Representaciones inspiradas en VAE para procesar texto a imagen

Laura, Miguel, César

Generación de texto a imágenes

Problema complejo

- De interés en AI
- Aplicaciones

GAN+descripciones

64 X 64 pixeles

Stack GAN

Stage I: bosquejos
Stage II: retoques
64 X 64 pixeles

Attention based RNN

Encoder
Decoder + Mecanismos de Atención
32 x32 pixeles

Proyecto DL

Problema

- Plantear modelo de DL para generar:
  - Imágenes a partir de texto
  - Conjunto de datos Flickr8

Conjunto de datos Flickr8

- **Imágenes:** 8,091 temas misceláneas (mascotas, paisajes, personas...)
- **Texto:** 5 descripciones en inglés.
- Mechanical Turk de Amazon, < 40 mil pares (texto, imagen)
- Usado originalmente para predecir texto a partir de imágenes

young girl with pigtails painting outside in the grass .

there is a girl with pigtails sitting in front of a rainbow painting .

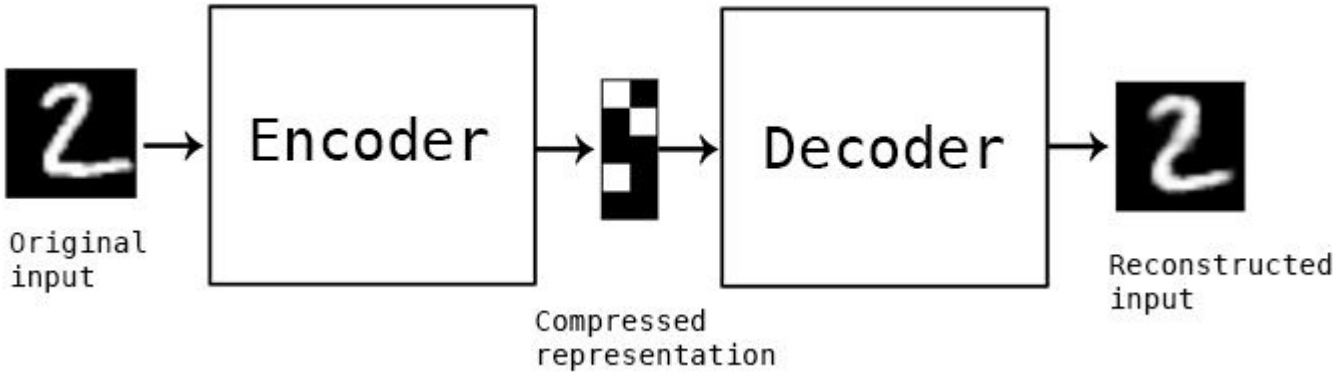
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .

a little girl is sitting in front of a large painted rainbow .

a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

Ideas (VAE):

- VAE codifica la representación de una imagen con los parámetros de una distribución de probabilidad de un espacio latente.



Ideas (inspiradas en VAE):

- Codificar la relación semántica texto-imágenes en espacio latente
- Ajustar parámetros de distribución normal multivariada como parte de entrenamiento para espacio latente.
- Emplear encaje pre-entrenado del texto de las descripciones de las imágenes para:
  - reducir dimensionalidad
  - considerar contexto
  - Reducir tiempo de entrenamiento

Global Vectors for Word Representation (GloVe)

- Representaciones vectoriales distribuidas de texto son útiles en NLP
  - Texto representado como punto de R^n
  - Textos “cercanos van a puntos cercanos de R^n
- **GloVe**
  - modelo pre-entrenado de Stanford University,
  - Relaciona probabilidades de co-ocurrencias de palabras y en documentos para codificar significado
- **GloVe.6B:**
  - Wikipedia 2014+Gigaword5;
  - 60 miles de millones de token,
  - Características: 50, 100, 200 y 300

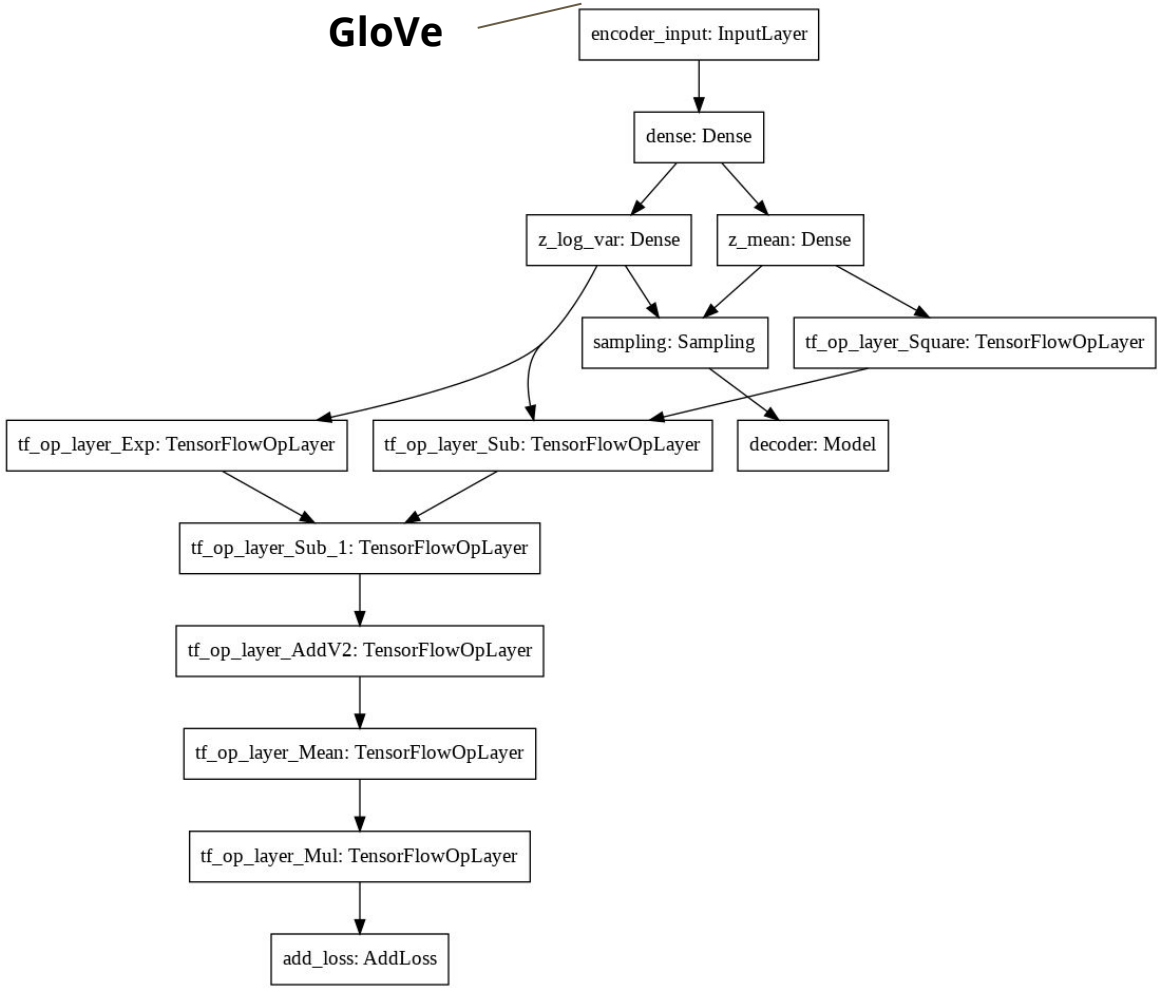
Arquitectura inspirada en VAE

Pérdida:

- Kullback - Liber
- (parámetros de distribución normal)
- MSE (imágenes)

RMSprop

Target:
64 x 64 pixeles 3 canales



Modelos inspirados en VAE

Modelo 1: LR “agresivo”

Modelo 2: Modelo 1 + LR “pasivo”

Modelo 3: Modelo 2 - nodos + Dropout

Param: ~ 7.6 millones

Datos: ~40 mil (80% train, 20% test)

Maldición de la dimensionalidad

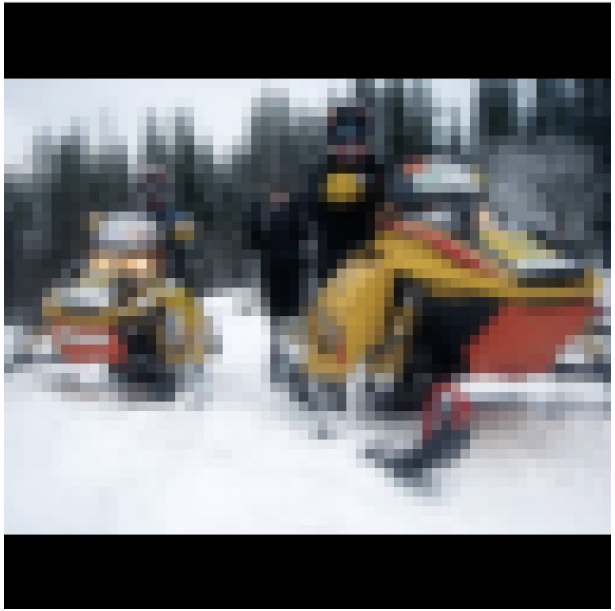
Limite RAM Google Colab

Table 2. Resultados de ajuste del modelo

Concepto	Modelo 1	Modelo 2	Modelo 3
Training Accuracy*	0.4477	0.4758	0.4781
Training Loss*	0.0684	0.0667	0.0667
Test Accuracy*	0.4961	0.4691	0.4555
Test Loss*	0.0679	0.0667	0.0667
Tiempo por época**	**$\sim$ 2 min**	**$\sim$ 2 min**	**< 2 min**

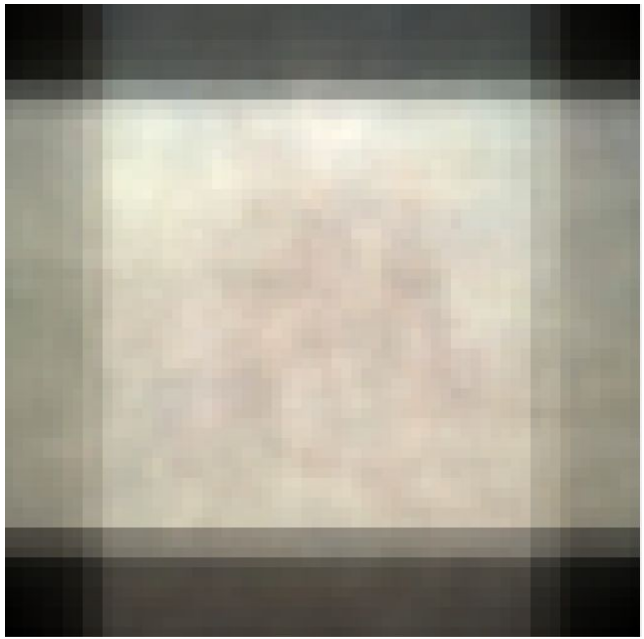
Pese a esfuerzos, no subió el desempeño

Resultados texto-imagen: Modelo 1



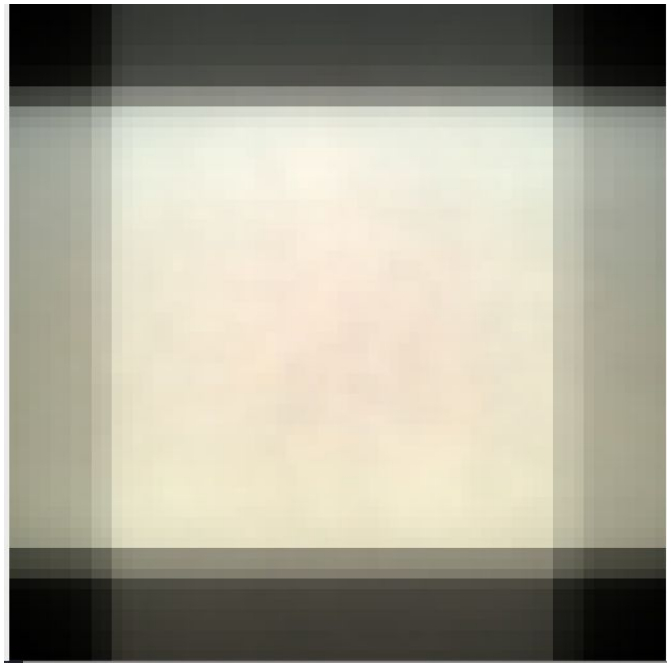
Two helmeted men sit on yellow snowmobiles while another man stands behind watching

Resultados texto-imagen: Modelo 2



The small child climbs on a red ropes on a playground

Resultados texto-imagen: Modelo 3



The brown and white dog with a green collar is biting a stick

¿Por qué es un problema tan difícil?

- Generalidad Texto - imágenes
- Necesidad de hardware
- Maldición dimensionalidad
- Explorar más arquitecturas

Imágenes de 32x32 pixeles obtenidas en 2020 por Zia et. al. para el conjunto de datos COCO

Caption: A person in blue and red ice climbing with two picks

Caption: A person climbing with two picks

Caption: A child in a red jacket playing street hockey guarding a goal

Caption: A child playing street hockey guarding a goal

Referencias principales

- H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 5908-5916, doi: 10.1109/ICCV.2017.629.

- Zia, T., Arif, S., Murtaza, S., and Ullah, M. A. (2020). Text-to-image generation with attention based recurrent neural networks.ArXiv, abs/2001.06658.