UPV

Image Captioning using pre-trained GPT-2 models

Author: Javier Garcia Gilabert
Tutor: Francisco Casacuberta Nolla

Degree in Data Science - Polytechnic University of Valencia - Course 2021-2022

Table of contents

1. Introduction
2. Motivation and objectives
3. State of the art
   - Computer vision
   - Language models
4. Image Captioning
   - State of the art architectures
5. Experimental framework
   - Dataset
   - Metrics
   - Proposals
6. Results
   - Comparison of models
   - Image examples
   - Textual cues
7. Conclusions and future work

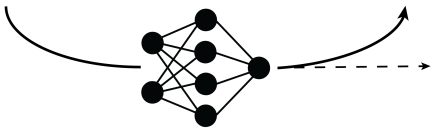
Introduction

For describing the image you will need to know:

- What everything is?
- Where is it?
- What is it doing?
- What is important to describe?
- Objects/colors/positions/relationships

. . .

Generated captions:
A group of kids pose in front of a pizza.
A team of pizza makers holding a fresh pizza.
Three people pose for the camera while holding a pizza.



- **Computer vision:** recognize objects.
- **Natural Language Processing:** generate fluent description.

Motivation and objectives

Applications of image captioning

- Image indexing
- Help for visually impaired people
- Medical image understanding
- Social Media

Problems in image captioning

- Object Hallucination: recognizing objects that are not in the scene.
- Exposure Bias Problem: LM models guess the next word based on the ground truth's previous words.

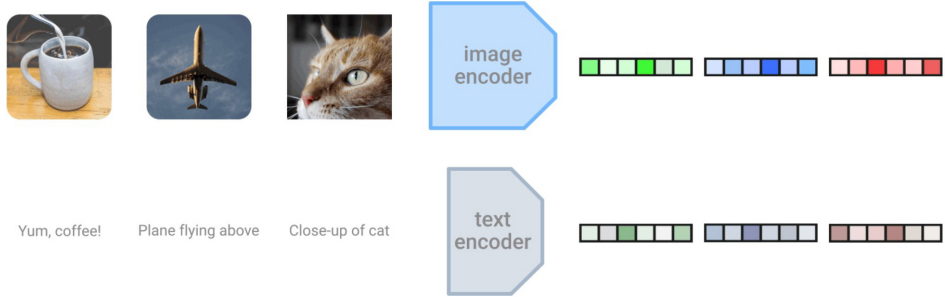
Objectives

1. Implement several strategies to improve the state of the art image captioning models.
   - Tackle the Exposure Bias Problem
   - Tackle the Object Hallucination problem
2. Study the state of the art models in image captioning.

State of the art

Computer vision

1. **Convolutional Neural Networks (CNN)** (LeCun et al., 1989)
2. **Vision transformers** (Dosovitskiy et al., 2020)
3. **Contrastive Language–Image Pre-training (CLIP)** (Radford et al., 2021): The model learns to associate each image with its text companion, while dissociating it from the other texts.



Images source: https://ai.googleblog.com/2022/04/locked-image-tuning-adding-language.html

State of the art

Language models

Objective: estimate the probability distributions of words in a vocabulary given a sequence of context words.

$$\mathrm{p}(w_i \mid w_1, \ldots, w_{i-1})$$

Neural network based language models: Recurrent Neural Networks (RNNs), Transformers

State of the art

Language models: Generative Pretrained Transformer 2 (GPT-2)

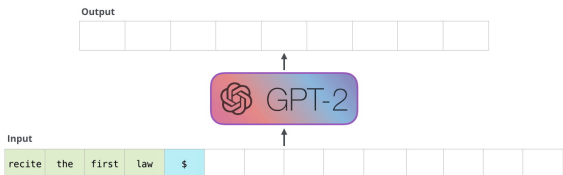
Outputs token probabilities given some context tokens

$$\mathrm{p}(w_i \mid w_1, \ldots, w_{t-1}), \quad i = 1, \ldots, |V|$$

model vocabulary size
50,257

0.19850038	aardvark
0.7089803	aarhus
0.46333563	aaron
	…
	…
	…
	…
	…
-0.51006055	zyzzyva

GPT-2 (Radford et al., 2019) is an autoregressive model:



Images source: https://jalammar.github.io/illustrated-gpt2/

Decoding methods:

- Greedy search: chooses the word with the highest probability.

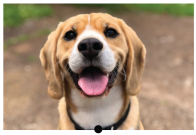
$$w_t = \arg\max_w \mathrm{p}(w \mid w_1, \ldots, w_{t-1})$$

- Beam search: keeps N hypotheses and selects the one with the greatest probability.

Image Captioning

State of the art architectures

CNN-based

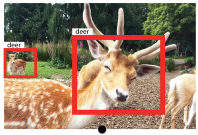


CNN

LM
e.g. Transformer, RNN

"A happy dog smiling"

Object detection + CNN



Object detector

CNN

LM
e.g. Transformer, RNN

"A happy deer standing in a field"

Pretrained Language Models

CLIP

Mapping Network

prefixes

Pretrained LM
e.g. GPT-2, GPT-3

"A picture of a sweet pangolin"

Experimental framework

Dataset

MSCOCO dataset (Lin et al., 2014)

1. 120K images, 5 captions per image
2. Human-curated captions
3. 80 categories of images

Karpathy splits (Karpathy and Fei-Fei, 2015):

118,000	5,000	5,000
Training	Validation	Test

A woman sitting at a desk in her work station.

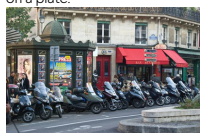
A pair of pizza rolls for sale are on a plate.

A dog holding a yellow frisbee in it's mouth.

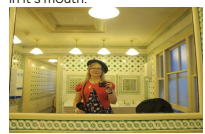
Sheep are separated in stalls two by two.

A red stop sign sitting in the middle of a street.

A very long sandwhich on a table.



Many small motorbikes are parked along the street.



A woman taking a selfie in front of a large mirror.

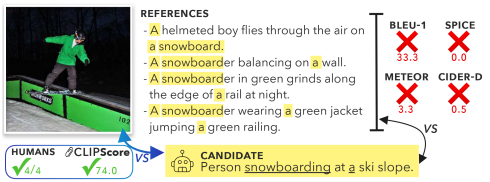
Two people use surfboards on a wave in the ocean.

A man pushes a cart with wheels with bananas.

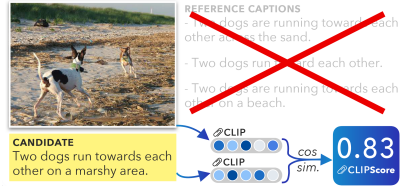
Experimental framework

Metrics

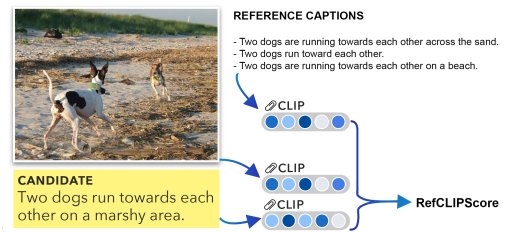
Reference based: `BLEU, METEOR, ROUGE, CIDER, SPICE`



Reference free: `CLIPScore`



Reference free + Reference based: `RefCLIPScore`



Images source: Hessel et al., 2021

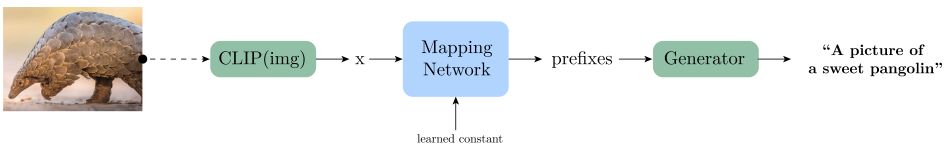
Experimental framework

Proposals

1. **CNN + Transformer.** The following CNNs are used:
   - EfficientNetB0 (Tan and Le, 2019).
   - VGG16 (Simonyan and Zisserman, 2014).
   - ResNet50 (He et al., 2016).

2. **Modifications on the ClipCap framework (Mokady et al., 2021)** to improve the generated caption by taking into account the CLIPScore metric during inference.

   Idea: By improving the captions using the CLIPScore metric, **the final caption will better describe the input image.**



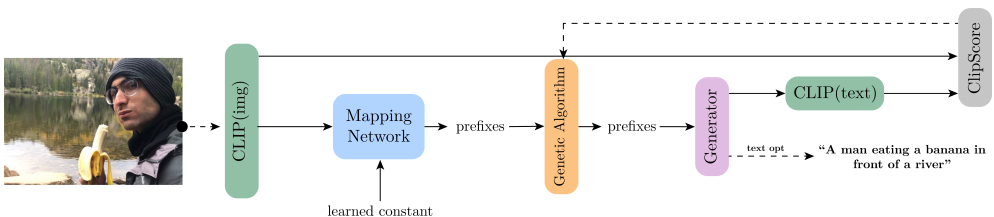
Experimental framework

Proposals: Genetic

The Transformer model outputs a set of vectors (*prefixes*) that tame the language model (GPT-2).

Idea: by optimizing the CLIPScore metric during inference, those prefixes can be enhanced and therefore, the candidate caption can be improved (based on Galatolo et al., 2021).

→ A Genetic Algorithm is used to optimize the prefixes by maximizing the CLIPScore metric.



Experimental framework

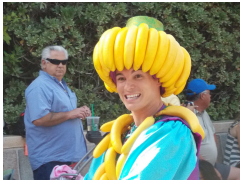
Proposals: Max Similarity

candidate caption

CLIP visual embedding

$$\max_{h \in H} \text{CLIP-S}(h, v)$$

set of hypotesis



H = { "A person wearing a banana headdress and necklace",
"A lady dressed in a blue and purple outfit wearing a hat made of fruit.",
"A person wearing a hat made out of yellow bananas.",
"Person with bananas on head and banana necklace.",
"A woman wears a hat that is made of bananas" }

- Beam search generates N candidate captions
- The candidate caption with the highest CLIPScore metric is chosen

N is a **hyperparameter**!

Experimental framework

Proposals: Greedy search based on CLIPScore

Greedy search: $w_t = \underset{i=1,\ldots,|V|}{\arg\max}\ \mathrm{p}(w_i \mid w_1, \ldots, w_{t-1})$

Proposed idea: select the next word that maximizes a linear combination between CLIPScore and the conditional probability $\mathrm{p}(w_i \mid w_1, \ldots, w_{t-1})$.

$$w_t = \underset{i=1,\ldots,N}{\arg\max}\ \beta\ \mathrm{p}(w_i \mid w_1, \ldots, w_{t-1}) + (1-\beta)\ \frac{1}{2.5}\ \texttt{CLIP-S}_i$$

Results

Comparison of models

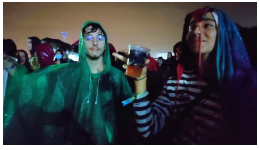
Results of the Karpathy split for several image captioning frameworks:

Model	METEOR	CIDEr	SPICE	CLIP-S	CLIP-S^{ref}
VinVL (Zhang et al., 2021)	**0.311**	**1.409**	**0.252**	0.760	0.820
tewel2021zero (Tewel et al., 2021)	0.115	0.146	0.055	**0.870**	0.790
CLIP-VL (Shen et al., 2021)	0.297	1.342	0.238	0.770	0.820
ClipCap (Mokady et al., 2021)	0.271	1.083	0.201	0.770	0.810
Ours; Genetic	0.225	0.696	0.164	0.814	0.827
Ours; Greedy Search	0.207	0.618	0.145	0.845	**0.842**
Ours; Max Similarity	0.260	0.932	0.197	0.799	0.830
Ours; VGG + Transformer	0.208	0.672	0.135	0.657	0.722
Ours; ResNet + Transformer	0.218	0.732	0.146	0.677	0.739
Ours; EfficientNetB0 + Transformer	0.228	0.807	0.159	0.700	0.759

Table: Comparison of models. We report supervised metrics (those that require human references): METEOR, ROUGE, CIDEr, SPICE. Finally, we report semantic relatedness to the image (CLIP-S), and to the human references (RefCLIPScore) based on CLIP's embeddings.

Results

Image examples



Beam search	A group of people with umbrellas in the air.	A man standing next to a woman holding a phone.	A group of people sitting next to a statue of a basilisk.
Greedy Search	A group of drinking fans outside a festival.	A couple of people holding hands talking on a political crisis.	A Christmas tree with a man and a woman reading a book.
Genetic	A group of people with different colors and sizes of glass bottles in the rain.	A picture of two people standing next to each other in a room.	A young boy reading a book in front of a statue of a man in a christmas costume.
Max Similarity	A group of people with umbrellas and a jug of water.	A man standing next to a woman holding a phone.	A group of people sitting next to a statue of a teddy bear.
EfficientNet + Transformer	A woman is holding a cell phone in her hand.	A man in a red shirt and a red tie.	A woman sitting on a bench with a cell phone.
VGG + Transformer	A woman is holding a cell phone to her ear.	A man in a suit and tie holding a cell phone.	A group of people sitting around a table with a cake.
ResNet + Transformer	A woman and a man are holding a wii remote.	A man and a woman are holding up a kite.	A woman and a man standing in front of a christmas tree.

Results

Textual cues: Adversarial pixel perturbations

Greedy Search
$\beta = 0.05$ $N = 30$

A boat owned by a criminal is at the bank of a harbor.

A white boat parked in a dock with a railing.

Conclusions and future work

Conclusions

- We have intoduced novel methods that improve the generated caption by using the CLIPScore metric. Objective 1
  - The Object Hallucination problem and the Exposure Bias Problem can be tackled by using the proposals based on improving the generated caption in testing time. Sub-objective 1
- Proposed methods outperformed the state of the art models in the RefCLIPScore metric.
- Several state of the art techniques to tackle image captioning have been reviewed. Objective 2

Future work

- Explore other pre-trained language models such as GPT-3 (Brown et al., 2020), GATO (Reed et al., 2022), Flamingo (Alayrac et al., 2022), etc.
- Determine whether the models have any biases.

UPV

Questions?

Javier Garcia Gilabert

Bibliography I

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation, 1*(4), 541–551.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *Computing research repository, abs/2010.11929*. https://arxiv.org/abs/2010.11929

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. Learning transferable visual models from natural language supervision. In: *Proceedings of the international conference on machine learning*. PMLR. 2021, 8748–8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll[illegible]r, P., & Zitnick, C. L. Microsoft coco: Common objects in context. In: *Proceedings of the european conference on computer vision*. Springer. 2014, 740–755.

Bibliography II

- Karpathy, A., & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015, 3128–3137.
- Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *Computing research repository, abs/2104.08718*. https://arxiv.org/abs/2104.08718
- Tan, M., & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the international conference on machine learning*. PMLR. 2019, 6105–6114.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Computing research repository, abs/1409.1556*.
- He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. 2016, 770–778.
- Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: CLIP prefix for image captioning. *Computing research repository, abs/2111.09734*. https://arxiv.org/abs/2111.09734

Bibliography III

Galatolo, F. A., Cimino, M. G. C. A., & Vaglini, G. (2021). Generating images from caption and vice versa via clip-guided generative latent space search. *Computing research repository, abs/2102.01645*. https://arxiv.org/abs/2102.01645

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., & Gao, J. Vinvl: Revisiting visual representations in vision-language models. In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2021, 5579–5588.

Tewel, Y., Shalev, Y., Schwartz, I., & Wolf, L. (2021). Zero-shot image-to-text generation for visual-semantic arithmetic. *Computing research repository, abs/2111.14447*. https://arxiv.org/abs/2111.14447

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K., Yao, Z., & Keutzer, K. (2021). How much can CLIP benefit vision-and-language tasks? *Computing research repository, abs/2107.06383*. https://arxiv.org/abs/2107.06383

Bibliography IV

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., & de Freitas, N. (2022). A generalist agent. https://doi.org/10.48550/ARXIV.2205.06175

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. https://doi.org/10.48550/ARXIV.2204.14198