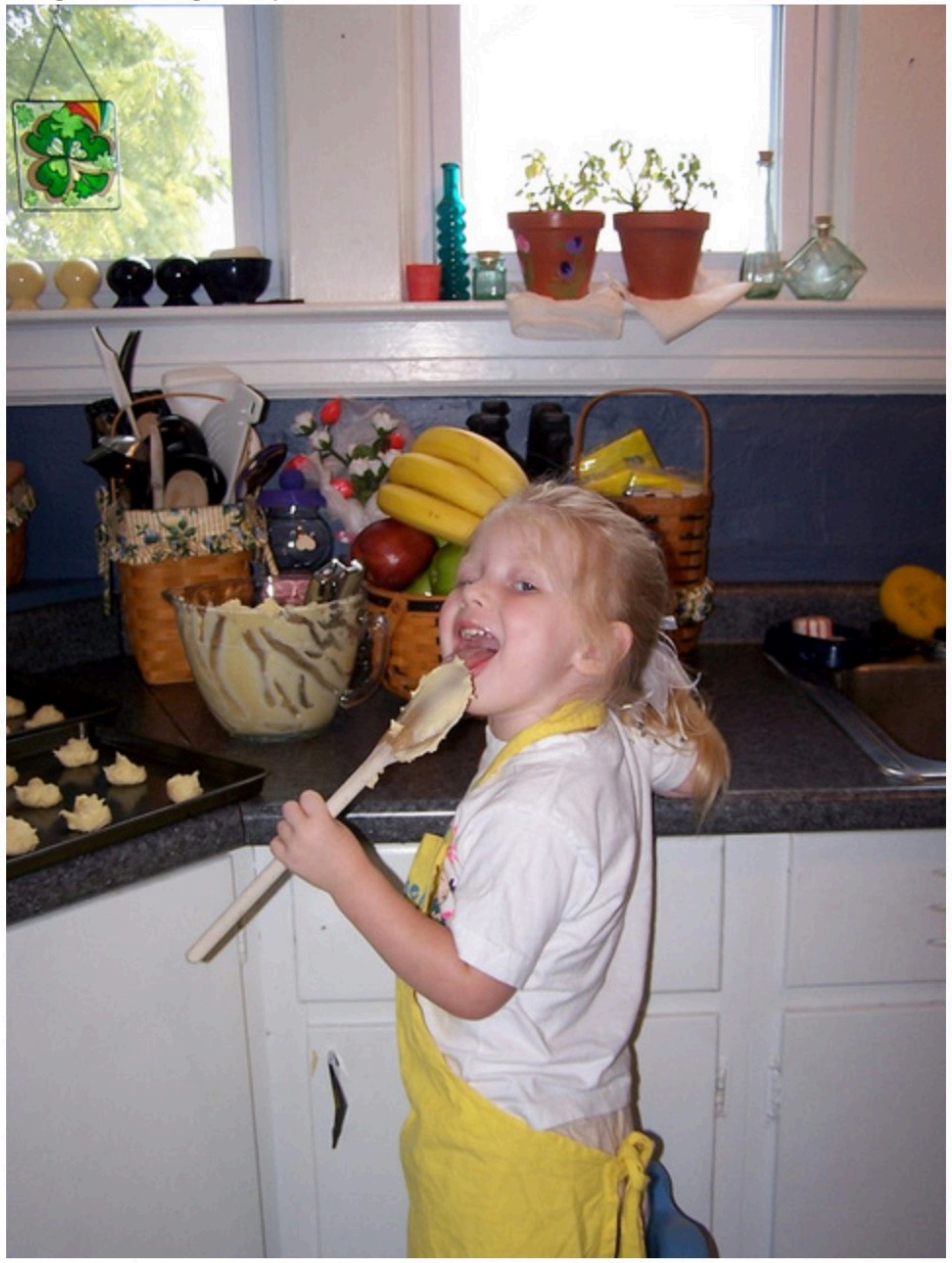
Grounding in Images

- How would you describe this image?
- What does the word "*spoon*" evoke?

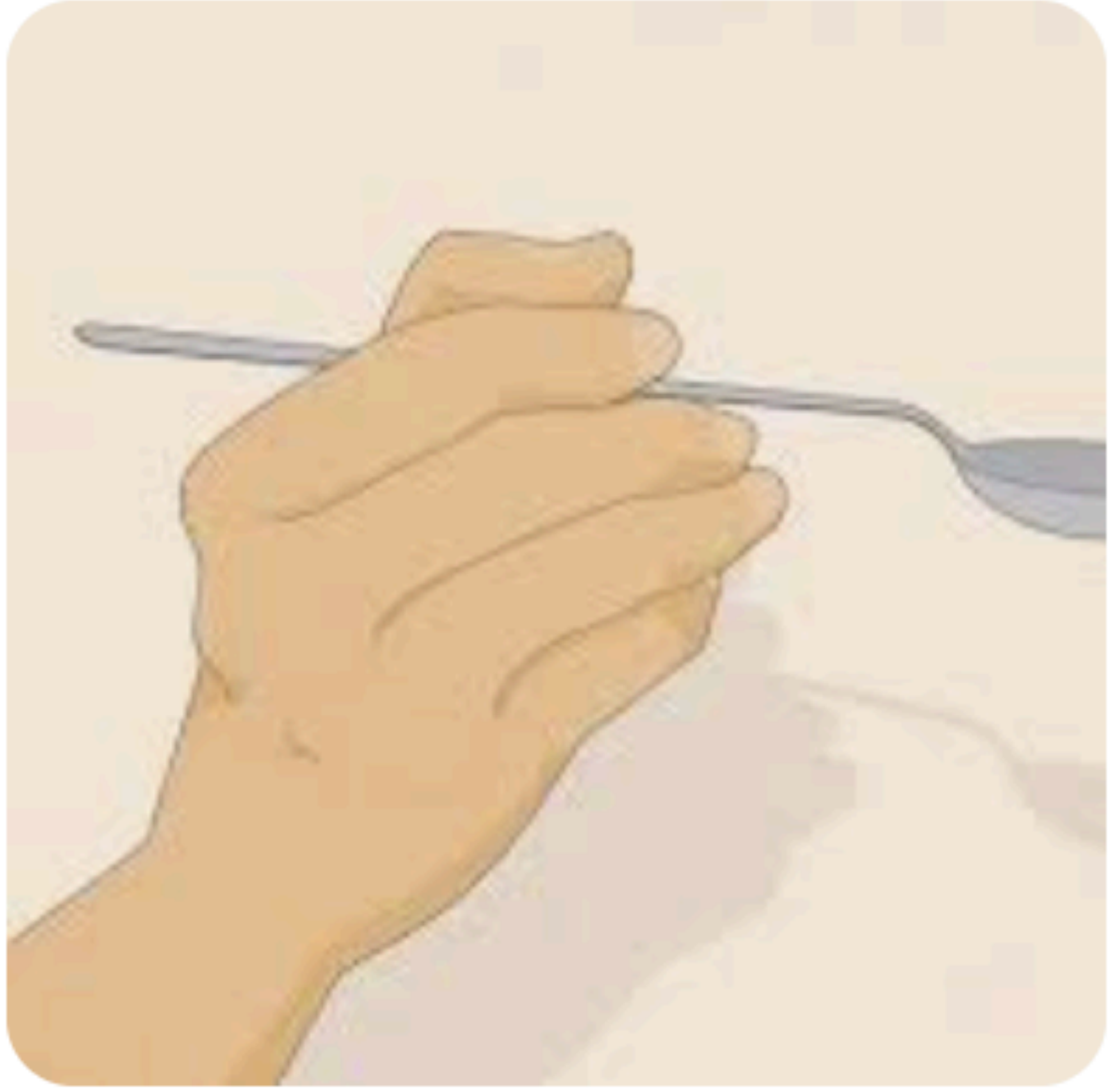


the girl is licking the spoon of batter

Grounding Spoon

Winco 0005-03 7 3/8" Dinner Spoon...

$7.16



wikiHow

How to Hold a Spoon: 13 Steps (...

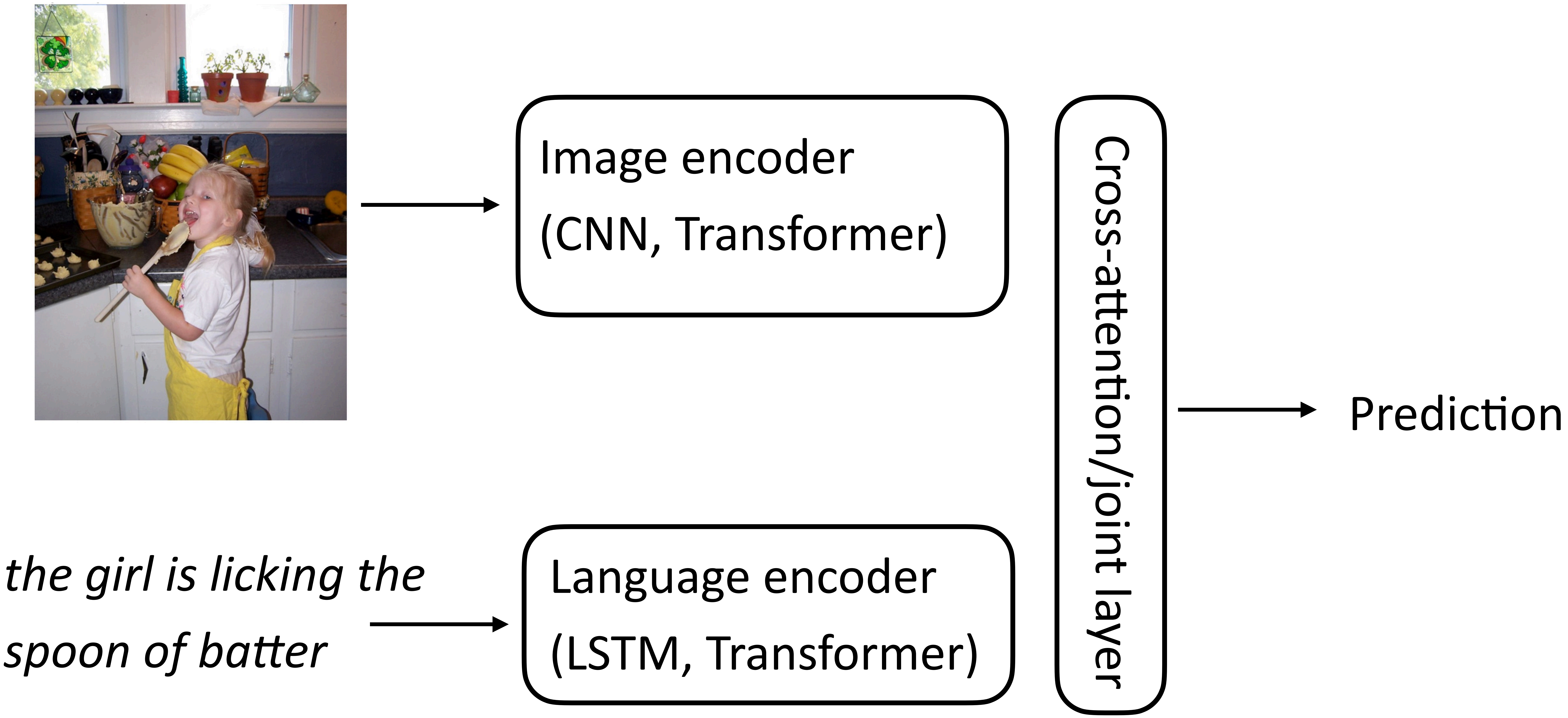
GO Indiegogo

Spoon that Elevates Taste ...

Grounding Language in Images

- Syntactic categories have some regular correspondences to the world:
  - Nouns: objects
  - Verbs: actions
  - Sentences: whole scenes or things happening
- Tasks:
  - Object recognition (pick out one most salient object or detect all of them)
  - Image captioning: produce a whole sentence for an image

Language-vision Models



the girl is licking the spoon of batter

Image encoder

(CNN, Transformer)

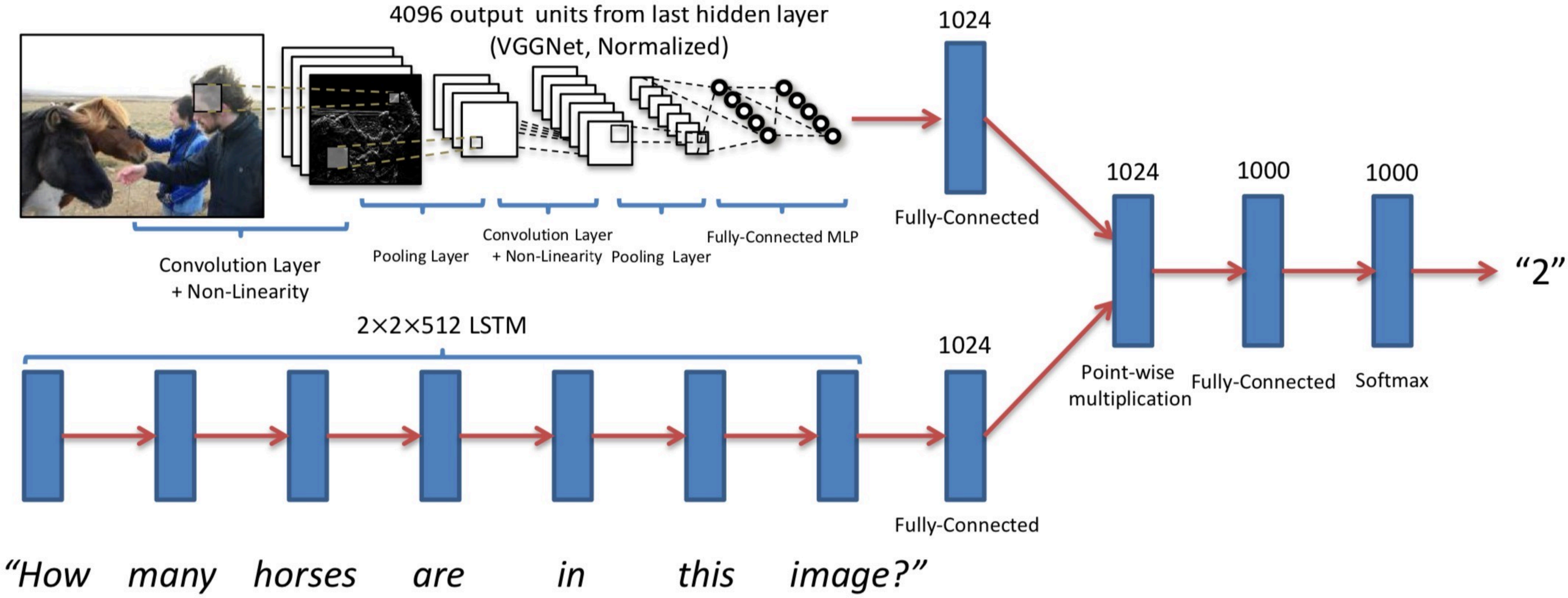
Language encoder

(LSTM, Transformer)

Cross-attention/joint layer

Prediction

Visual Question Answering



Agrawal et al., 2015

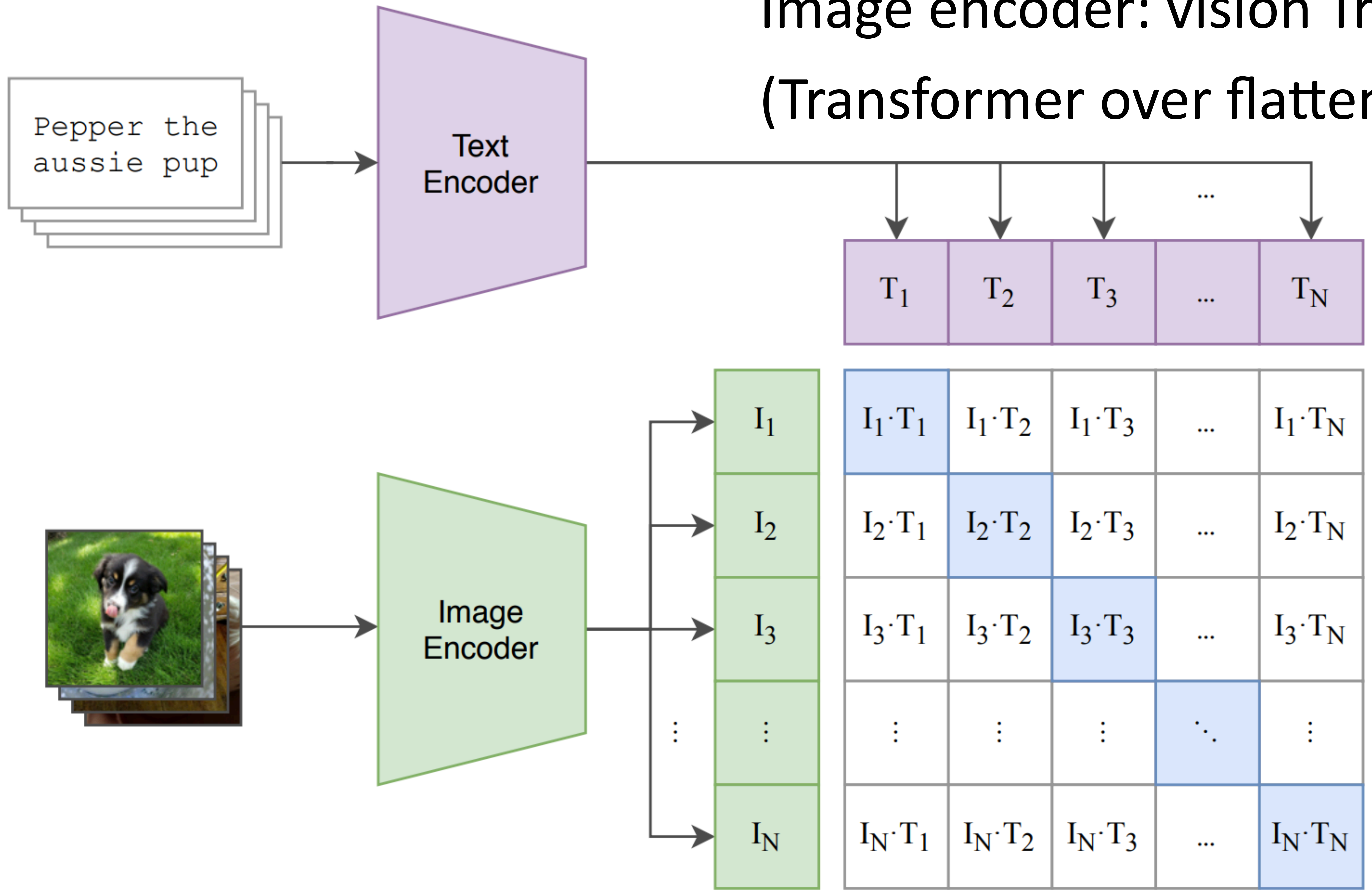
Language-vision Pre-training: CLIP

Text encoder: Transformer

Image encoder: vision Transformer

(Transformer over flattened patches)

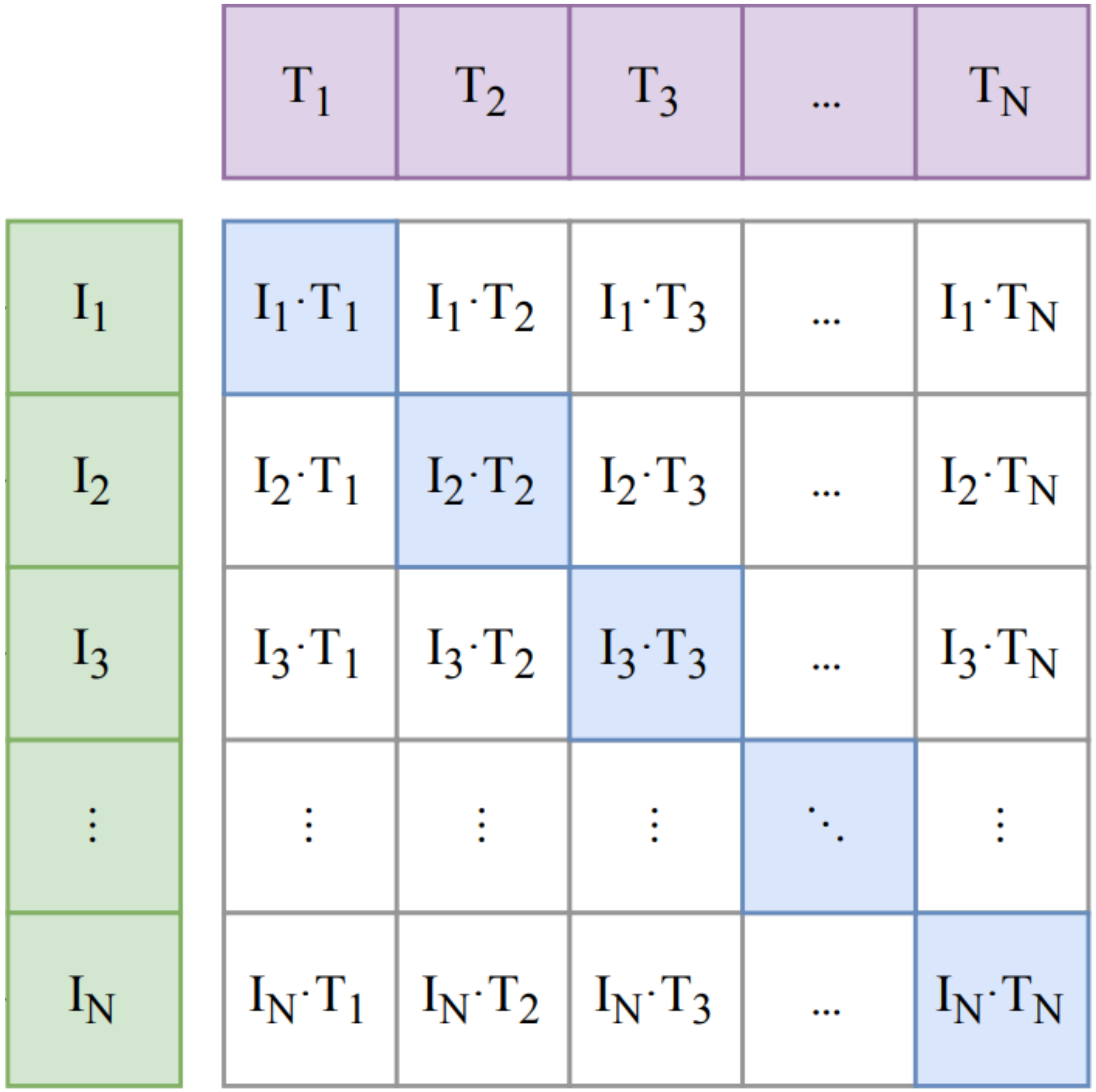
(1) Contrastive pre-training



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Radford et al., 2021

Language-vision Pre-training: CLIP



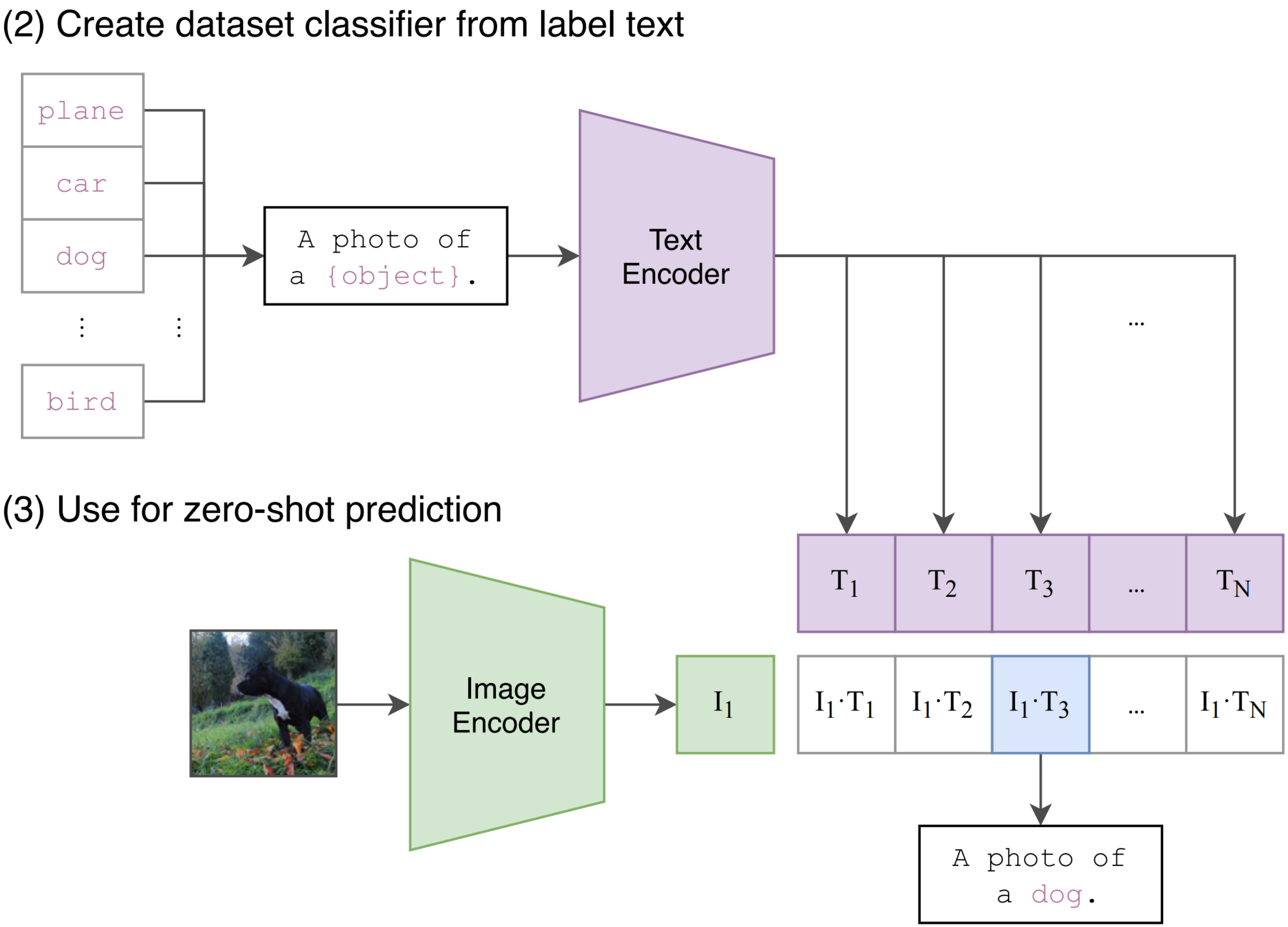
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

- Contrastive objective: each image should be more similar to its correspond caption than to other captions

$$
\begin{aligned}
\text{maximize } & \text{softmax}(I_1^{T} T_i)[1] \\
& + \text{softmax}(I_2^{T} T_i)[2] \\
& + \ldots
\end{aligned}
$$

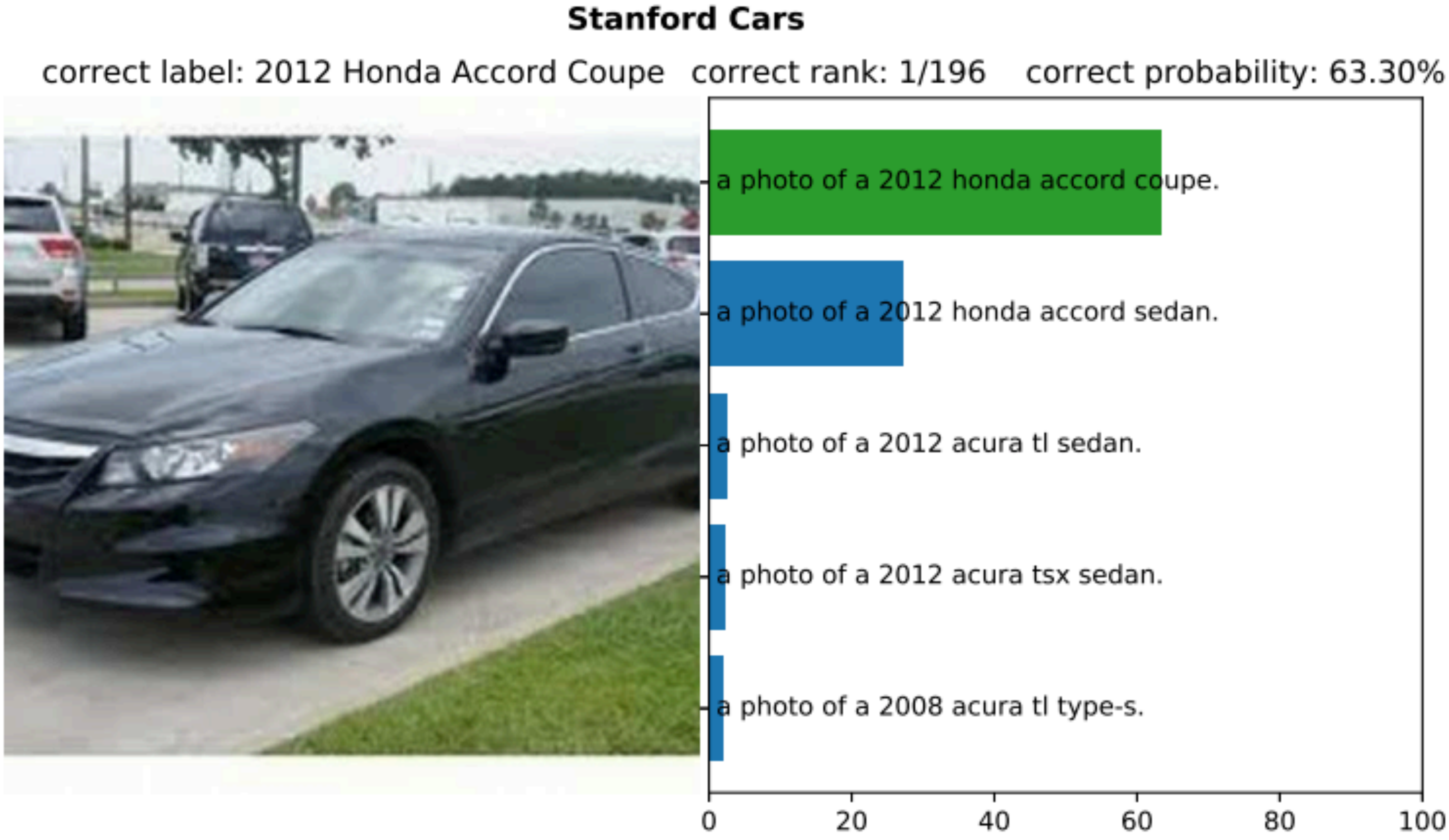
Radford et al., 2021

Language-vision Pre-training: CLIP

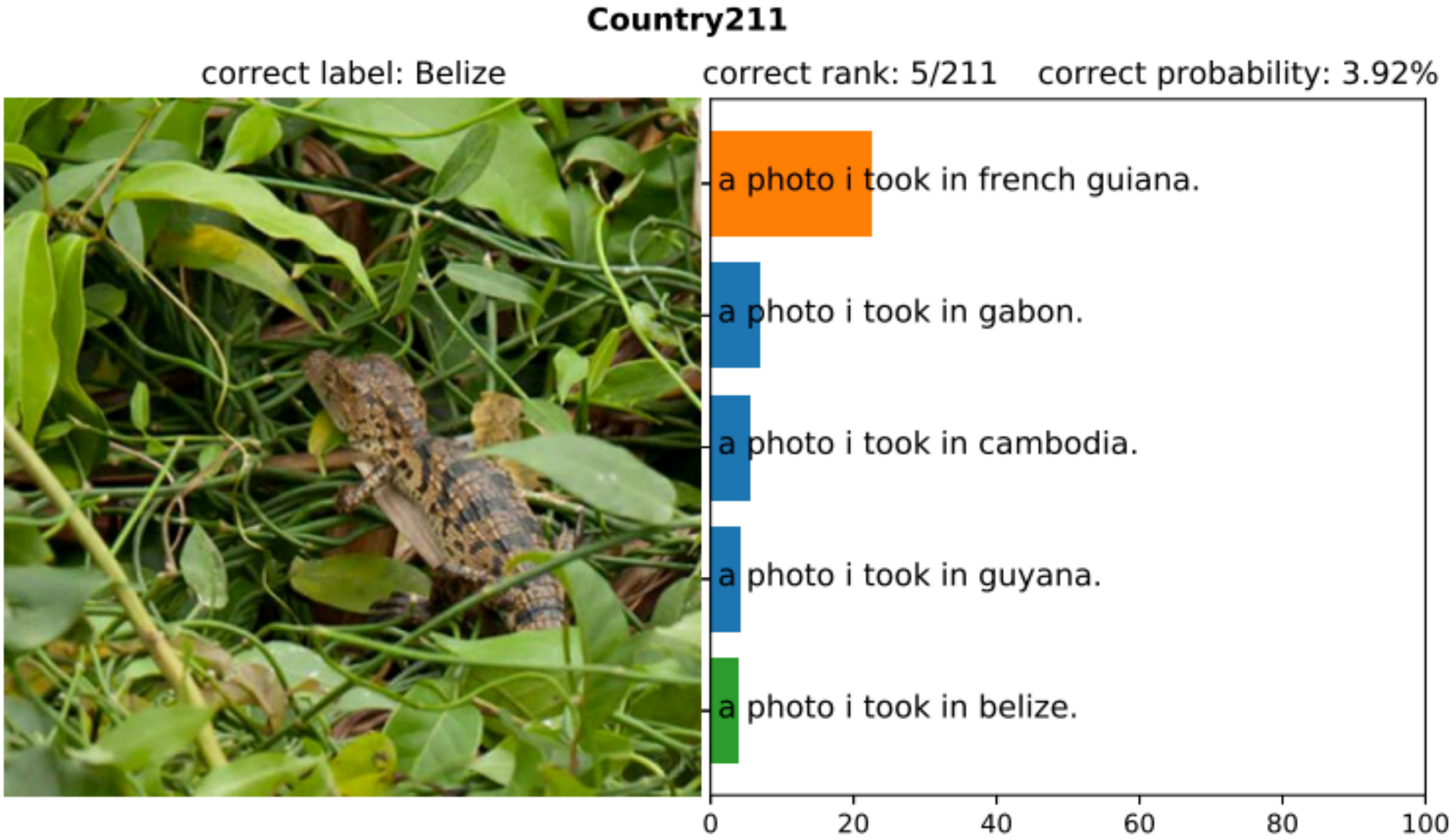


Radford et al., 2021

CLIP: Zero-shot Results



CLIP: Zero-shot Results


Country211
correct label: Belize
correct rank: 5/211
correct probability: 3.92%
a photo i took in french guiana.
a photo i took in gabon.
a photo i took in cambodia.
a photo i took in guyana.
a photo i took in belize.