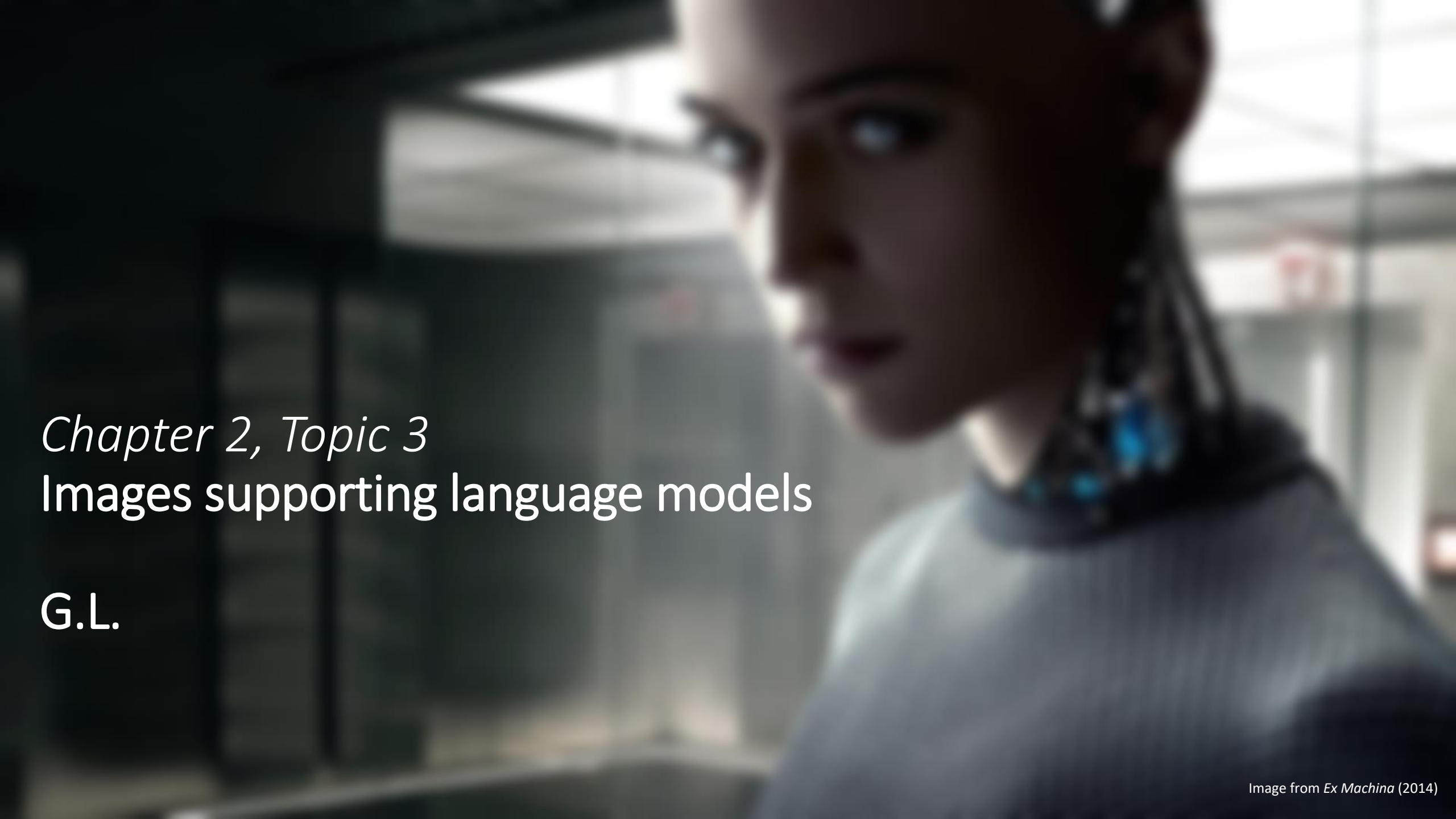


Chapter 2, Topic 3

Images supporting language models

G.L.

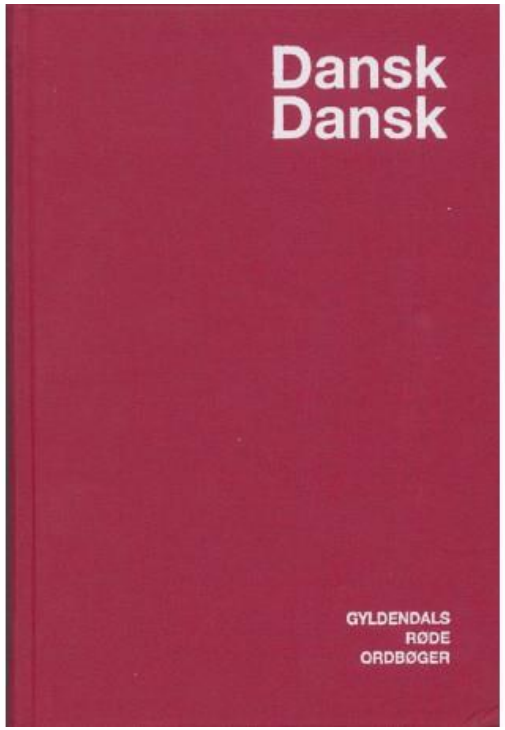
Image from *Ex Machina* (2014)

Index

Problem definition and scope

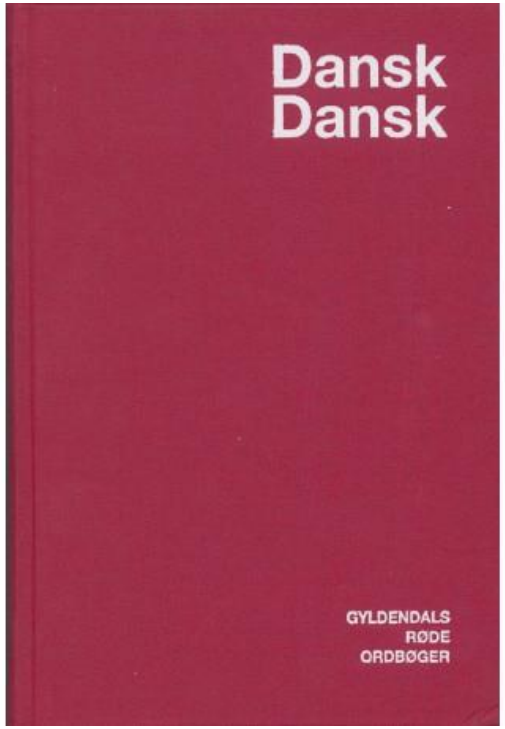
What is the meaning of the
Danish word **øl**?

Problem definition and scope



What is the meaning of the Danish word **øl**?

Problem definition and scope



What is the meaning of the Danish word **øl**?

Problem definition and scope

The Symbol Grounding Problem

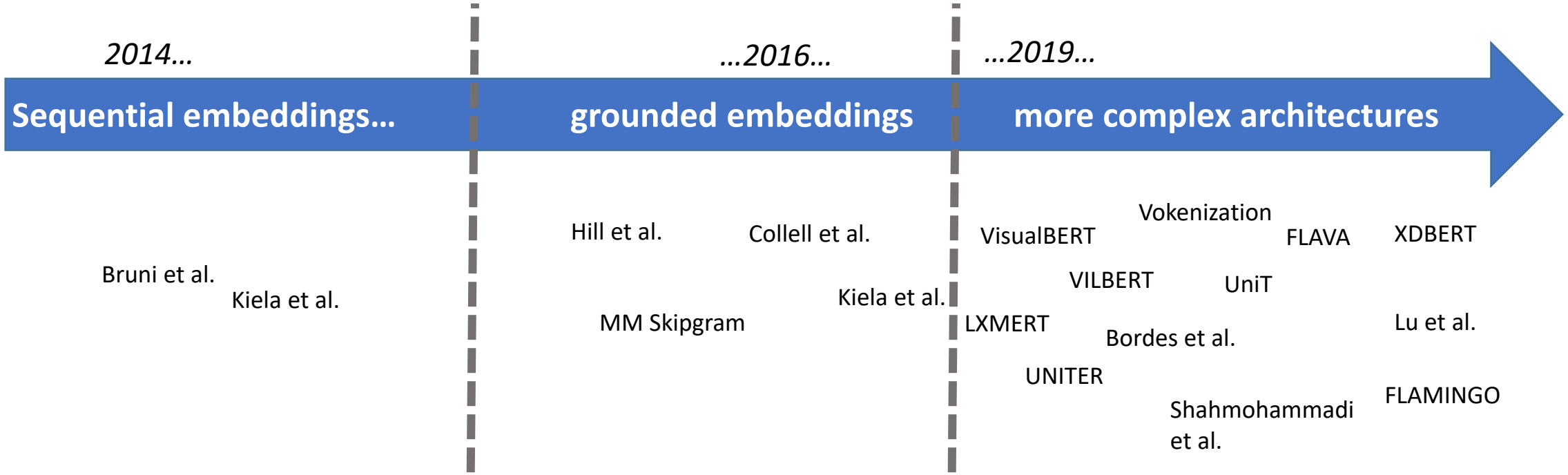
It is possible to understand the meaning of a word only if the word is put (grounded) in a context, a perceptual space, other than that of written language

Object of the chapter

How could visual elements help generate word representations capable of capturing their meaning?

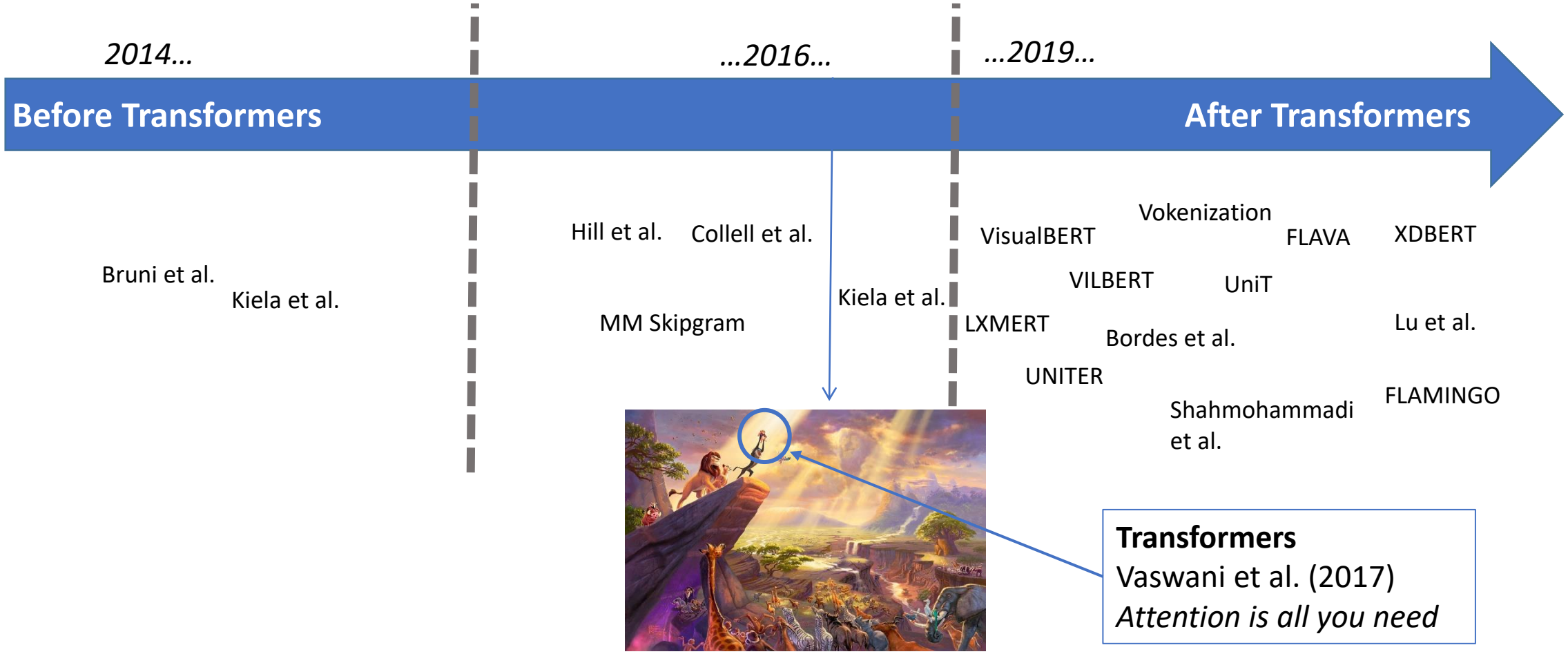
How to get text and images together? Historical evolution (1/2)

Option 1 (“architecture view”)



How to get text and images together? Historical evolution (2/2)

Option 2 (“evaluation view”)

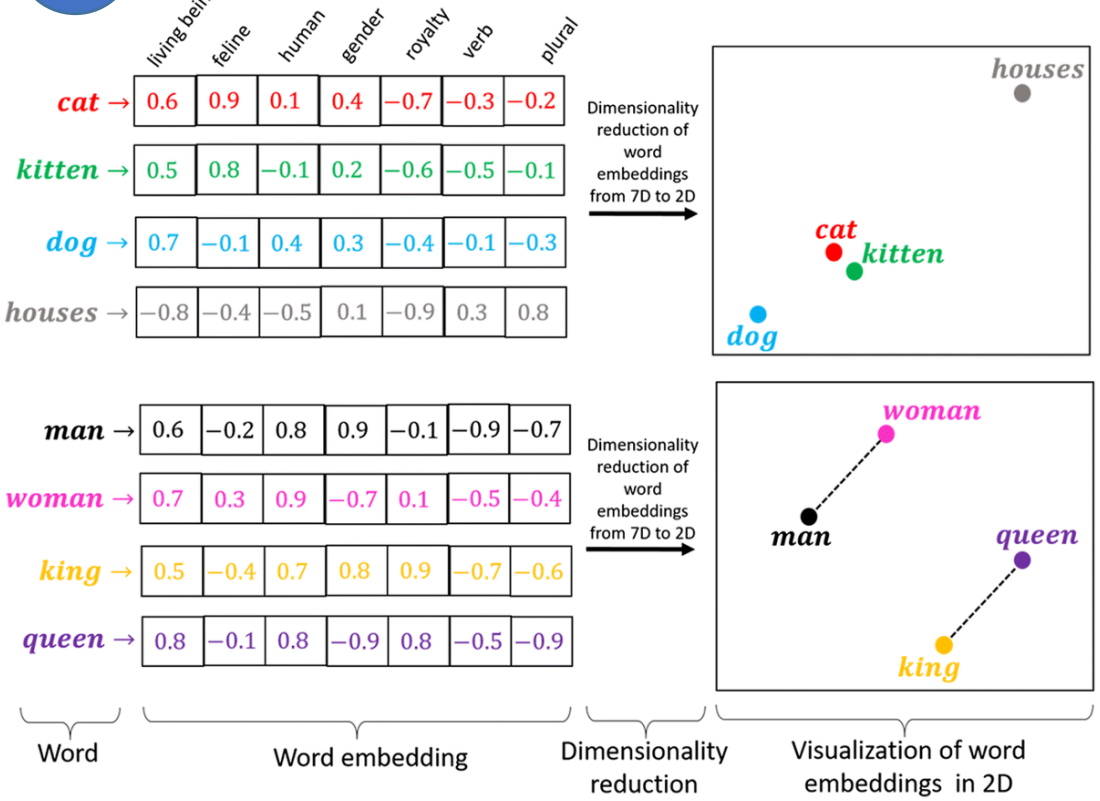


Starting point: Pure language models

1 **Tokens generation**

„This is a very sunny day“ → Tokenizer → [This is] [a very] [sunny day] → **Tokens**

2 **Word embeddings**



Credit: Hariom Gautam

Skipgram (Word2vec)
Neighbouring words

$$NLL = -\sum_{t=1}^{N} \sum_{\substack{i \in \{-c, \ldots c\} \\ i \neq 0}} log\, p\left(w_{[t+i]} \middle| w_{[t]}\right)$$

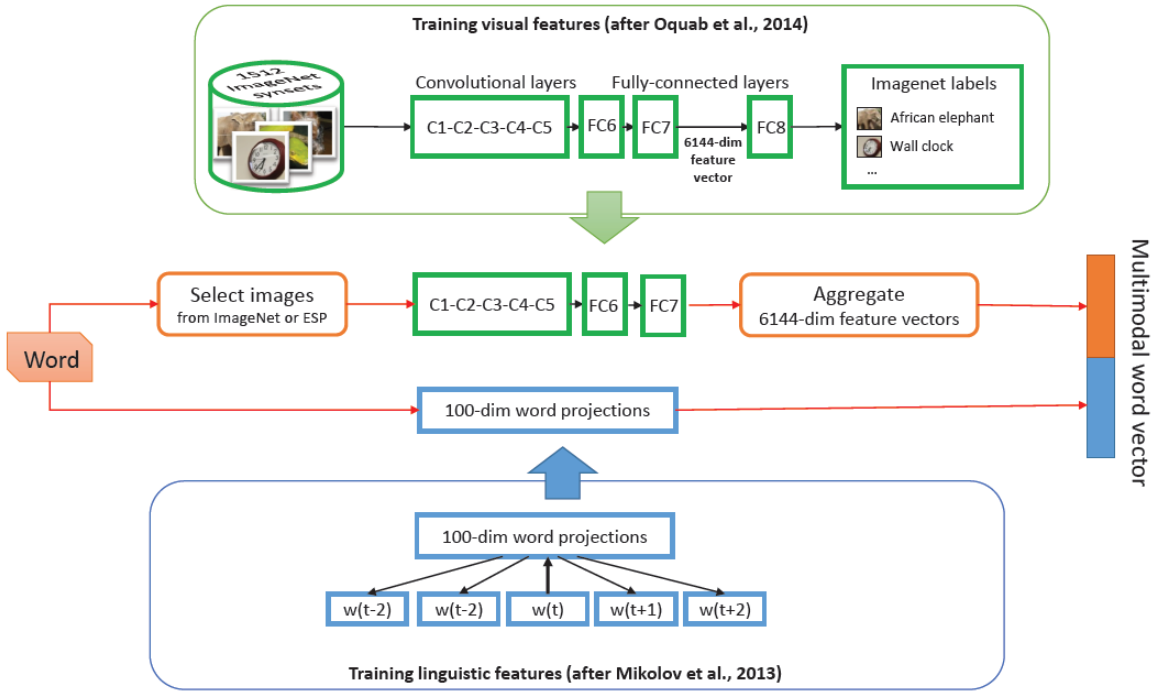
GloVe
Words co-occurrence

$$J = \sum_{i,j=1}^{V} f\left(X_{i,j}\right)\left(w_i^T \widetilde{w_j} + b_i + \widetilde{b_j} - log X_{ij}\right)$$

BERT
More than one embedding per word

- Masked Language Modelling (**MLM**)
- Next Sentence Prediction (**NSP**)

Sequential embeddings: Kiela and Bottou (2014)



Source: Kiela and Bottou (2014)

	Linguistic representation	Visual representation
	• Skipgram	Each image is processed by the 8-layer CNN. Two ways to aggregate features vectors: • **CNN-Mean** • **CNN-Max**
Data	• Text8 Corpus • British National Corpus	• ImageNet • ESP Game

Multimodal representation

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} || (1 - \alpha) \times \vec{v}_{vis}$$

Hyperparameters	• α: relative weight of the linguistic modality

Sequential embeddings: Other papers & drawbacks

Honourable mention(s)

- **Bruni et al. (2014)**: the „father of concatenation in this field“
- **Silberer et al. (2014)**: stacked autoencoders to learn higher-level embeddings from textual and visual inputs

Drawbacks

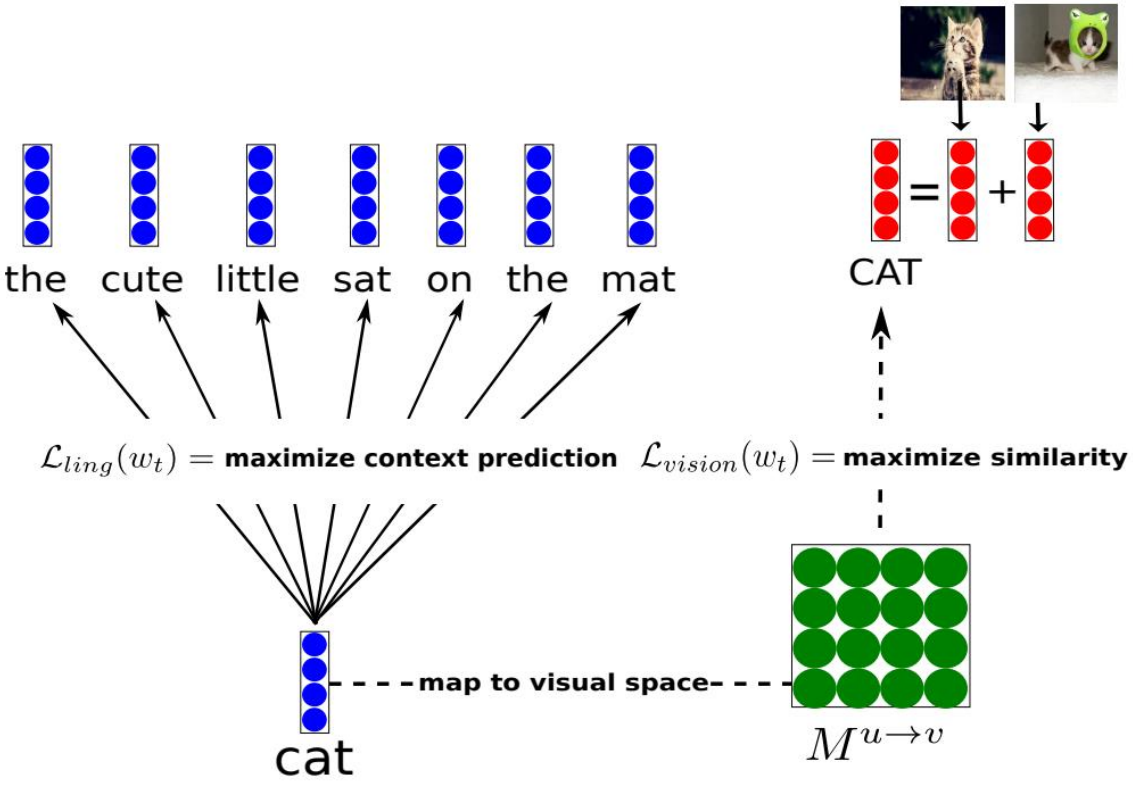

No interaction between modalities during training


1:1 correspondence between modalities assumed

How to lift the 1:1 restriction?


Ground language in vision!

Grounding language in vision: Lazaridou et al. (2015) (1/3)



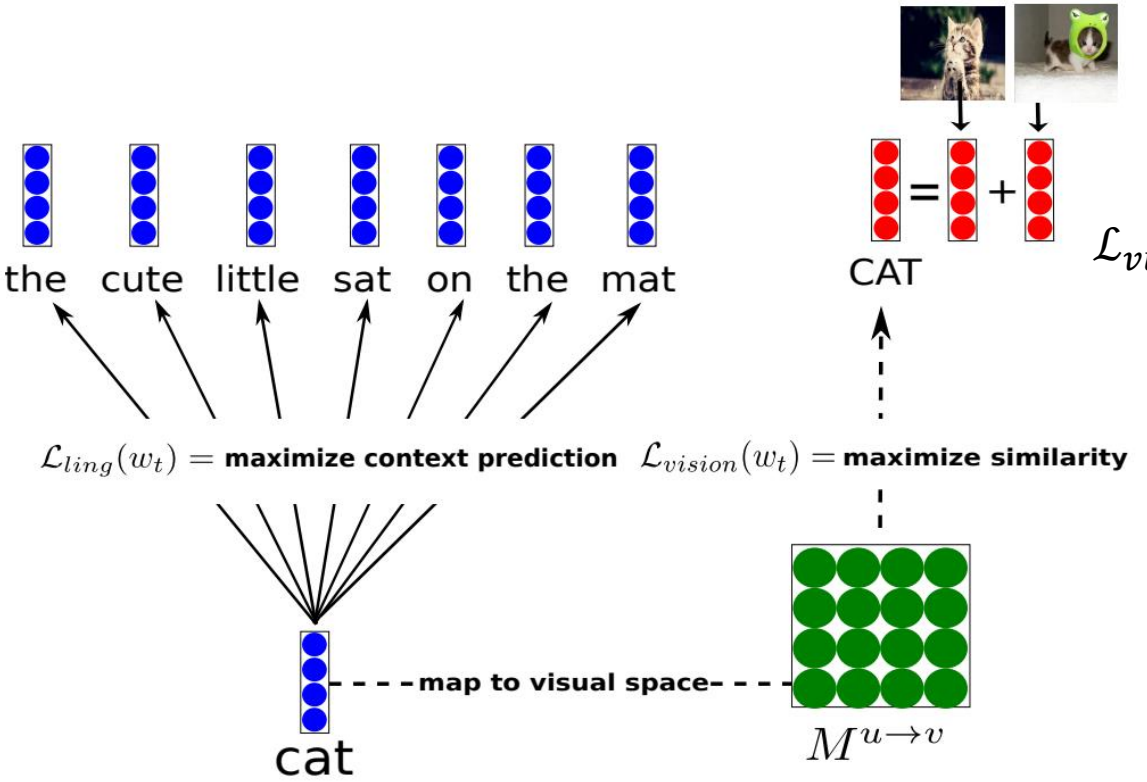
	Linguistic representation	Visual representation
	• Skipgram	• Sample 100 images for each word and extract a 4096-d array with a CNN • average the vectors of 100 pictures associated to each word to get visual representation
Data	• Wikipedia 2009	• ImageNet

Multimodal objective function

$$\frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_{ling}(w_t) + \mathcal{L}_{vision}(w_t))$$

Source: Lazaridou et al. (2015). MMSKIP-GRAM-B.

Grounding language in vision: Lazaridou et al. (2015) (2/3)



Source: Lazaridou et al. (2015). MMSKIP-GRAM Model B.

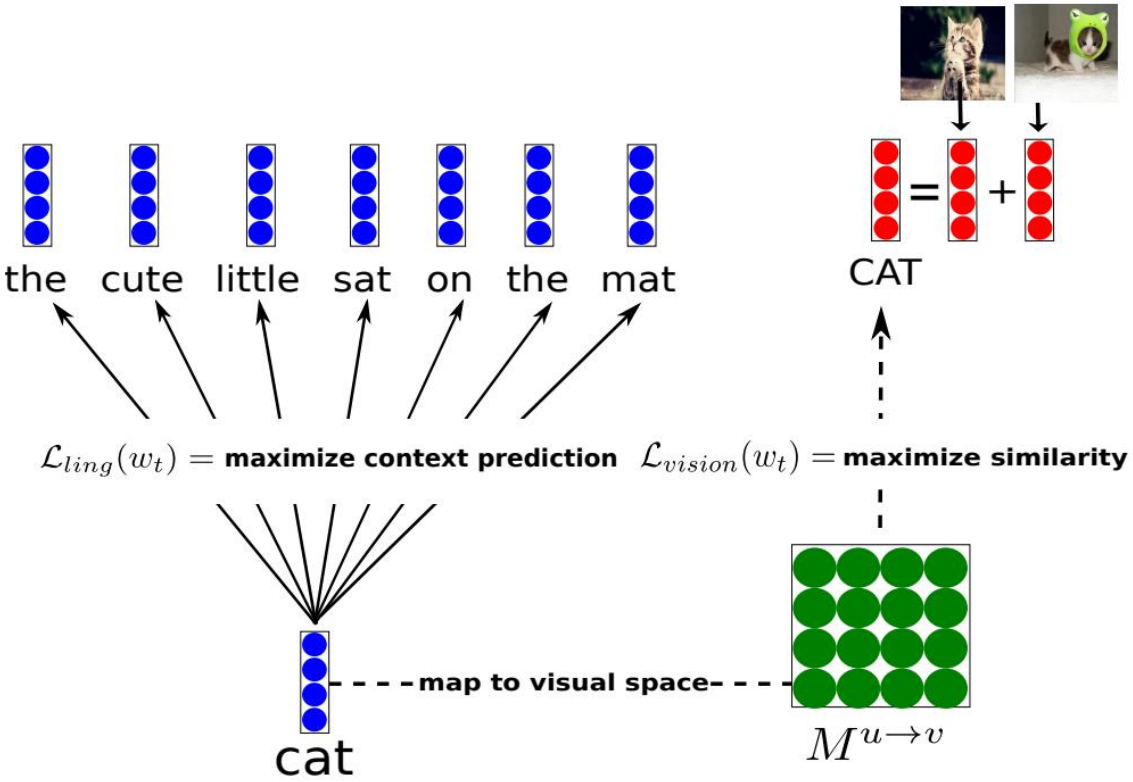
Multi-modal Skipgram Model A

$$\mathcal{L}_{visionA} = - \sum_{w' \backsim P_n(w)} max\left(0, \gamma - cos\left(u_{w_t}, v_{w_t}\right) + cos\left(u_{w_t}, v_{w'}\right)\right)$$

- **Goal**: align the textual representation u_{w_t} to the visual one v_{w_t} (which is fixed)
- **Max-margin** objective function

Hyperparameters	• k: number of negative samples • γ: margin

Grounding language in vision: Lazaridou et al. (2015) (3/3)



Source: Lazaridou et al. (2015). MMSKIP-GRAM Model B.

Multi-modal Skipgram Model B

- **Same** objective as Model A
- New layer: estimation of a a **cross-modal mapping** M from textual to visual space; $z_{w_t} = M^{u \to v} u_{w_t}$ instead of u_{w_t}
- No more 1:1 correspondence needed

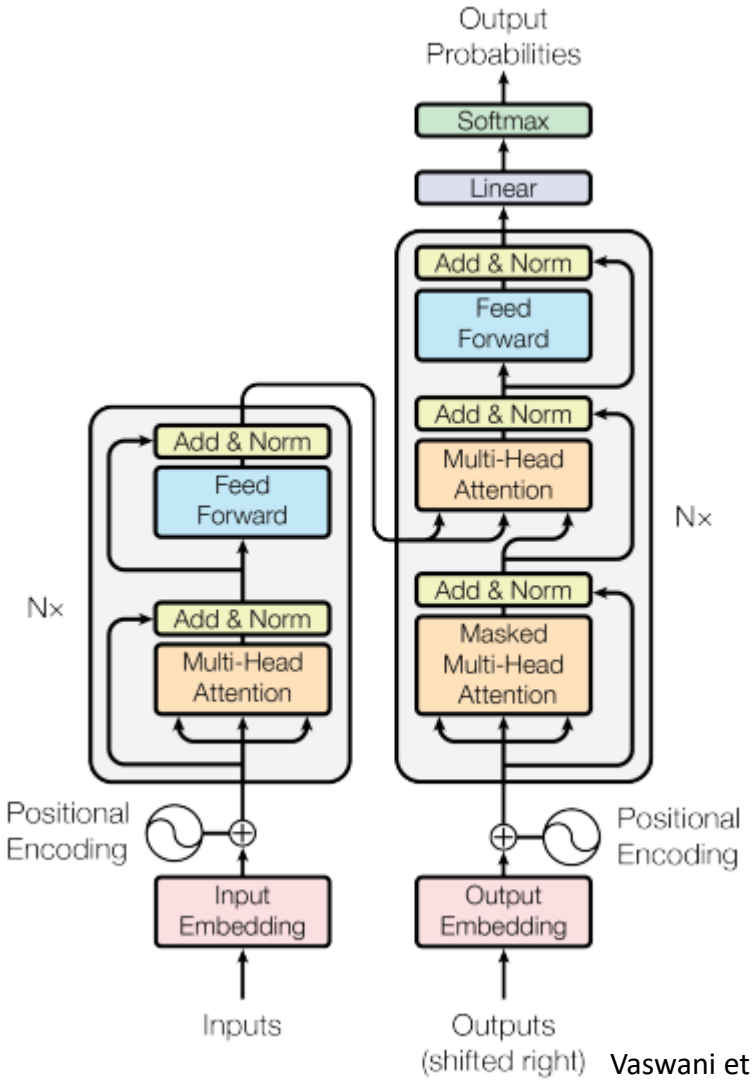
Hyperparameters	• k: number of negative samples • γ: margin • λ: regularization parameter

Honourable mention(s)

- **Collell et al. (2017)**: linear and CNN mapping from textual to visual space
- **Kiela et al. (2018)**: Bidirectional LSTM and linear mapping from text into the visual space
- **Bordes et al. (2020)**: Trained on a grounded objective function which contains parameters also from the textual modality

TRANSFORMERS

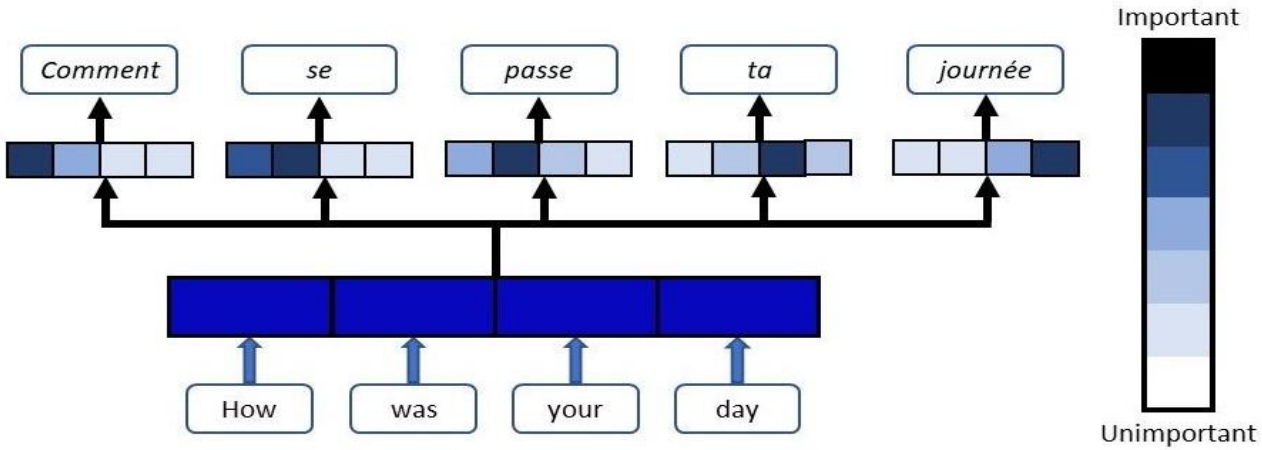
Transformers (1/2): In a nutshell



Vaswani et al. (2017)

Why are transformers better than RNNs in modelling natural language?

- **Self attention**
- **Parallel** input processing
- **Positional encoding**

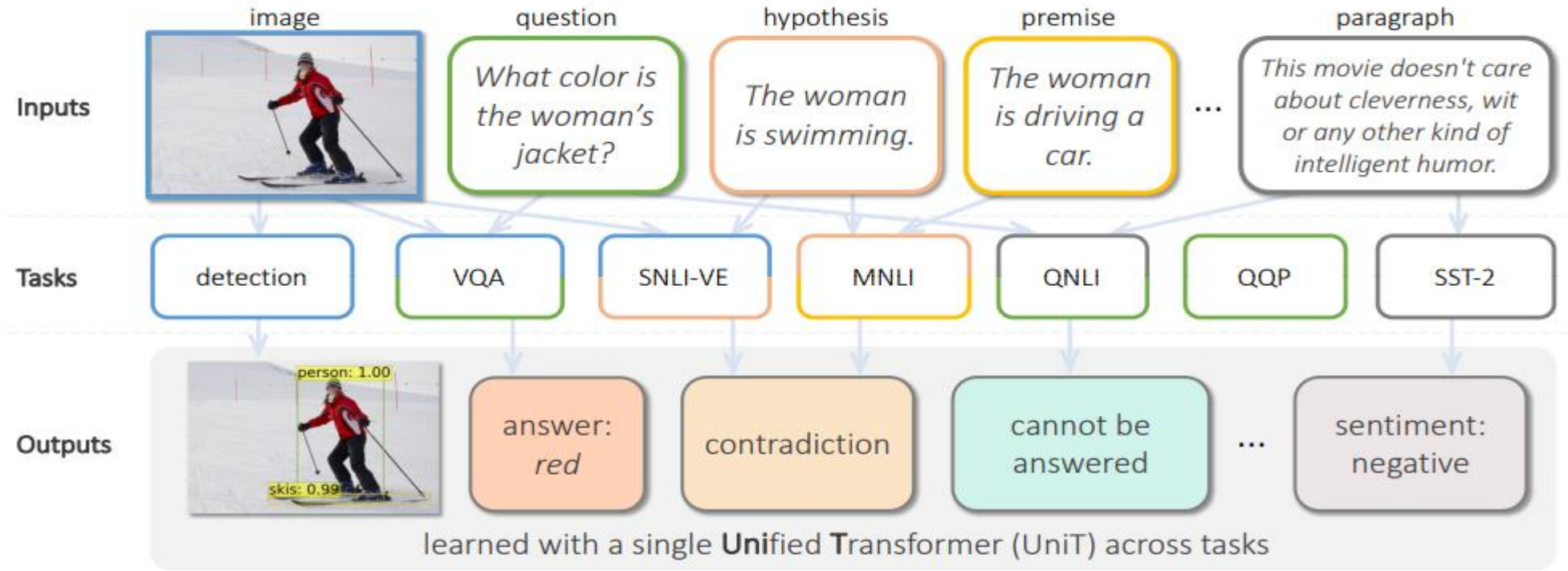


Credit: Gabriel Loye

Transformers (2/2): „Universal“ vocation

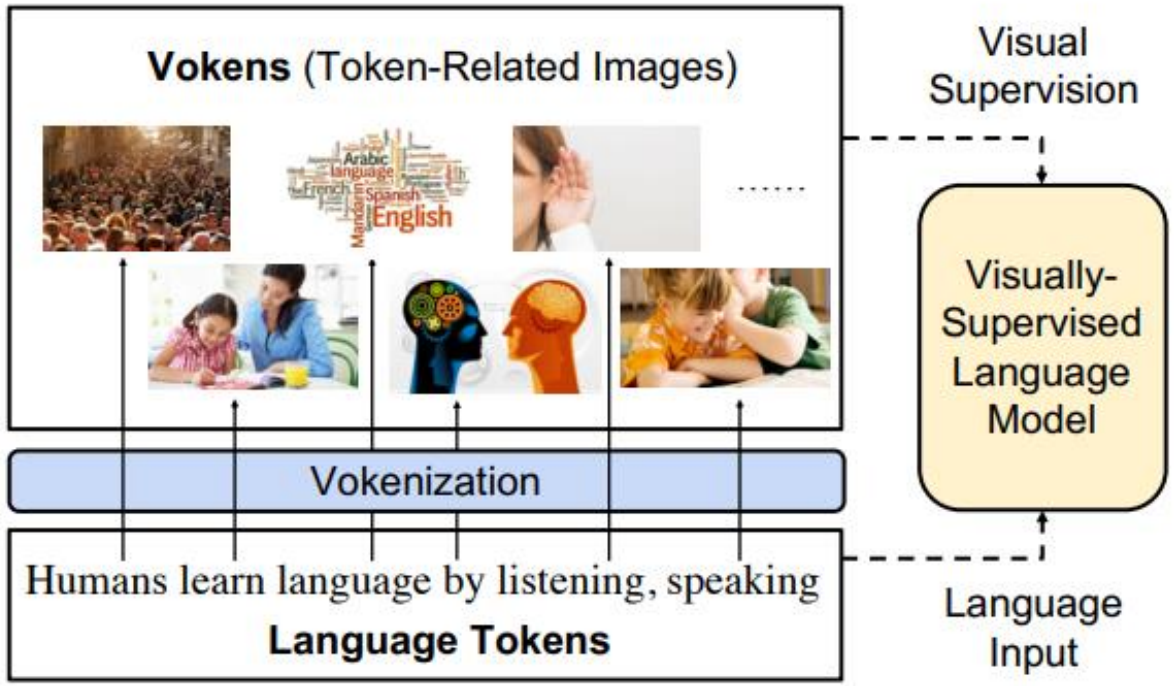
- Focus on **downstream** tasks
- Pre-training and transfer knowledge
- Universal models

Examples: Uniter, Vilbert, VisualBERT, Flamingo, UFO, FLAVA, UniT



Source: Hu et al. (2021), UniT

Vokenization (1/6): Visually supervised models



Source: Tan et al. (2020)

IDEA: Visually-supervised language model („visual pointing“) to imitate human language understanding

Supposing a dataset with tokens w_i and associated vokens v_i **existed**...

... **Voken Classification Task** performed during pre-training:

$$\mathcal{L}_{VOKEN-CLS}(s) = -\sum_{i=1}^{l} log\ p_i(v(w_i; s)|s)$$

$$p_i(v|s) = softmax_v\{W\mathrm{h}_i + b\}$$

→ **integrable** in the pretraining framework of **any** language model

But...

Vokenization (2/6): Challenges in creating vokens

...This dataset of tokens-vokens does **not** exist...

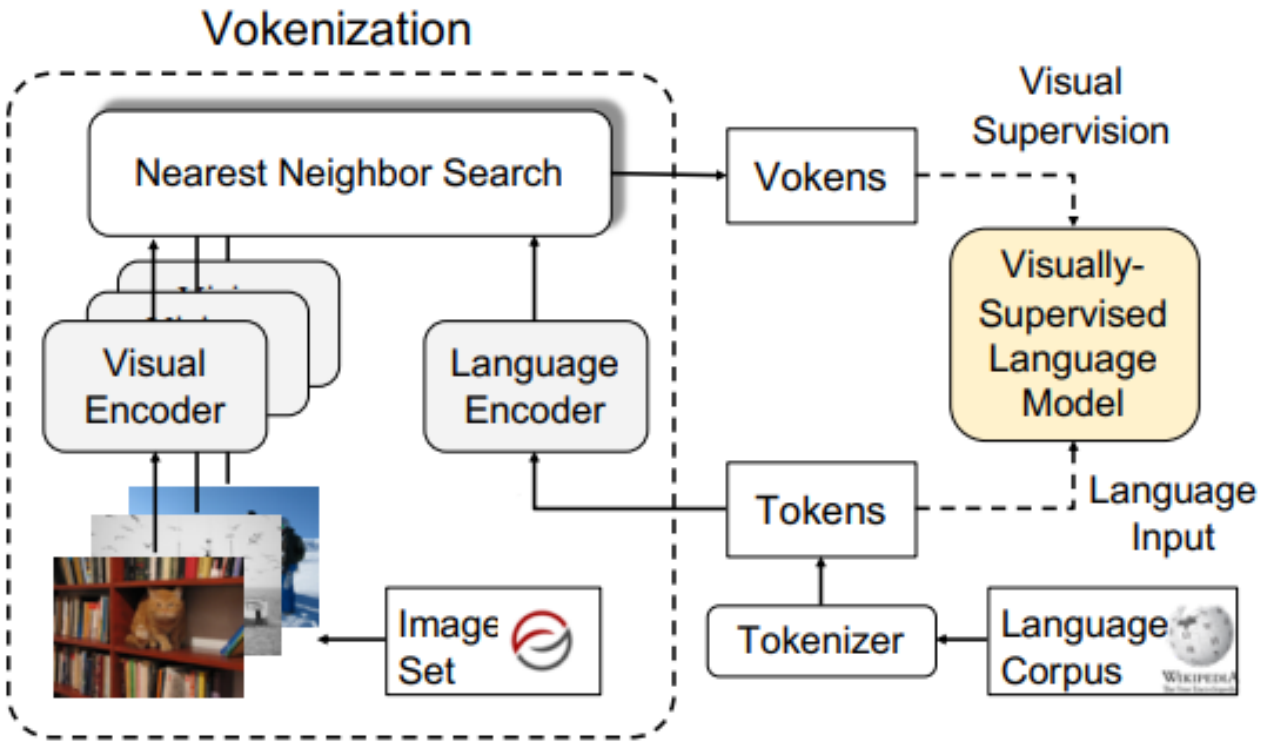
QUESTION: Would it possible to use **image-captioning datasets** to get the Vokens?

Dataset	# of Tokens	# of Sents	Vocab. Size	Tokens #/ Sent.	1-Gram JSD	2-Gram JSD	Grounding Ratio
MS COCO	7.0M	0.6M	9K	11.8	0.15	0.27	54.8%
VG	29.2M	5.3M	13K	5.5	0.16	0.28	57.6%
CC	29.9M	2.8M	17K	10.7	0.09	0.20	41.7%
Wiki103	111M	4.2M	29K	26.5	0.01	0.05	26.6%
Eng Wiki	2889M	120M	29K	24.1	0.00	0.00	27.7%
CNN/DM	294M	10.9M	28K	26.9	0.04	0.10	28.3%

Source: Tan et al. (2020). **Grounding Ratio** defined as the proportion of tokens from a language corpora which have more than 100 occurrences in the image dataset

Low Grounding Ratio for image-captioning datasets...

Vokenization (3/6): Vokenizer – a framework for Vokens generation



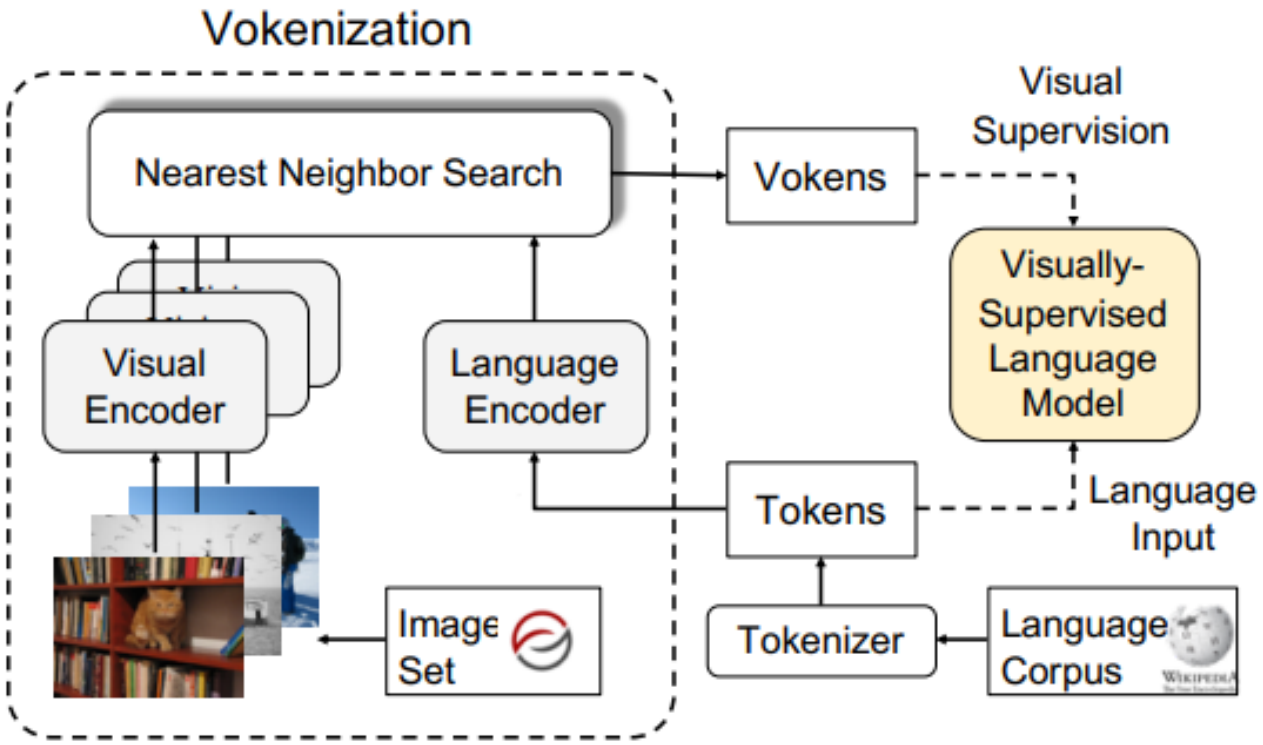
IDEA: learn a vokenizer capable of generating images for large language corporas

Find the image $x \in X$ maximizing the **relevance scoring function**

$$v(w_i; s) = \underset{x \in X}{\operatorname{argmax}} r_{\theta^*}(w_i, x, s)$$

Source: Tan et al. (2020)

Vokenization (4/6): Vokenizer – Relevance Scoring Function factorization



Source: Tan et al. (2020)

$$r_\theta(w_i, x, s) = \mathrm{f}_\theta(w_i; s)^\intercal \mathrm{g}_\theta(x)$$

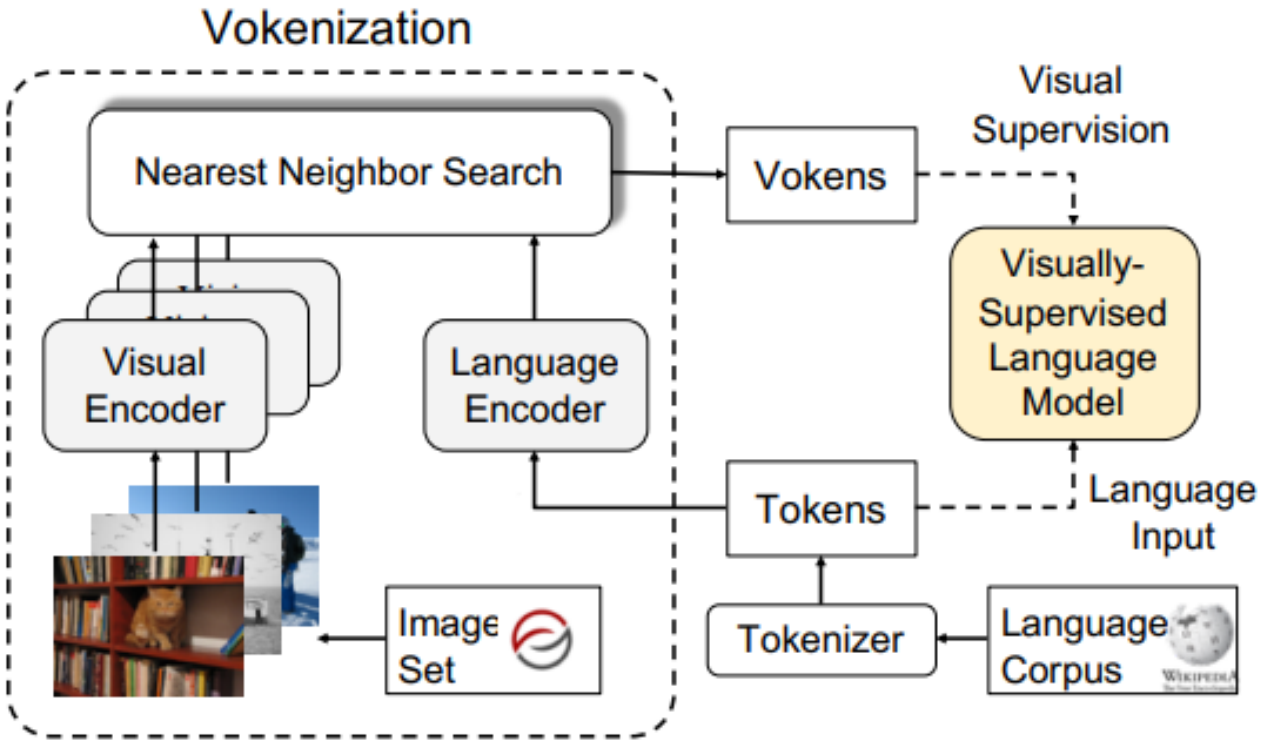
Language feature representation

Visual feature representation

For both representations, *separately*:

1. Encoder with a pretrained BERT (language) or ResNeXt (visual)
2. A multi-layer perceptron is applied to down-project the output of the encoders
3. L2-Normalization (norm-1 vectors)

Vokenization (5/6): Vokenizer – Relevance Scoring Function factorization



Source: Tan et al. (2020)

$$r_\theta(w_i, x, s) = \mathrm{f}_\theta(w_i; s)^\intercal \mathrm{g}_\theta(x)$$

Language feature representation

Visual feature representation

Training: how to estimate **Θ**

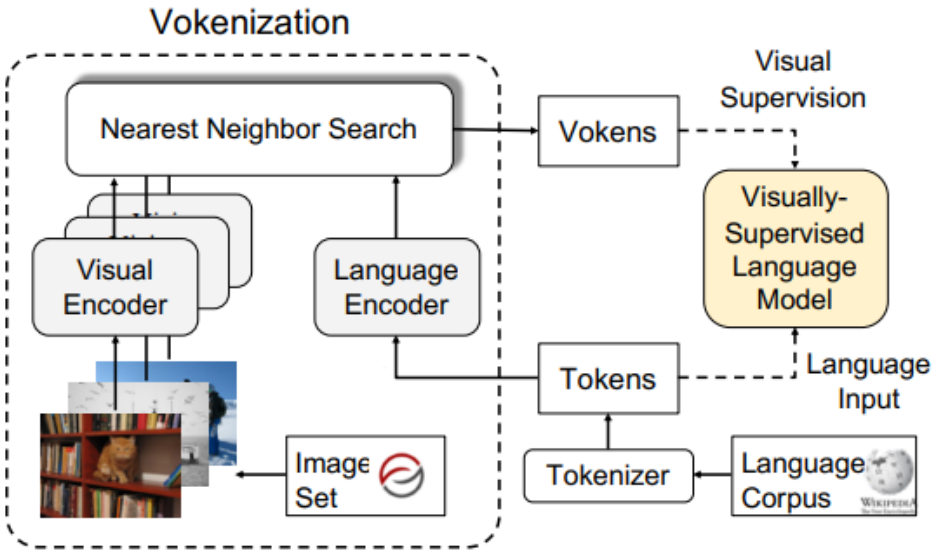
$$\mathcal{L}_\theta(s, x, x') = \sum_{i=1}^{l} max\big(0, M - r_\theta(w_i, x, s) + r_\theta(w_i, x', s)\big)$$

- Minimize the hinge loss

How to maximize the Relevance Scoring function?

- RSF is factorized as inner product of feature representations…
- …Maximum inner product ↔ Nearest Neighbour (Vectors have 1-norm)

Vokenization (6/6): Vokenizer – Revokenization



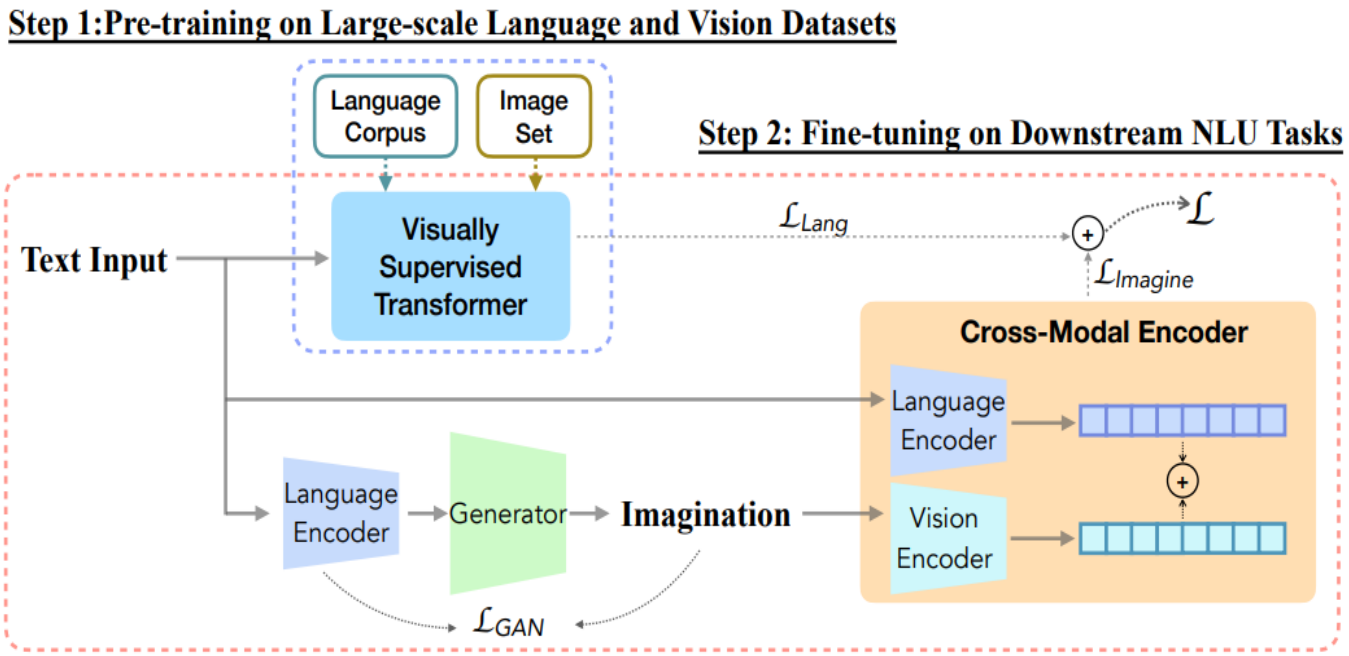
Source: Tan et al. (2020)

PROBLEM
Vokenizer depends on the tokenizer of the language encoder

REVOKENIZATION

1. Build the Vokenizer with a tokenizer T1, which generates tokens w_i
2. Generate vokens v_i
3. Given a tokenizer T2, which generates tokens u_i, find the tokens which generates tokens w_i which overlaps the most with u_i
4. Assign to u_i the voken associated to w_i

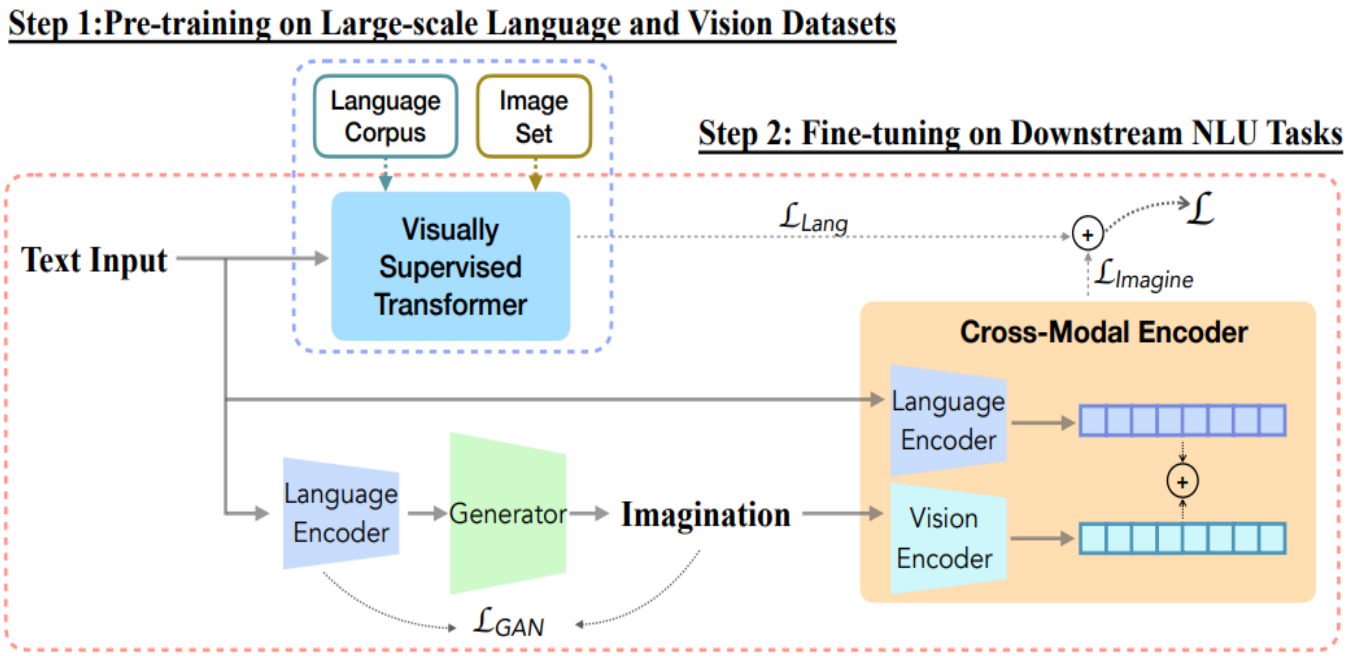
Imagination augmented NLU: Lu et al. (2022) (1/3)



Source: Lu et al. (2022)

	Linguistic representation	Visual representation
	• Any language model (in paper BERT-base and RoBERTa	• Pretrained visually supervised transformer → Vokenizer
Data	• Wikipedia	• MS COCO

Imagination augmented NLU: Lu et al. (2022) (1/2)



Source: Lu et al. (2022)

Multimodal representation

The framework **iACE** is composed of two modules:

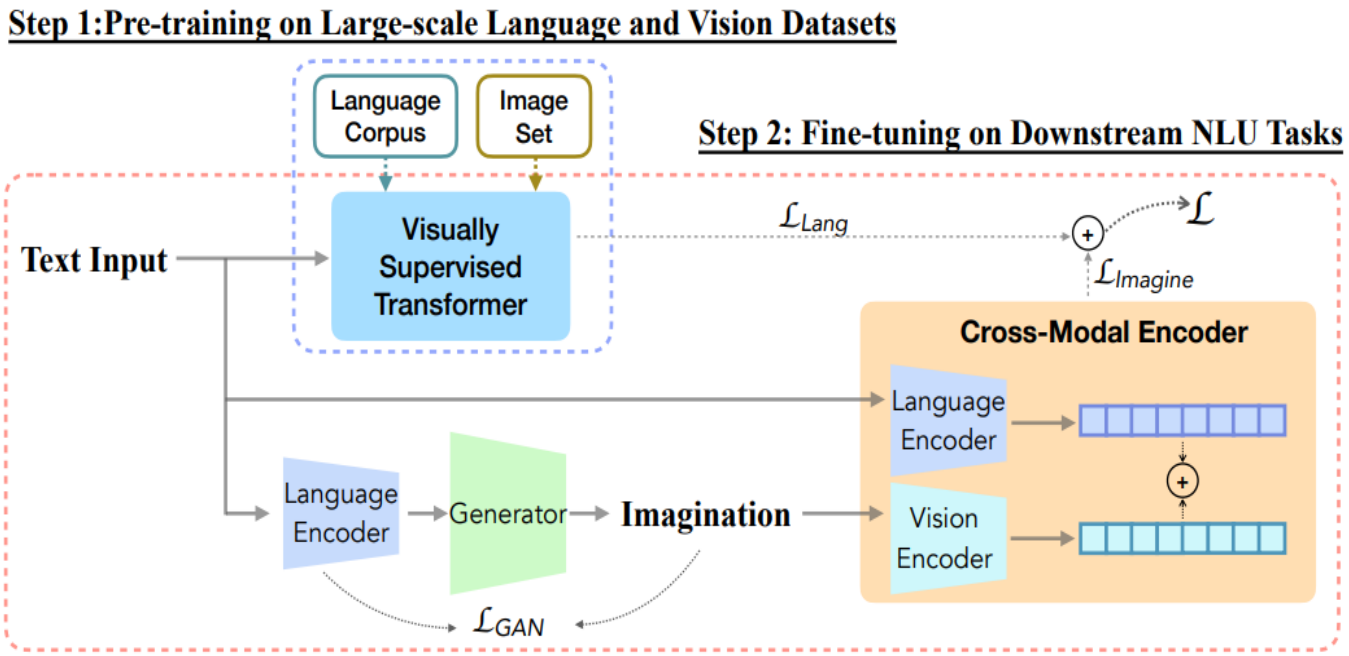
- **Imagination generator G**: for each text input, VQGAN+CLIP generates an "imagined visual“

→ minimize the distance between textual t and visual v embeddings

$$\mathcal{L}_{GAN} = 2[arcsin(\frac{1}{2}||t - v||)]^2$$

- **Imagination augmented cross-modal encoder**: CLIP

Imagination augmented NLU: Lu et al. (2022) (2/2)



Source: Lu et al. (2022)

2-Step learning procedure

Step 1:
Pretraining a visually supervised Transformer (Vokenization)

Step 2 (GLUE is used for downstream tasks):
Minimize of the joint cross-entropy loss function:

$$\mathcal{L} = \lambda \mathcal{L}_{Imagine} + (1 - \lambda)\mathcal{L}_{Lang}$$

$$\mathcal{L}_{Imagine} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \ log\ p_k\left(d_j(t,v)\middle|D\right)$$

$$\mathcal{L}_{Lang} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \ log\ p_k(d_j(t)|D)$$

Evaluation – Before Transformers (1/2)

Focus on **intrinsic** evaluation : alignment with similarity judgments by human ?

- Semantic *similarity*, e.g. “pasta is similar to rice”
- Semantic *relatedness*, e.g. “Michelin-Star is related to restaurant” (or “pizza is **NOT** related to pineapple”)
- (*Visual* similarity: cucumbers look like zucchinis)

Benchmarks

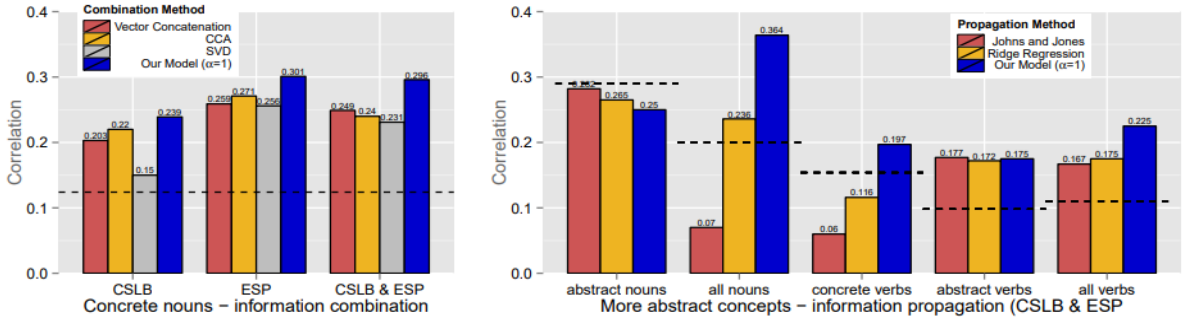
- MEN
- WordSim353
- SimLex999

Evaluation

Word pairs (w1,w2) → Cosine Similarity $\frac{w_1 \cdot w_2}{||w_1||\,||w_2||}$ → Correlation with human ratings

Evaluation – Before Transformers (2/2)

- Multimodal representation show generally improvements over pure language models in the pre-Transformers Age… **But…**



For abstract pure language models (dashed lines) still perform better

Source: Hill et al. (2014)

Model	MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK
STb-1024	70.3	68.0	87.5	85.5	69.7/80.6	78.3	67.3	76.6
STb-2048	73.1	**75.7**	88.3	86.5	71.6/**81.7**	79.0	71.0	78.8
2×STb-1024	71.4	74.7	88.2	86.6	71.3/80.7	75.8	69.4	78.3
Cap2Cap	71.4	74.7	86.7	**86.7**	70.3/79.8	76.1	68.5	78.2
Cap2Img	72.1	75.5	86.9	86.0	**72.3**/81.1	77.7	71.4	81.2
Cap2Both	71.6	74.4	86.5	85.5	71.4/79.5	78.5	71.3	**81.7**
GroundSent-Cap	73.1	73.0	**88.6**	86.6	70.8/81.2	79.4	70.7	79.1
GroundSent-Img	72.5	74.9	88.4	85.7	71.3/81.2	79.4	70.5	79.7
GroundSent-Both	**73.3**	75.2	87.5	86.6	69.9/79.9	**80.3**	**72.0**	78.1

Dataset	Concreteness
MR	2.3737 ± 0.965
CR	2.4714 ± 1.025
SUBJ	2.4510 ± 1.007
MPQA	2.3158 ± 0.834
MRPC	2.5086 ± 0.987
SST	2.7471 ± 1.142
SNLI	3.1867 ± 1.309
SICK	3.1282 ± 1.372

More “concrete” datasets benefit more from grounded representations

Source: Kiela et al. (2018)

Evaluation – After Transformers (1/4)

- Focus on **extrinsic** evaluation/downstream tasks

Benchmarks

- GLUE
- SQuAD
- SWAG

Evaluation

- Each subset of each benchmark-dataset has usually train/validation/test splits

Evaluation – After Transformers (2/4)

Model	Init. with BERT?	Diff. to BERT Weight	SST-2	QNLI	QQP	MNLI
ViLBERT (Lu et al., 2019)	Yes	0.0e-3	90.3	89.6	88.4	82.4
VL-BERT (Su et al., 2020)	Yes	6.4e-3	90.1	89.5	88.6	82.9
VisualBERT (Li et al., 2019)	Yes	6.5e-3	90.3	88.9	88.4	82.4
Oscar (Li et al., 2020a)	Yes	41.6e-3	87.3	50.5	86.6	77.3
LXMERT (Tan and Bansal, 2019)	No	42.0e-3	82.4	50.5	79.8	31.8
$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019)	-	0.0e-3	90.3	89.6	88.4	82.4
$\text{BERT}_{\text{BASE}}$ + Weight Noise	-	6.5e-3	89.9	89.9	88.4	82.3

Method	SST-2	QNLI	QQP	MNLI	SQuAD v1.1	SQuAD v2.0	SWAG	Avg.
$\text{BERT}_{\text{6L/512H}}$	88.0	85.2	87.1	77.9	71.3/80.2	57.2/60.8	56.2	75.6
$\text{BERT}_{\text{6L/512H}}$ + Voken-cls	89.7	85.0	87.3	78.6	71.5/80.2	61.3/64.6	58.2	76.8
$\text{BERT}_{\text{12L/768H}}$	89.3	87.9	83.2	79.4	77.0/85.3	67.7/71.1	65.7	79.4
$\text{BERT}_{\text{12L/768H}}$ + Voken-cls	**92.2**	**88.6**	**88.6**	**82.6**	**78.8/86.7**	68.1/71.2	**70.6**	**82.1**
$\text{RoBERTa}_{\text{6L/512H}}$	87.8	82.4	85.2	73.1	50.9/61.9	49.6/52.7	55.1	70.2
$\text{RoBERTa}_{\text{6L/512H}}$ + Voken-cls	87.8	85.1	85.3	76.5	55.0/66.4	50.9/54.1	60.0	72.6
$\text{RoBERTa}_{\text{12L/768H}}$	89.2	87.5	86.2	79.0	70.2/79.9	59.2/63.1	65.2	77.6
$\text{RoBERTa}_{\text{12L/768H}}$ + Voken-cls	**90.5**	**89.2**	**87.8**	**81.0**	**73.0/82.5**	**65.9/69.3**	**70.4**	**80.6**

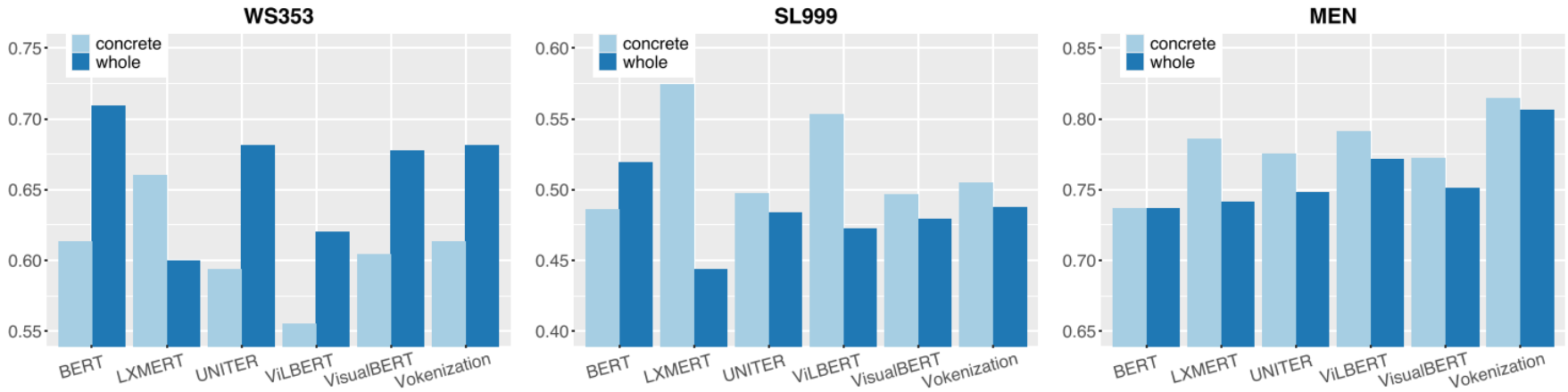
Extrinsic evaluation

- Pure language models such as BERT perform better than universal transformers models…
- Additional training tasks such as the case of Vokenization can be beneficial

Source: Tan et al. (2020)

Evaluation – After Transformers (3/4)

model	*input*	*Spearman ρ correlation (layer)*				
		RG65	WS353	SL999	MEN	SVERB
BERT-1M-Wiki*	L	0.7242 (1)	0.7048 (1)	0.5134 (3)	–	0.3948 (4)
BERT-Wiki *ours*	L	0.8107 (1)	0.7262 (1)	0.5213 (0)	0.7176 (2)	0.4039 (4)
GloVe	L	0.7693	0.6097	0.3884	0.7296	0.2183
BERT	L	0.8124 (2)	**0.7096 (1)**	**0.5191 (0)**	0.7368 (2)	**0.4027 (3)**
LXMERT	LV	0.7821 (27)	0.6000 (27)	0.4438 (21)	0.7417 (33)	0.2443 (21)
UNITER	LV	0.7679 (18)	0.6813 (2)	0.4843 (2)	0.7483 (20)	0.3926 (10)
ViLBERT	LV	0.7927 (20)	0.6204 (14)	0.4729 (16)	0.7714 (26)	0.3875 (14)
VisualBERT	LV	0.7592 (2)	0.6778 (2)	0.4797 (4)	0.7512 (20)	0.3833 (10)
Vokenization	L_V	**0.8456 (9)**	0.6818 (3)	0.4881 (9)	**0.8068 (10)**	0.3439 (9)



Intrinsic evaluation

- Pure linguistic models still better…
- Visual supervision (e.g. Vokenization) can improve performance

Source: Pezzele (2021)

Evaluation – After Transformers (4/4)

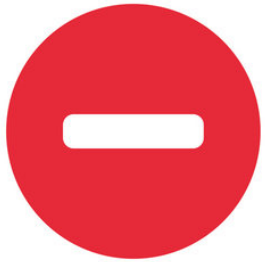
Few-shot learning in downstream tasks – low resources setting

	SST-2			QNLI			QQP			MNLI		
Extreme Few-shot	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%
$VOKEN(Bert_{base})$	54.70	77.98	80.73	50.54	51.60	61.96	44.10	60.65	65.46	37.31	54.62	58.79
$iACE(Bert_{base})$	**77.98**	**80.96**	**81.42**	**51.64**	**58.33**	**64.03**	**49.36**	**63.67**	**71.17**	**40.07**	**56.49**	**59.57**
$VOKEN(Roberta_{base})$	70.99	71.10	77.86	54.37	62.23	65.78	62.32	67.25	70.18	48.59	49.76	58.23
$iACE(Roberta_{base})$	**75.34**	**78.66**	**83.60**	**54.79**	**65.03**	**65.83**	**65.43**	**68.11**	**70.77**	**48.94**	**52.74**	**59.39**
Normal Few-shot	1%	3%	5%	1%	3%	5%	1%	3%	5%	1%	3%	5%
$VOKEN(Bert_{base})$	81.40	86.01	84.75	64.17	77.36	80.19	72.55	78.37	80.50	60.45	62.73	72.35
$iACE(Bert_{base})$	**82.45**	**87.04**	**86.47**	**65.09**	**79.54**	**80.52**	**74.31**	**78.69**	**80.52**	**62.15**	**70.43**	**73.73**
$VOKEN(Roberta_{base})$	83.78	84.08	87.61	75.00	81.16	81.23	73.14	79.09	79.63	63.51	70.68	74.02
$iACE(Roberta_{base})$	**83.83**	**84.63**	**89.11**	**79.35**	**81.41**	**81.65**	**73.72**	**79.38**	**79.81**	**65.66**	**70.76**	**74.10**

Source: Lu et al. (2022)

Was it worth? (1/2): Main take-aways

- Concrete words
- Visual supervision (Vokenizer)
- Generative models („imagination“)
- Zero-shot learning



- Abstract words
- Universal models

Was it worth? (2/2): Fields of application

Promising… but still in early development…

Machine Translation

- Ive et al. (2019): "Distilling translations with visual awareness."
- Huang et al. (2020): "Unsupervised multimodal neural machine translation with pseudo visual pivoting."
- Zhang et al. (2020): "Neural machine translation with universal visual representation."

Dialogue generation – conversational agents

- Yang et al. (2021): "Open domain dialogue generation with latent images."
- Liang et al. (2021): "Maria: A visual experience powered conversational agent."

All (relevant) models (1/3)…

Year	Paper	#citat.	(intern)	Inserted il	Fokus Vc	Language model (LM)	LM-Pretraining sources	Incorporation of visual elements (IMG)	IMG-Pretraining sources	Multimodal representation and model description	Testset/Fine-tuning	Baseline(s)/model settings/comparison to other models	Results
2014	Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. "Multimod	882		Y,s1	w	Distributional model expressed as a matrix with rows as "semantic vectors" representing the meaning of a set of target words. The model is based on co-occurrence counts of words (as a result, the matrix is a squared one)	- ukWaC, 1.9B tokens - Wackypedia, 820M tokens	(i) Local descriptors to extract low-level visual features (ii) Assign local descriptors to cluster of visual words (bag of words) to build the vector representation of an image (iii) Sum up visual words co-occurrence to across all images/instances to get co-occurrence counts related to a target word (the resulting matrix is a squared one)	ESP-Game dataset, 100K images	Only words for which there is a related image are considered. Two steps to build multimodal representation: (i) Textual and visual matrices are concatened and projected into a common latent multimodal space with a singular value decomposition. From this matrix, the "textual **mixed** matrix" and the "visual **mixed** matrix" are extracted (ii) Association between words is assessed with cosine similarity Two fusion methods to estimate similarity of pairs: - Feature level fusion: linear combination of textual and visual mixed matrix and then similarity estimation - Scoring level fusion: word similarity computed on both textual and visual mixed matrices separately and then the final score is a linear cobination of the two In both methods the weights in the linear	- WordSim353 - MEN	- Text mixed embeddings only - Visual mixed embeddings only - Equally weighted versions of fusion and scoring level fusion model settings - Several "fine tuned" versions of fusion and scoring level fusion model settings	▲Multimodal word representations enhance performance of purely textual or visual embeddings • No alternative model used as a mean of comparison
2014	Hill, Felix, and Anna Korhonen. "Learning abstract concept e	91	Y,j,e	Y, s2	w	Skipgram	400m word Text8 Corpus	Mapping of words w to a bag of perceptual features b(w), extracted from external sources and encoded in an associative array P. Generation of pseudosentences based on these perceptual features to be fed into the language model	- ESP-Game (100K images) - CSLB Property Norms	Extension of the Skipgram injecting perceptual information by generating pseudo-sentences based on a bag-of-visual-words. A hyperparameter α controls the level of perceptual information relative to linguistic input	USF Dataset	- Concatenation of linguistic and perceptual features - Canonical Correlation Analysis applied on vectors of both modalities - SVD of matrix of concatenated multimodal representations	▲Concepts, which can **directly** be represented in the perceptual modality (e.g. concrete verbs and nouns) ▲Propagation of perceptual input from concrete concepts (nouns and verbs) to enhance the representation of abstract verbs, those for which no direct representation in the visual space is available ▼Abstract nouns (for which is more difficult to find a concrete visual representation) are still more efficiently learned from language-only models
2014	Douwe Kiela and Léon Bottou. 2014. Learning Image Embed	248	c	Y,s1	w	Skipgram	- Text8 Corpus (400M words) - British National Corpus (100M words)	Seventh layer of a CNN to extract 6144-d features vectors for images, obtained in two ways: - CNN-Mean (average of all features vectors representing images) - CNN-Max (component-wise maximum of all features vectors)	- ImageNet (12.5M images) - Esp-Game (100K images)	Concatenation of visual and textual embeddings	- MEN - WordSim353 (it captures not only "relatedness" but also "similarity"	- Skipgram (text-only baseline) - Embeddings - visual only	▲CNN-Mean better on MEN: averaging might capture relatedness better. CNN-Max better on WordSim353
2015	Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "C	276	Y,j	Y,s2	w	Skipgram	Wikipedia 2009, 800M Tokens	Visual information for 5,100 words with an entry in ImageNet, occur >500 times in the text corpus and have a concreteness score ≥0.5; sample 100 images for each word and extract a 4096-d array with a CNN; average the vectors of 100 pictures associated to each word to get visual representation	ImageNet	The objective function is a linear composition of the language objective L-ling from the Skipgram and a visual objective L-vision. For the L-vision objective two variants are proposed: - MM Skipgram A (MMSA): aligning vectors of visual and linguistic representations (1:1 correspondence assumed) - MM Skipgram B (MMSB): estimate a cross-modal mapping matrix from linguistic onto visual representations	- MEN - SemSim - VisSim	- Kiela and Bottou (2014) - Bruni et al. (2014) - Silberer & Lapata (2014) - Skipgram (text-only baseline) - Embeddinggs - visual only - Concatenation - SVD	▲Both MMSA and MMSB better than simpler models (linguistic/vision only, concatenation SVD) • MMSA and MMSB competitive in relatedness and visual similarity, despite having often less training data than other models • Visual grounding less effective with abstract words

All (relevant) models (2/3)…

2017	Collell, Guillem, Ted Zhang, and Marie-Francine Moens. "Im	80	Y,j,H	Y, s2	w	300-d GloVe	Common Crawl corpus, 840B tokens, 2.2M words	To extract visual features, the last hidden layer of a CNN is taken. For each concept, two different ways to combine the extracted visual features: - Averaging (averaging of al features vectors) - Maxpooling (component-wise maximum)	Imagenet	Mapping from language to vision. No need of 1:1 correspondence between linguistic and visual inputs. Two different mappings are considered: - Linear (MAP-Clin) - Neural Network (MAP-Cnn)	- MEN - WordSim353 - SemSim - Simlex999 -SimVerb3500 - VisSim	- Kiela and Bottou (2014) - Lazaridou et al. (2015) - Silberer & Lapata (2014) - GloVe (text-only baseline) - Concatenation	▲Outperformance in all instances where words have associated images in the training set ▼Perfomance on the zero-shot learning still inferior in many instances to the textual baselines
2018	Kiela, Douwe, et al. "Learning visually grounded sentence rep	63	Y,j,H	Y, s2	s	- GloVe for word embeddings - Bidirectional LSTM for sentece representation		Image features obtained from the final layer of a ResNet-101	COCO	Word embeddings space are projected to a ground space with a linear mapping. Linear mapping and Bi-LSTM are trained jointly. 3 methods to ground sentences in images, captions or both: - Cap2Img: predict latent features of an image from its caption by mapping the (final) hidden state h(T) of the Bi-LSTM to the latent repreentation of the image. A ranking loss is to be minimized - Cap2Cap: given the caption pair (x,y) describing the same image, the goal is to maximize the joint. probability p(y|x). Negative log-likelyhood as loss. - Cap2Both: Goal is to minimize the two loss functions above In another setting, grounded and sentence-only (Skipthought) representations are concatenated with layer normalization to get the final sentence representations. Goal is to include information on less concrete concepts which are not likely to be represented in image-captioning databases but	Intrinsic evaluation of word embeddings: - MEN - SimLex 999 - Rare Words - WordSim-353 Extrinsic evaluations: '- Movie review Sentiment (MR) - Product reviews (CR) - Subjectivity classification (SUBJ) - Opinion polarity (MPQA) - Paraphrase identification (MSRP) - Sentiment classification (SST) - SNLI (Entailment) - SICK (Entailment)	- Skipthought (text-only baseline)	▲Word embeddings are of higher quality than those obtained with GloVe, measured on the following similarity benchmarks: MEN, SimLex999, Rare Words and WordSim-353 ▲In extrinsic evaluations, grounding increases performace but it is not clear which one of the three grounding strategies considered is dominant - Performance seems to be driven in a smaller amount of instances by a larger number of parameters rather than effectiveness of grounding - Performance is better when dataset have a higher level of concreteness
2020	Bordes, Patrick, et al. "Incorporating visual semantics into s	12	Y,s,j,H	Y, s3 before transformer s or s2	s	- Skipthought	Toronto Book Corpus: 11M books, 74M ordered sentences, 13 words per sentence on average	MS COCO: 118K/5K/41K (train/val/test) images		The objective function is composed of: - a textual objective Lt - a grounding objective Lg, which among ist parameters has also those of the textual objective, which in turn profit from both objective functions. Lg is not applied directly on the sentecence embeddings; it is trained on an intermediate space called the "grounded space". The sentence embeddings are projected to the gronded space with the projection function beign a multi-layer perceptron. The goal is to move away from the 1:1 correspondence between textual and visual space. Lg the can be decomposed in two components, whose individual contribution is controlled by two hyperparameters: - Cluster Information (Cg): sentences associated with the same image(s) sould be similar. The visual space is thus used to asses sentence similarity. The Max-margin ranking loss is used - Perceptual information (Pg): similarity between sentences in the grounded space should be correlated with similarity between corresponding images in the visual space. The loss is based on the negative Pearson correlation. Model scenarios include many compositions of the above mentioned elements	Intrinsic evaluation of word embeddings: - STS - SICK Extrinsic evaluations: '- Movie review Sentiment (MR) - Product reviews (CR) - Subjectivity classification (SUBJ) - Opinion polarity (MPQA) - Paraphrase identification (MSRP) - Sentiment classification (SST) - SNLI (Entailment) - SICK (Entailment)	- Skipthought (text-only baseline) For extrinsic evaluations: - Kiros et al. (2014) - Kiela et al. (2018) - Lazaridou et al. (2015) - crossmodal - Collell et al. (2017) - sequential/concatenation	▲Word embeddings are better than the textual benchmark for data with a high level of concreteness and are similar in performance with respect to more abstract concepts ▲Projections on the grounded space are more effective than cross-modal projection and concatenation - Not always best performance on entailment tasks (benchmarks SNLI, SICK)

All (relevant) models (3/3)…

2020	Tan, Hao, and Mohit Bansal. "Vokenization: Improving lang	43	Y,j,H	Y, s3		- BERT, but it can be adapted to any language model (through Revokenization)	- English Wikipedia	- ResNeXt	- MS COCO	Language model with visual supervision. Each token in a sentence obtains a corresponding image (voken) assigned from a **finite** set of images. The voken is the image which maximize a Relevance Scoring Function between a token and all images in the aforementioned fine set of images. With this token-voken pairs a voken classification pretraining task is performed that can be built in pure language models alongside other pretraining tasks such MLM or next-sentence prediction.	- GLUE (only SST-2, QNLI, QQP, MNLI) - SQuAD - SWAG	- BERT (various versions) - VilBert - VL-BERT - VisualBERT - Oscar - LXMERT	▲ Improvement over the purely self-supervised language model on multiple language tasks
2021	Hu, Ronghang, and Amanpreet Singh. "Unit: Multimodal mul	37	Y,j	Y, s3	NOT ALL DA	BERT-base with a learned task-specific vector (to capture task-specific information) as additional input, which is positioned at the beginning of the embedded token sequence	(pretrained version)	CNN (ResNet-50) to extract visual features map + transformer encoder to ecode the features map to a set of hidden states. A learned task-task specific vector (to capture task-specific information) is concatenated to the beginning of the visual feature list before entering the encoder (architecture inspired by DETR)	- MS COCO - Visual Genome	To both modalities is then applied a domain agnostic transformer architecture. As input the transformer takes the hidden states of either language or visual encoders or concatenation of both together with a task specific query embedding sequence. Self attention is applied in each layer among decoder hidden states and cross attention is applied to the encoded input modalities. The output is a set of decoded hidden states to which a task-specific head is applied (two-layer MPLP with GeLU activation and cross entropy loss). Training is done jointly on multiple tasks. At each training iteration, a task is randomly selected. Three settings (third one is the model described above): 1.) Single-task training: each model is trained separately on each task 2.) Multi-task training with separate decoders: a specific decoder for each task and jointly trained on all tasks 3.) Multi-task training with shared decoder. In this settig there are still task-specif heads for each	Extrinsic evaluation.: GLUE: - QNLI - QQP - MNLI - SST2	- BERT (text-only baseline)	▲ Model setting (1), single task training, outperforms all other settings and is comaparable to the text-only baseline ▼ PThis Model setting (3), domain-agnostic, multi-task training with shared decoder acorss modalities exhibits a lower performance compared to domain-specific transformer models like BERT, the text-only baseline
2021	Shahmohammadi, Hassan, Hendrik Lensch, and R. Harald B	5	J,H	Undecided, s	w	- GloVe, 300d, 2.2M words - fastText, 300d, 2M	pretrained	Image vectors obtained by transferring the penultimate layer of pretrained Inception-V3 trained on ImageNet. A neural network with one hidden layer and tanh activation is used to to project the image vectors into the initial hidden state of the GRUs employed in the model	- MS COCO	Given embeddings originating from a pretrained text-only model, the goal is to generate a mapping matrix M to ground word embeddings visually (the mapping matrix is used in both directions, to map text to grounded space and to map grounded embeddings back to the textual space) This is obtained by performing three different tasks: (i) Next word prediction with a GRU, given previous words in the sentecnce provided as image caption, together with the related image embedding vector (ii) Same as (i) but the sentence is provided backwards to another GRU (iii) Binary classification task if the representation of a given sentence in the grounded space obtained from (i) and (ii) matched the associated	Limited to intrinsic evaluation: - MEN - SimLex999 - Rare Words - MTurk771 - WordSim353 - SimVerb3500	- GloVe (text-only baseline) - fastText (text-only baseline) - Collell et al. (2017) - Park & Myaeng (2017) - Kiros et al. (2018) - Kiela et al. (2018)	▲ Textual baselines and related models are outperformed and the model seems to improve the textual vector space by aligning it with real-world relations from the images (similarity appears to be favoured by the model over relatedness) ▲ Embeddings related to less concrete words exhibit good quality compared to baselines
2022	Hsu, Chan-Jan, Hung-yi Lee, and Yu Tsao. "XDBERT: Distill	0	Y,j	Y,s3	w	TO BE COMPLETED						
2022	Lu, Yujie, et al. "Imagination-Augmented Natural Language Understanding." *arXiv preprint arXiv:2204.08535* (2022).	0	Y,j	Yes, s3		- BERT-base - RoBERTa	Same datasets as in VOKENIZATION paper	Same datasets as in VOKENIZATION paper	Same datasets as in VOKENIZATION paper	The framewor iACE is composed of two modules: 1.) Imagination generator G: for each text input, VQGAN generates an "imagined visual" and CLIP is used to test how the generated image corresponds to the text and encodes the text and the image in a cross-modal embedding space and	From GLUE and SWAG: - SST-2 - QNLI - QQP - MultiNLI - MRPC	- BERT (text-only baseline) - RoBERTa (text-only baseline) With and w/o Vokenization	▲ Better performance of iACE over visually supervised transformers (VOKEN) in **all** instances of few-shots learning. Imagination can help existing language models to perform better in a setting with small training set (which means "less human annotated data").

The End

Image from *Ex Machina* (2014)

References (1/3)

- Agirre, Eneko, et al. "A study on similarity and relatedness using distributional and wordnet-based approaches." (2009).
- Bordes, Patrick, et al. "Incorporating visual semantics into sentence representations within a grounded space." arXiv preprint arXiv:2002.02734 (2020).
- Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. "Multimodal distributional semantics." Journal of artificial intelligence research 49 (2014): 1-47.
- Collell, Guillem, Ted Zhang, and Marie-Francine Moens. "Imagined visual representations as multimodal embeddings." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017.
- Crowson, Katherine, et al. "Vqgan-clip: Open domain image generation and editing with natural language guidance." arXiv preprint arXiv:2204.08583 (2022).
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
- Harnad, Stevan. "The symbol grounding problem." Physica D: Nonlinear Phenomena 42.1-3 (1990): 335-346.
- Hill, Felix, and Anna Korhonen. "Learning abstract concept embeddings from multi-modal data: Since you probably can’t see what I mean." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.
- Hill, Felix, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." Computational Linguistics 41.4 (2015): 665-695.
- Hu, Ronghang, and Amanpreet Singh. "Unit: Multimodal multitask learning with a unified transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

References (2/3)

- Huang, Po-Yao, et al. "Unsupervised multimodal neural machine translation with pseudo visual pivoting." arXiv preprint arXiv:2005.03119 (2020).
- Ive, Julia, Pranava Madhyastha, and Lucia Specia. "Distilling translations with visual awareness." arXiv preprint arXiv:1906.07701 (2019).
- Kiela, Douwe, et al. "Learning visually grounded sentence representations." arXiv preprint arXiv:1707.06320 (2017).
- Kiela, Douwe, et al. "Learning visually grounded sentence representations." arXiv preprint arXiv:1707.06320 (2017).
- Kiela, Douwe, and Léon Bottou. "Learning image embeddings using convolutional neural networks for improved multi-modal semantics." Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP). 2014.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).
- Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. "Combining language and vision with a multimodal skip-gram model." arXiv preprint arXiv:1501.02598 (2015).
- Liang, Zujie, et al. "Maria: A visual experience powered conversational agent." arXiv preprint arXiv:2105.13073 (2021).
- Lu, Yujie, et al. "Imagination-Augmented Natural Language Understanding." arXiv preprint arXiv:2204.08535 (2022).
- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

References (3/3)

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
- Pezzelle, Sandro, Ece Takmaz, and Raquel Fernández. "Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation." Transactions of the Association for Computational Linguistics 9 (2021): 1563-1579.
- Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).
- Silberer, Carina, and Mirella Lapata. "Learning grounded meaning representations with autoencoders." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.
- Tan, Hao, and Mohit Bansal. "Vokenization: Improving language understanding with contextualized, visual-grounded supervision." arXiv preprint arXiv:2010.06775 (2020).
- Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
- Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).
- Yang, Ze, et al. "Open domain dialogue generation with latent images." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 16. 2021.
- Zellers, Rowan, et al. "Swag: A large-scale adversarial dataset for grounded commonsense inference." arXiv preprint arXiv:1808.05326 (2018).
- Zhang, Zhuosheng, et al. "Neural machine translation with universal visual representation." International Conference on Learning Representations. 2019.