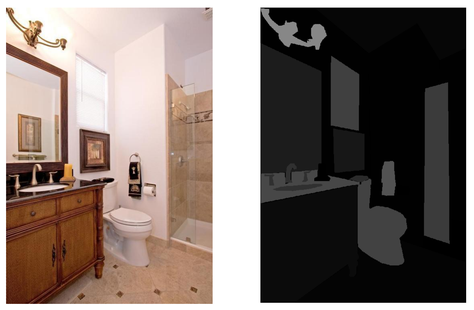
Les Données

L'apprentissage et l'évaluation se font sur le dataset de segmentation semantique **ADE20k2016.**

=> 20k exemples d'apprentissage.

=> 2k exemples de validation.

Contient N=150 catégories sémantiques qui incluent des éléments comme le ciel, l'herbe, la voiture ...



Exemple d'une image et sa carte d'annotation sémantique

Préparation

- Standardisation de la taille des images

Le modèle OASIS

OASIS reprend l'architecture GAN mais **redéfinit le discriminateur D comme un modèle de segmentation** où les cartes sémantiques sont données en supervision, de sorte à prédire pour chaque pixel sa classe de segmentation. Aux N classes sémantiques est ajoutée une dernière classe correspondant aux pixels générés par le générateur G (pour un total de N+1 classes).

Pondération par classe C

$$\mathcal{L}_D = -\mathbb{E}_{(x,t)}\left[\sum_{c=1}^{N} \alpha_c \sum_{i,j}^{H\times W} t_{i,j,c} \log D(x)_{i,j,c}\right] - \mathbb{E}_{(z,t)}\left[\sum_{i,j}^{H\times W} \log D(G(z,t))_{i,j,c=N+1}\right]$$

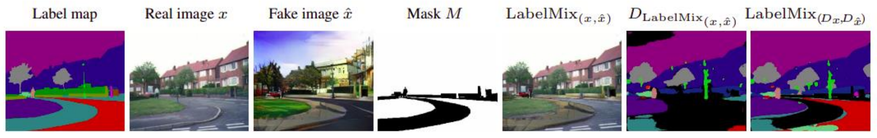
Label de classe

Afin d'encourager le discriminateur à se concentrer sur les différences de contenu et de structure entre les fausses et les vraies classes, une régularisation *LabelMix* est ajoutée à la fonction de coût:

Masque binaire respectant les frontières de la segmentation

$$\mathcal{L}_{cons} = \left\| D_{\text{logits}}\Big(\text{LabelMix}(x, \hat{x}, M)\Big) - \text{LabelMix}\Big(D_{\text{logits}}(x), D_{\text{logits}}(\hat{x}), M\Big)\right\|^2$$

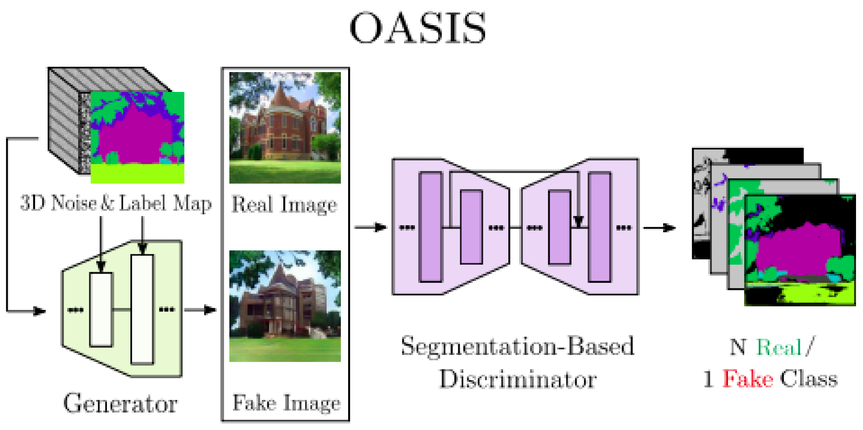
$$\text{LabelMix}(x, \hat{x}, M) = M \odot x + (1 - M) \odot \hat{x}$$



Pour profiter du feedback des informations sémantiques de la part du discriminateur, la fonction de coût du générateur est définie comme suit :

$$\mathcal{L}_G = -\mathbb{E}_{(z,t)}\left[\sum_{c=1}^{N} \alpha_c \sum_{i,j}^{H\times W} t_{i,j,c} \log D(G(z,t))_{i,j,c}\right]$$

Architecture du Modèle



Résultats

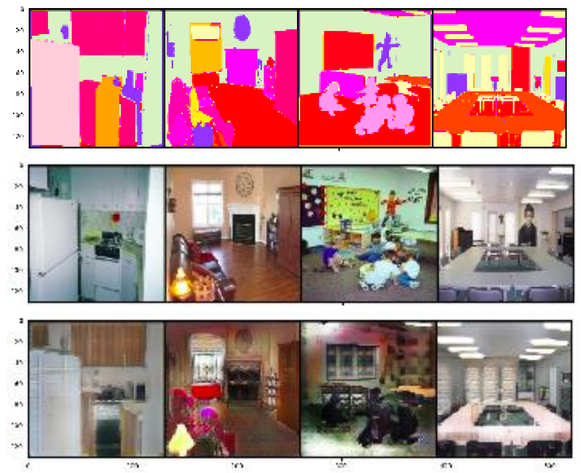


Figure 1: Images générées par notre modèle après 38 epochs. ⬆
L'apprentissage est plus rapide pour les classes de texture « uniforme » (mobilier) et plus lent pour les textures complexes (humains)…

⬇ **Figure 2**: Exploration de l'espace latent. Contrôle de la luminosité.

Table: Scores obtenus par notre modèle. Comparaison avec l'existant.

	FID – Fréchet Inception Distance	mIOU - mean Intersection Over Union
SPADE	33.9	38.5
OASIS	**28.3**	**48.8**
OASIS – *50 epochs / 200 (notre implémentation)*	99.0	**2.3**

Conclusion

- Le modèle présente de très bons résultats en terme de qualité, diversité des images générées.
- Article bien écrit et détaillé. Un solide protocole d'évaluation a été menée par les auteurs.
- Un modèle peu novateur concernant l'architecture.

Références

[1] Edgar Schönfeld~Edgar_Schönfeld1, Vadim Sushko, Dan Zhang, Juergen Gall,
You Only Need Adversarial Supervision for Semantic Image Synthesis, 2021
https://openreview.net/forum?id=yvQKLaqNE6M