Redes Neuronales para Lenguaje Natural

2023

Grupo de Procesamiento de Lenguaje Natural
Instituto de Computación

Multimodalidad

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015, June). Draw: A recurrent neural network for image generation. In International conference on machine learning (pp. 1462-1471). PMLR.

...

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. arXiv preprint arXiv:2304.08485.

Motivación

Hasta ahora vimos como usar modelo de redes neuronales para procesar y generar **texto**.

Resulta interesante combinar con **otras modalidades de datos**:
- imágenes, imágenes médicas, sonidos, etc

Las distintas modalidades pueden ser **representadas** en espacios vectoriales y ser utilizadas conjuntamente por modelos de redes neuronales.

Redes Convolucionales

Para procesar **imágenes** vimos el uso de **Redes Convolucionales**.

Capas de convolución y pooling realizan **sucesivas transformaciones** a la imágen de entrada hasta obtener la salida (ej. clasificación del imágen)

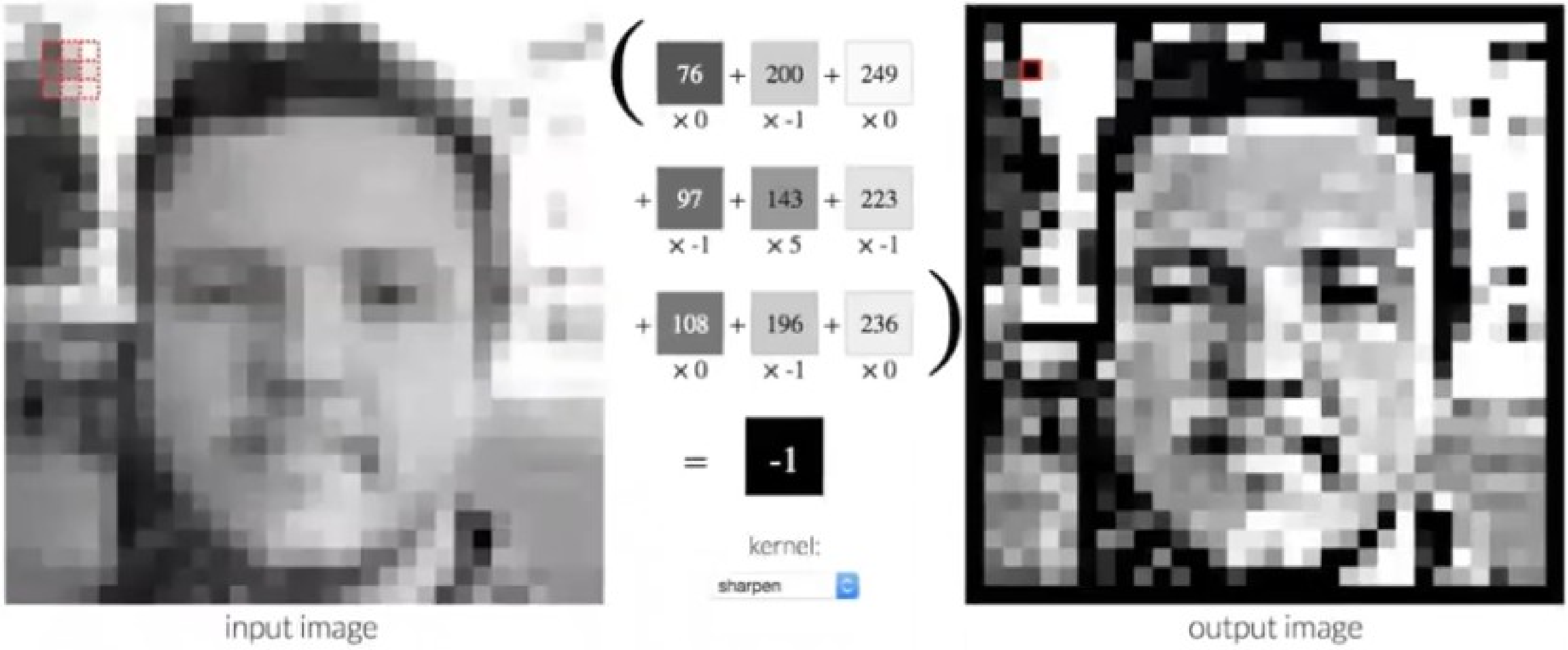
Por otro lado, se obtienen **representaciones vectoriales** de las imágenes que reflejan aspectos semánticos (resultado del aprendizaje).

Vamos con un repaso

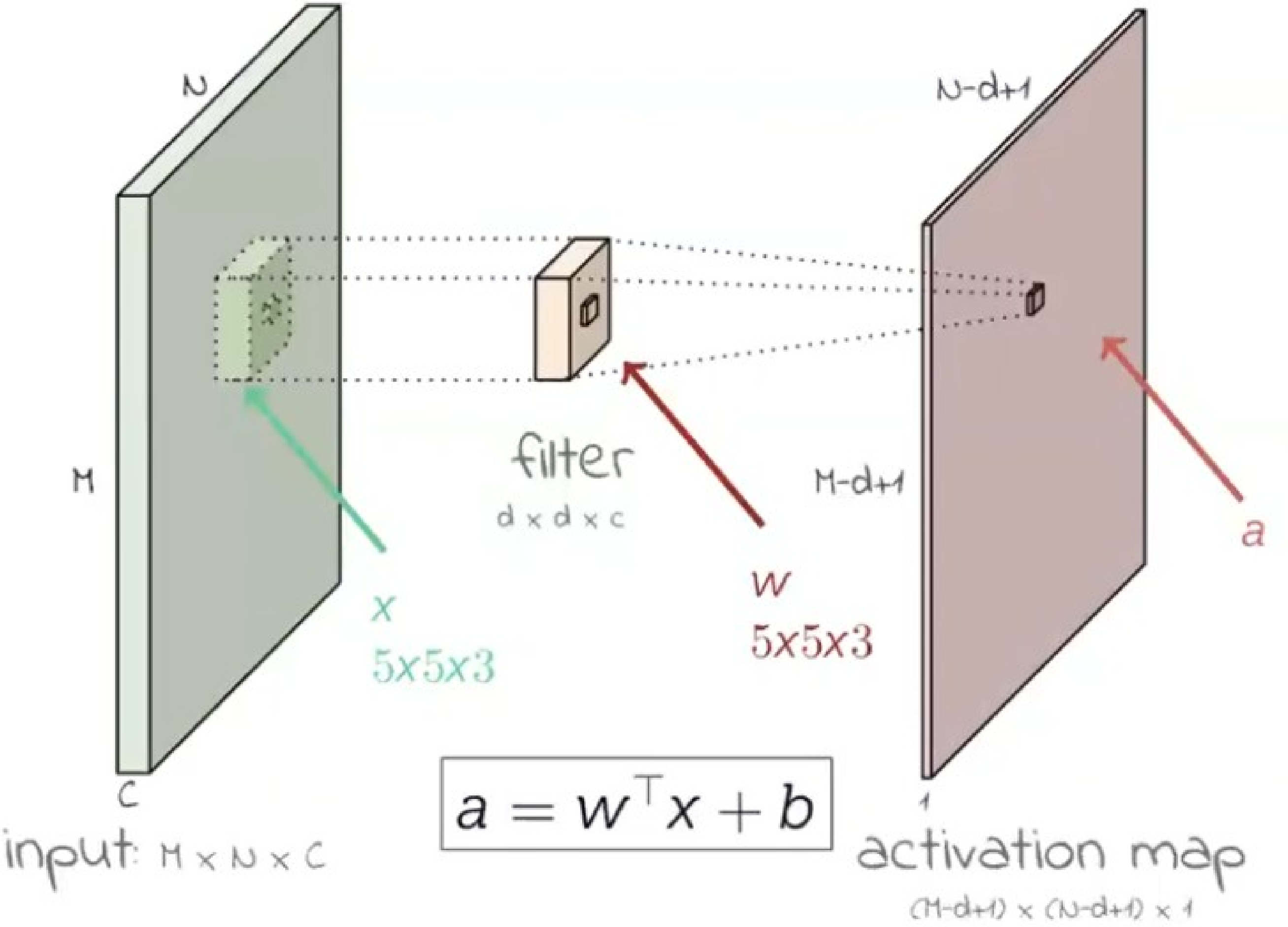
Convolución en imágenes

Image Kernels, Explained Visually *por Victor Powell*

`setosa.io/ev/image-kernels`



Capa convolucional (convolución)

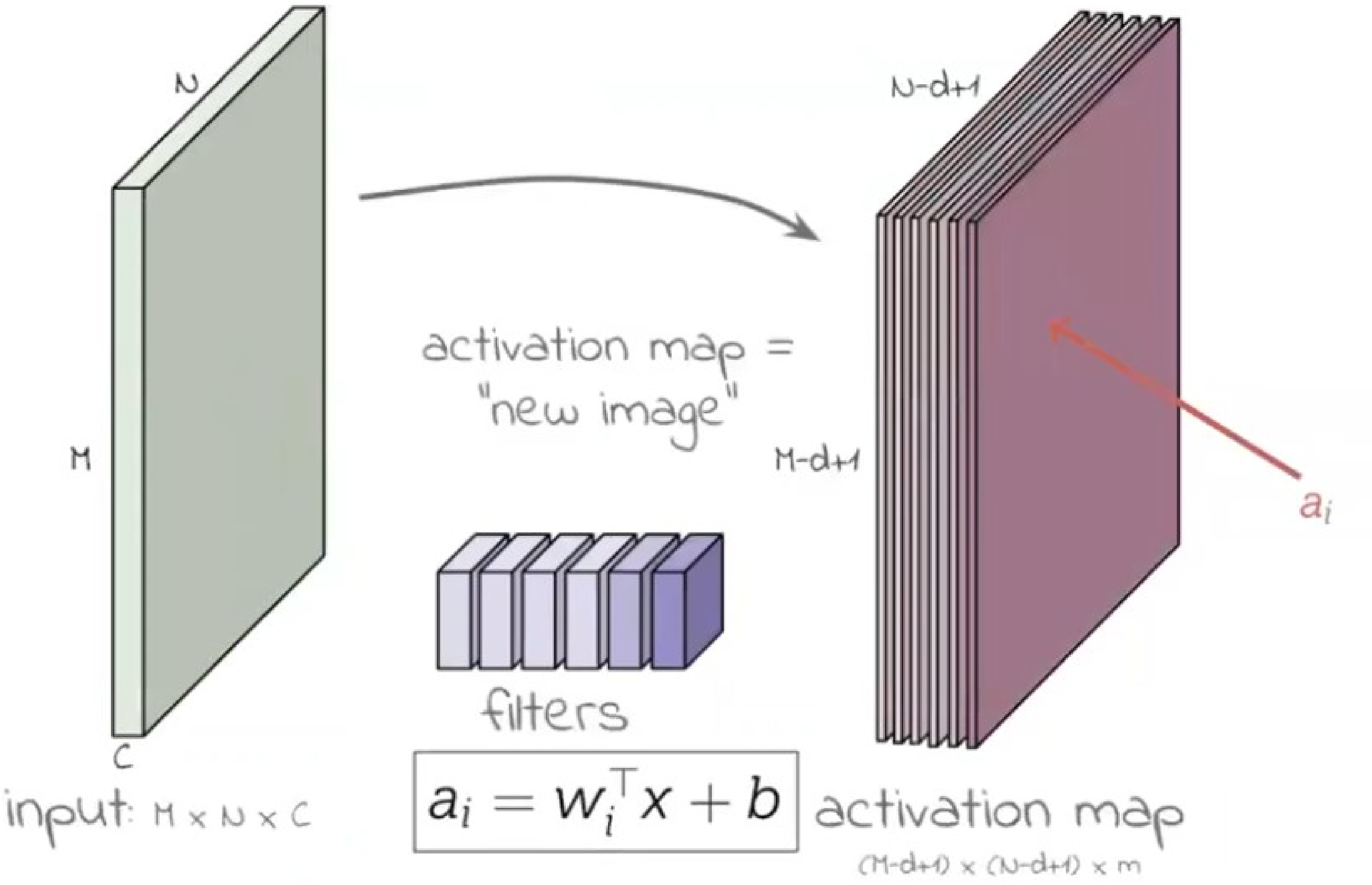


El **tamaño de la salida** queda definido por la entrada, el filtro y el stride:

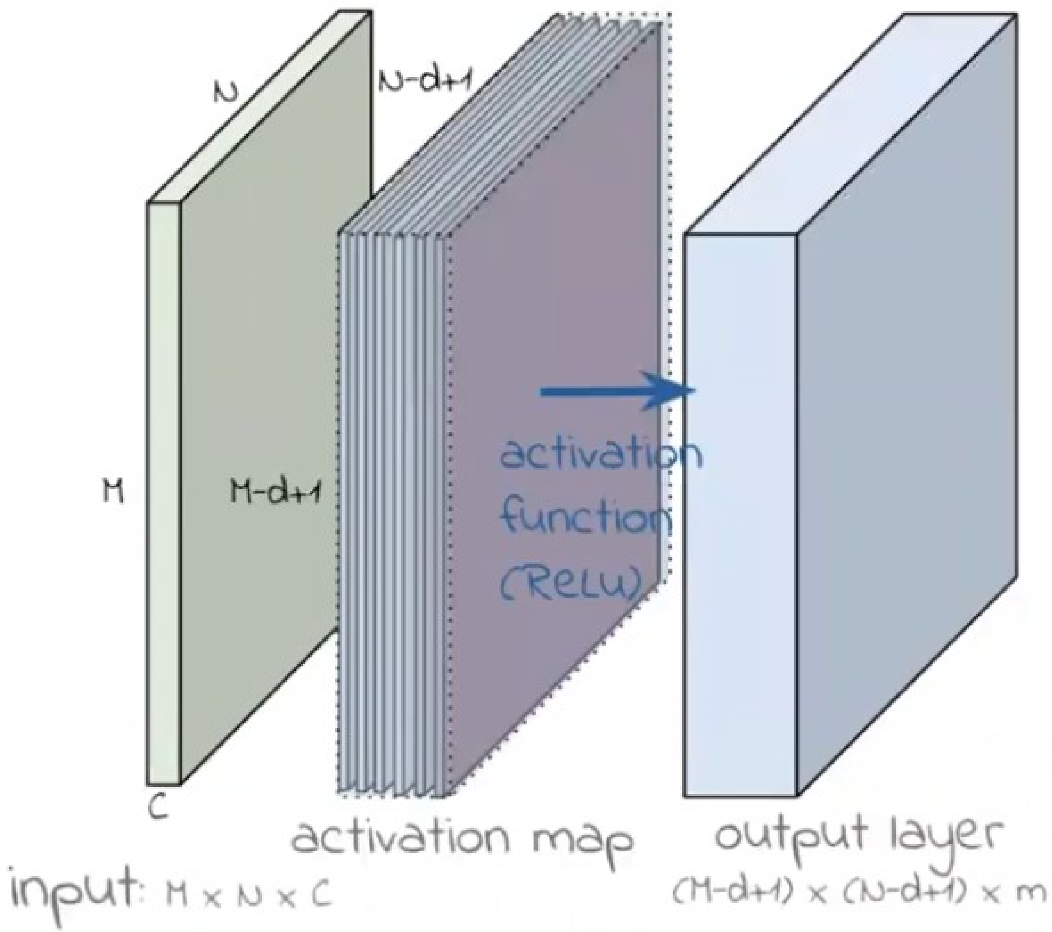
(N-F)/stride+1

En la práctica se usa zero-padding si se quiere preservar el tamaño.

Capa convolucional (múltiples filtros)

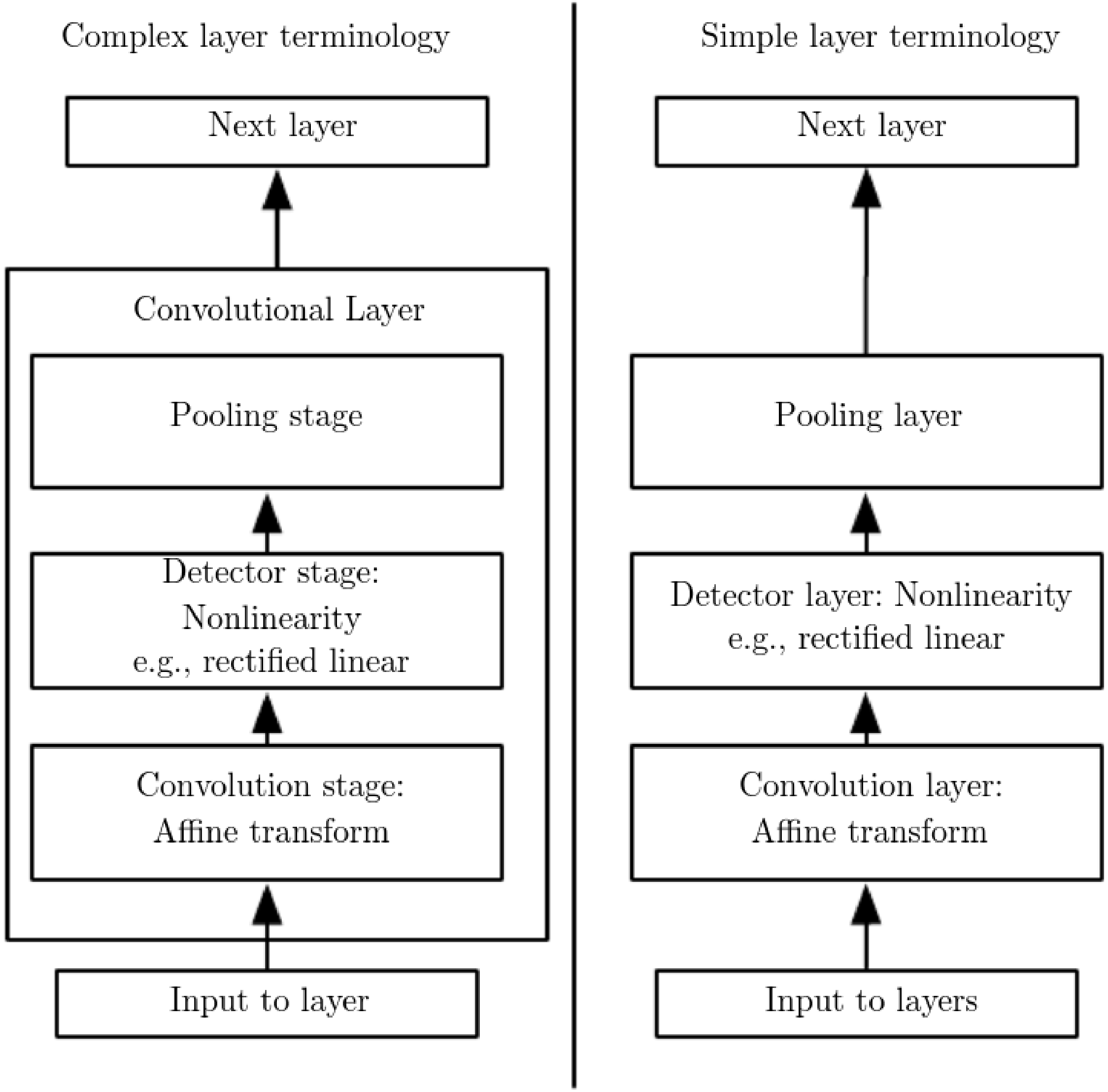


Capa convolucional (conv + activación)



Convolución + Pooling

Reduce la cantidad de parámetros, mejora los resultados



ImageNet (2009)

Idea que comienza en 2006 por Fei-Fei Li

1000 clases de imágenes, más de 1 millón de imágenes

AlexNet (2012)

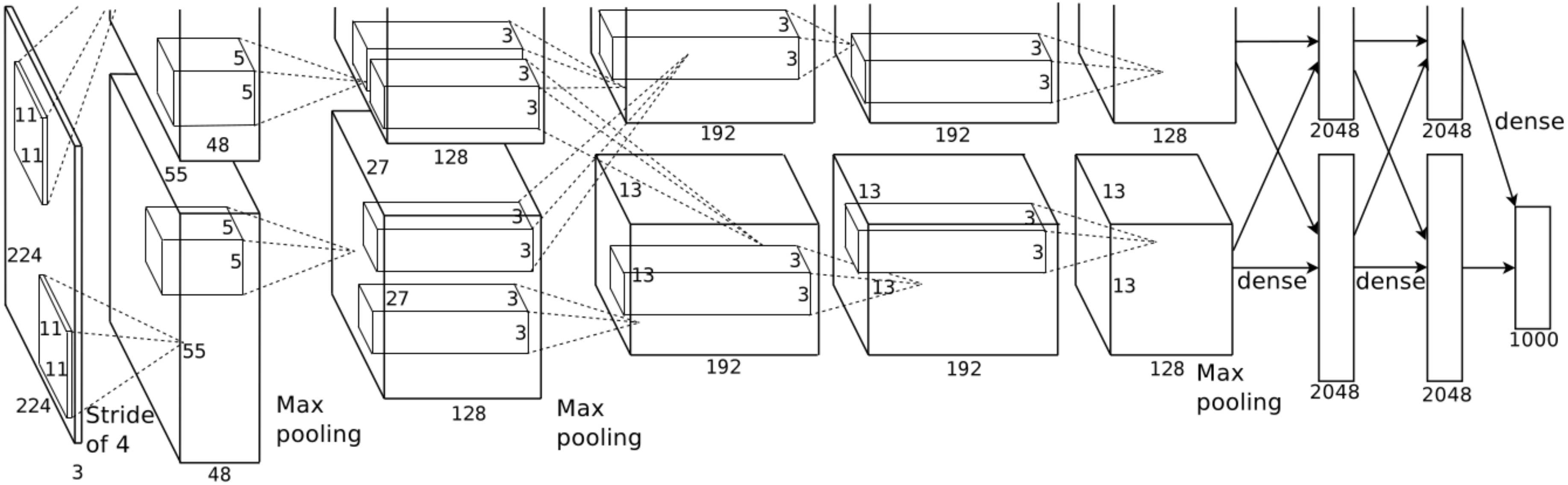
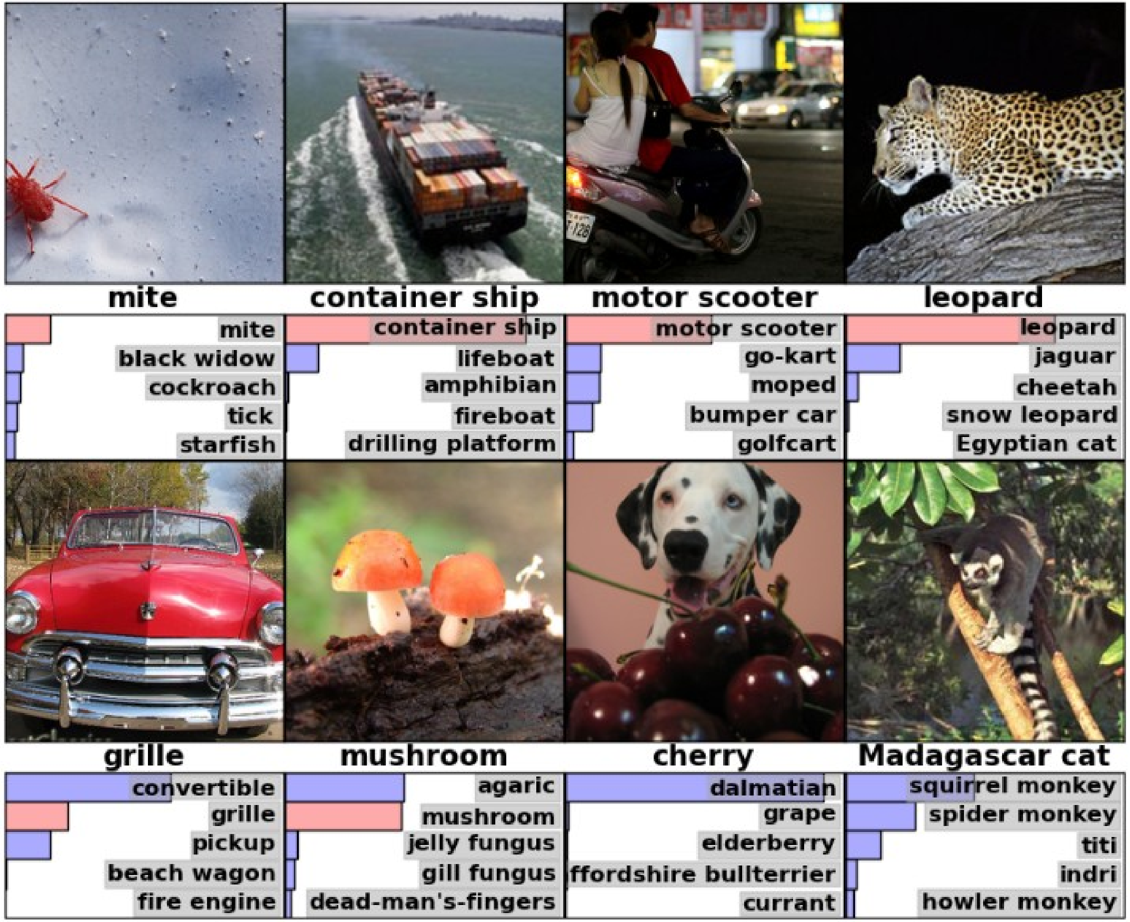


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

AlexNet (salida clasificación)



Visualización t-SNE de imágenes (salida CNN)

Visualización cercanía Euclidea de salida de CNN

Show and Tell: A Neural Image Caption Generator (2014)

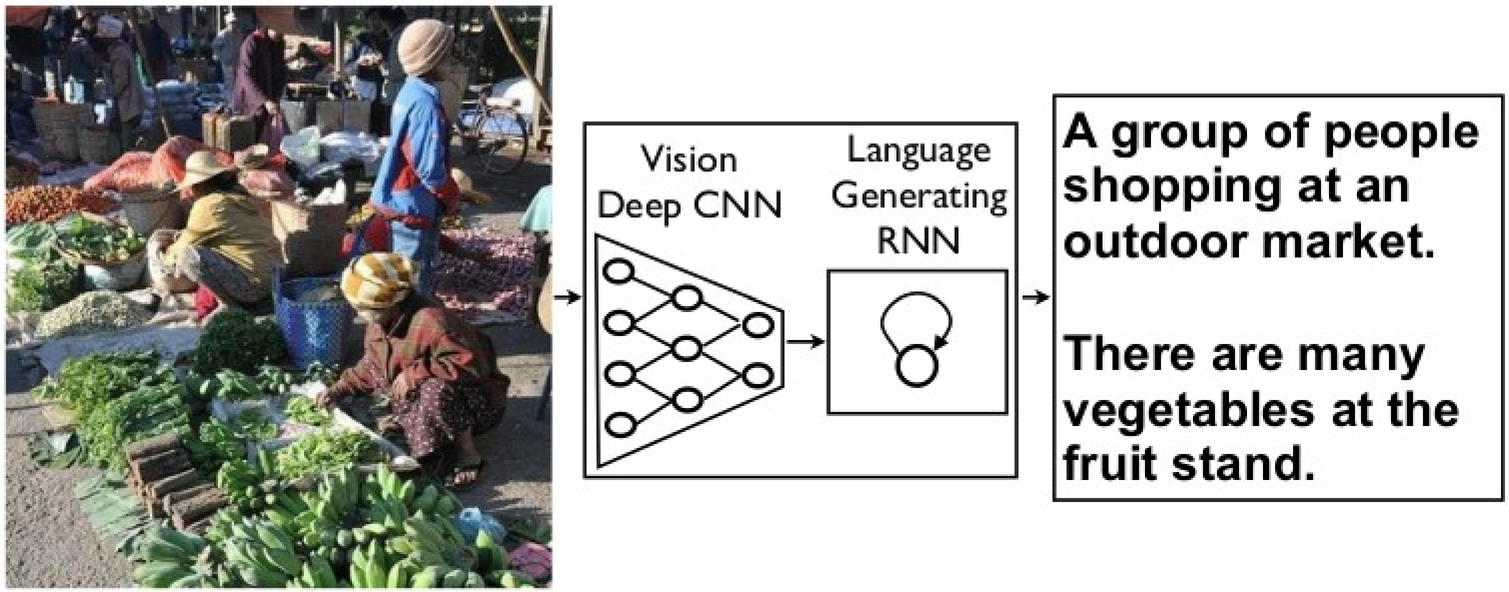
Generar descripción de una imágen.

CNN como encoder de la imágen y **RNN** para decodificar la descripción

El modelo es entrenado para maximizar la probabilidad de una secuencia de palabras que describe a la imágen.

- MSCOCO >80k-40k-40k / SBU 1M-?-1k
- flickr8k / flickr30k
- Pascal VOC 2008 0-0-1k

Show and Tell: A Neural Image Caption Generator (2014)



Show and Tell: A Neural Image Caption Generator (2014)

El entrenamiento del modelo consiste en

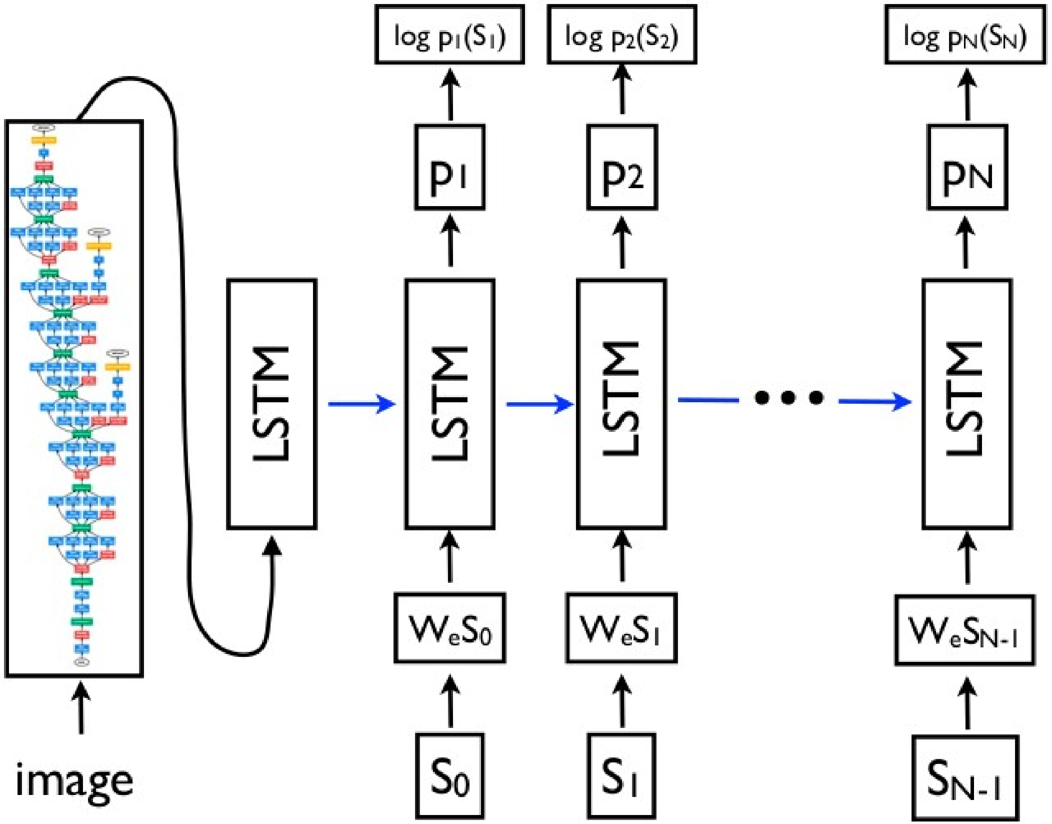
$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

donde I es una imágen del dataset y S una secuencia de palabra que la describe.

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

Show and Tell: A Neural Image Caption Generator (2014)

La imágen (salida de la CNN) se toma como el primer elemento de la secuencia




A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator filled with lots of food and drinks.
A herd of elephants walking across a dry grass field.
A close up of a cat laying on a couch.
A red motorcycle parked on the side of the road.
A yellow school bus parked in a parking lot.
Describes without errors
Describes with minor errors
Somewhat related to the image
Unrelated to the image

Show, attend and tell: Neural image caption generation with visual attention (2015)

Se propone utilizar un conjunto de representaciones de la imágen que refieran a regiones distinas.

Se utilizan las **capas cercanas a la entrada de la CNN** (campo receptivo local)

- otros modelos usaban capas cercanas a la salida de la red encoder “global” de la imágen

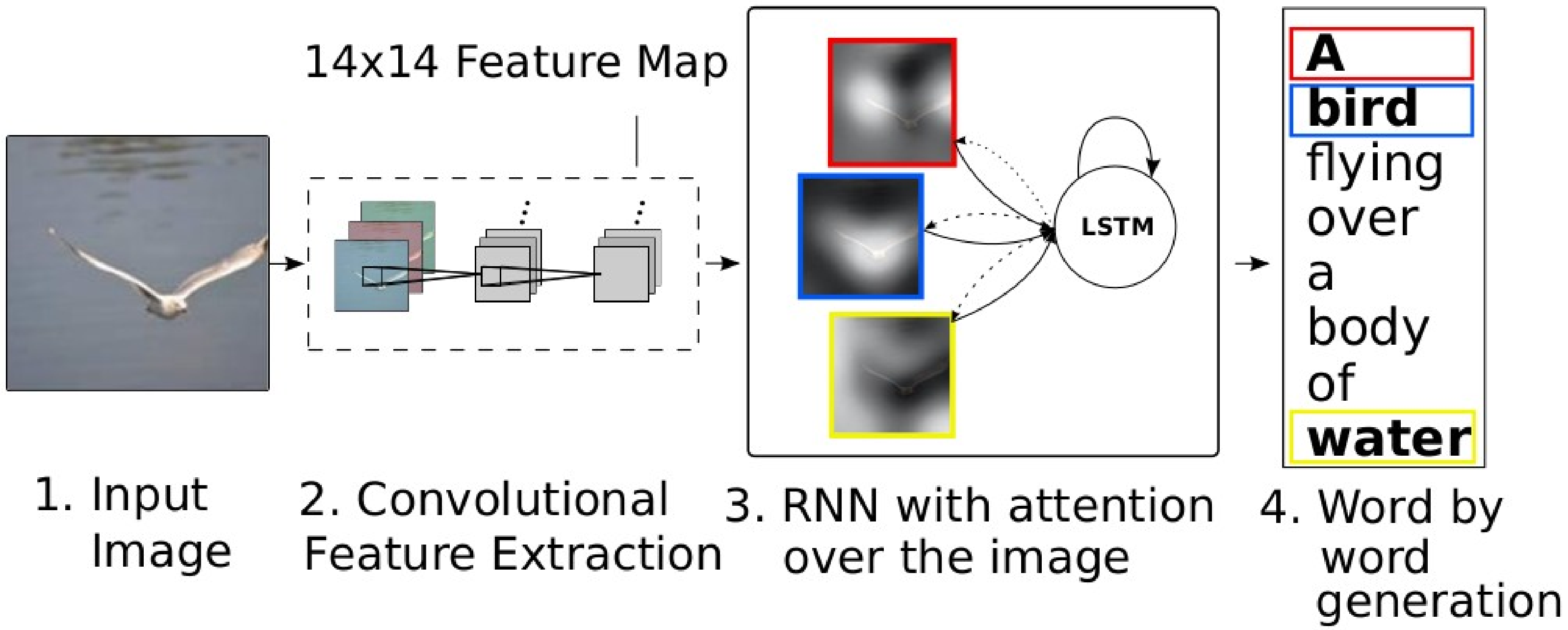
Proponen dos mecanismo de atención

- Hard Attention
- Soft Attention

Show, attend and tell: Neural image caption generation with visual attention (2015)

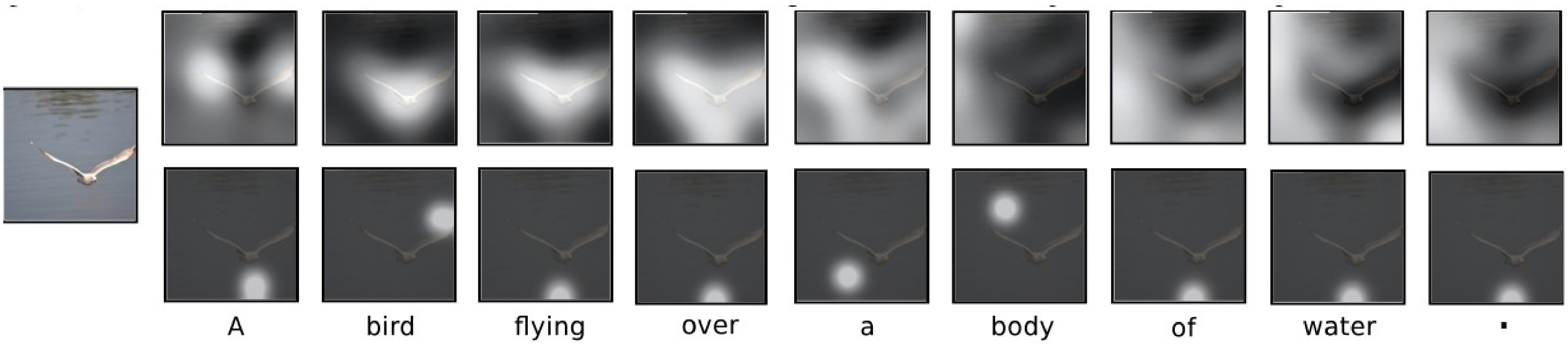
CNN como encoder de **L vectores de la imágen** (regiones) y

RNN para decodificar la descripción, con **atención** en los L vectores



Show, attend and tell: Neural image caption generation with visual attention (2015)

soft attention



hard attention

Show, attend and tell: Neural image caption generation with visual attention (2015)

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

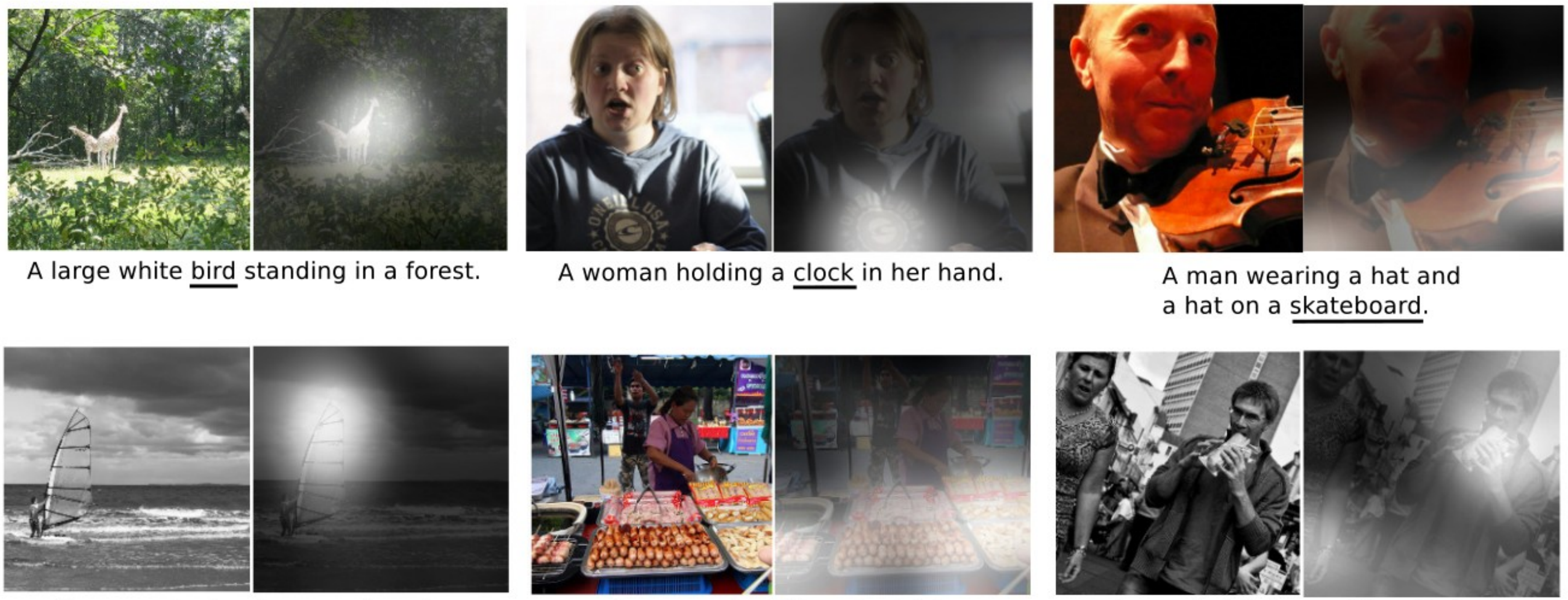
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Show, attend and tell: Neural image caption generation with visual attention (2015)

Algunos errores que muestran cierta coherencia en el mecanismo de atención



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.

A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Show, attend and tell: Neural image caption generation with visual attention (2015)

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

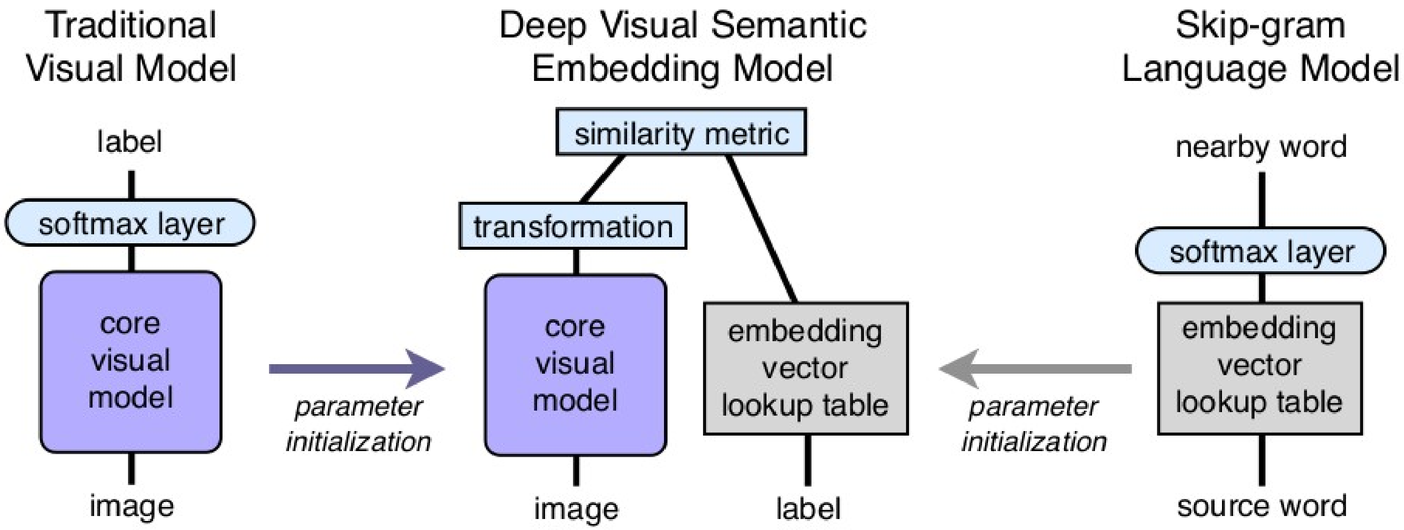
DeViSE: A Deep Visual-Semantic Embedding Model (NIPS 2013)

El encoder CNN de los modelos anteriores está limitado a un número fijo de categorías (ej. 1000)

Proponen mejorar esta situación usando texto

Obtienen mejoras en el la tarea inicial (1000-class ImageNet)
Además, “mejores” errores y resultados en 10000 clases (zero-shot)

DeViSE: A Deep Visual-Semantic Embedding Model (NIPS 2013)



Encoder de imágenes pre-entrenado (CNN lower layers)

Modelo conjunto

Modelo word embedding pre-entrenado

DeViSE: A Deep Visual-Semantic Embedding Model (NIPS 2013)

Por cada ejemplo del conjunto de entrenamiento minimizan:

$$loss(image, label) = \sum_{j \neq label} \max[0, margin - \vec{t}_{label} M \vec{v}(image) + \vec{t}_j M \vec{v}(image)]$$

La etiqueta **j** es tomada random entre clases de imágenes posibles

DeViSE: A Deep Visual-Semantic Embedding Model (NIPS 2013)

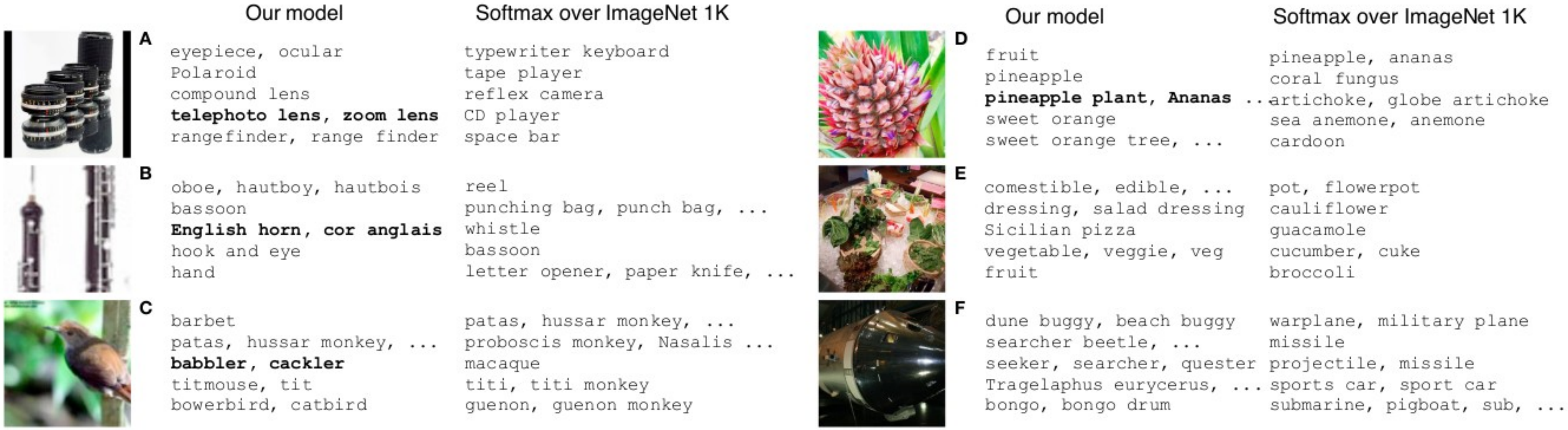


Figure 2: For each image, the top 5 zero-shot predictions of DeViSE+1K from the 2011 21K label set and the softmax baseline model, both trained on ILSVRC 2012 1K. Predictions ordered by decreasing score, with correct predictions in bold. Ground truth: (a) *telephoto lens, zoom lens*; (b) *English horn, cor anglais*; (c) *babbler, cackler*; (d) *pineapple, pineapple plant, Ananas comosus*; (e) *salad bar*; (f) *spacecraft, ballistic capsule, space vehicle*.

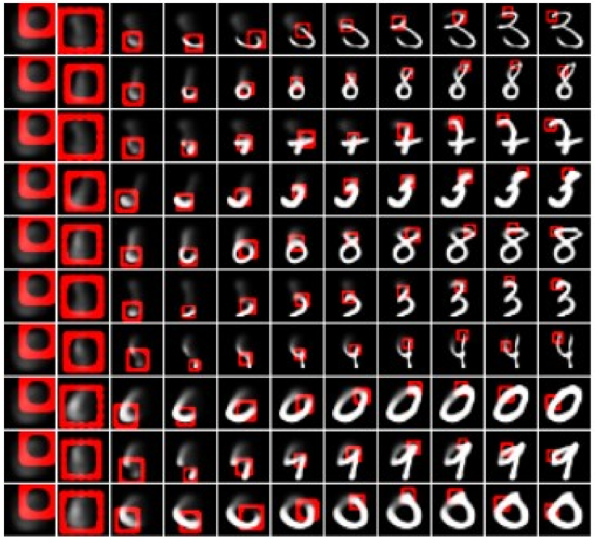
Image generation (LSTM)

DRAW: A Recurrent Neural Network For Image Generation

Karol Gregor KAROLG@GOOGLE.COM
Ivo Danihelka DANIHELKA@GOOGLE.COM
Alex Graves GRAVESA@GOOGLE.COM
Danilo Jimenez Rezende DANILOR@GOOGLE.COM
Daan Wierstra WIERSTRA@GOOGLE.COM
Google DeepMind

Abstract

This paper introduces the *Deep Recurrent Attentive Writer* (DRAW) neural network architecture for image generation. DRAW networks combine a novel spatial attention mechanism that mimics the foveation of the human eye, with a sequential variational auto-encoding framework that allows for the iterative construction of complex images. The system substantially improves on the state of the art for generative models on MNIST, and, when trained on the Street View House Numbers dataset, it generates images that cannot be distinguished from real data with the naked eye.



Time ⟶

1. Introduction

Zero-Shot Text-to-Image Generation

Aditya Ramesh [1] **Mikhail Pavlov** [1] **Gabriel Goh** [1] **Scott Gray** [1]
Chelsea Voss [1] **Alec Radford** [1] **Mark Chen** [1] **Ilya Sutskever** [1]

Abstract

Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.

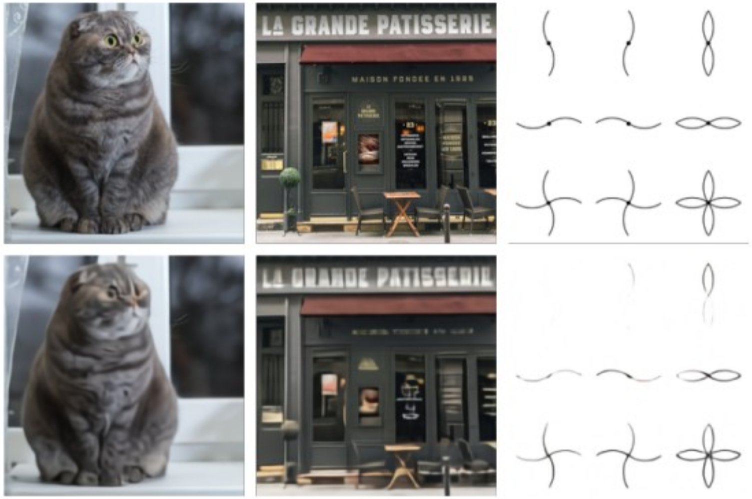


Figure 1. Comparison of original images (top) and reconstructions from the discrete VAE (bottom). The encoder downsamples the spatial resolution by a factor of 8. While details (e.g., the texture of the cat's fur, the writing on the storefront, and the thin lines in the illustration) are sometimes lost or distorted, the main features of the image are still typically recognizable. We use a large vocabulary size of 8192 to mitigate the loss of information.

1. Introduction

Modern machine learning approaches to text to image synthesis started with the work of Mansimov et al. (2015), who showed that the DRAW Gregor et al. (2015) generative

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh* OpenAI aramesh@openai.com

Prafulla Dhariwal* OpenAI prafulla@openai.com

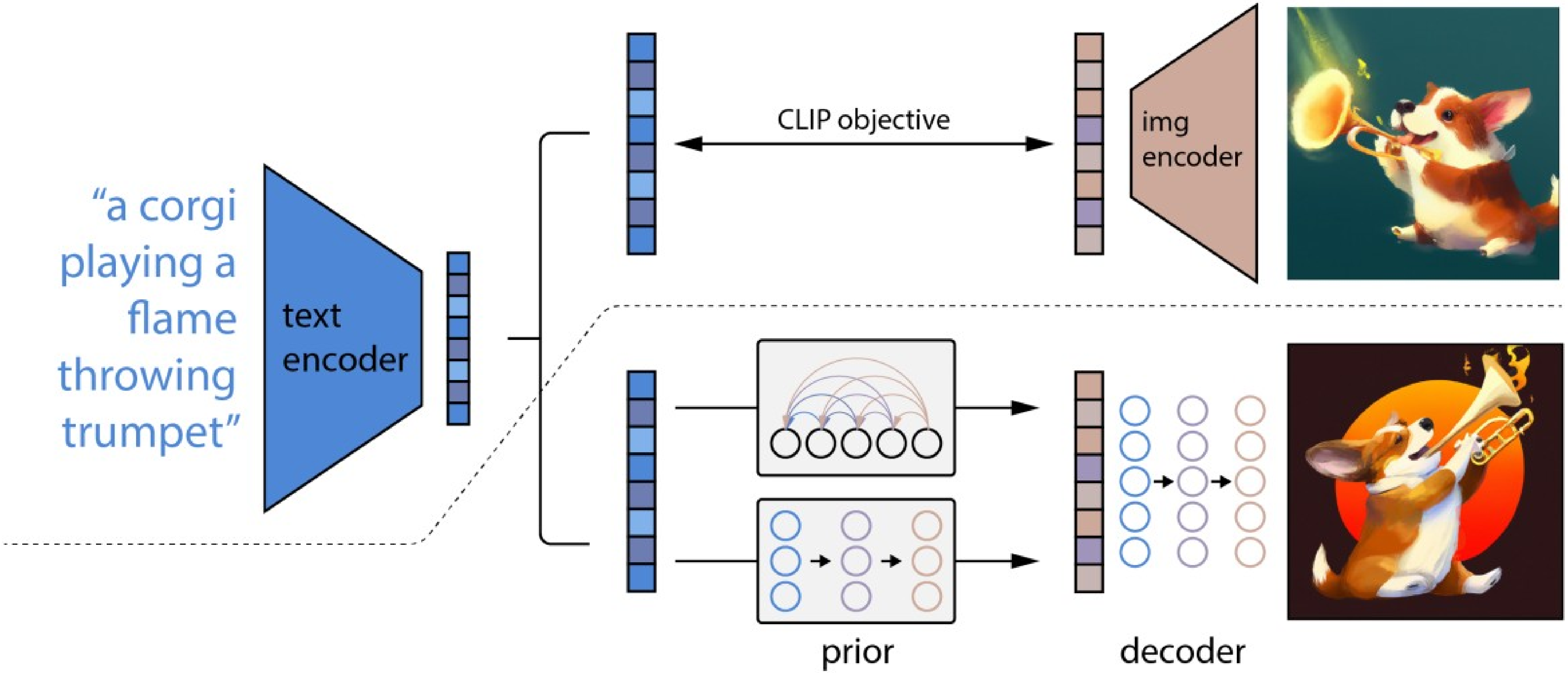
Alex Nichol* OpenAI alex@openai.com

Casey Chu* OpenAI casey@openai.com

Mark Chen OpenAI mark@openai.com

Abstract

Contrastive models like CLIP have been shown to learn robust representations of images that capture both semantics and style. To leverage these representations for image generation, we propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. We show that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity. Our decoders conditioned on image representations can also produce variations of an image that preserve both its semantics and style, while varying the non-essential details absent from the image representation. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion. We use diffusion models for the decoder and experiment with both autoregressive and diffusion models for the prior, finding that the latter are computationally more efficient and produce higher-quality samples.


"a corgi playing a flame throwing trumpet"
text encoder
CLIP objective
img encoder
prior
decoder

Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

Abstract

Instruction tuning large language models (LLMs) using machine-generated instruction-following data has improved zero-shot capabilities on new tasks, but the idea is less explored in the multimodal field. In this paper, we present the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. By instruction tuning on such generated data, we introduce LLaVA: **L**arge **L**anguage **a**nd **V**ision **A**ssistant, an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language understanding. Our early experiments show that LLaVA demonstrates impressive multimodel chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images/instructions, and yields a 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset. When fine-tuned on Science QA, the synergy of LLaVA and GPT-4 achieves a new state-of-the-art accuracy of 92.53%. We make GPT-4 generated visual instruction tuning data, our model and code base publicly available.

The primary goal is to effectively leverage the capabilities of both the pre-trained LLM and visual model. The network archtecture is illustrated in Figure 1. We choose LLaMA as our LLM $f_\phi(\cdot)$ parameterized by ϕ, as its effectiveness has been demonstrated in several open-source language-only instruction-tuning works. [43, 45, 34].

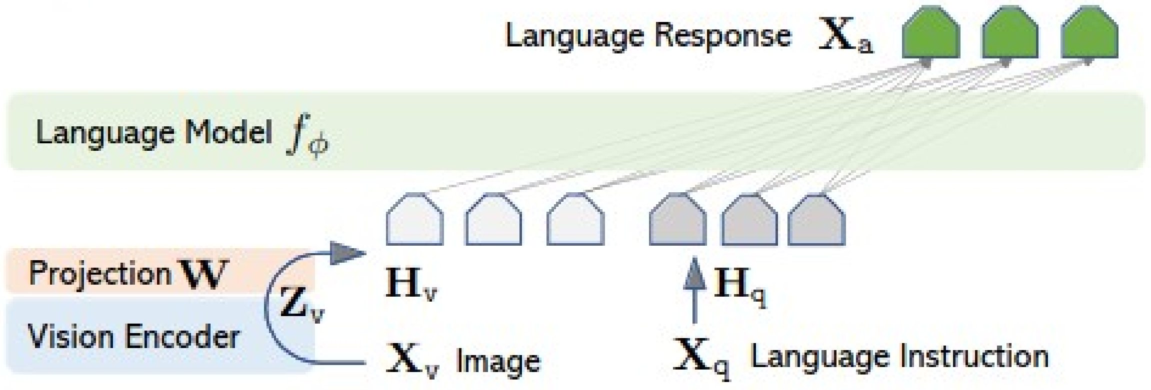


Figure 1: LLaVA network architecture.

Visual input example, Extreme Ironing:	
	
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [32]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.