Emergent ImageNation

Image emergence in visual referential games

Nicolo’ Brandizzi

Institute for Logic, Language and Computation
&
Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti

Background

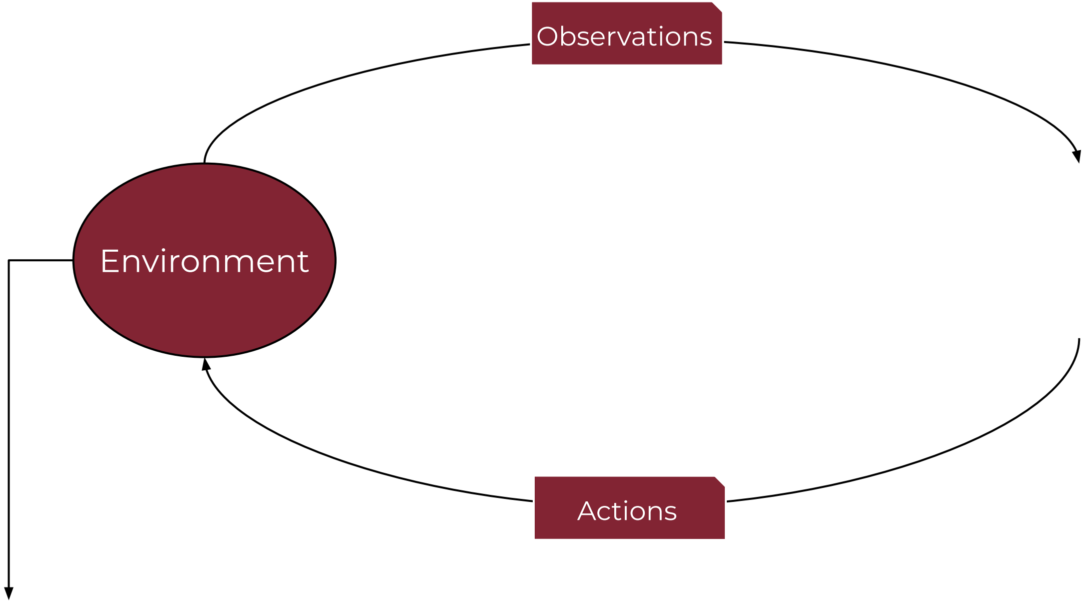
- **Reinforcement Learning**
- Emergent Communication
- Referential Games
- Architectures & Frameworks:
  - Image captioning
  - DALL-E: Creating Images from Text

Reinforcement Learning

Environment

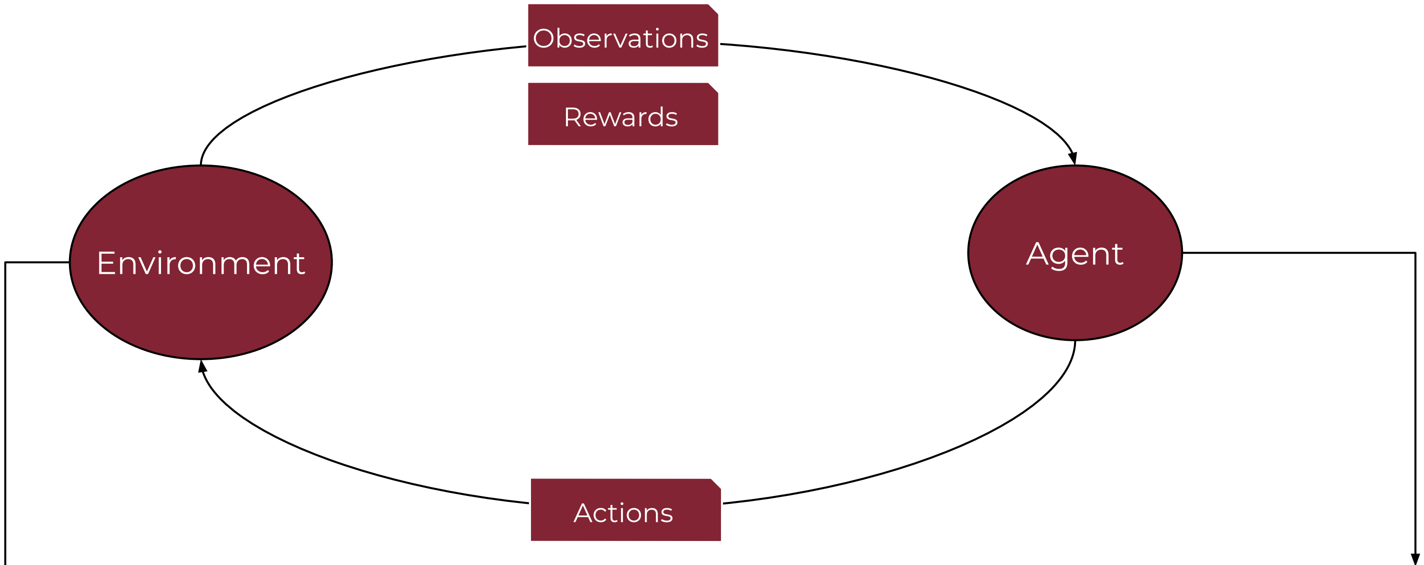
Usually a game
Its state is expressed through *observations* and can be changed with admissible *actions.*

Reinforcement Learning



Usually a game
Its state is expressed through *observations* and
can be changed with admissible *actions.*

Reinforcement Learning



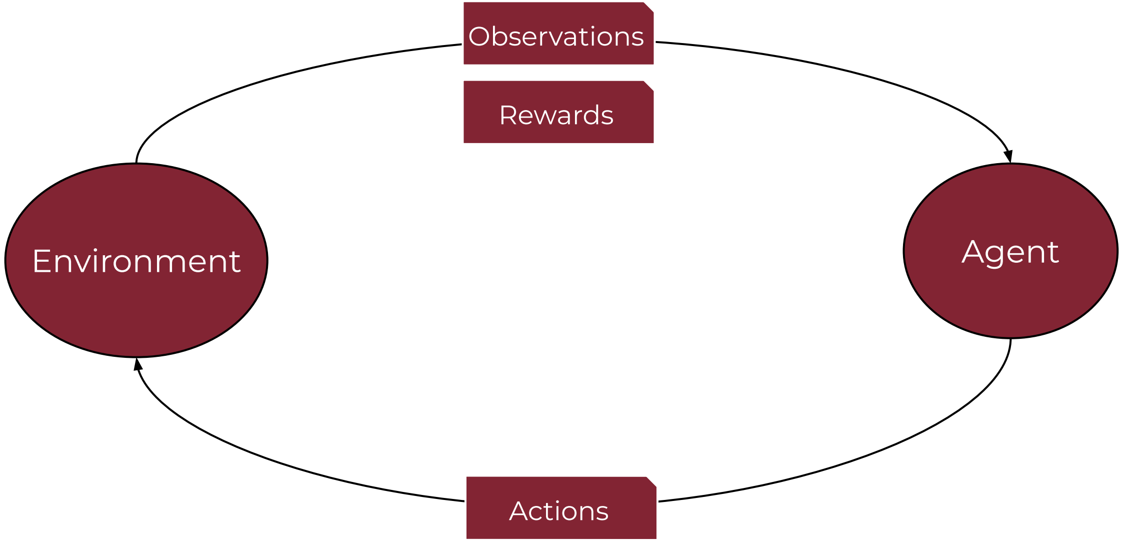
Usually a game
Its state is expressed through *observations* and can be changed with admissible *actions.*

Learns a policy that maximizes *rewards* by choosing *actions*

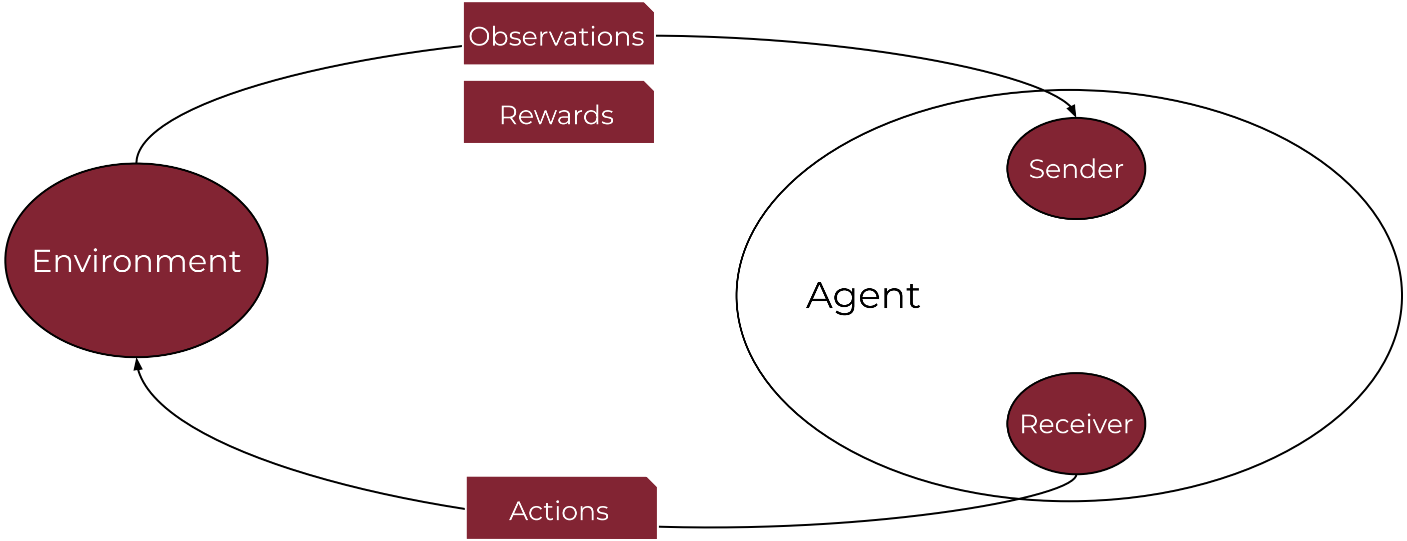
Background

- Reinforcement Learning
- **Emergent Communication**
- Referential Games
- Architectures & Frameworks:
    - Image captioning
    - DALL-E: Creating Images from Text

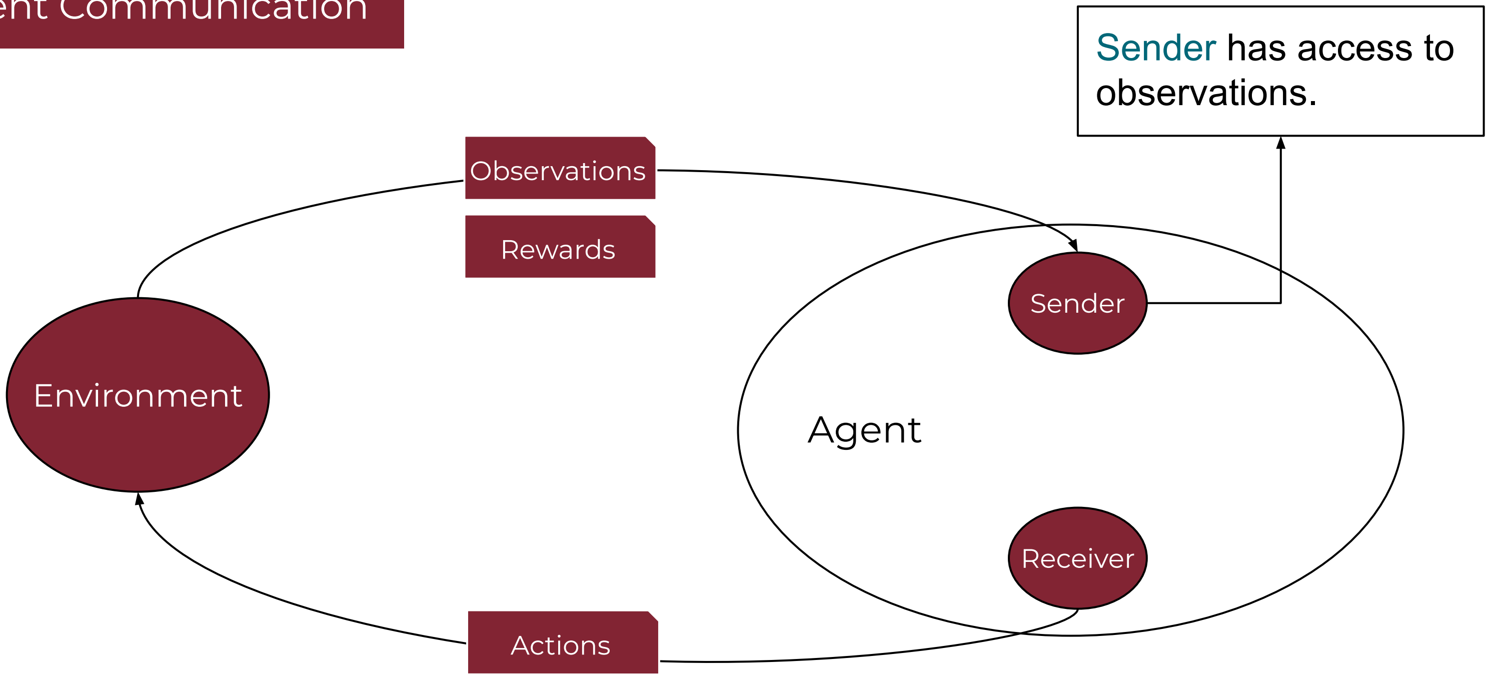
Reinforcement Learning



Emergent Communication

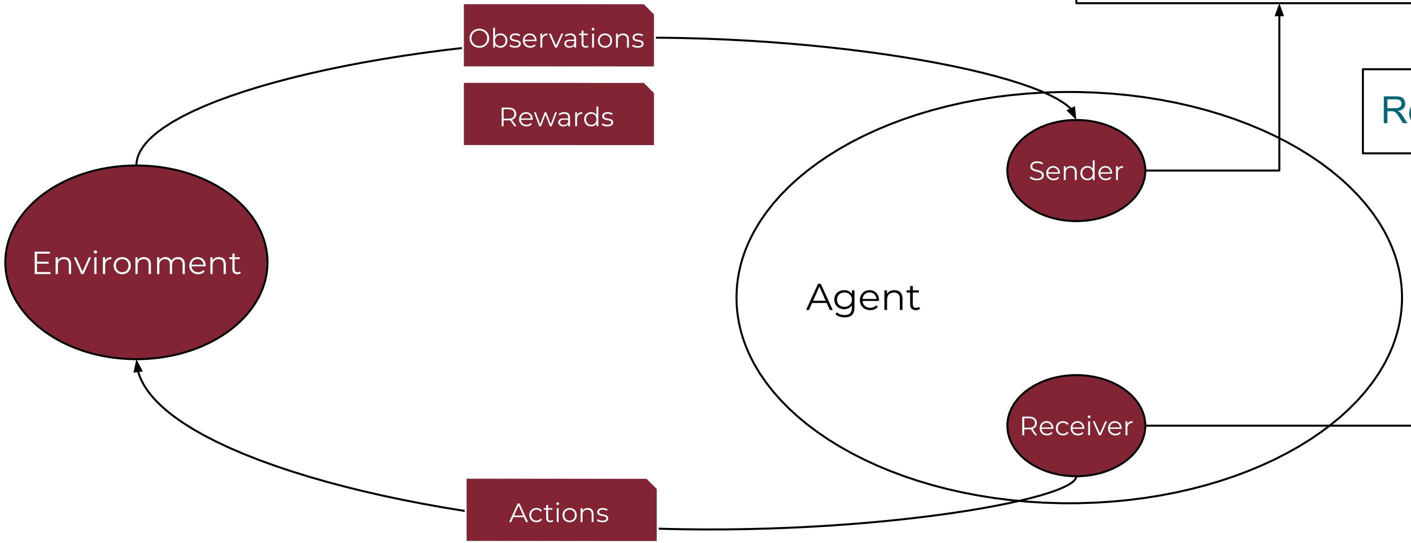


Emergent Communication



Sender has access to observations.

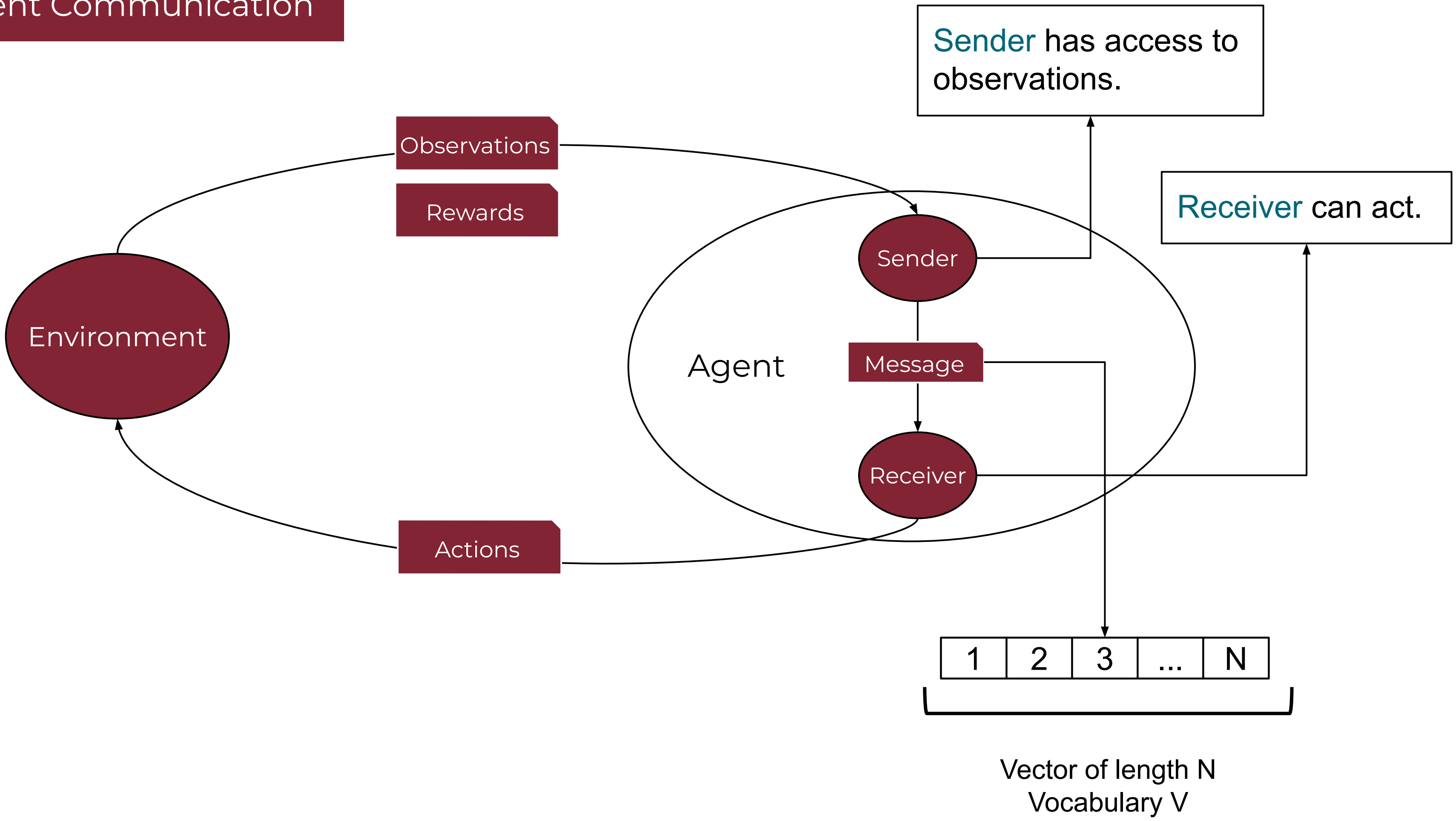
Emergent Communication



Sender has access to observations.

Receiver can act.

Emergent Communication



Observations

Rewards

Environment

Actions

Agent

Sender

Message

Receiver

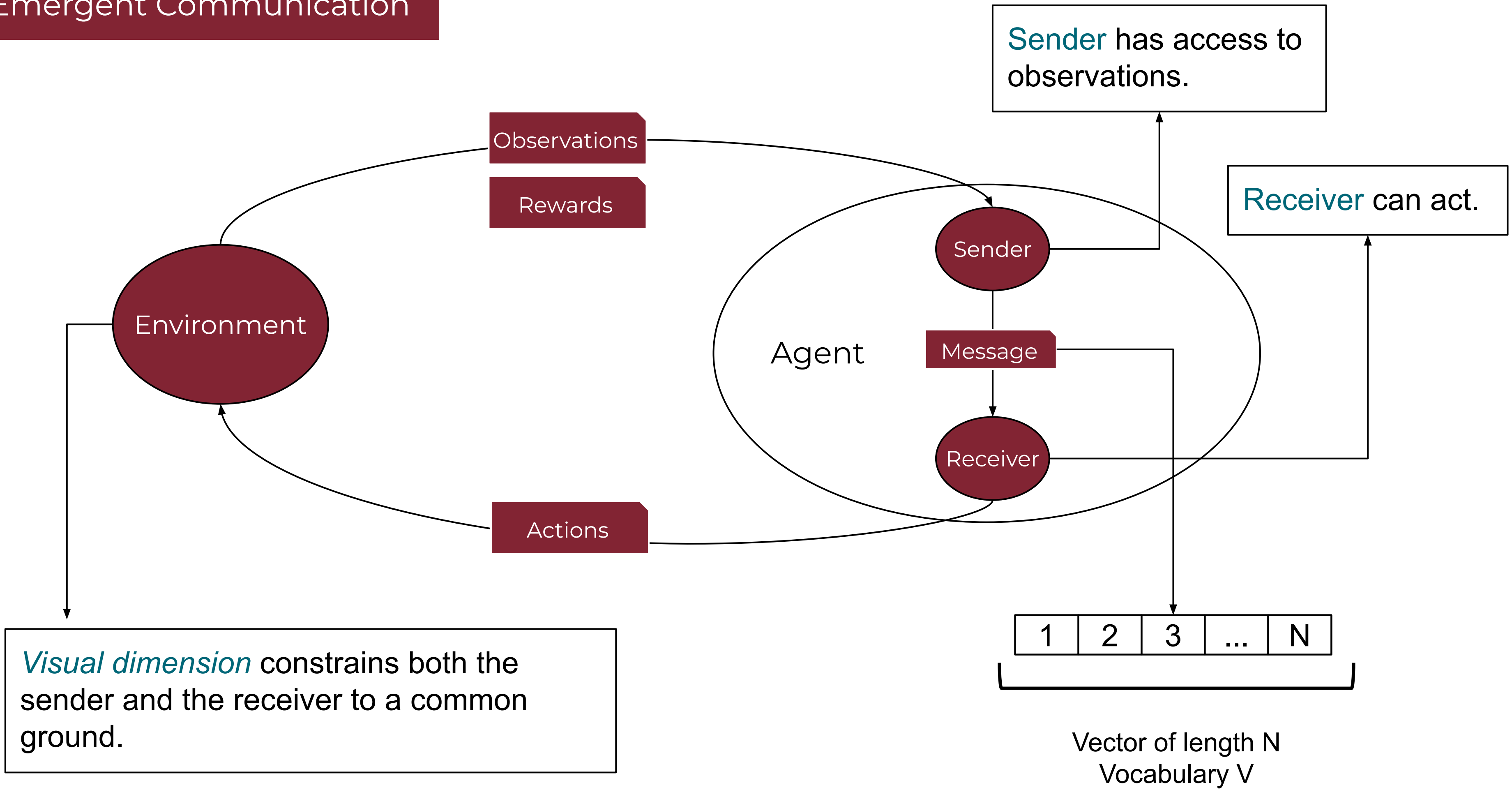
Sender has access to observations.

Receiver can act.

1	2	3	...	N

Vector of length N
Vocabulary V

Emergent Communication



Sender has access to observations.

Receiver can act.

Visual dimension constrains both the sender and the receiver to a common ground.

Vector of length N
Vocabulary V

But why?

As humans, we do not learn languages by reading wikipedia.
Learning a language involves interaction with other humans in a shared environment.

Cooperation:

- Cooperative agents
- Explainable decision making
- Cooperation in mixed human-robot teams

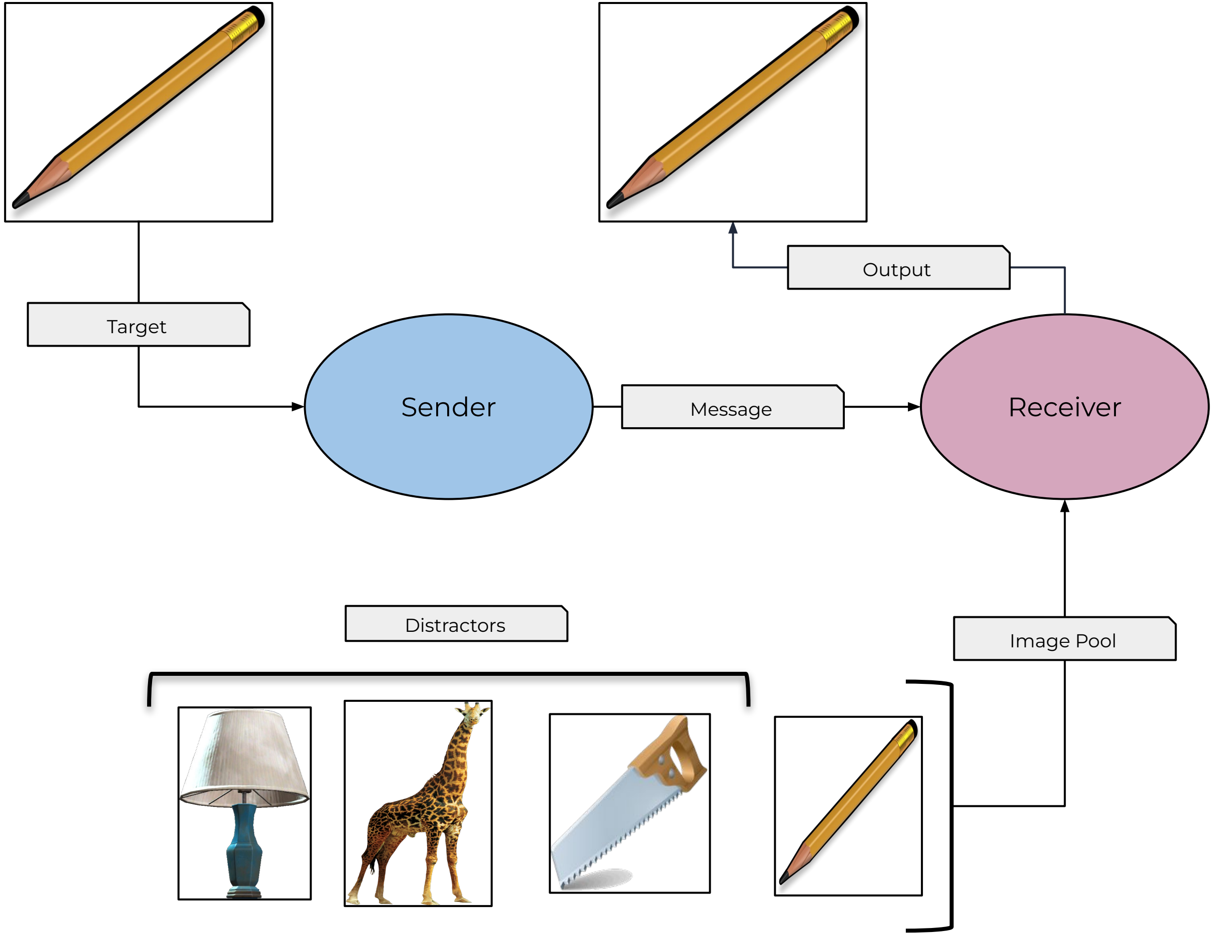
Language:

- Study language evolution in humans
- Emerge natural language properties in machines
- Understand learning difference between machines and humans

Background

- Reinforcement Learning
- Emergent Communication
- **Referential Games**
- Architectures & Frameworks:
  - Image captioning
  - DALL-E: Creating Images from Text

Referential Game [1]



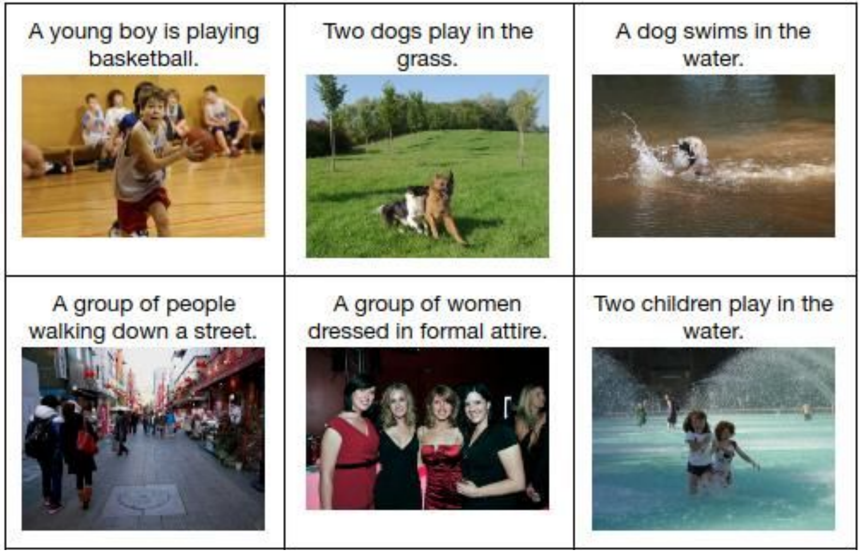
Background

- Reinforcement Learning
- Emergent Communication
- Referential Games
- **Architectures & Frameworks:**
  - Image captioning
  - DALL-E: Creating Images from Text

Image captioning

Proprieties:

- Active field since 2014
- Strong and defined benchmarks
- Datasets and Pretrained models available



A young boy is playing basketball.

Two dogs play in the grass.

A dog swims in the water.

A group of people walking down a street.

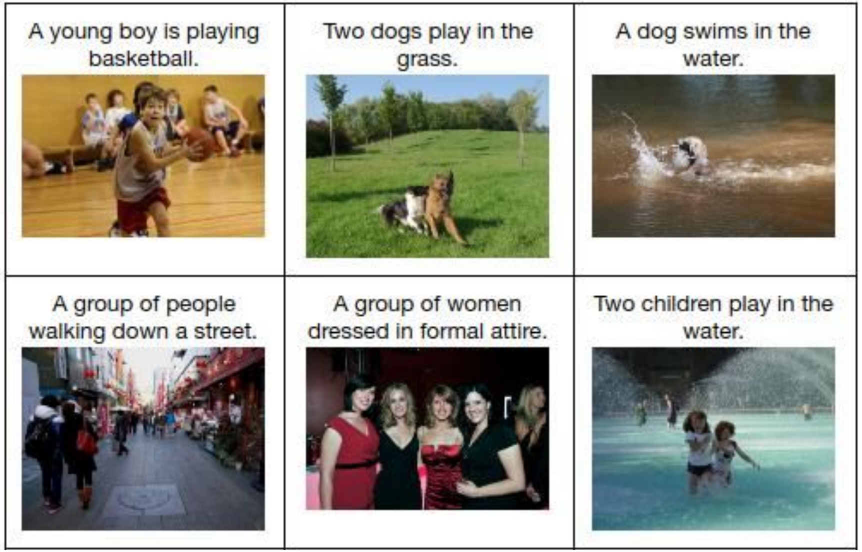
A group of women dressed in formal attire.

Two children play in the water.

Image captioning

Proprieties:

- Active field since 2014
- Strong and defined benchmarks
- Datasets and Pretrained models available



A young boy is playing basketball.

Two dogs play in the grass.

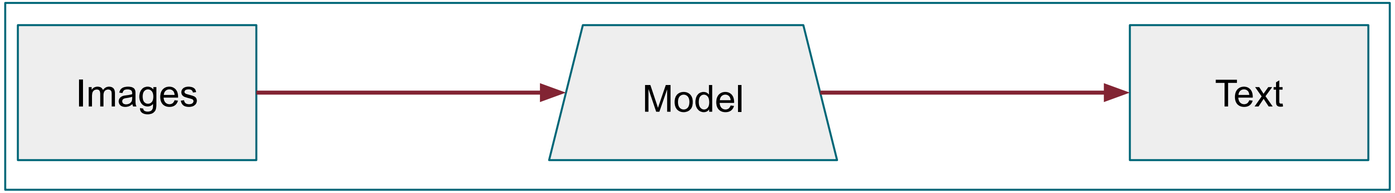
A dog swims in the water.

A group of people walking down a street.

A group of women dressed in formal attire.

Two children play in the water.

Pipeline



Images

Model

Text

Image captioning

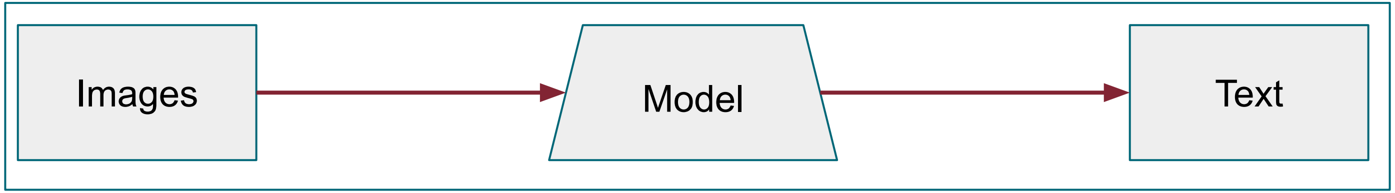
Datasets:

- COCO captions [6]
- SCICAP [7]
- VizWiz [11]
- Flickr30k [13]

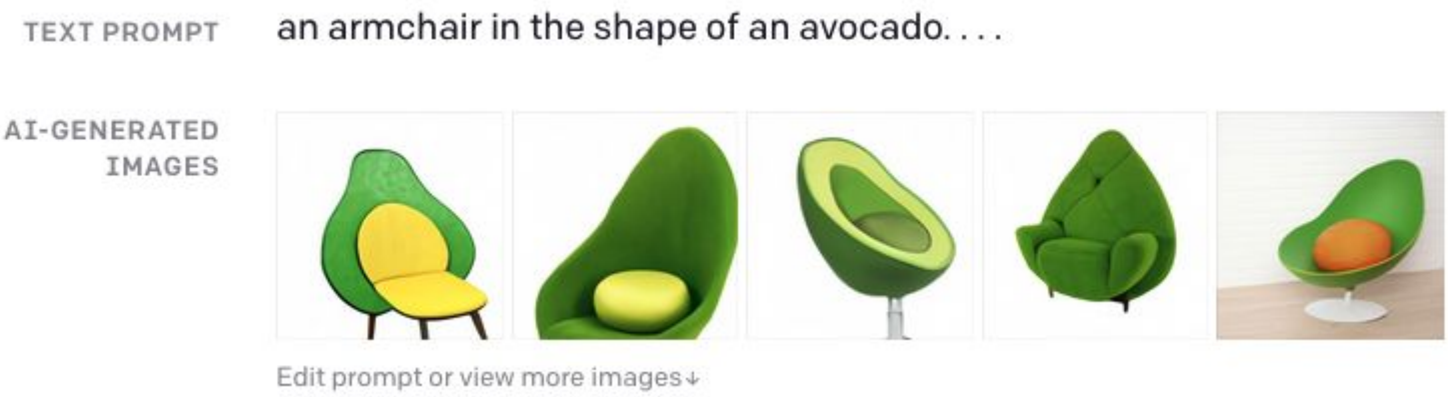
Architectures:

- CNN+LSTM [7]
- LEMON [8] : CNN+Attention
- BLIP [9] : Visual Transformer + Encoder Decoder
- M2 [10] : Transformer
- …

Pipeline



DALL-E [2]



TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

Architectures:

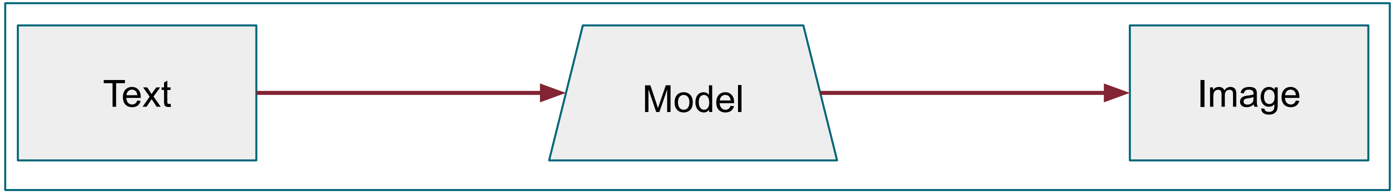
- GPT-3 [3]
- VQ-VAE [4]
- CLIP reranker [5]

Proprieties:

- 12-billion parameters
- Input text - output image
- Full dataset not disclosed

DALL-E [2]

Pipeline



Architectures:

- GPT-3 [3]
- VQ-VAE [4]
- CLIP reranker [5]

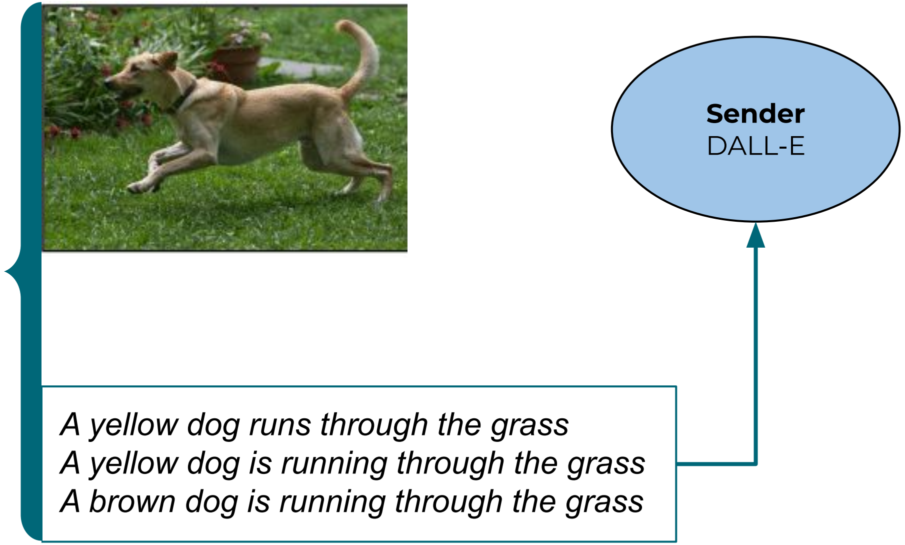
Proprieties:

- 12-billion parameters
- Input text - output image
- Full dataset not disclosed

Emergent ImageNation [EmIn]

- **Framework**
- Research lines:
    - Training for dalle
    - Population of speakers/listeners
    - Communicating through images
- Code is available
- Bibliography

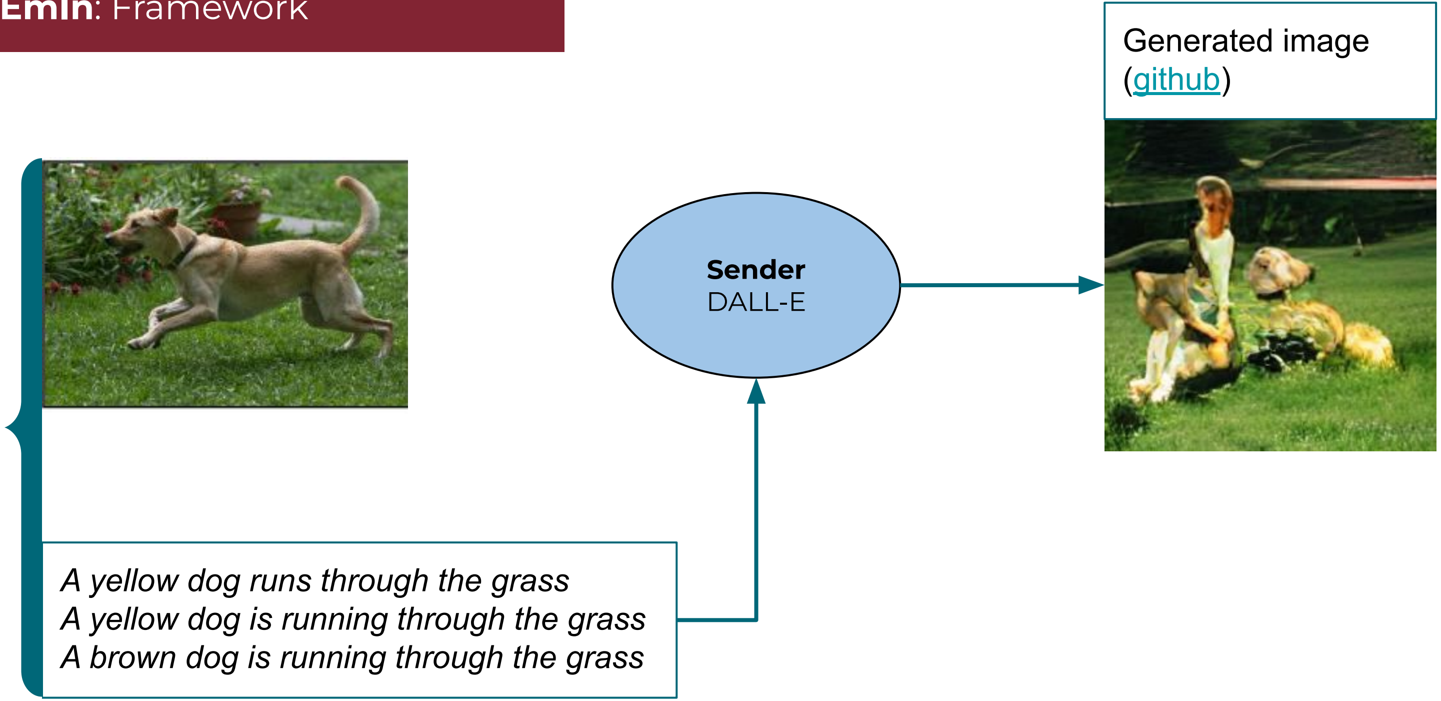
EmIn: Framework



A yellow dog runs through the grass
A yellow dog is running through the grass
A brown dog is running through the grass

Sender
DALL-E

EmIn: Framework

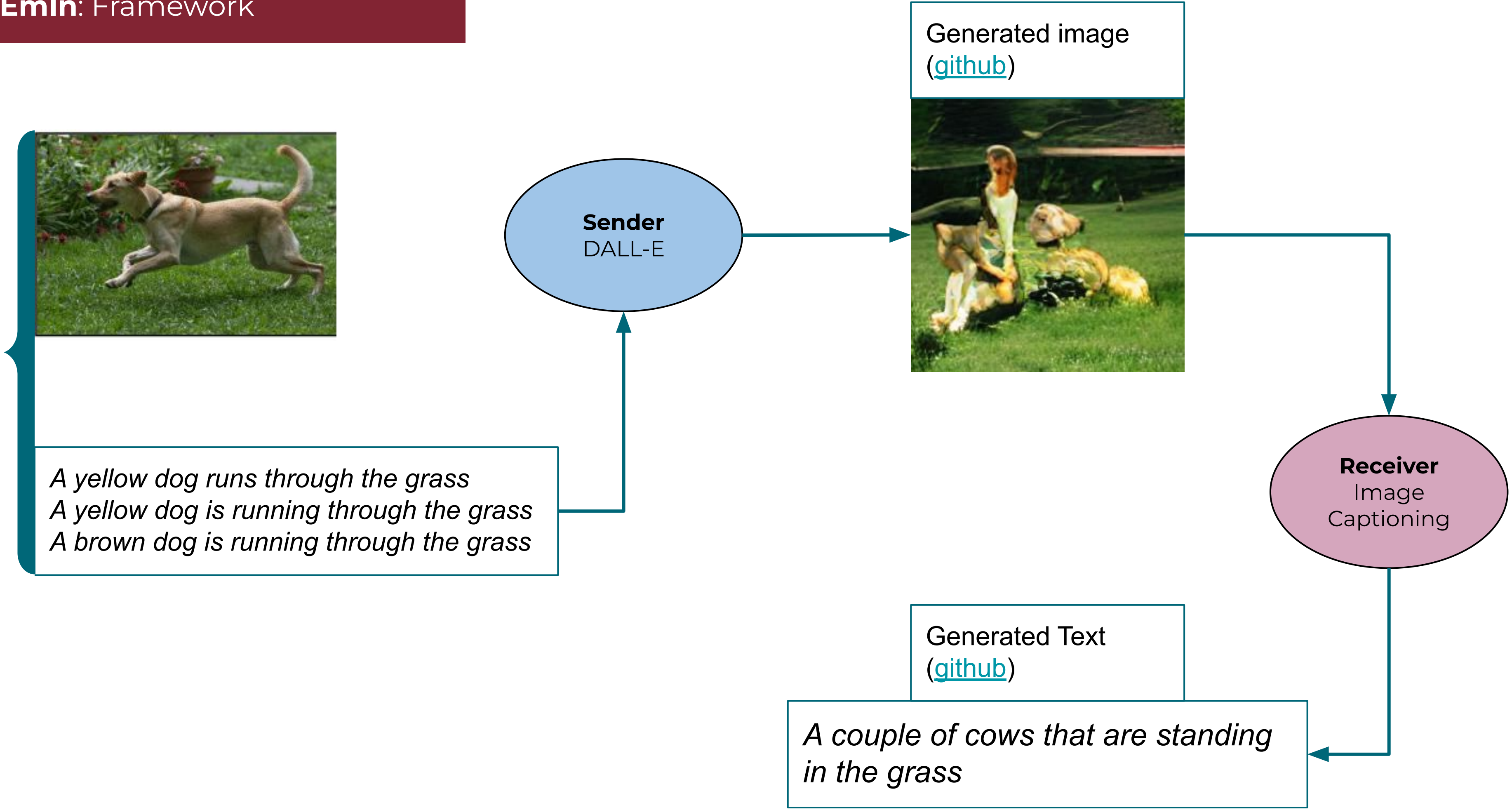


A yellow dog runs through the grass
A yellow dog is running through the grass
A brown dog is running through the grass

Sender
DALL-E

Generated image
(github)

EmIn: Framework



A yellow dog runs through the grass
A yellow dog is running through the grass
A brown dog is running through the grass

Sender
DALL-E

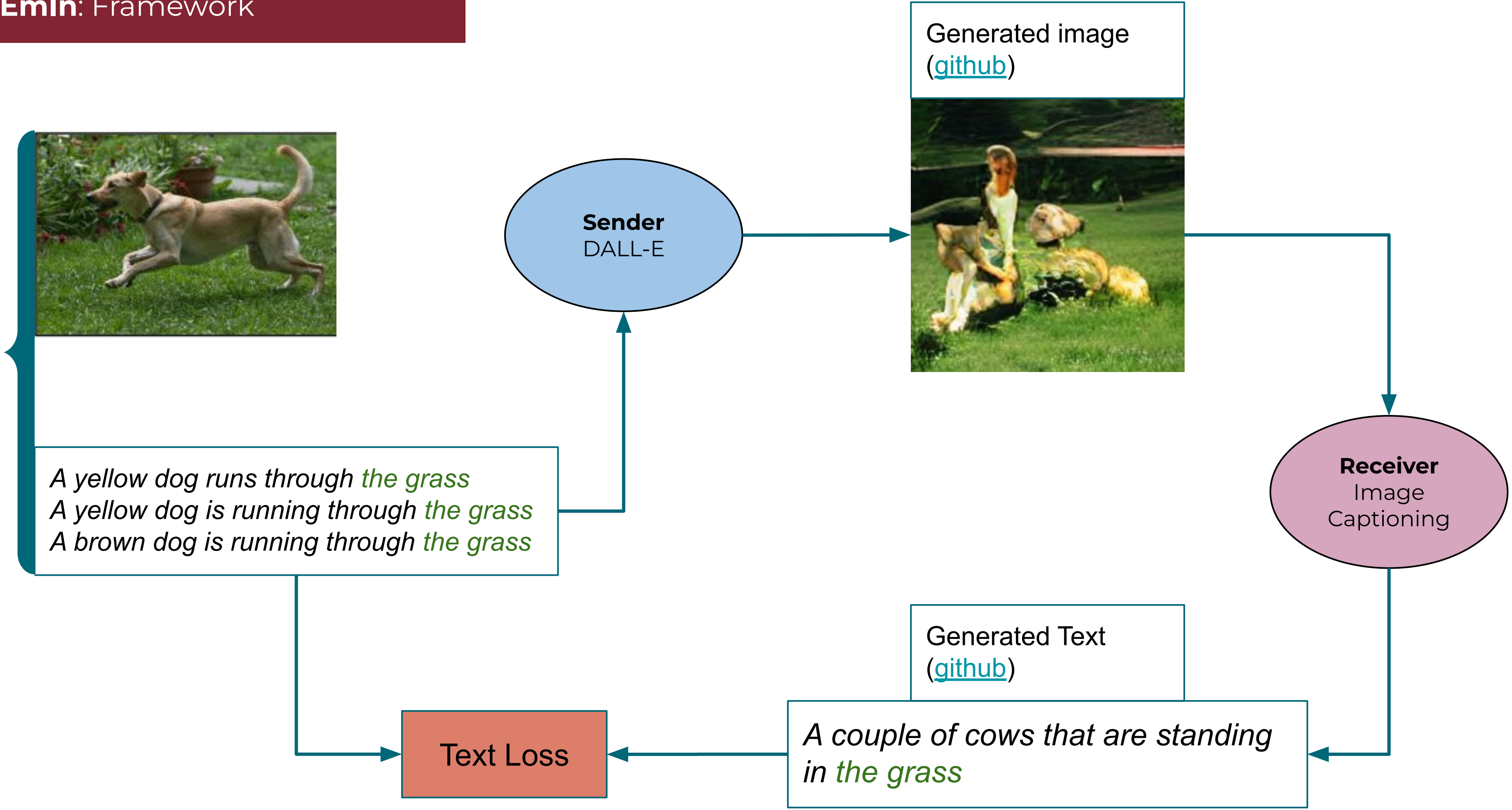
Generated image
(github)

Receiver
Image
Captioning

Generated Text
(github)

A couple of cows that are standing in the grass

EmIn: Framework



Generated image (github)

A yellow dog runs through the grass
A yellow dog is running through the grass
A brown dog is running through the grass

Sender
DALL-E

Receiver
Image Captioning

Generated Text (github)

A couple of cows that are standing in the grass

Text Loss

EmIn: Framework

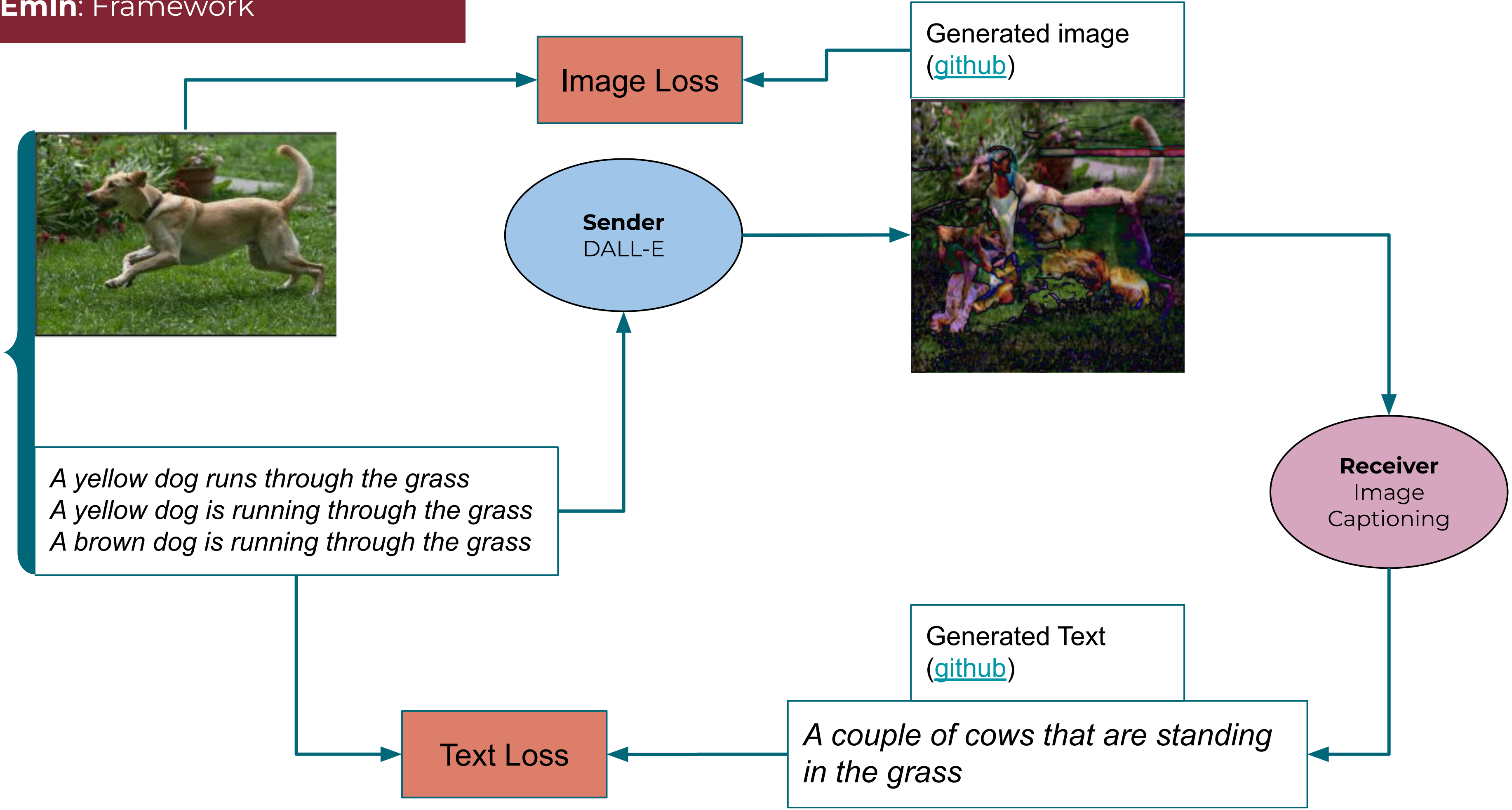


Image Loss

Generated image (github)

Sender
DALL-E

A yellow dog runs through the grass
A yellow dog is running through the grass
A brown dog is running through the grass

Receiver
Image Captioning

Generated Text (github)

A couple of cows that are standing in the grass

Text Loss

EGG library

EGG 🐣: Emergence of lanGuage in Games

Repository:

- 86 Fork
- 227 Star
- >15 Papers based on EGG

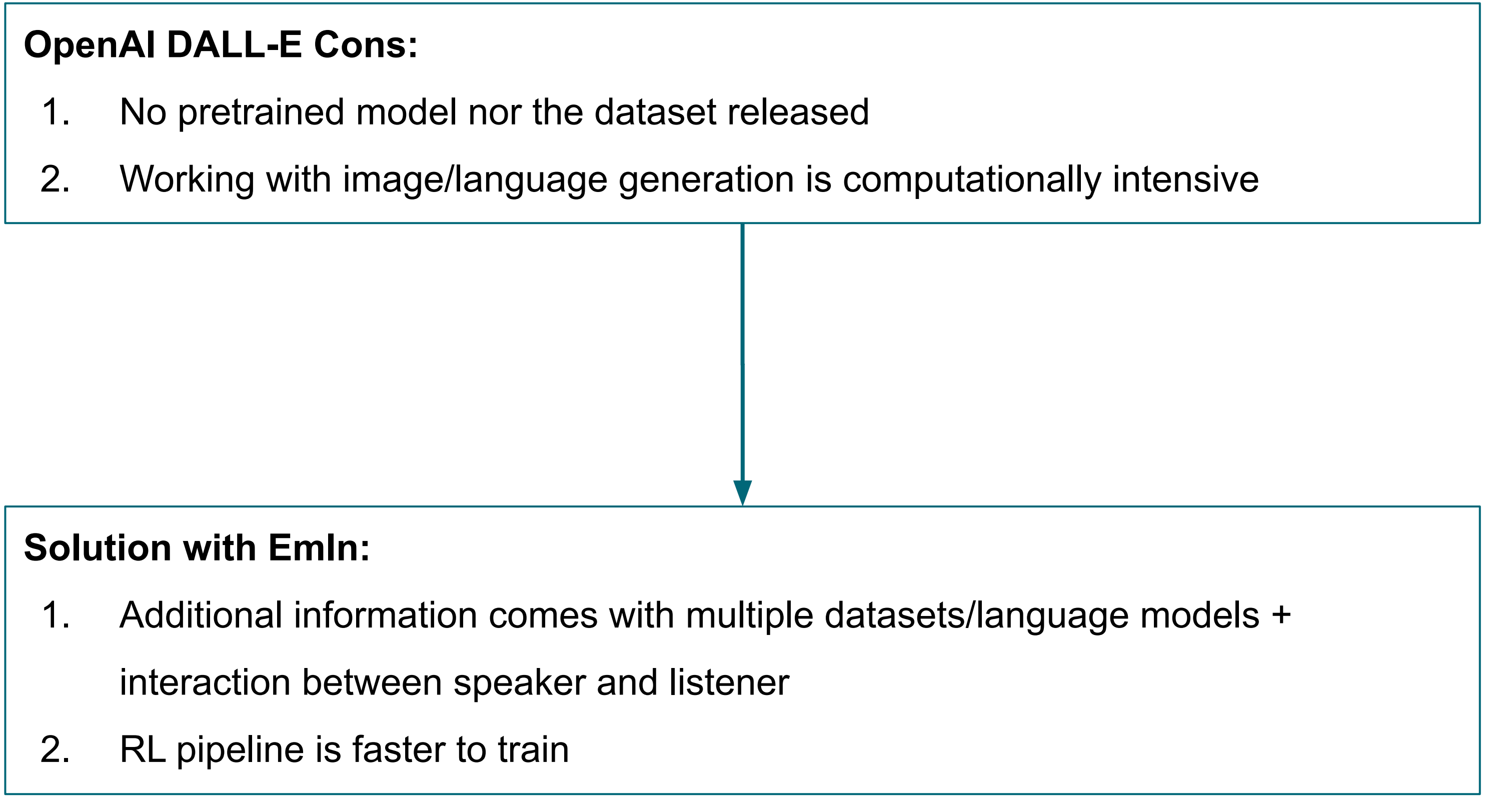
Features:

- Discrete / continuous Communication
- Single pair/ population of agents
- Optimization with Reinforce or Gumbel-Softmax
- Distributed training
- Cuda-aware command for grid-search

Emergent ImageNation [EmIn]

- Framework
- Research lines:
    - **Training for dalle**
    - Population of speakers/listeners
    - Communicating through images
- Code is available
- Bibliography

Objective : DALL-E train



OpenAI DALL-E Cons:

1. No pretrained model nor the dataset released
2. Working with image/language generation is computationally intensive

Solution with EmIn:

1. Additional information comes with multiple datasets/language models + interaction between speaker and listener
2. RL pipeline is faster to train

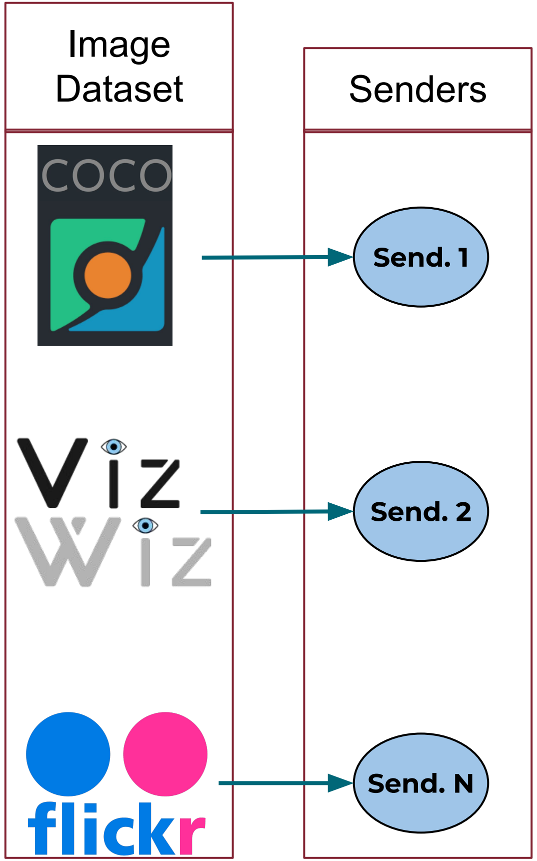
Emergent ImageNation [EmIn]

- Framework
- Research lines:
  - Training for dalle
  - **Population of speakers/listeners**
  - Communicating through images
- Code is available
- Bibliography

Objective : Population

Image Dataset
COCO
VizWiz
flickr

Objective : Population



Objective : Population

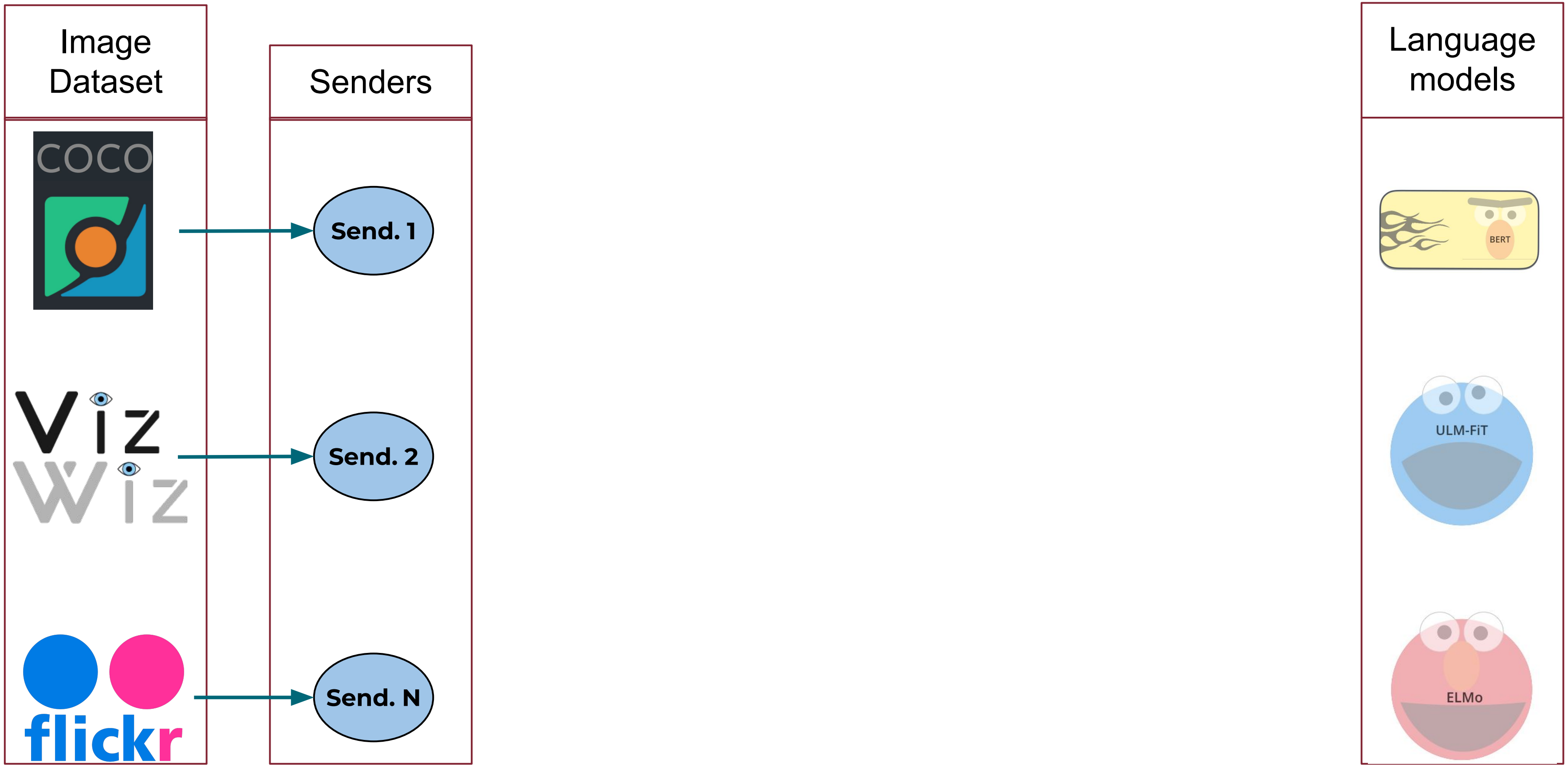


Image Dataset

Senders

Language models

Send. 1

Send. 2

Send. N

Objective : Population

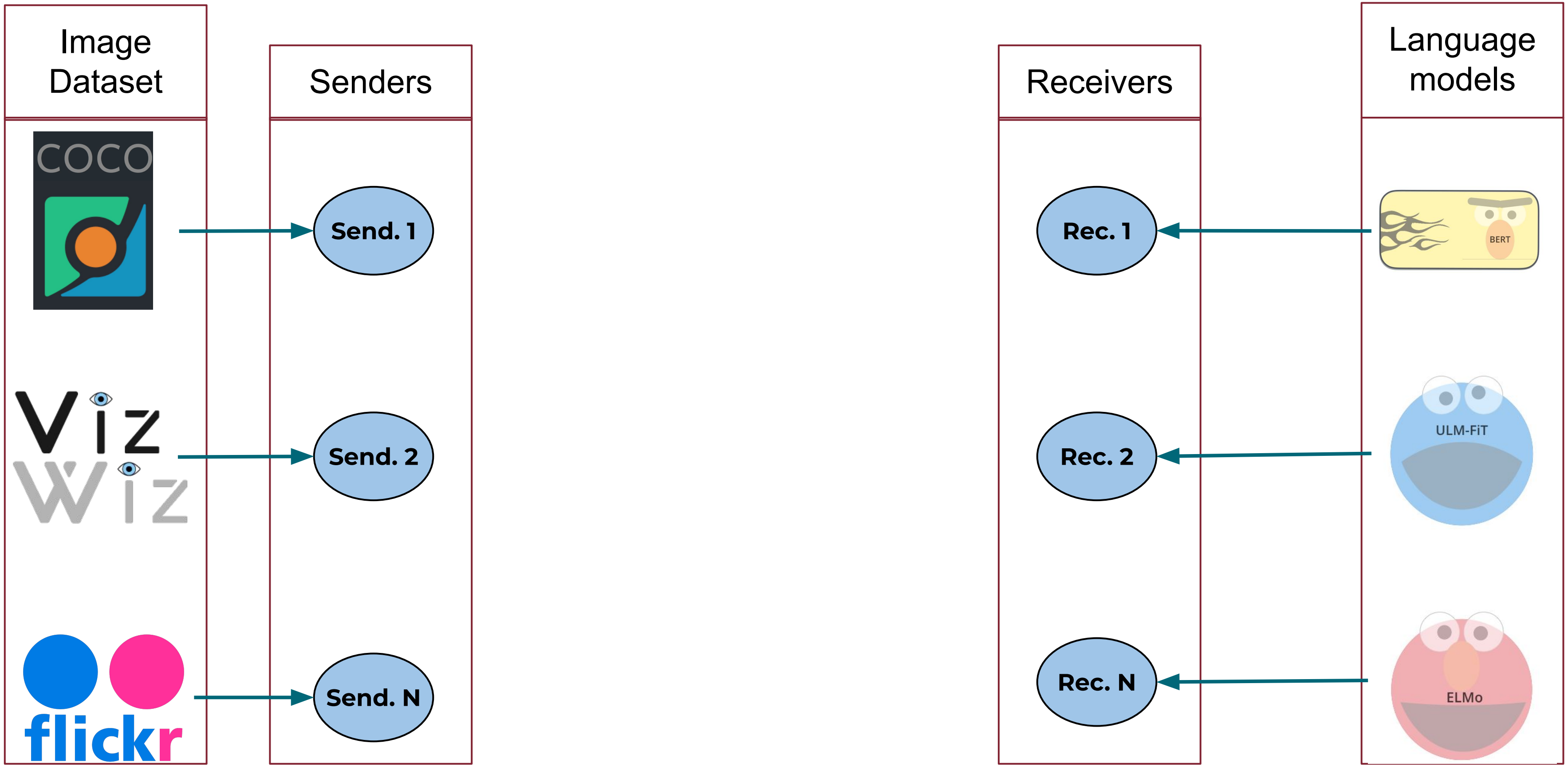


Image Dataset

Senders

Receivers

Language models

Send. 1

Send. 2

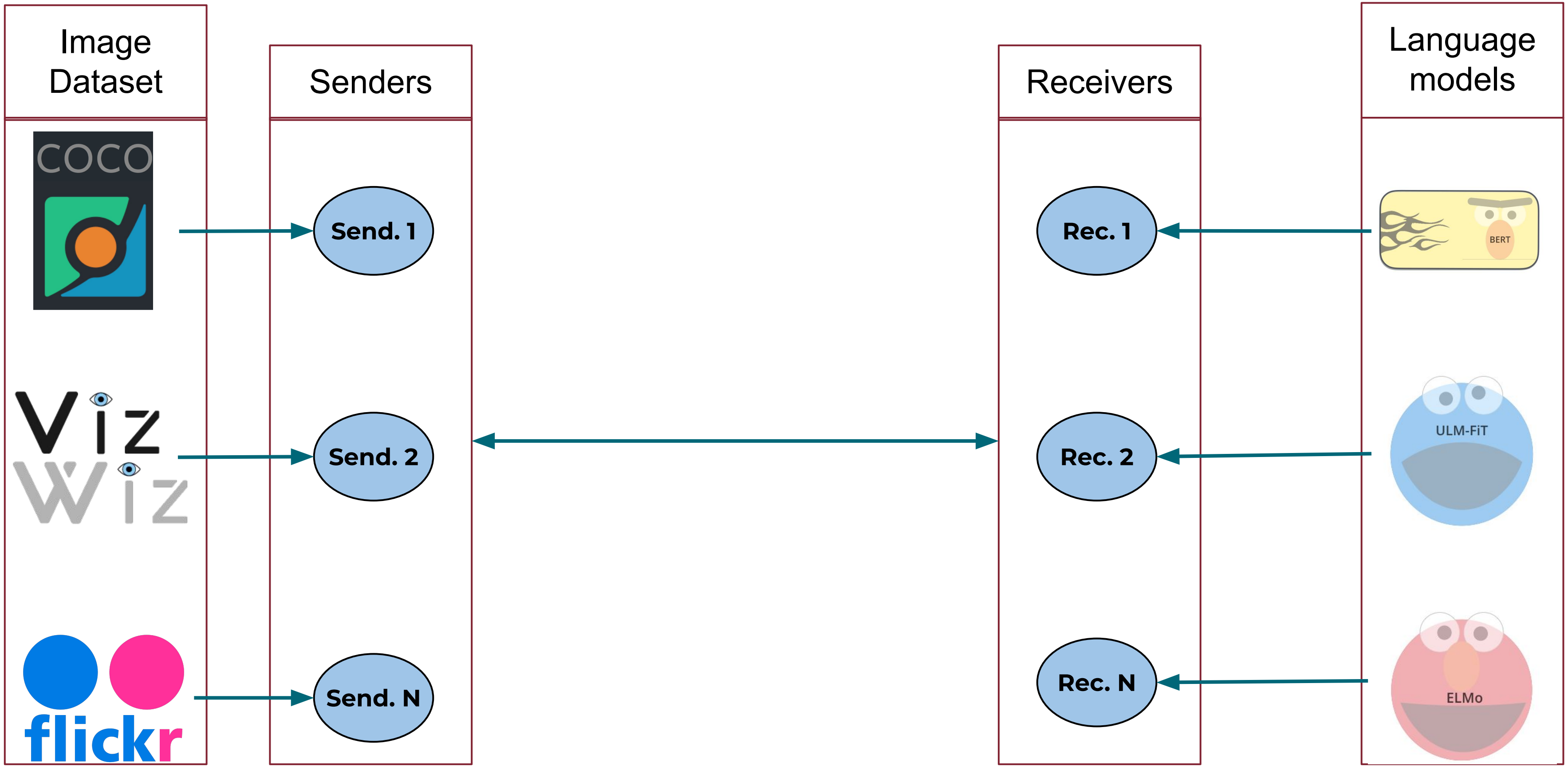
Send. N

Rec. 1

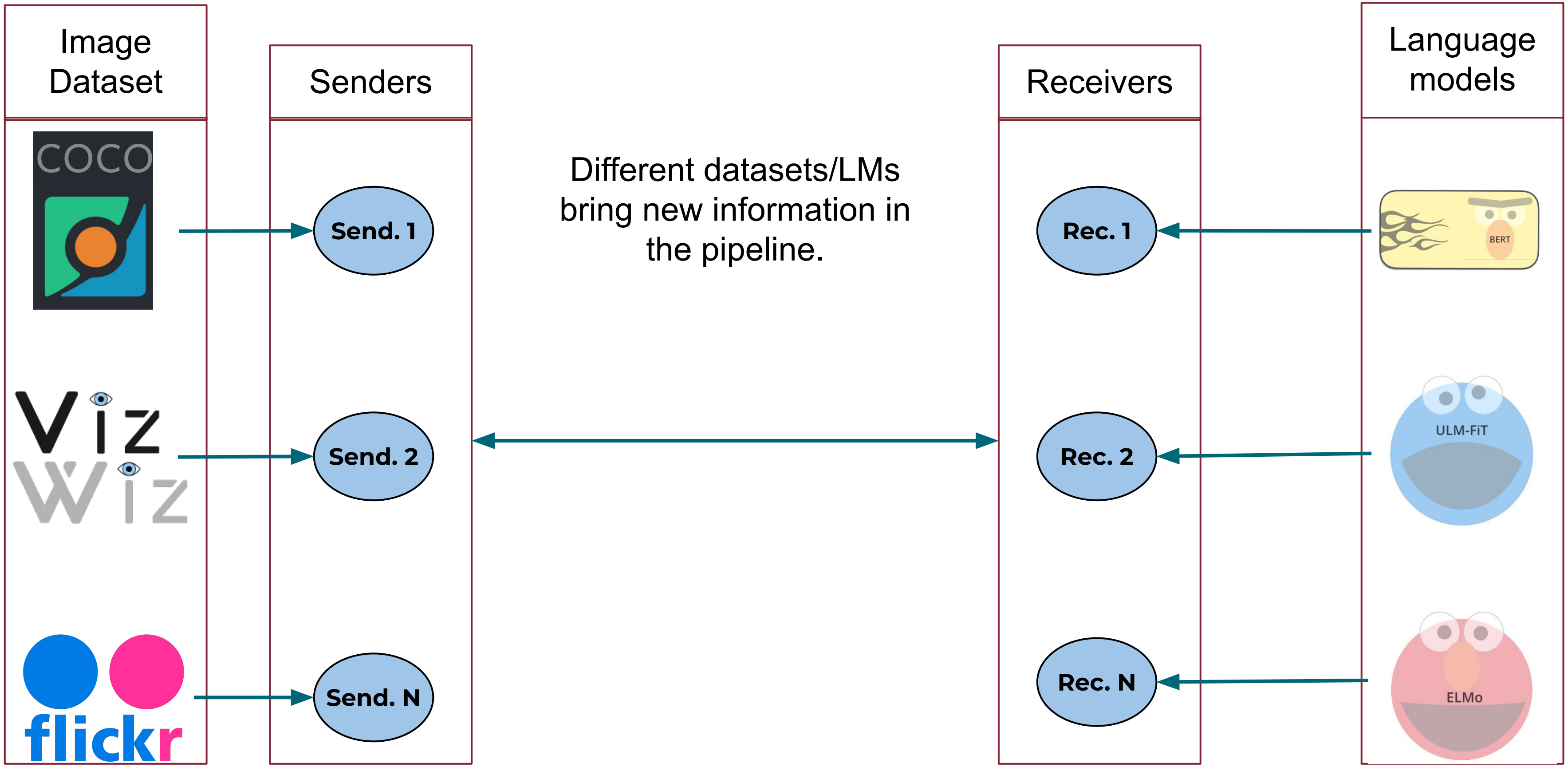
Rec. 2

Rec. N

Objective : Population

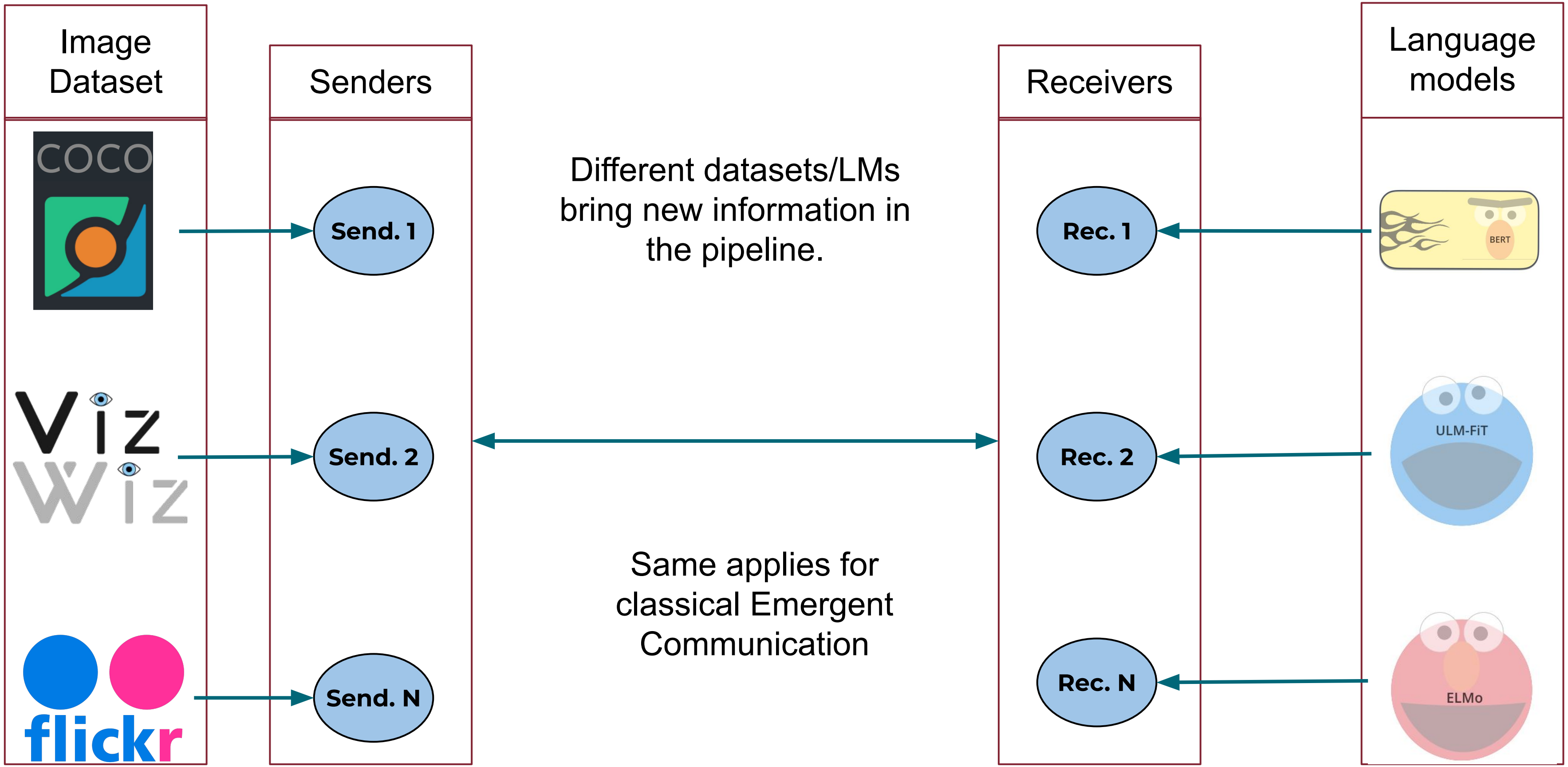


Objective : Population



Different datasets/LMs bring new information in the pipeline.

Objective : Population



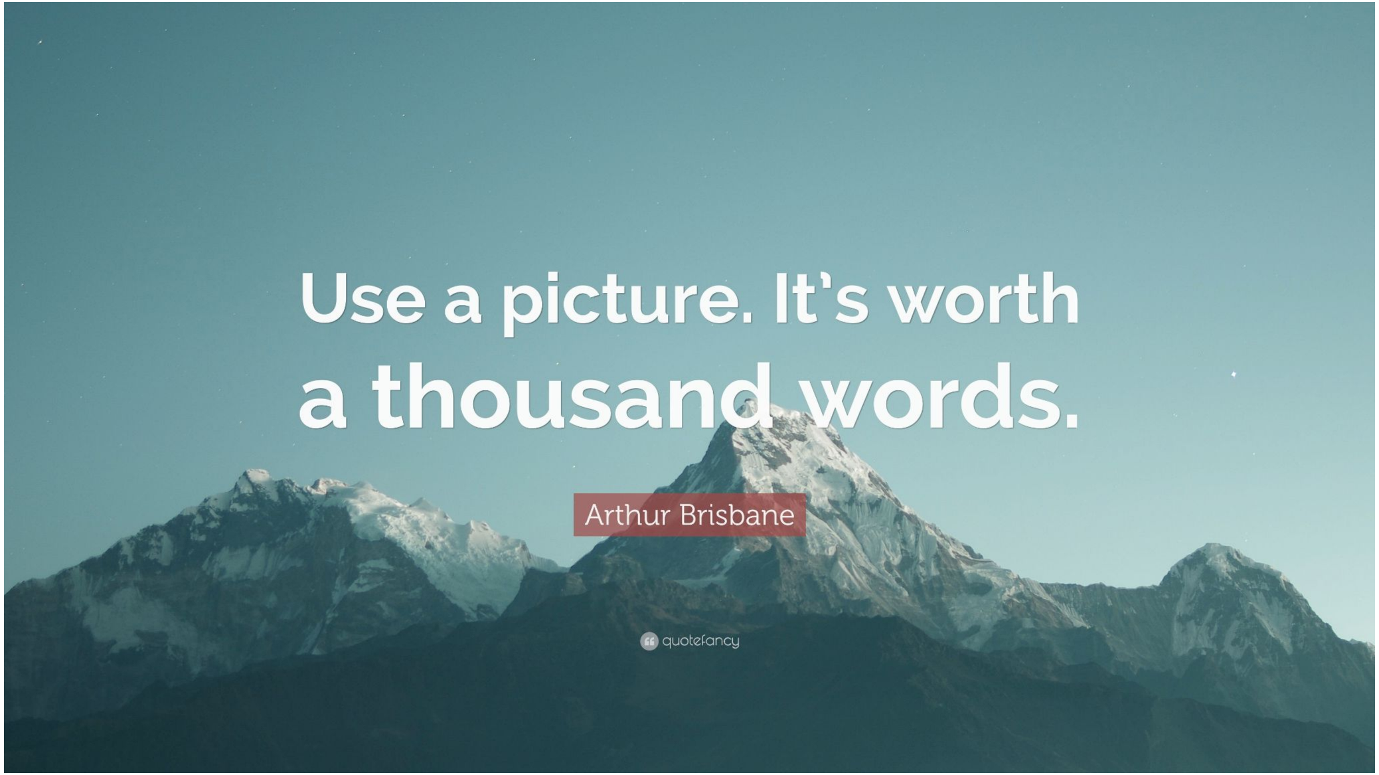
Different datasets/LMs bring new information in the pipeline.

Same applies for classical Emergent Communication

Emergent ImageNation [EmIn]

- Framework
- Research lines:
  - Training for dalle
  - Population of speakers/listeners
  - **Communicating through images**
- Code is available
- Bibliography

Communication



“70 to 93 percent of all communication is nonverbal”
[12]

Lost in translation

Pacioccone

(Italian) A plump person, with a jovial and good-natured appearance. A lover of the quiet life.

Tartle

(Scottish) If you've ever been talking to someone you've been introduced to before but their name has completely disappeared from your brain then you've tartled.

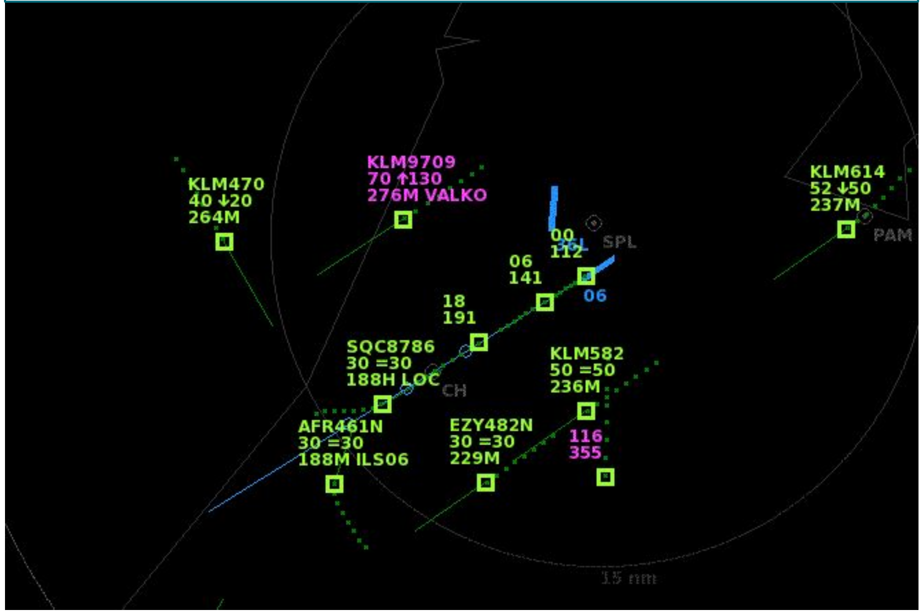
Sobremesa

(Spanish) The moment after eating a meal when the food is gone but the conversation is still flowing at the table.

Augmenting text with images

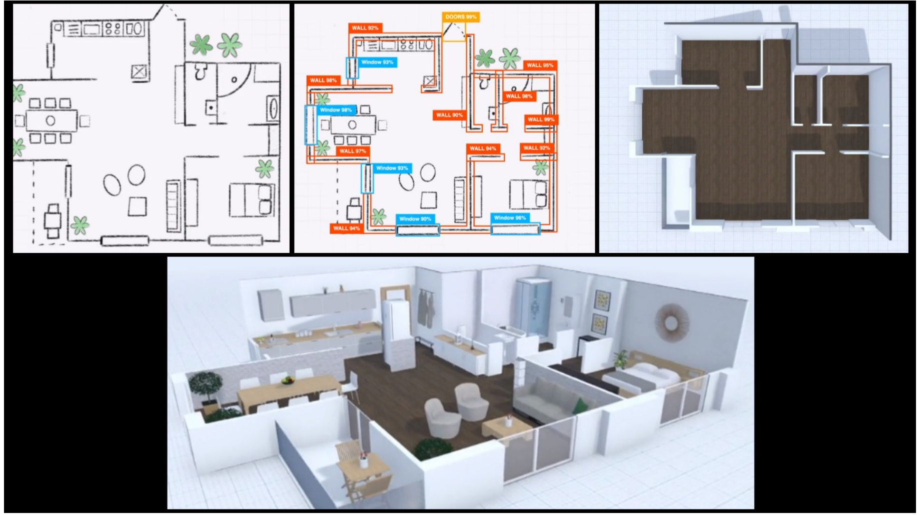
Air traffic is based on communication protocols with thousand of messages every hour.
Generating informative images aids the general organization.

Air traffic



Projecting interior design structure trough natural text prompts.

Interior design architecture



Emergent ImageNation [EmIn]

- Framework
- Research lines:
  - Training for dalle
  - Population of speakers/listeners
  - Communicating through images
- **Code is available**
- Bibliography

Code available

https://github.com/nicofirst1/Emergent-ImageNation

Bibliography

[1] David Lewis. Convention: A philosophical study. John Wiley & Sons, 2008.
[2] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning. PMLR, 2021.
[3] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
[4] Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." Advances in neural information processing systems 30 (2017).
[5] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
[6] Wang, Peng, et al. "Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework." arXiv preprint arXiv:2202.03052 (2022).
[7] Hsu, Ting-Yao, C. Lee Giles, and Ting-Hao'Kenneth Huang. "SciCap: Generating Captions for Scientific Figures." arXiv preprint arXiv:2110.11624 (2021).
[8] Hu, Xiaowei, et al. "Scaling up vision-language pre-training for image captioning." arXiv preprint arXiv:2111.12233 (2021).
[9] Li, Junnan, et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." arXiv preprint arXiv:2201.12086 (2022).
[10] Cornia, Marcella, et al. "Meshed-memory transformer for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[11] Gurari, Danna, et al. "Captioning images taken by people who are blind." European Conference on Computer Vision. Springer, Cham, 2020.
[12] Mehrabian, Albert. Silent messages. Vol. 8. No. 152. Belmont, CA: Wadsworth, 1971.
[13] Young, Peter, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions." Transactions of the Association for Computational Linguistics 2 (2014): 67-78.