Image Captioning

Danna Gurari
University of Colorado Boulder
Fall 2022

https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html

Review

- Last week
  - Visual question answering applications
  - Visual question answering datasets
  - Visual question answering evaluation
  - Mainstream challenge 2015 winner: baseline approach
  - Mainstream challenge 2019 winner: transformer-based approach
  - Programming tutorial
- Assignments (Canvas)
  - Lab assignment 4 due next week
- Questions?

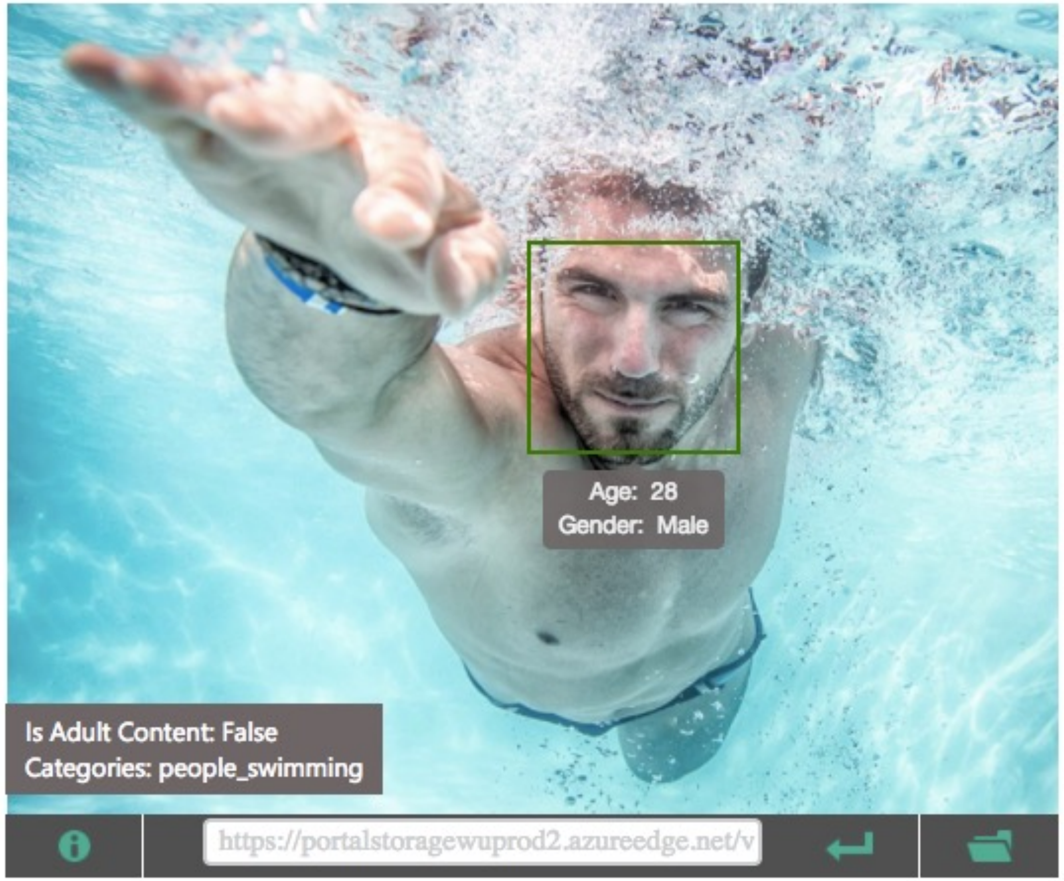
Today’s Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners

Today’s Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners

A “Human-Like” Description



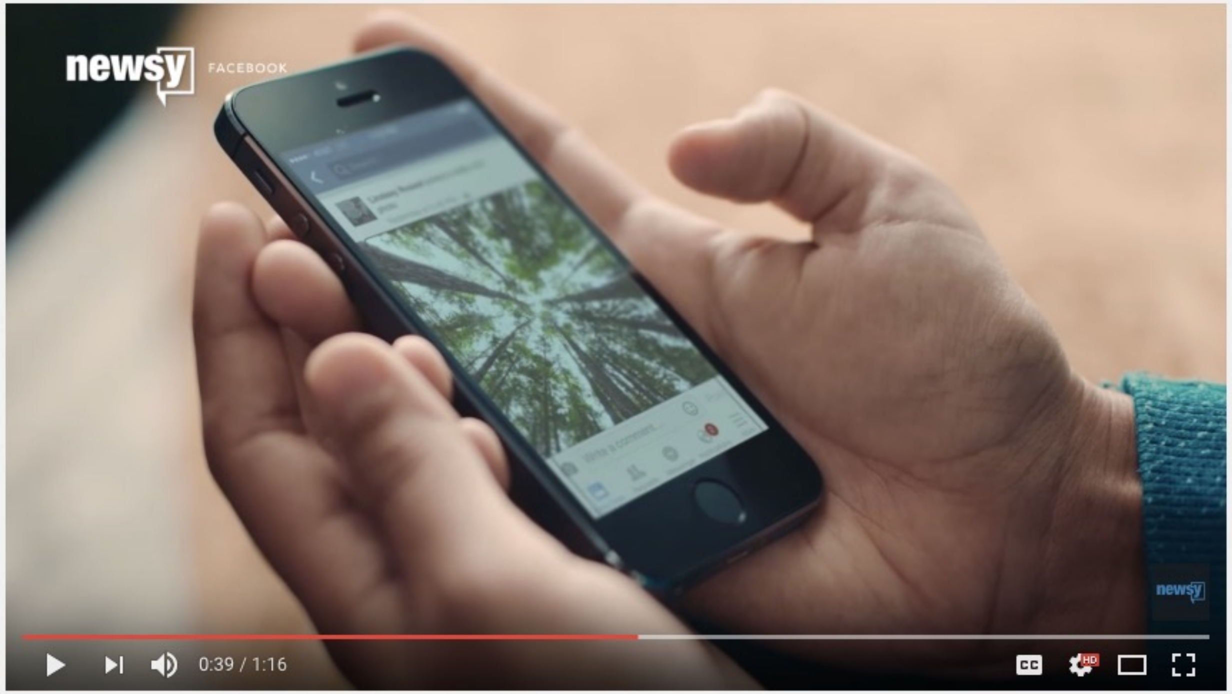
Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108693093019 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	jpeg
Image Dimensions	1500 x 1155
Clip Art Type	0 Non-clipart
Line Drawing Type	0 Non-LineDrawing
Black & White Image	False

Captions: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

Visual Assistance for People with Visual Impairments

Facebook



Facebook's New AI Tool Is Helping Blind Users 'See' Photos - Newsy

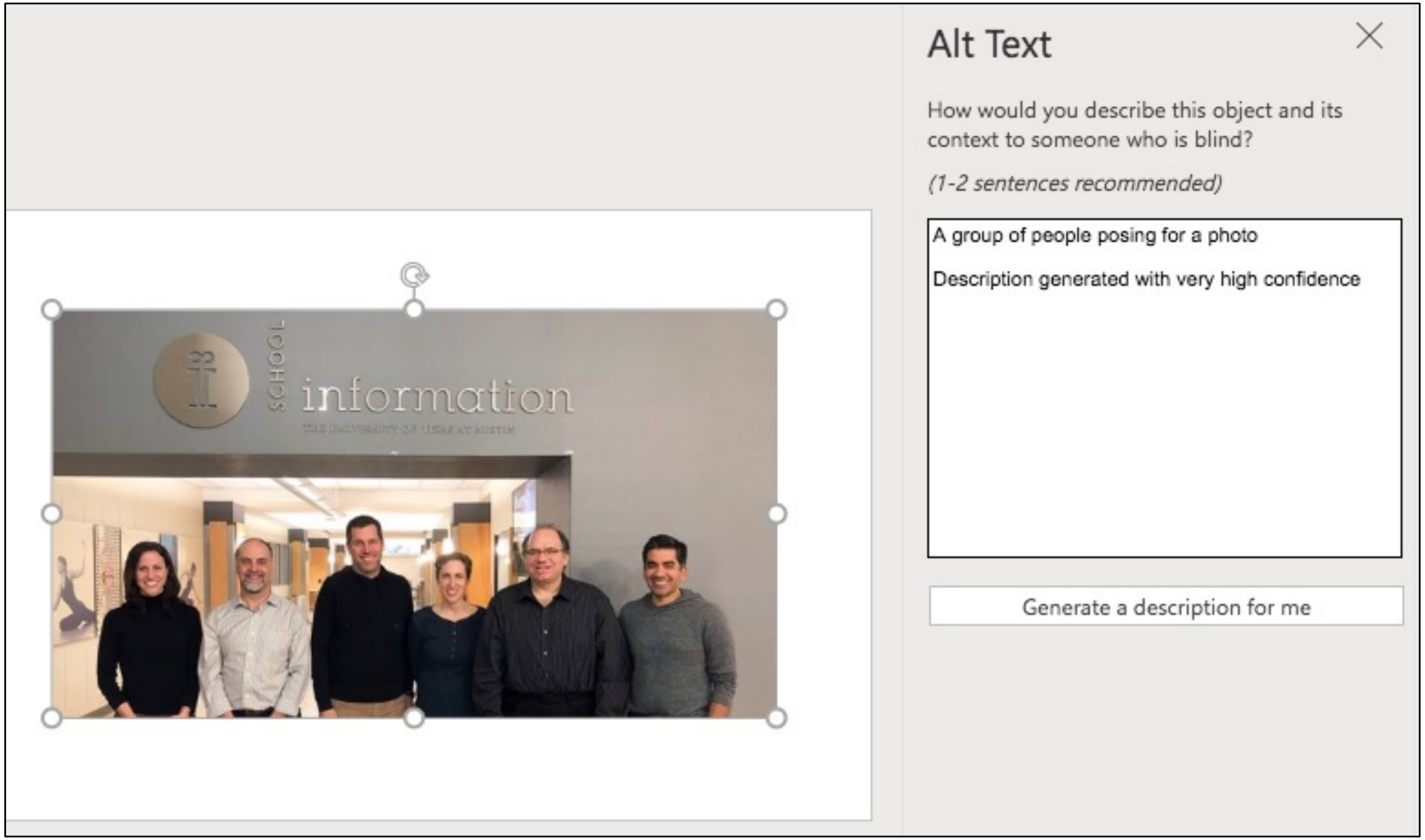
https://www.youtube.com/watch?v=Tjugc8a836Q

Microsoft

Saqib Shaikh : Microsoft Developer Can 'See' Using Artificial Intelligence Headset

https://www.youtube.com/watch?v=R2mC-NUAmMk

Alt Text for People with Visual Impairments



e.g., Microsoft Power Point (Office 365 demo)

Image Captioning for Newspaper Articles

BBC NEWS UK EDITION

Last Updated: Friday, 23 September 2005, 18:00 GMT 19:00 UK

E-mail this to a friend   Printable version

Apple-growing Pole feels squeeze

By Oana Lungescu

Brussels correspondent

As part of our ongoing season "Who Runs Your World," the BBC took a Polish apple-grower to meet the movers and shakers of Brussels.

Aid for farmers has become one of the most hotly disputed issues in the European Union.

When Poland was one of the 10 countries to join the EU last year, six million farmers feared they would be put out of business by their heavily-subsidised colleagues in western Europe.

Marcin and Florent face intense competition from outside Europe

But it has been a good year for Polish farmers.

Those who feared the worst from EU membership are now the first to harvest its fruit. With more than 1.6bn euros (£1.1bn) in subsidies so far, farmers have practically doubled their income.

Marcin Hermanowicz grows apples on 30 hectares of land, in a region south of Warsaw often described as Europe's largest orchard.

Feng and Lapata. Automatic Image Annotation Using Auxiliary Text Information. Proceedings of ACL-08: HLT. 2008

Aiding Tourism with Captioned Images

Figure 7: Tourists from three different tour groups at the Salt Lake of Uyuni in Bolivia

Figure 3: Examples for people shots
(Peruvian Children, Korean Guards, Russian Singers)

Figure 8: The Cathedral of Cuzco, Peru, in different viewing angles (right, left and front)

Figure 4: Examples for animal photos
(Humpback Whale, Kangaroos, Galapagos Giant Turtle)

Grubinger et al. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. 2006

Describing and Responding to Images Posted to Social Media with "Personality"

Standard captioning output: A plate with a sandwich and salad on it.

Our model with different personality traits (215 possible traits, not all shown here):

Sweet	That is a lovely sandwich.
Dramatic	This sandwich looks so delicious! My goodness!
Anxious	I'm afraid this might make me sick if I eat it.
Sympathetic	I feel so bad for that carrot, about to be consumed.
Arrogant	I make better food than this
Optimistic	It will taste positively wonderful!
Money-minded	I would totally pay $100 for this plate.

Shuster et al. Engaging Image Captioning via Personality. 2019

Describing Products

Title: Stand Collar A-Line Dress

Fashion Caption: A pearly button accents the stand collar that gives this so-simple, yet so-chic A-line dress its retro flair

Color: Black and ivory

Meta: - 33" petite length (size 8P) - Hidden back-zip closure - Stand collar - Cap sleeves - Side-seam pockets – A-Lined - 63% polyester, 34% rayon, 3% spandex - Dry clean or hand wash, dry flat - Imported – Dress

Image Caption: A person in a dress

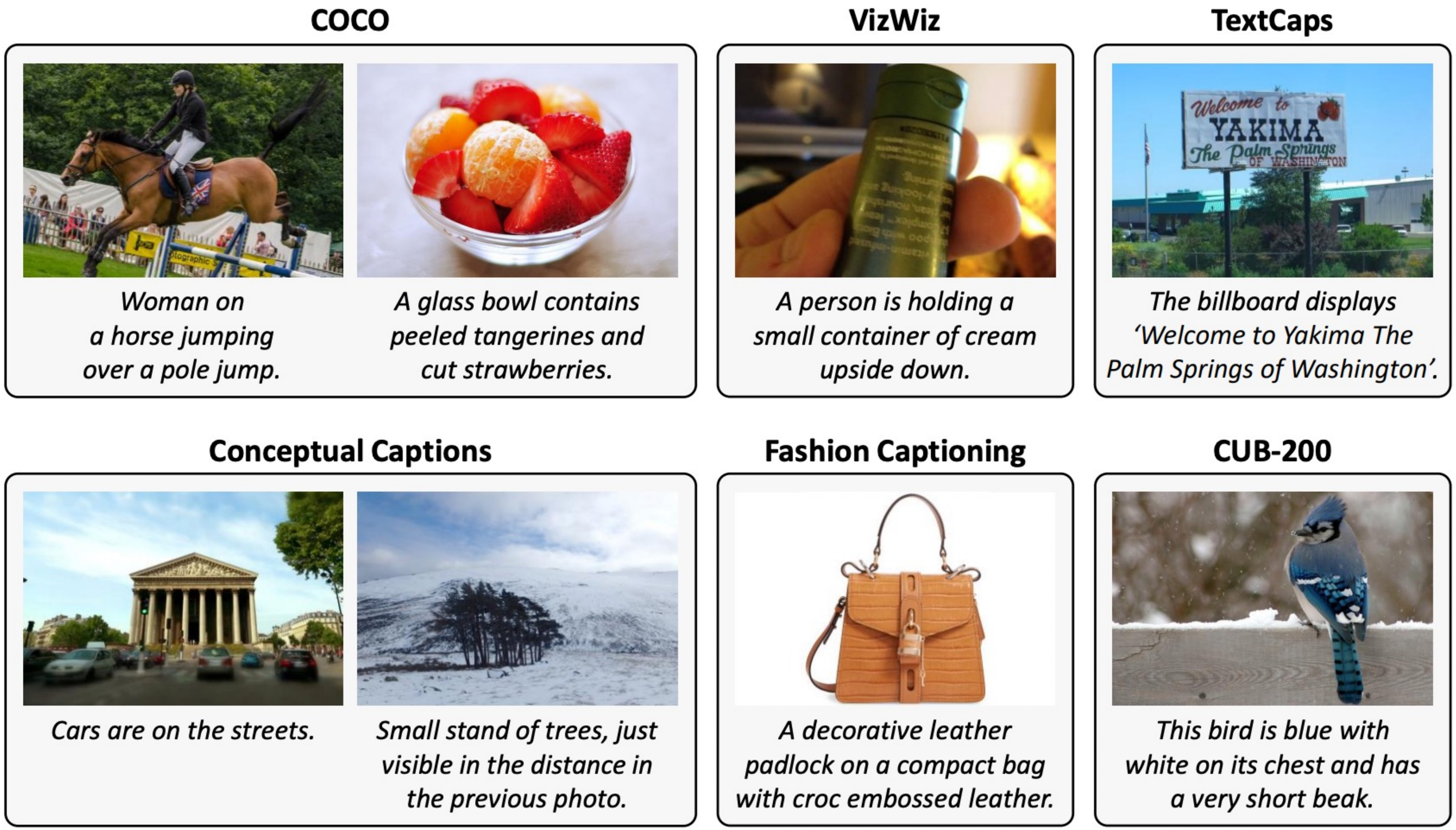
Yang et al. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. ECCV 2020

What are other potential applications for image captioning?

Today's Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners

Sample of Existing Dataset Challenges



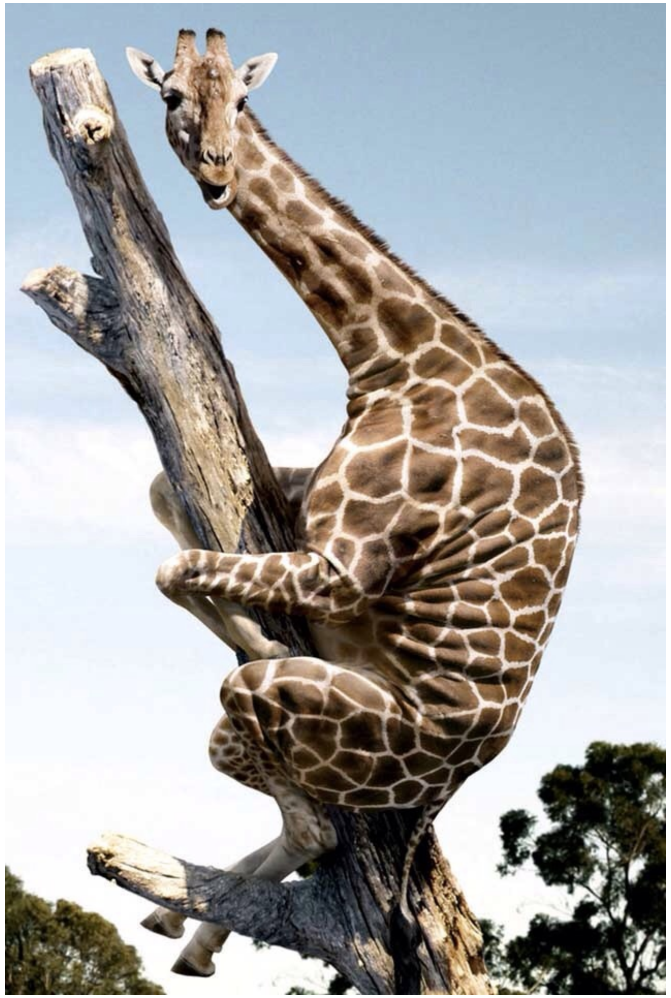
Stefanini et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. arXiv 2021

Sample of Existing Dataset Challenges

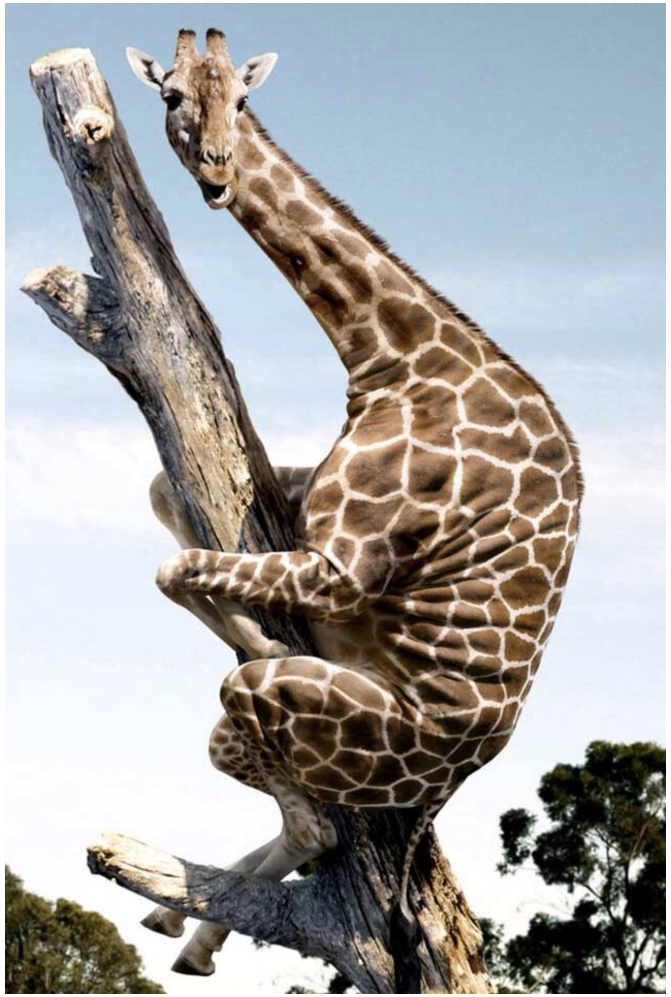
	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
COCO [128]	Generic	132K	5	27K (10K)	10.5
Flickr30K [129]	Generic	31K	5	18K (7K)	12.4
Flickr8K [19]	Generic	8K	5	8K (3K)	10.9
CC3M [130]	Generic	3.3M	1	48K (25K)	10.3
CC12M [131]	Generic	12.4M	1	523K (163K)	20.0
SBU Captions [4]	Generic	1M	1	238K (46K)	12.1
VizWiz [132]	Assistive	70K	5	20K (8K)	13.0
CUB-200 [133]	Birds	12K	10	6K (2K)	15.2
Oxford-102 [133]	Flowers	8K	10	5K (2K)	14.1
Fashion Cap. [134]	Fashion	130K	1	17K (16K)	21.0
BreakingNews [135]	News	115K	1	85K (10K)	28.1
GoodNews [136]	News	466K	1	192K (54K)	18.2
TextCaps [137]	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives [138]	Generic	849K	1/5	16K (7K)	41.8

Stefanini et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. arXiv 2021

Challenge: What Instructions Should Be Provided When Collecting Captions from Human Annotators?



Class Task: How Would You Describe This Image?



Form: https://forms.gle/Nbue5HcdP9Dib8Co8

VLT2K

Guidelines and Examples:

Read these guidelines carefully. You must write exactly two sentences.

1. Describe the action being performed and mention the person performing the action and all objects involved in the action.
2. Describe any objects in the image that are not directly involved in the action.

A man is reading a newspaper.
It is cloudy and there are
skyscrapers in the background.

A boy is typing on a laptop.
There is a brown bookshelf
behind him and a bright window.

A man is talking on the telephone.
There is a red lampshade and
three red chairs in the background.

D. Elliott and F. Keller. Image description using visual dependency representations. EMNLP 2013.

Flickr8K and 30K

Guidelines:

- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
Okay: This describes the dog, but it ignores the hat.

(4) The red hat is adorned with gold sequins.
Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
Bad: This is speculation.

(6) The dog.
Very Bad: This could describe any image of any dog.

Hodosh, Young, and Hockenmaier. Framing image description as a ranking task: Data, models, and evaluation metrics. JAIR 2013

MSCOCO

Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

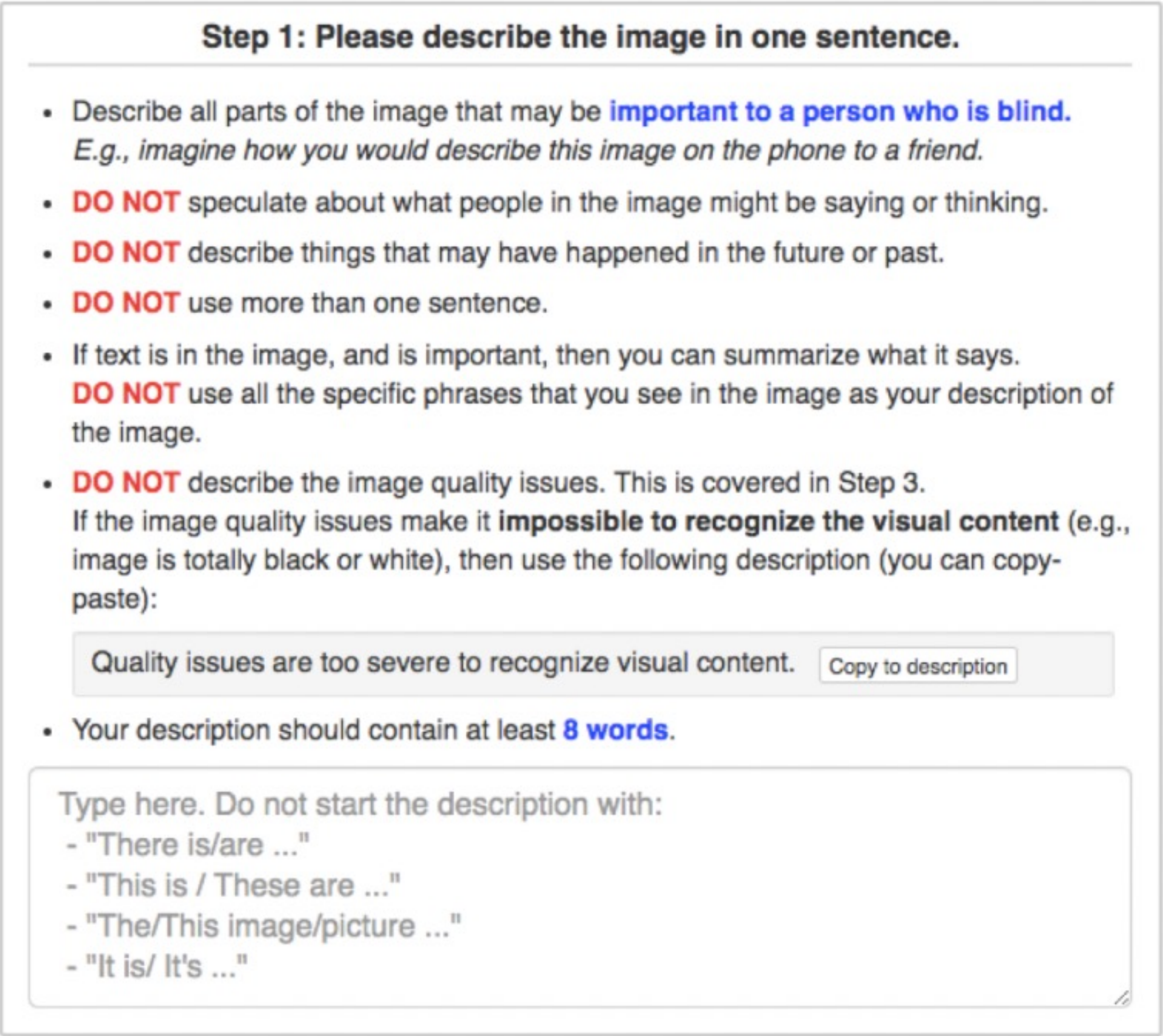
Please describe the image:

Enter description here

prev next

Chen et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

VizWiz



Step 1: Please describe the image in one sentence.

- Describe all parts of the image that may be **important to a person who is blind.** *E.g., imagine how you would describe this image on the phone to a friend.*
- **DO NOT** speculate about what people in the image might be saying or thinking.
- **DO NOT** describe things that may have happened in the future or past.
- **DO NOT** use more than one sentence.
- If text is in the image, and is important, then you can summarize what it says. **DO NOT** use all the specific phrases that you see in the image as your description of the image.
- **DO NOT** describe the image quality issues. This is covered in Step 3. If the image quality issues make it **impossible to recognize the visual content** (e.g., image is totally black or white), then use the following description (you can copy-paste):

Quality issues are too severe to recognize visual content. Copy to description

- Your description should contain at least **8 words**.

Type here. Do not start the description with:
- "There is/are ..."
- "This is / These are ..."
- "The/This image/picture ..."
- "It is/ It's ..."

Gurari et al. Captioning Images Taken by People Who Are Blind. ECCV 2020

Personality-Captions

215 personalities selected from this list: http://ideonomy.mit.edu/essays/traits.html

Comment on an Image

Description

In this task, you will be shown 5 images, and will write a comment about each image. The goal of this task is to write something about an image that someone else would find engaging.

STEP 1

With each new photo, you will be given a **personality trait** that you will try to emulate in your comment. For example, you might be given "**snarky**" or "**glamorous**". The personality describes ***YOU***, not the picture. It is *you* who is snarky or glamorous, not the contents of the image.

STEP 2

You will then be shown an image, for which you will write a comment *in the context of your given personality trait*. Please make sure your comment has at least **three words**. Note that these are *comments*, not captions.

E.g., you may be shown an image of a tree. If you are "**snarky**", you might write "What a boring tree, I bet it has bad wood;" or, if you were "**glamorous**", you might write "What an absolutely beautiful tree! I would put this in my living room it's so extravagent!"

Image

Your assigned personality is:

Adventurous

Reminder - please do not write anything that involves any level of discrimination, racism, sexism and offensive religious/politics comments, otherwise the submission will be rejected.

K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. CVPR 2019.

Today’s Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners

Group Discussion: How Would You Evaluate Captions from an Algorithm?

FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "giraffe", "animal", "mammal", "standing", "field", "top", "branch", "bird", "eating", "head", "grazing", "neck", "water", "large", "man", "grassy", "tall", "group", "dirt", "zoo"], "captions": [{ "text": "a giraffe standing in the dirt", "confidence": 0.982929349 }] }

Evaluation: Human Judgments

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic Description from Images: A Survey of Models, Datasets, and Measures. JAIR 2016.

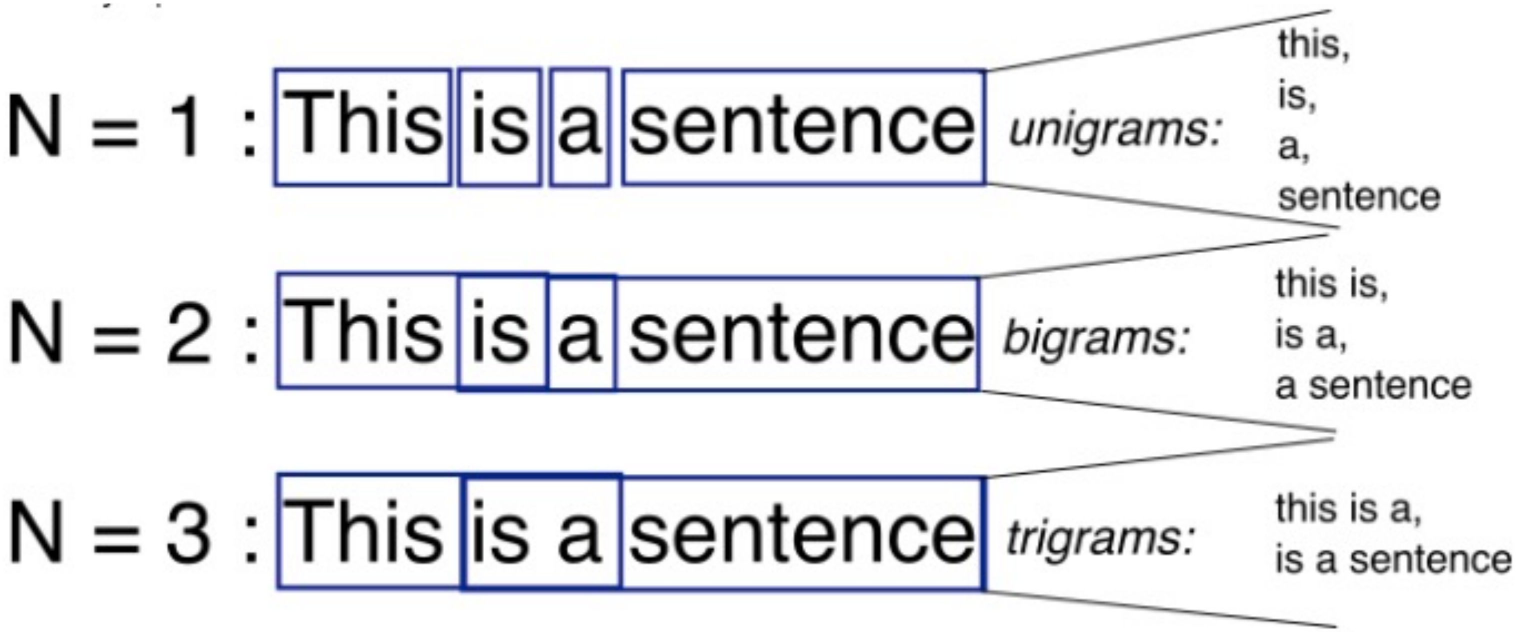
Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compute similarities of n-grams between a predicted caption and each ground truth caption



http://recognize-speech.com/language-model/n-gram-model/comparison

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- **CIDEr**
- SPICE

Idea: measure similarity of a predicted caption to how most people describe an image based on *n*-grams unique to the image

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- **CIDEr**
- SPICE

What content do most people describe in this image?

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very similar

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very different

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compare scene graph of prediction to scene graph of ground truth

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- **SPICE**

What is the meaningful semantic content in these captions?

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Meaningful semantic content in these captions:

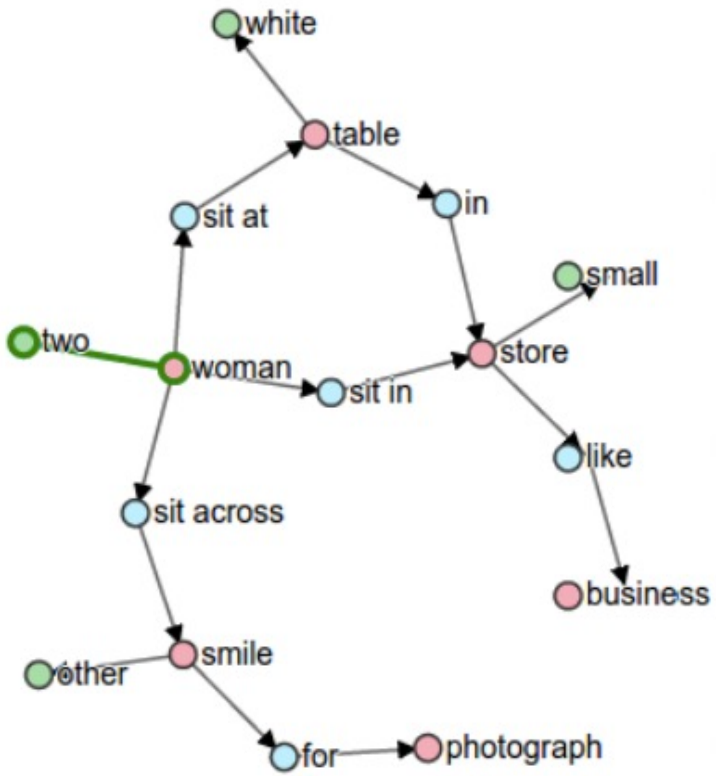
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

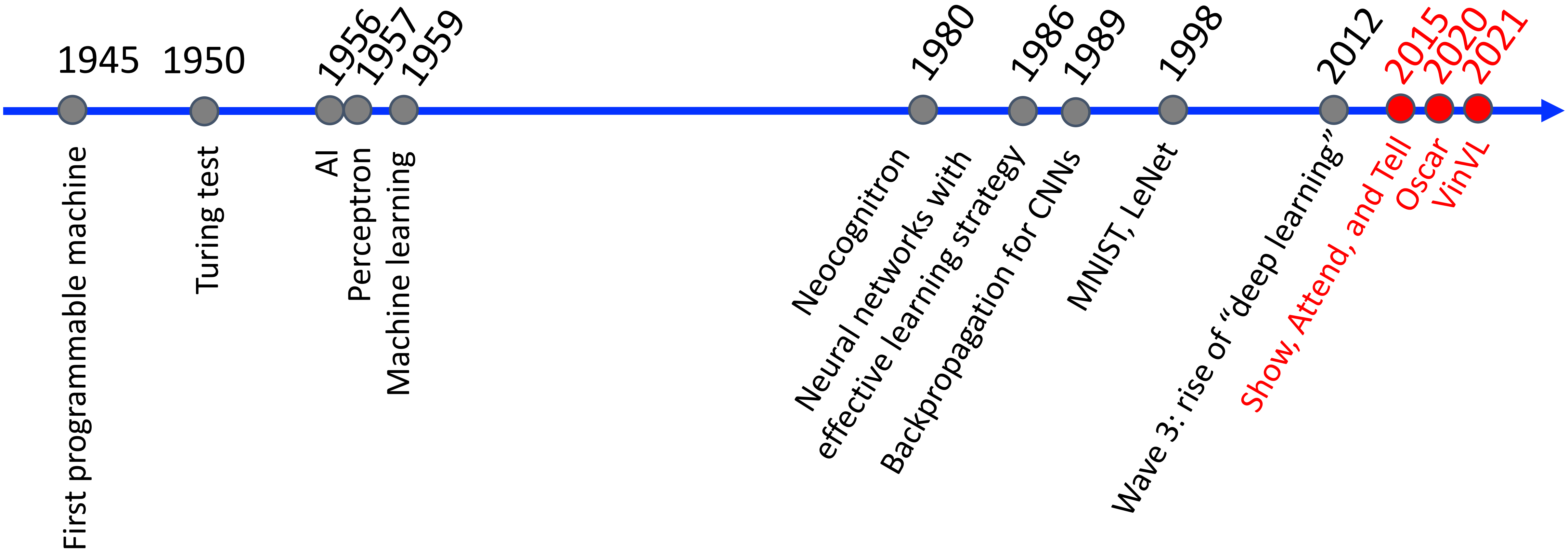


P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

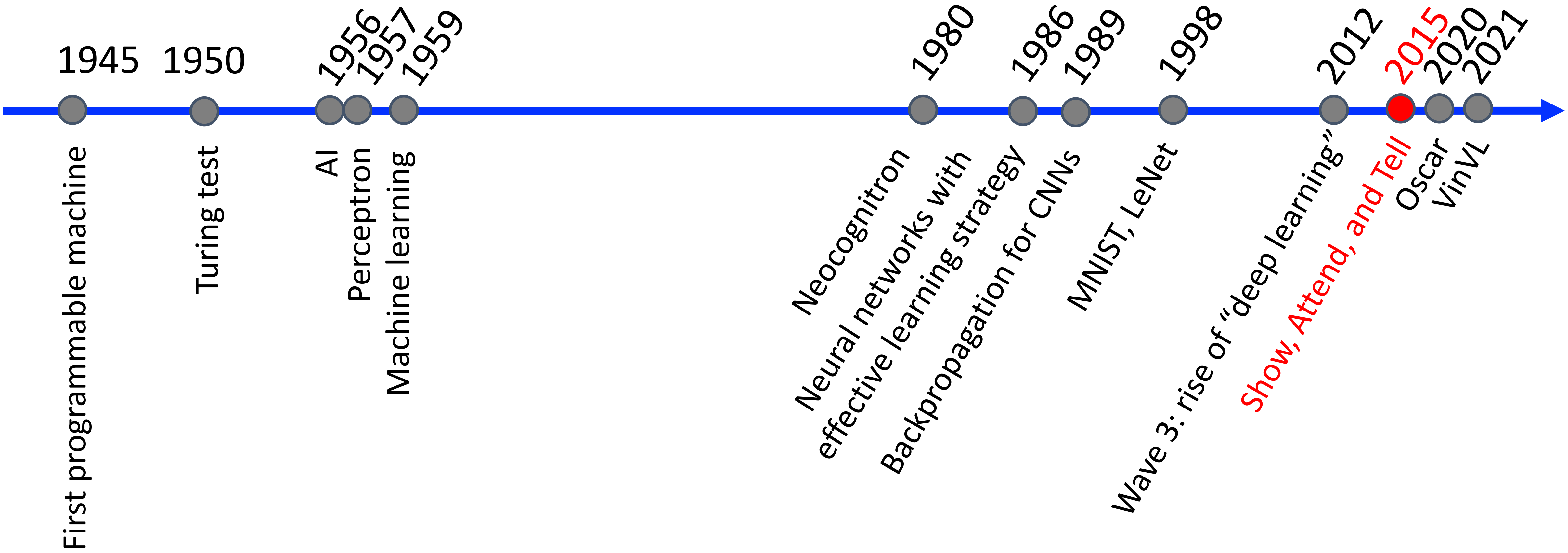
Today’s Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners

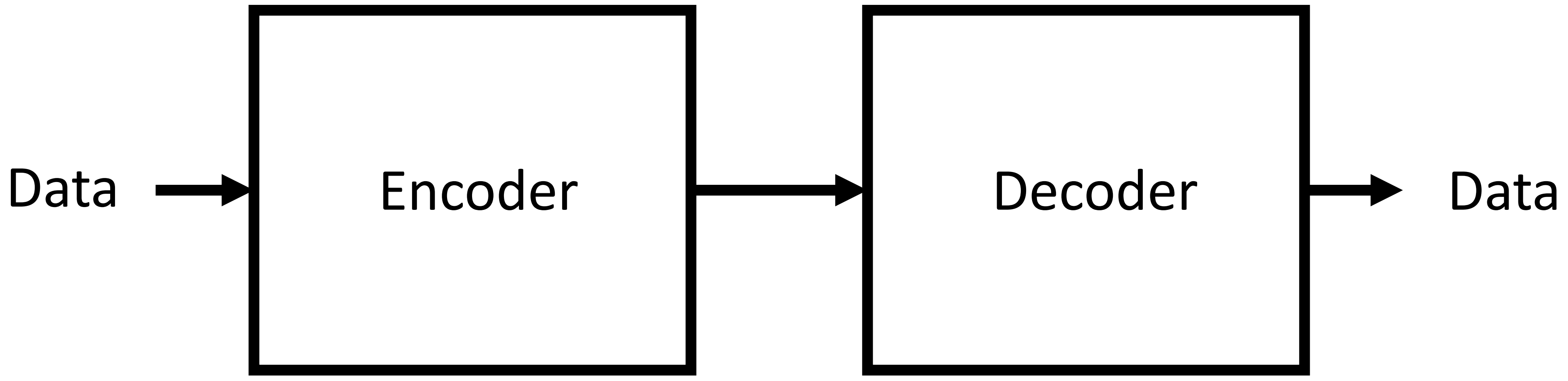
Historical Context



Historical Context



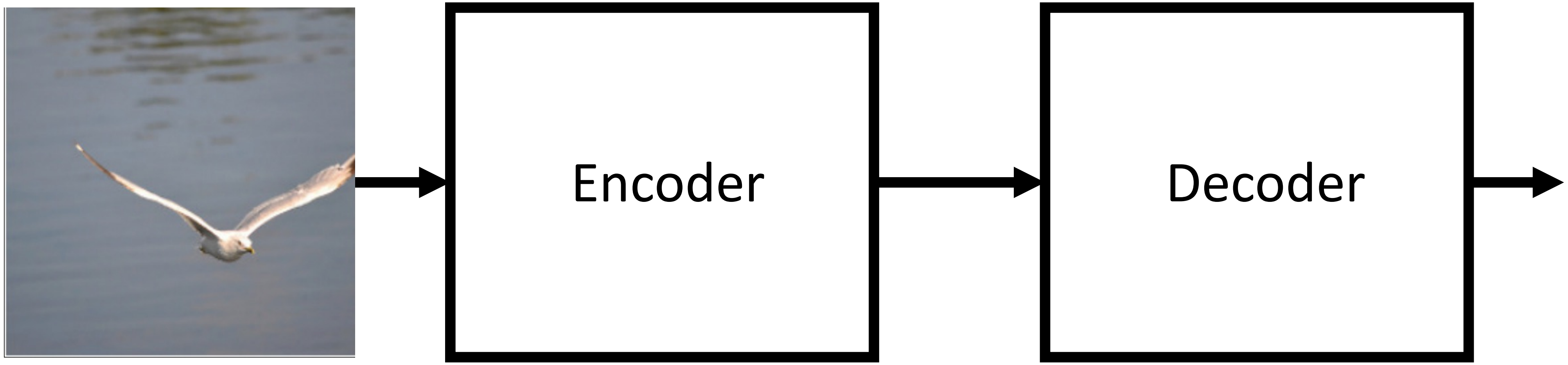
Idea: Treat Problem Like Machine Translation



Idea: Treat Problem Like Machine Translation

Image

Caption



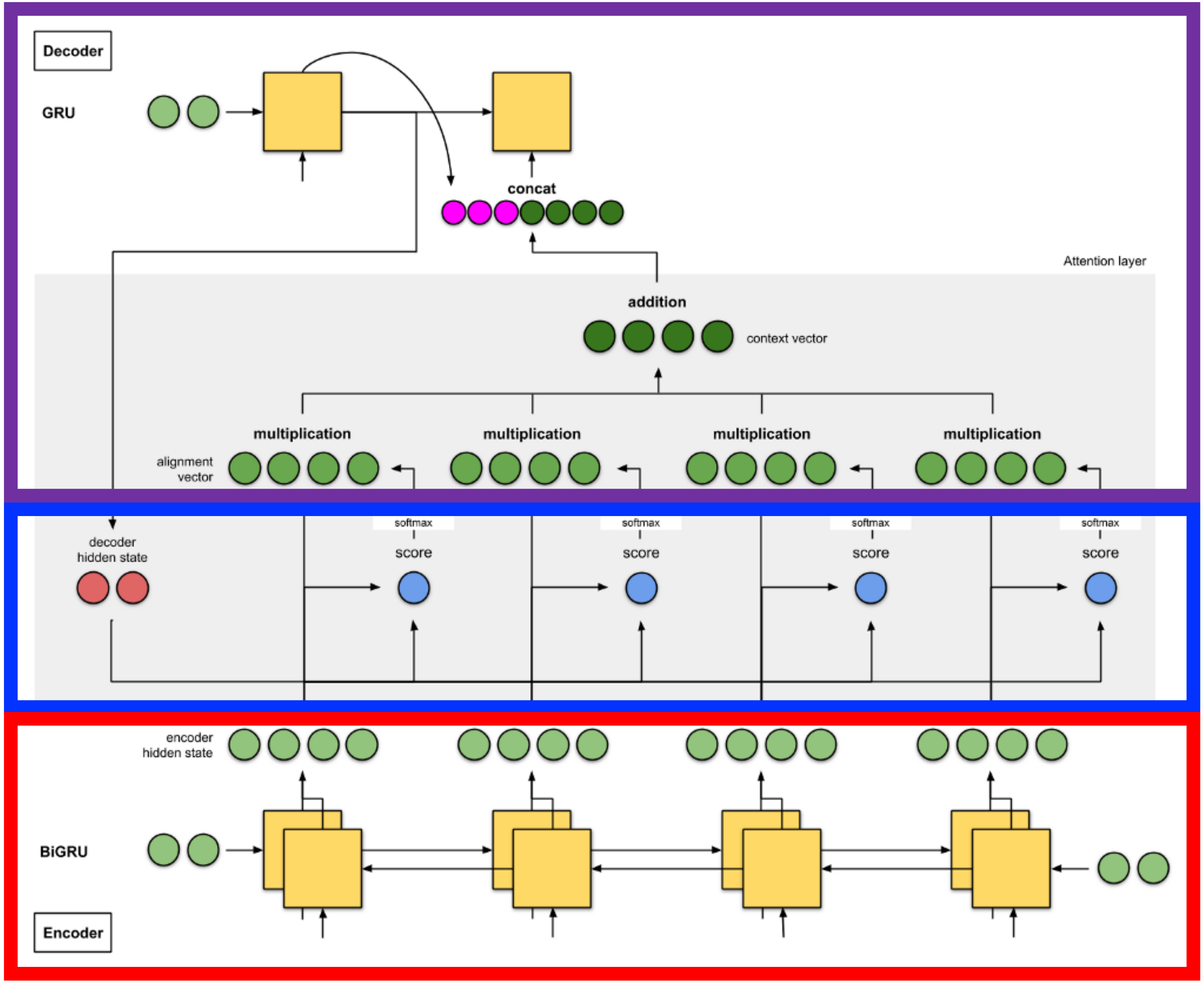
A bird flying over
a body of water.

Recall Solution:

3. At each decoder time step, a prediction is made based on the weighted sum of the inputs

2. At each decoder time step, attention weights are computed that determine each input's relevance for the prediction

1. Encoder produces hidden state for every input

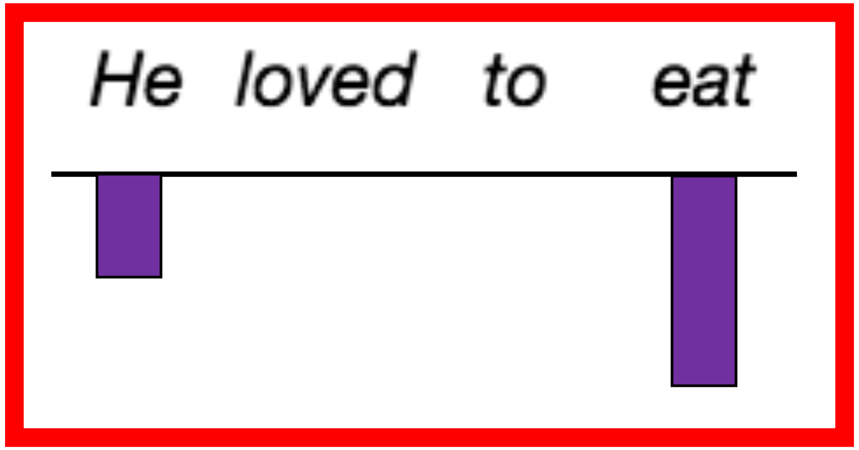


https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Recall Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., "soft attention" uses a weighted combination of the input

Input



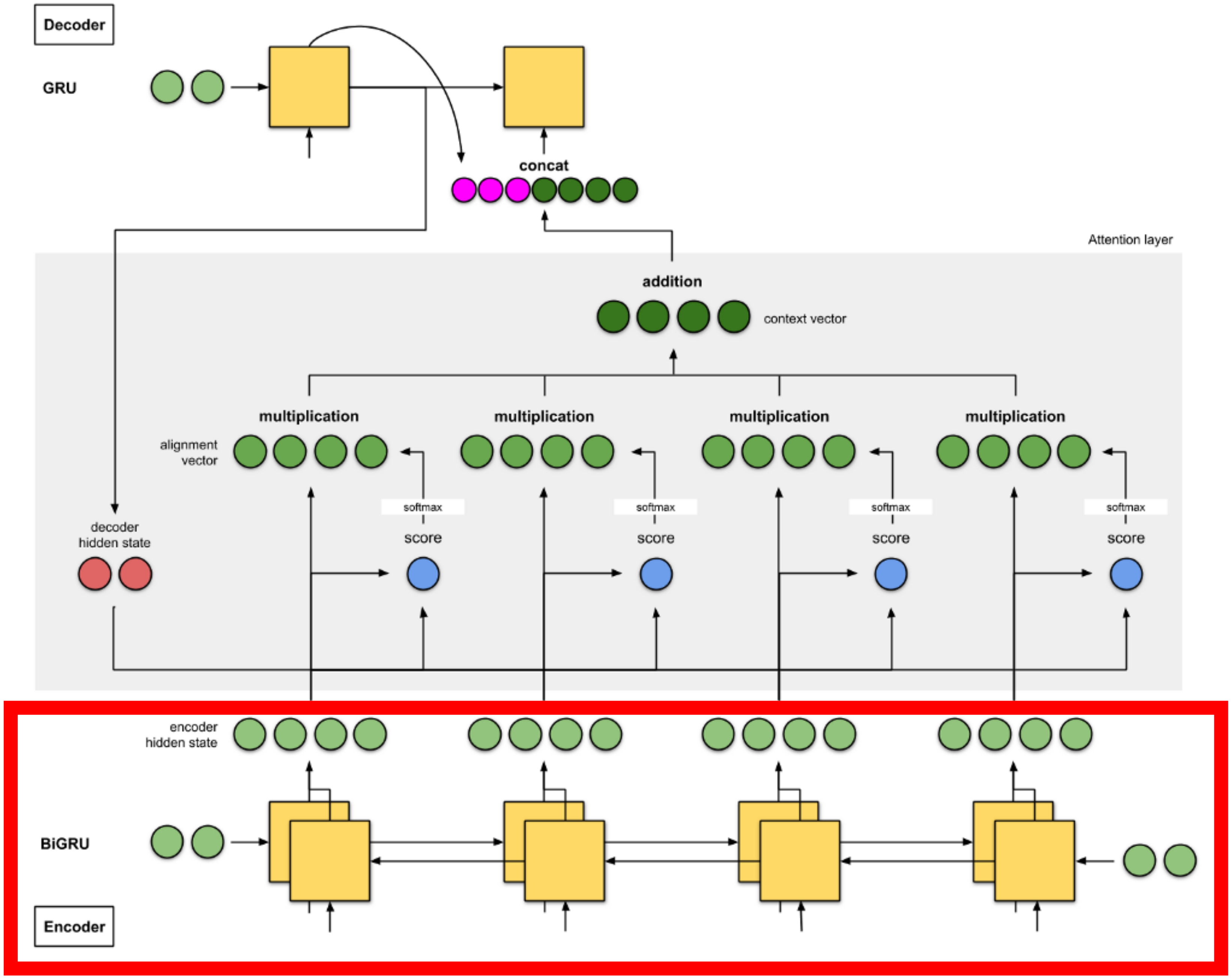
Model learns how to weight each input!

Target

Er liebte zu essen

t = 1 t = 2 t = 3 t = 4

Approach: Key Difference



1. Input represents an image

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3

Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input

Input

Target

A	bird	is	flying…
t = 1	t = 2	t = 3	t = 4

Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., "soft attention" uses a weighted combination of the input

Input

Target

A

t = 1

Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input

Input

Target

A bird

t = 1 t = 2

Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., "soft attention" uses a weighted combination of the input

Input

Target

A bird is

t = 1 t = 2 t = 3

Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input

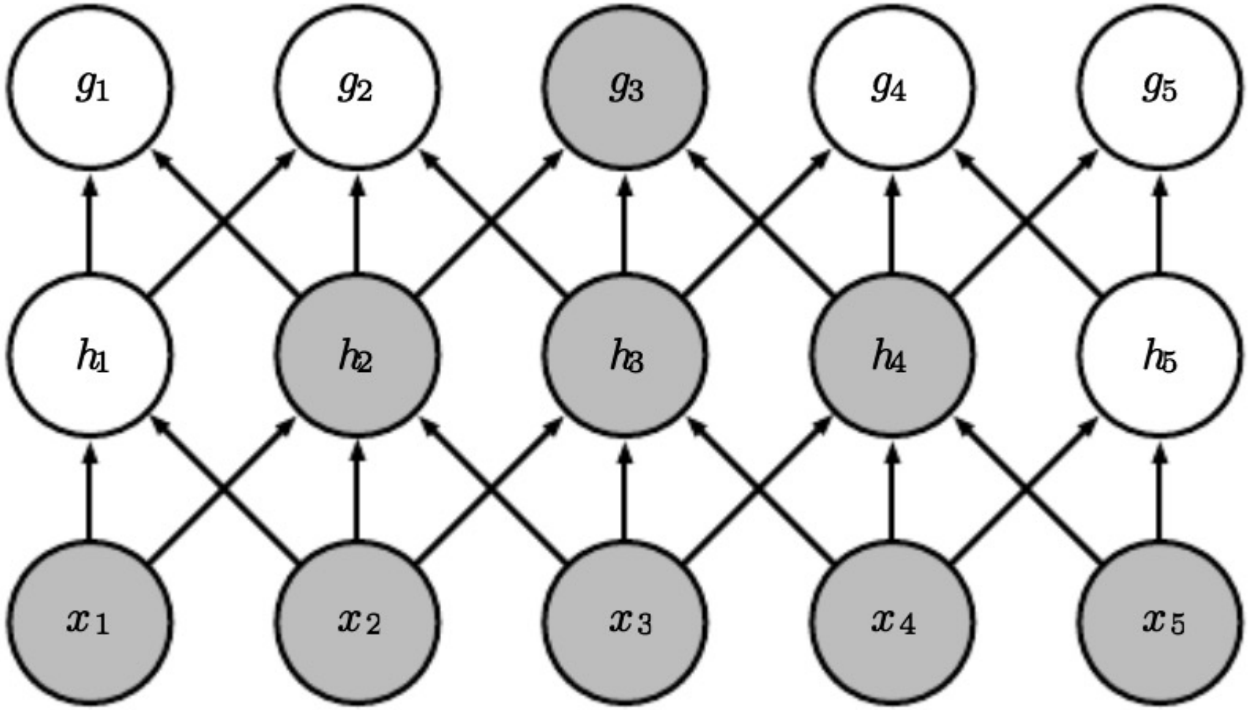
Input

Target

A	bird	is	flying	...
t = 1	t = 2	t = 3	t = 4	

Input Representation: Idea

Use convolutional layer that map to **regions of the input (e.g., pixel) space**; e.g., 1rst layer with *h* values

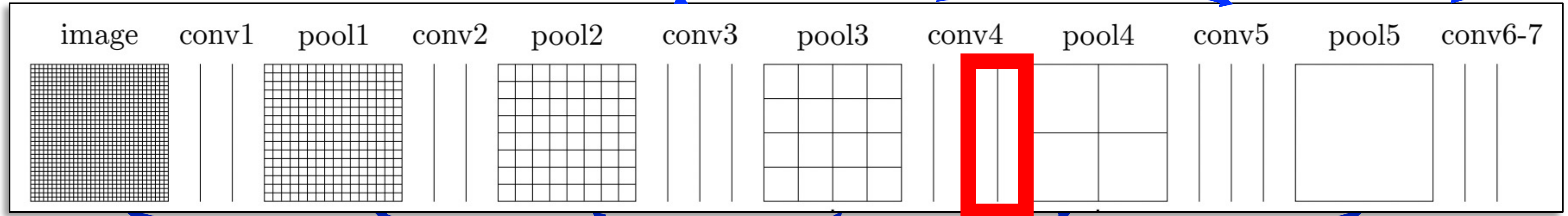


https://www.deeplearningbook.org/contents/convnets.html

Input Representation: Implementation

Each line represents a convolutional layer

VGG16:



Grids reflect relative spatial coarseness at each layer

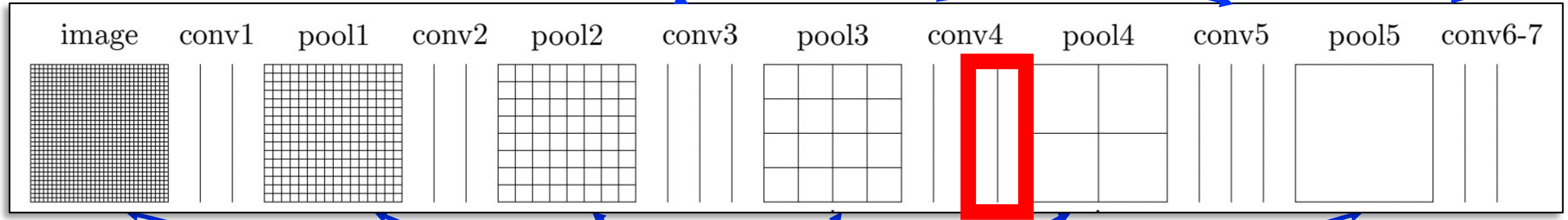
Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Input Representation: Implementation

Each line represents a convolutional layer

The size of each attended to region is therefore the size of the original image observed at this layer

VGG16:



Grids reflect relative spatial coarseness at each layer

Long, Shelhamer, and Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

Experimental Results

State-of-the-art performance on three dataset challenges

(Flickr8k, Flicker30k, and MS COCO)

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

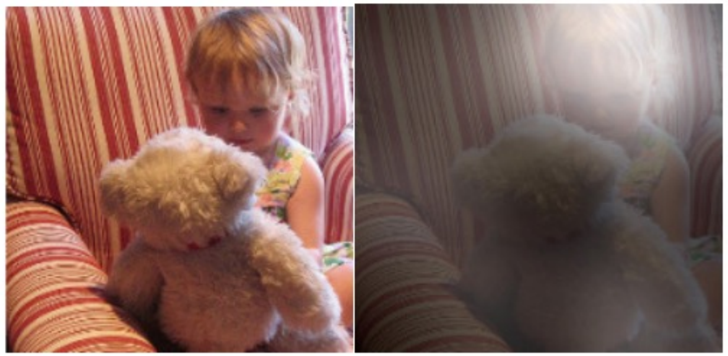
Experimental Results: Visualizations

Examples where correct content was attended to when predicting the word:

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.

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Experimental Results: Visualizations

Examples where incorrect content was attended to when predicting the word:

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.

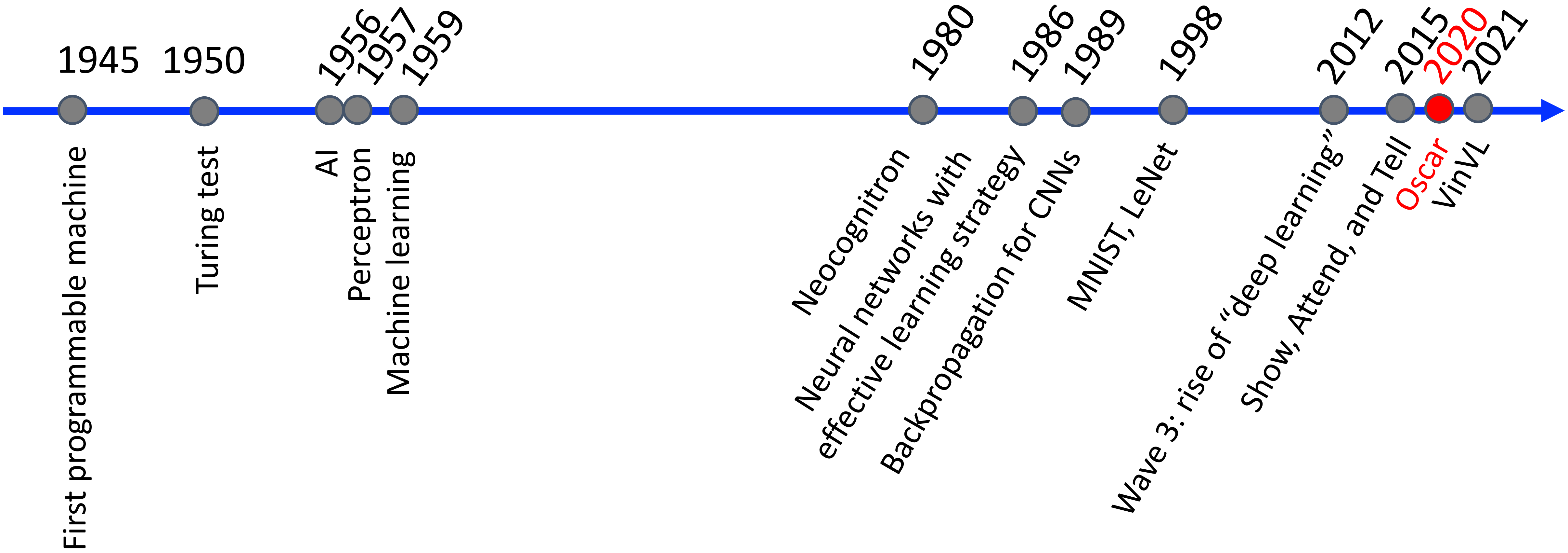
A person is standing on a beach
with a surfboard.

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

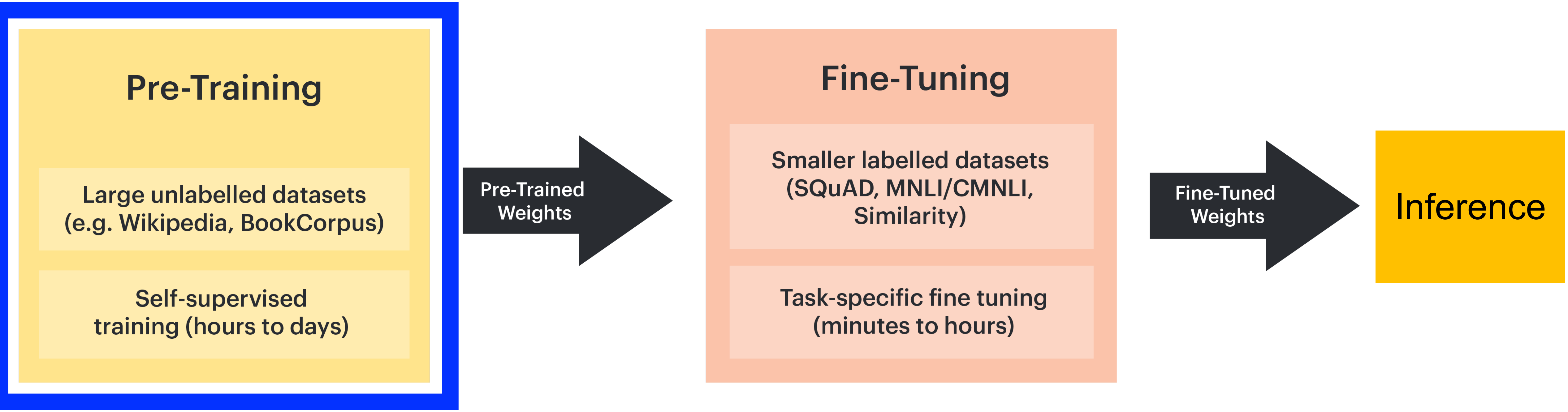
A man is talking on his cell phone
while another man watches.

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

Historical Context



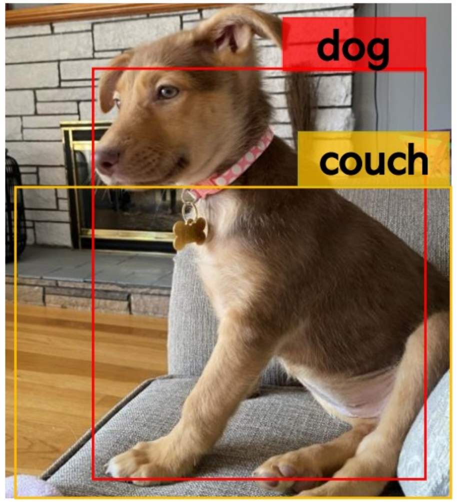
Oscar: Transformer Design



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Novelty: Adds **Explicit** Alignment Between Visual and Textual Concepts

- **Idea**: rather than have algorithm learn alignment between text and features describing image regions, align them explicitly
- **Motivating observations**: often, salient objects are mentioned in image descriptions and can be located by object detection algorithms

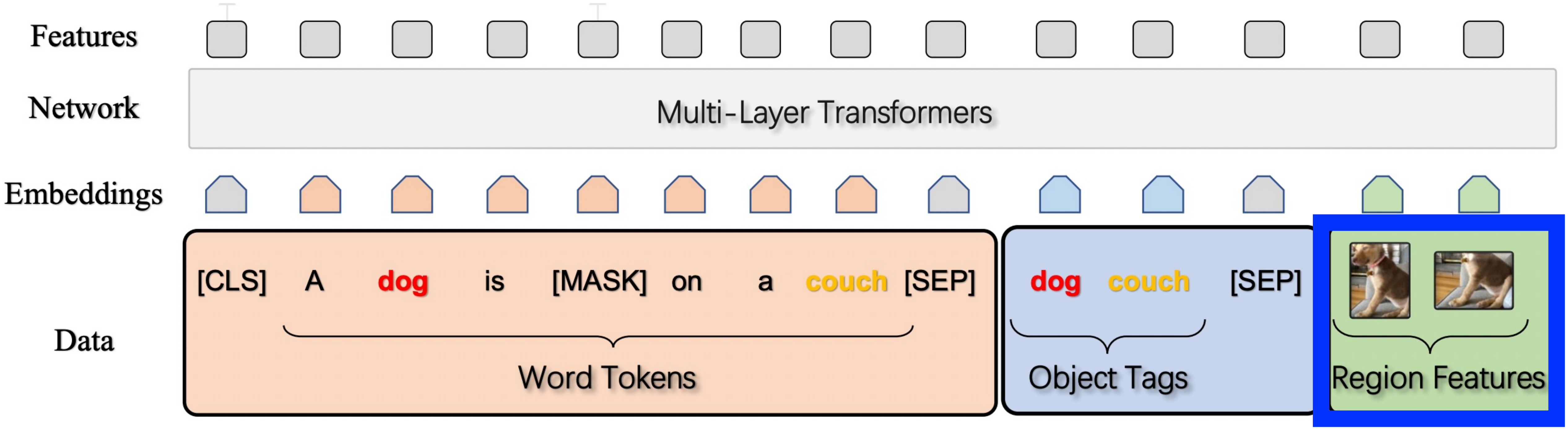


A **dog** is sitting on a **couch**

VS

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

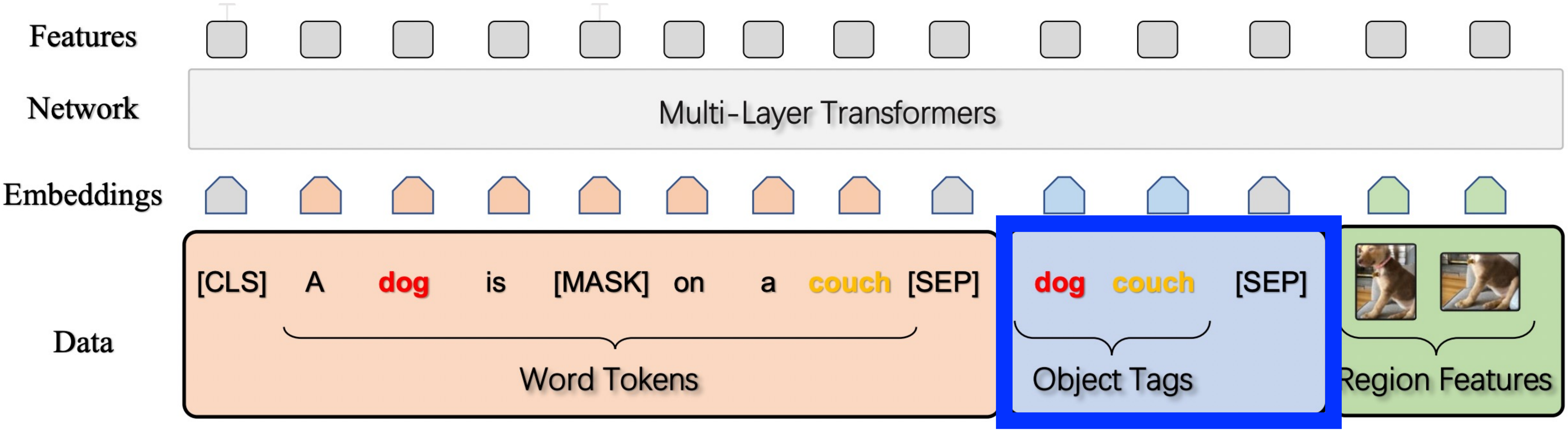
Oscar: Architecture



Like LXMERT, each image is represented as a description of objects detected with Faster R-CNN using features from Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: Architecture

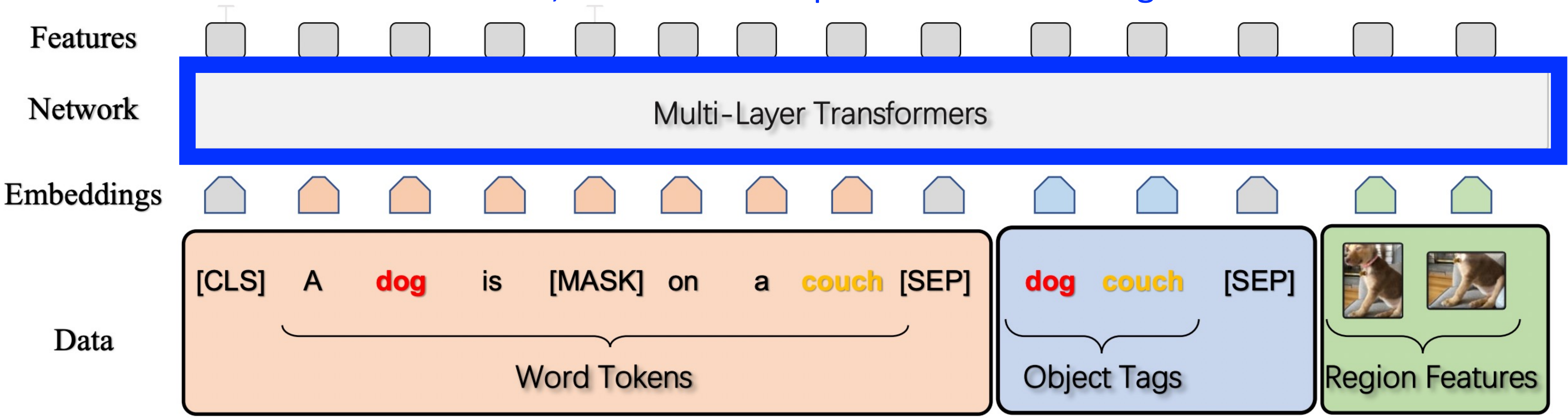


Novelty is to incorporate tags predicted by Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

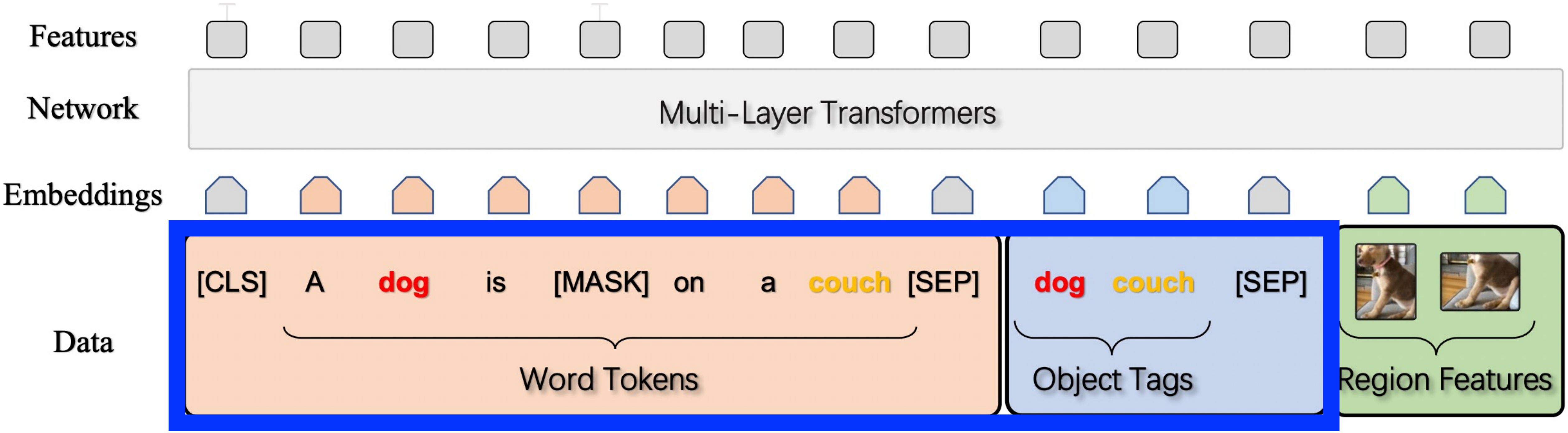
Oscar: Architecture

Uses BERT architecture, initialized with pretrained BERT weights



Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

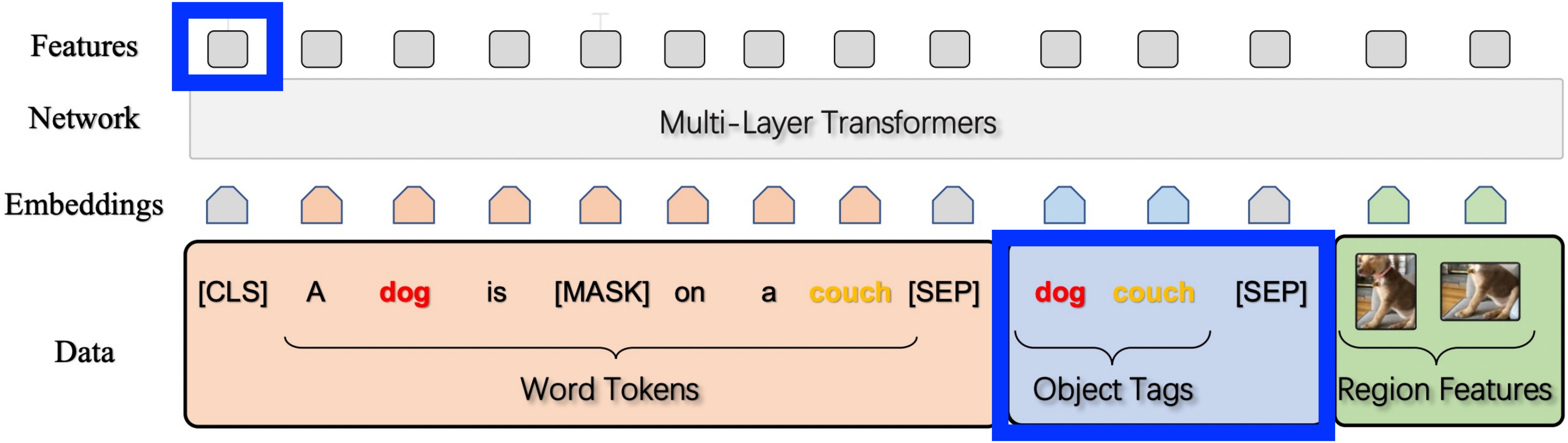


Like BERT, predict randomly masked tokens based on surrounding words, tags, and image information

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

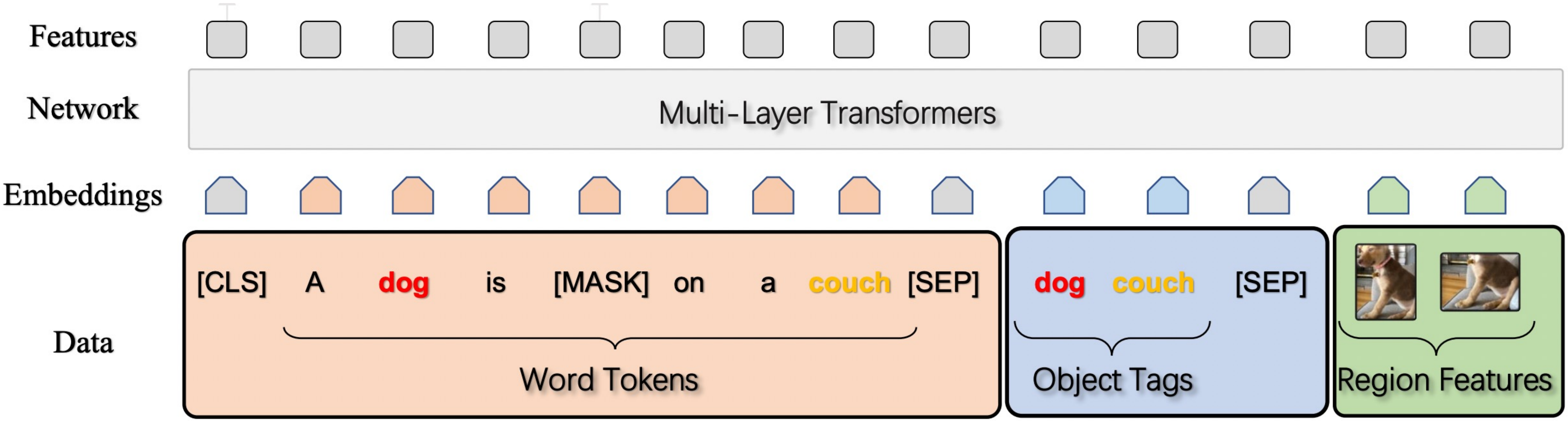
Fully-connected layer added to enable binary classification based on the fused vision-language token representation



Goal is to determine whether tags are original when 50% of tags are replaced with randomly selected tag sequence in the dataset

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

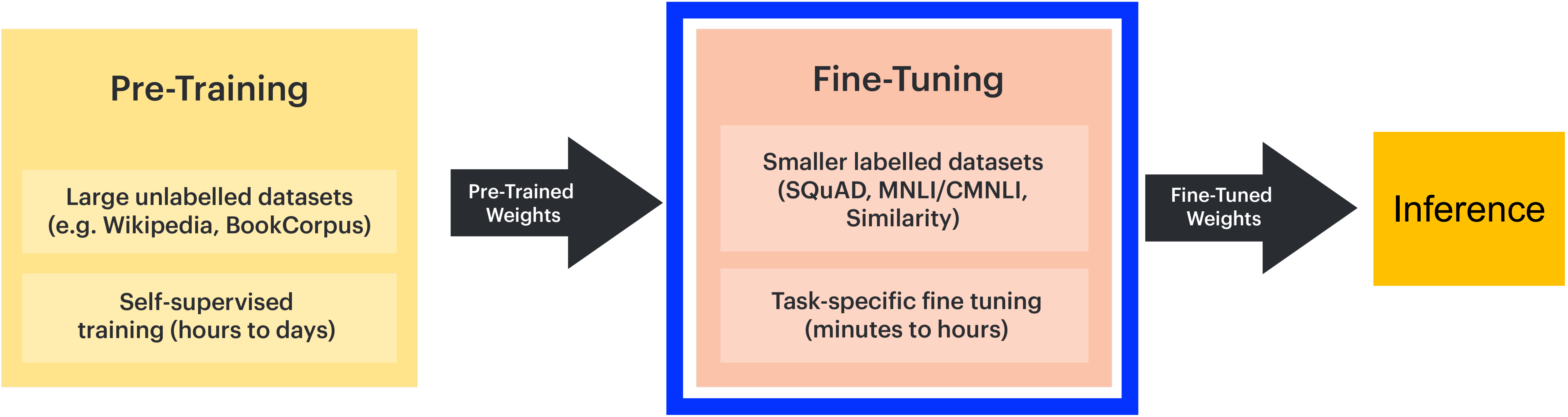
Oscar: 2 Pretraining Dataset



6.5 million text-tag-image triplets derived from existing V+L datasets

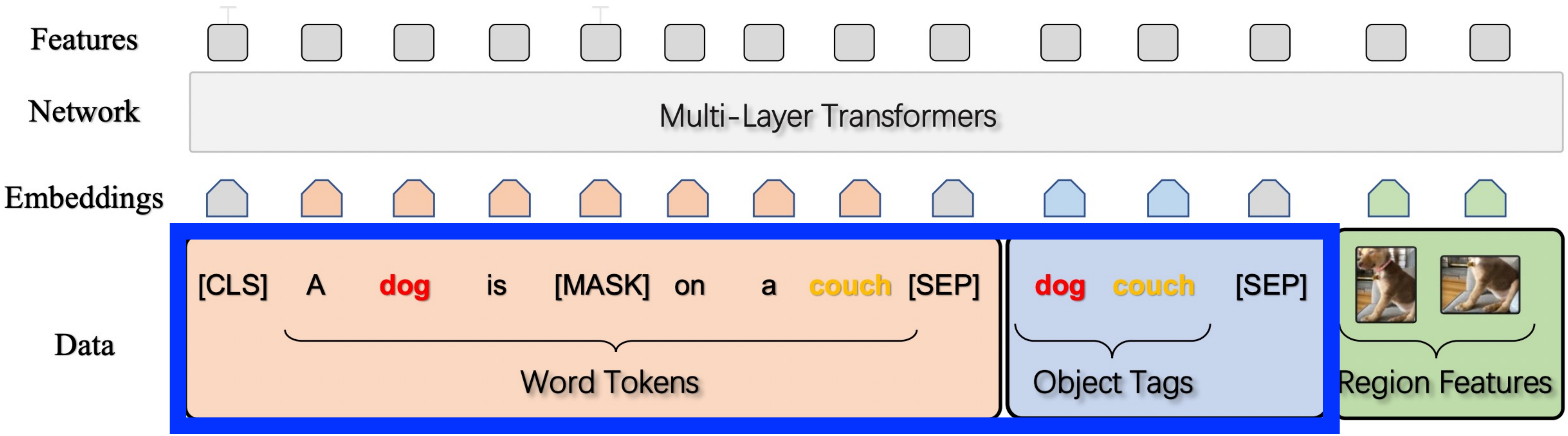
Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: Transformer Design



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

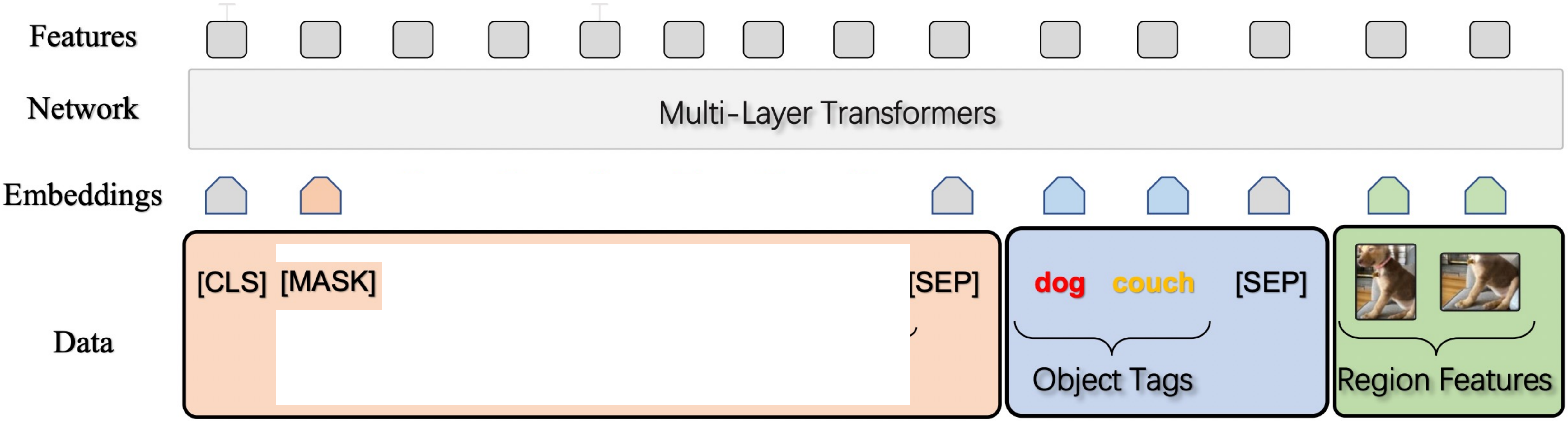
Oscar: 2 Fine-Tuning Task (Masked Token Loss)



Similar to pre-training, predict randomly masked tokens based on surrounding words, tags, and image information (on COCO dataset)

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

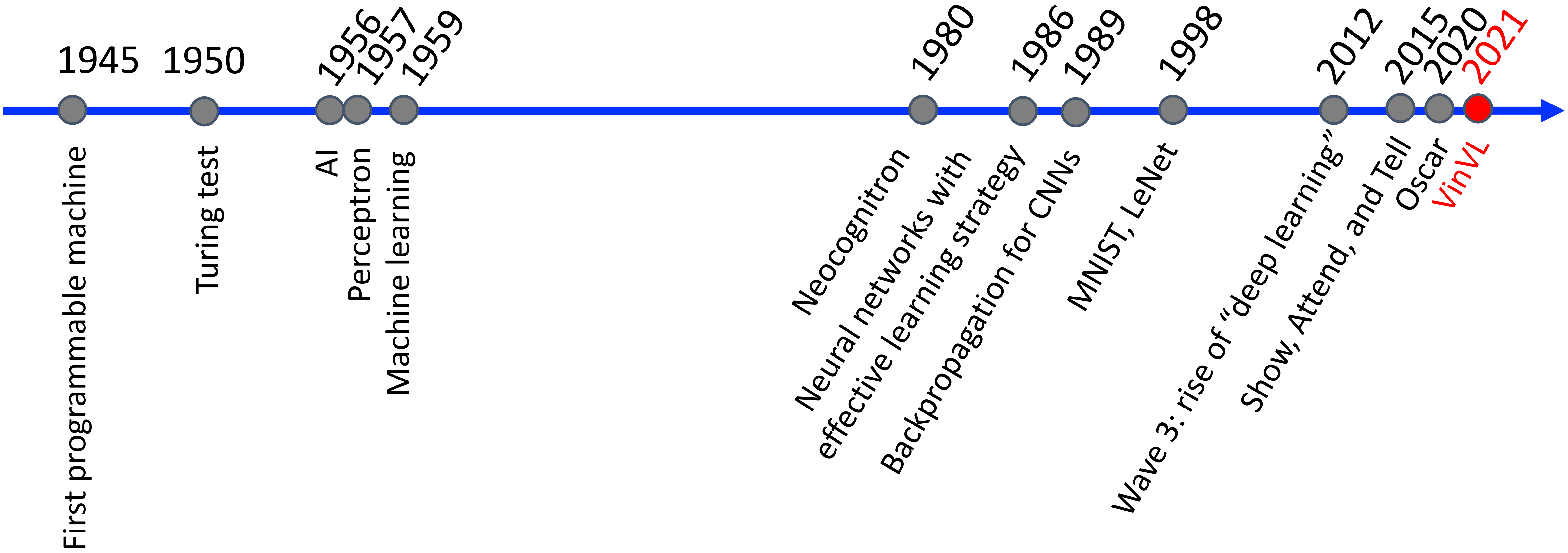
Oscar: Inference Time



Repeatedly predict a new [MASK] token, incorporating the predicted word into the sequence, until [STOP] is predicted.

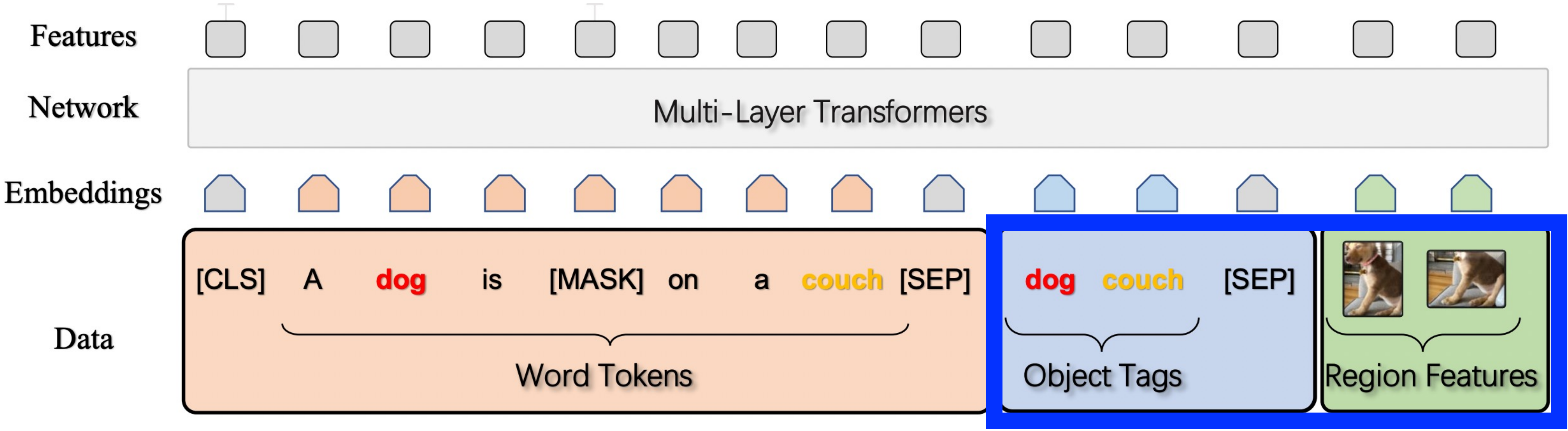
Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Historical Context



Idea: Oscar + Improved Visual Representation

VinVL Architecture: Oscar + New Object Detector

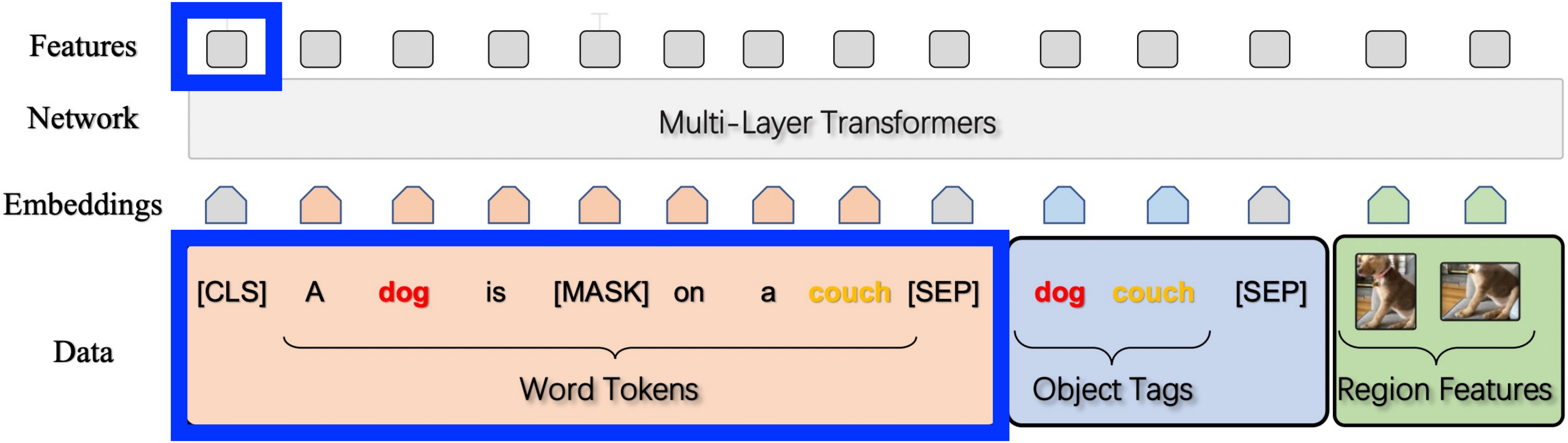


Improved object detector to predict more diverse categories and train larger models on larger datasets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

VinVL: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

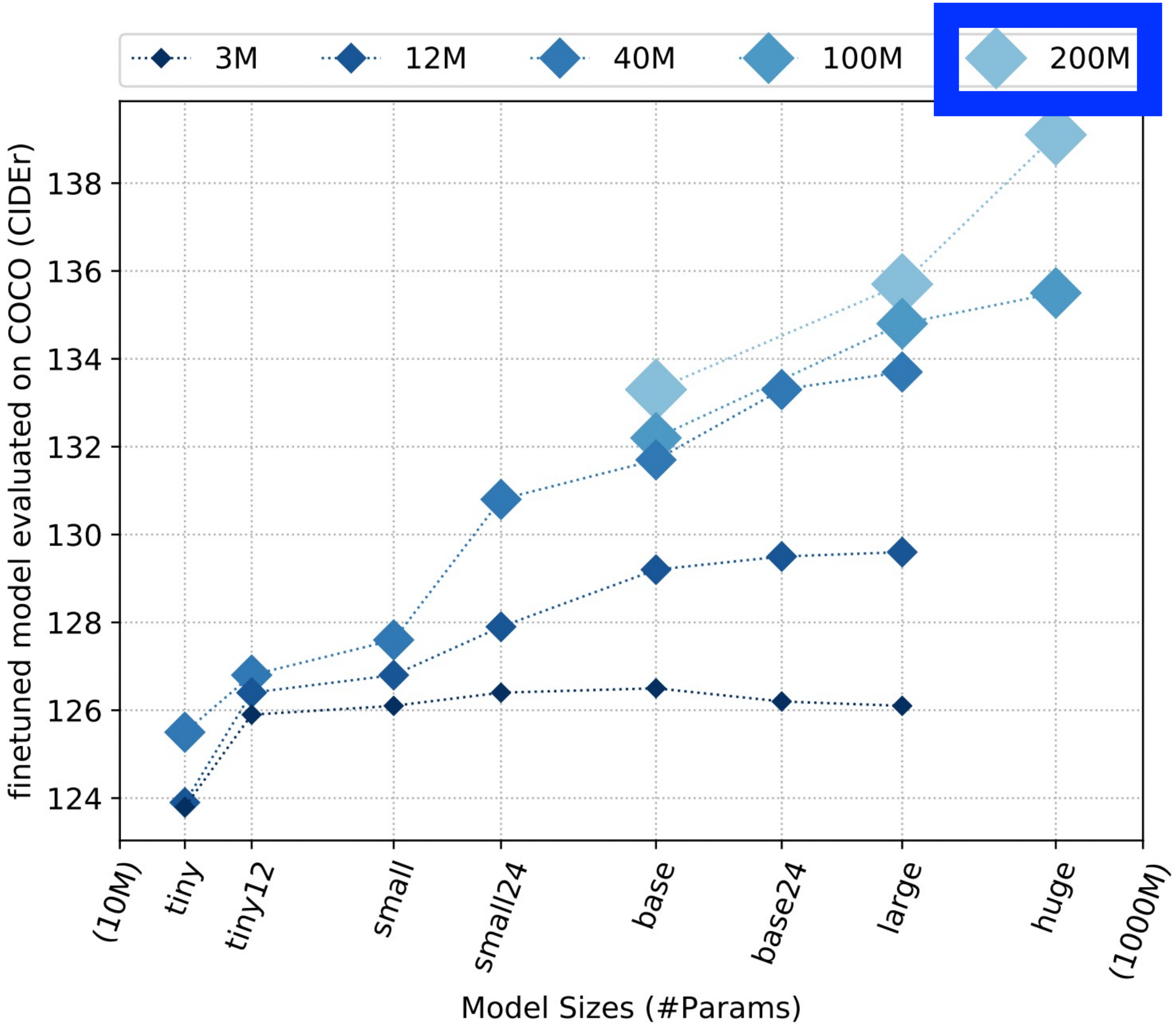
Fully-connected layer added to enable 3-way classification based on the fused vision-language token representation



Trained on 8.85 million text-image pairs to decide whether either captions or answers are corrupted (50% are not) for caption-tags-image triplets and question-answer-image triplets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

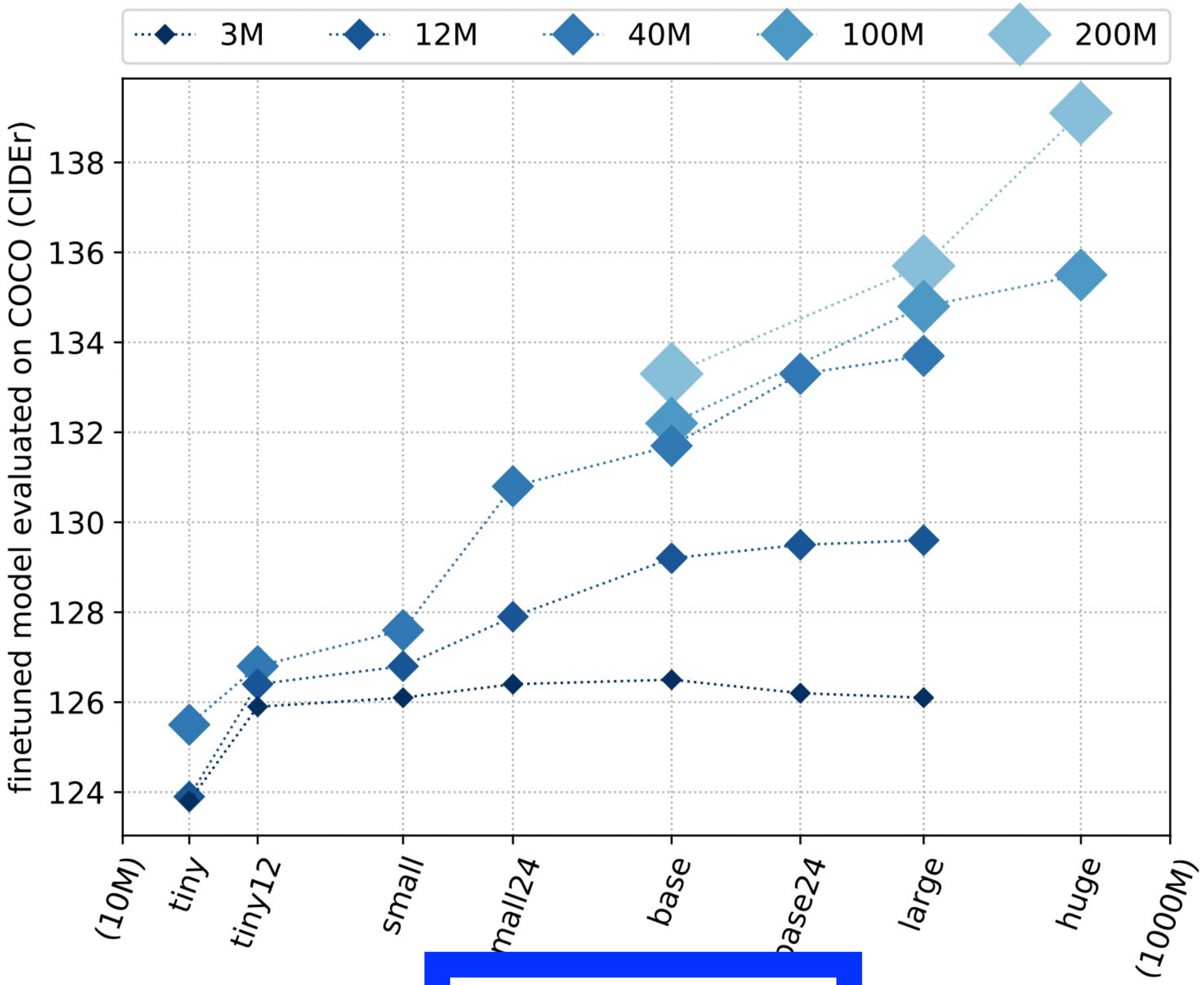
VinVL: Influence of Model and Dataset Sizes



200M images, each with 1 alt text description, collected from Internet

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

VinVL: Influence of Model and Dataset Sizes



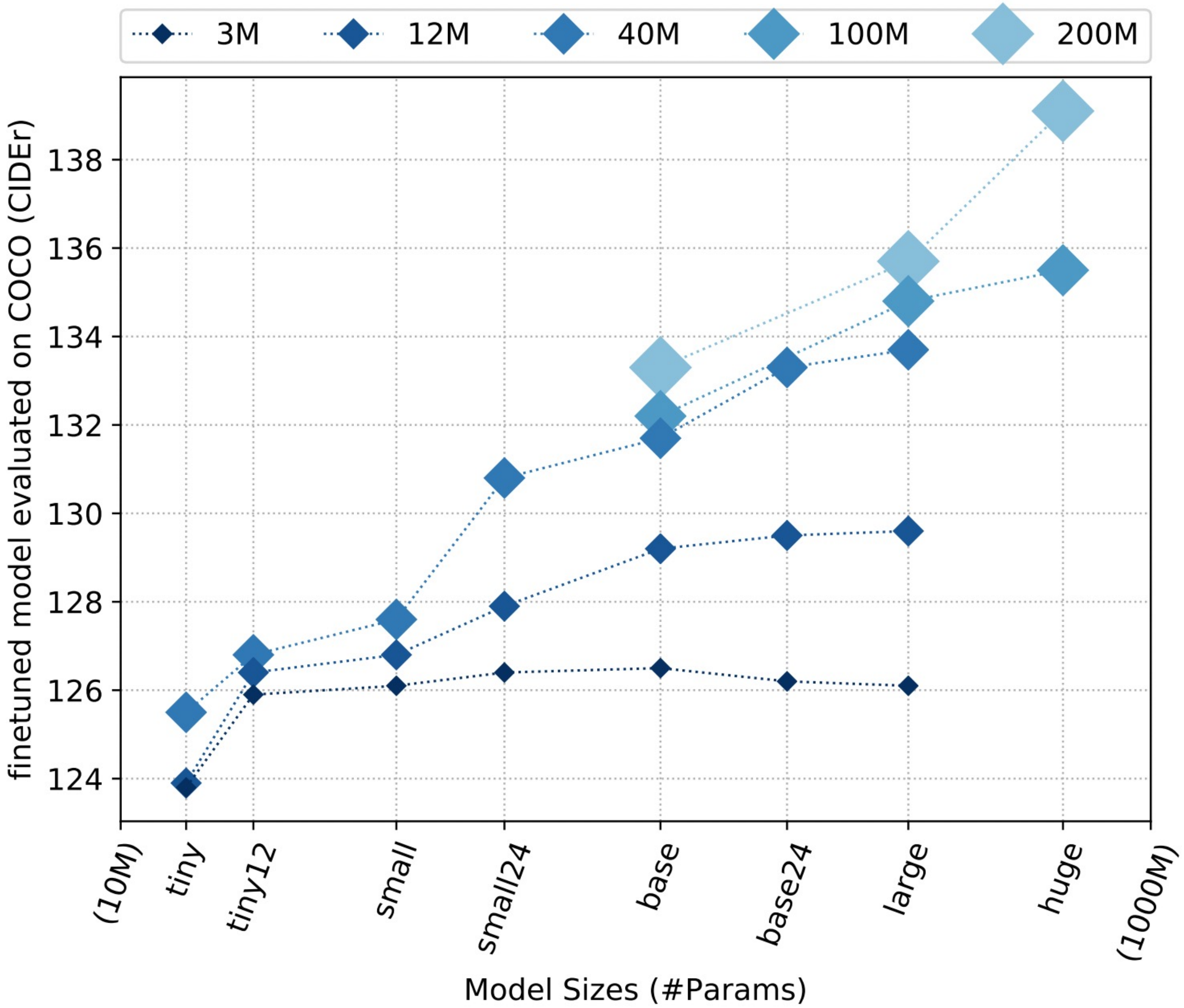
Model Sizes (#Params)

8 model sizes tested on COCO dataset

Model	Layers	Width	MLP	Heads	Param (M)
tiny	6	256	1024	4	13.4
tiny12	12	256	1024	4	18.1
small	12	384	1536	6	34.3
small24	24	384	1536	6	55.6
base	12	768	3072	12	111.7
base24	24	768	3072	12	196.7
large	24	1024	4096	16	338.3
huge	32	1280	5120	16	675.4

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

VinVL: Influence of Model and Dataset Sizes

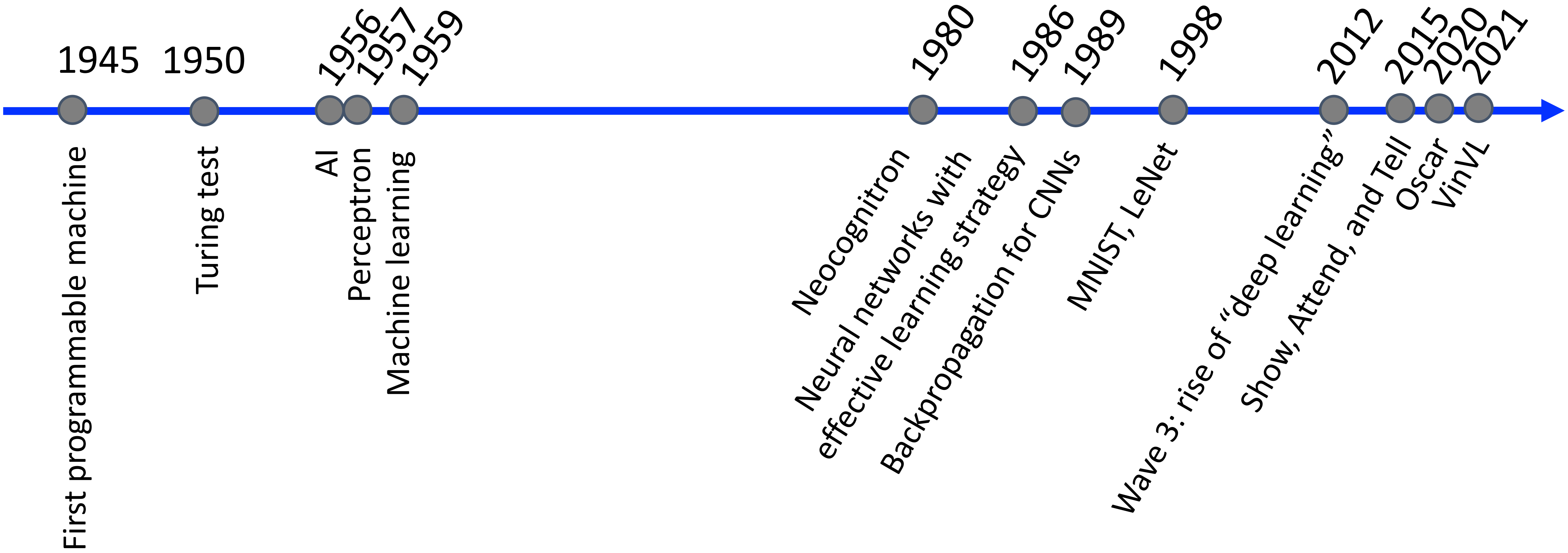


What trend(s) do you observe?

The trends of improved performance for large models and training datasets is generally observed for transformers

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Historical Context



Today’s Topics

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winners


The End