NETWORKS FOR OTHER DATA TYPES

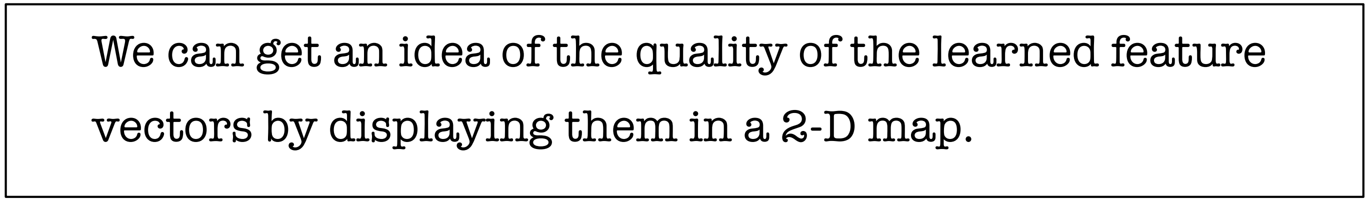
Various types of data

- 2D signals, fixed size, e.g., images
  - Grayscale image → 2-tensor
  - RGB image → 3-tensor
  - 2D CNNs
- 1D signals (temporal signals), variable length, e.g., acoustic signals
  - 1D CNNs
- 3D data, fixed size, e.g., video clip, CT images
  - 3-tensor
  - 3D CNNs (3D convolution)
- Sequential data
  - Sentence = Sequence of words
- Graphs
- Sets (of elements)
  - Order-less

- CNNs---1D, 2D, 3D
  - Applicable also to variable size input
- RNNs
  - Designed for variable length sequence
  - LSTM/Gated RNN
  - Autoregressive model
- Attention mechanisms
  - Transformer
    - Natural input type is a set
    - Applicable to sequential data
- Nets for graphs
  - Graph convolutional networks
- Nets for sets
  - PointNet/Deep sets

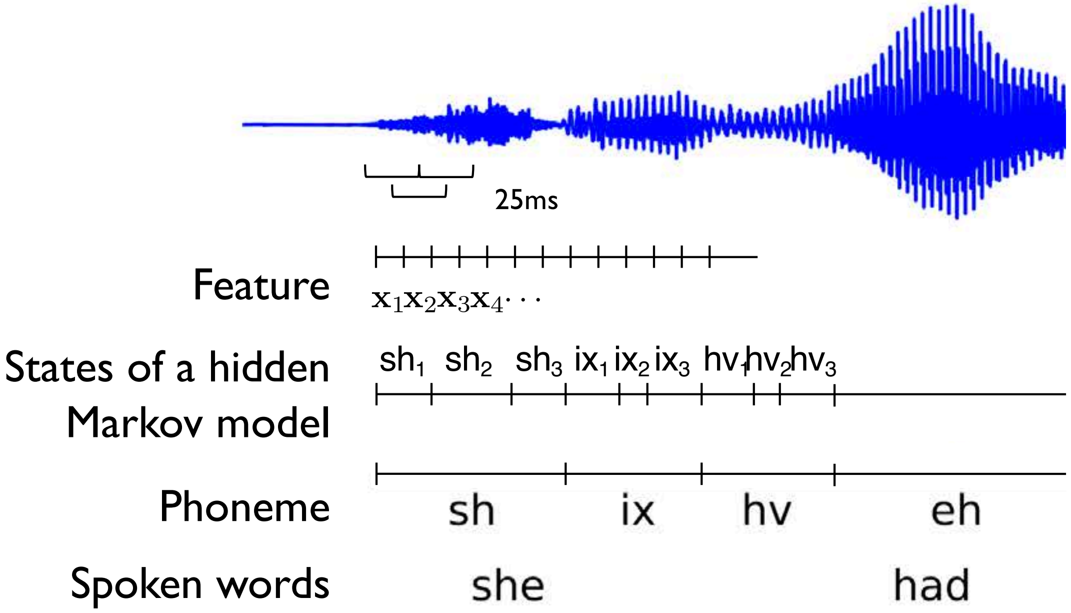
Sequential data

- A sequence of something that can have a *variable length*
  - One sequence is treated as a sample
- E.g., A *sentence*, which consists of words; there is an order in them



We can get an idea of the quality of the learned feature vectors by displaying them in a 2-D map.

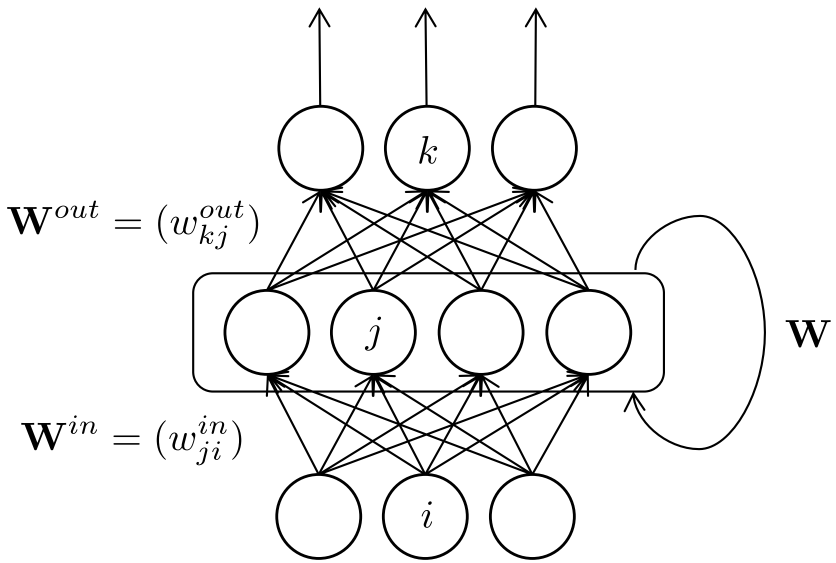
- E.g., Acoustic signals
  - sampled at a fixed frequency; the sampled values are quantized



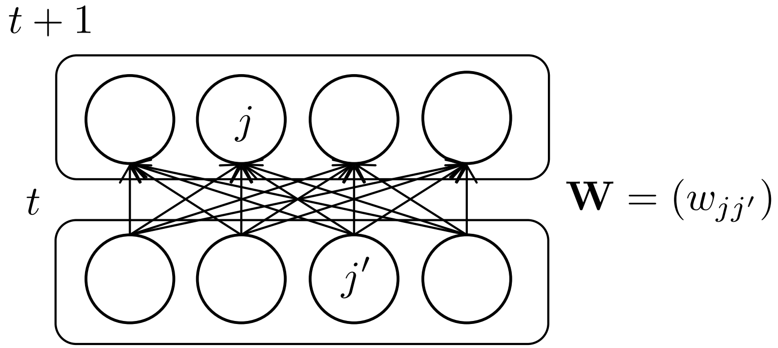
25ms
Feature
$\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3\mathbf{x}_4\cdots$
States of a hidden Markov model
sh_1 sh_2 sh_3 ix_1 ix_2 ix_3 hv_1 hv_2 hv_3
Phoneme
sh
ix
hv
eh
Spoken words
she
had

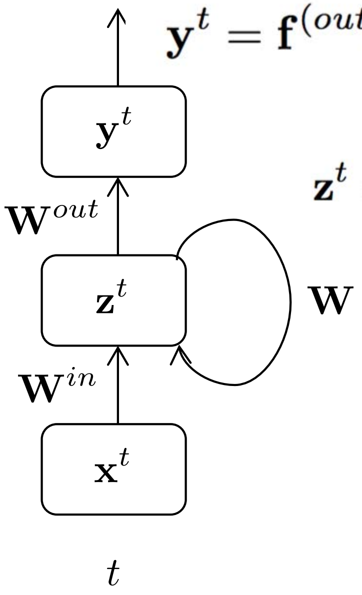
Recurrent Neural Networks (RNNs)

- Notion of time step *t*
- At each time step *t*, **x** is input to the net
- The net output **y** at *t*
- Memorize the activation at its intermediate layer(s) and transfer to the next time step *t*+1



k
$\mathbf{W}^{out} = (w_{kj}^{out})$
j
$\mathbf{W}$
$\mathbf{W}^{in} = (w_{ji}^{in})$
i



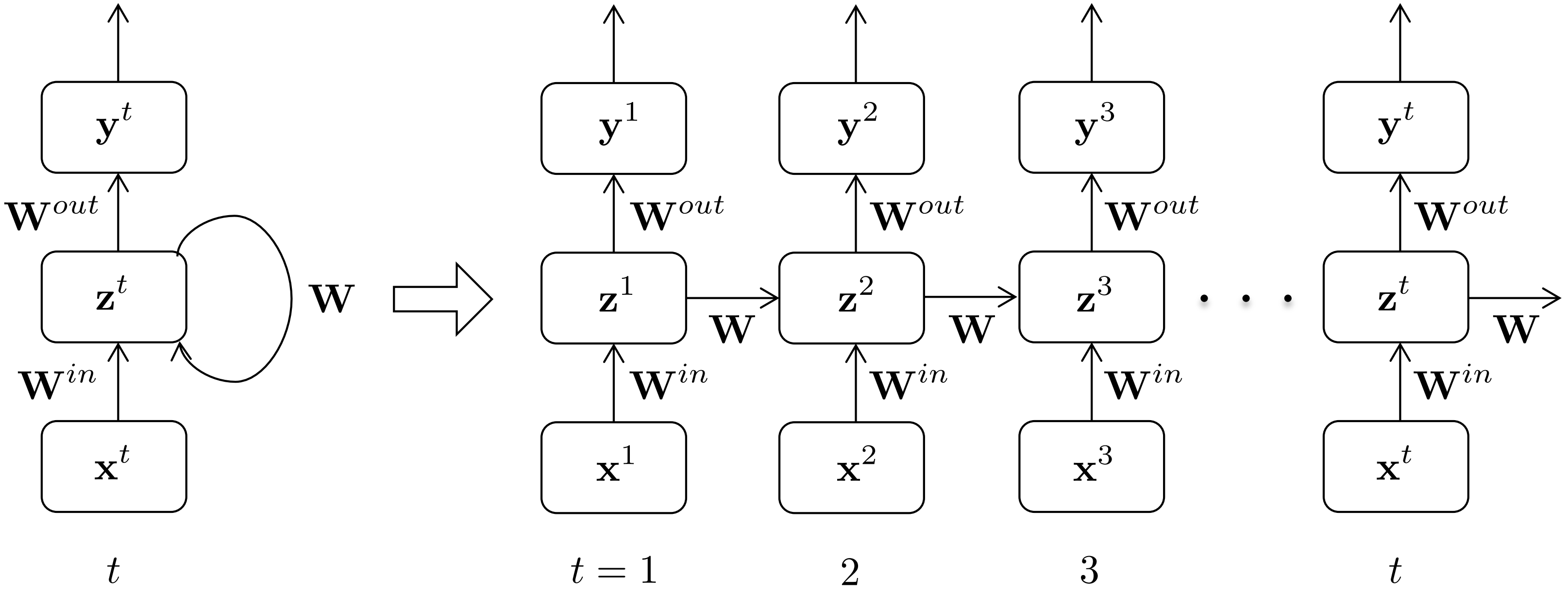


$$\mathbf{y}^t = \mathbf{f}^{(out)}(\mathbf{v}) = \mathbf{f}^{(out)}(\mathbf{W}^{(out)}\mathbf{z}^t)$$

$$\mathbf{z}^t = \mathbf{f}(\mathbf{W}^{(in)}\mathbf{x}^t + \mathbf{W}\mathbf{z}^{t-1})$$

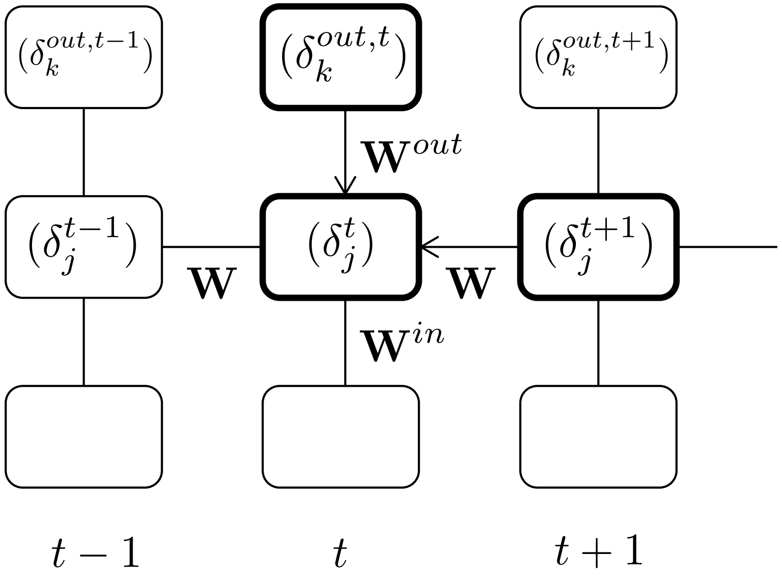
Expanding an RNN in the temporal direction

- RNNs are nothing but deep feed-forward networks



Computing gradients (deltas) for RNNs

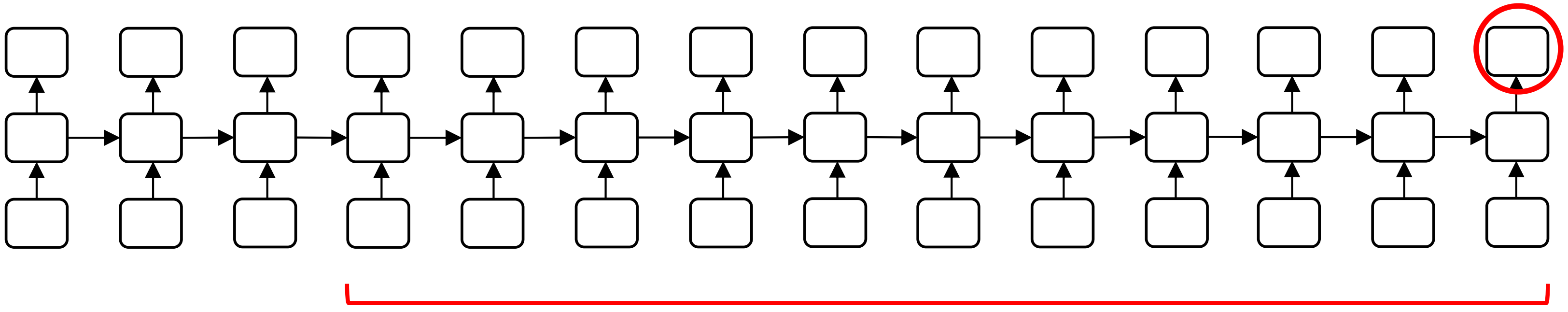
- Basically the same as in FF nets
  - Back propagation of δ's is given as follows:
  - Called BPTT (Back Propagation Through Time)



$$\delta_j^t = \left(\sum_k w_{kj}^{out} \delta_k^{out,t} + \sum_{j'} w_{j'j} \delta_{j'}^{t+1} \right) f'(u_j^t)$$

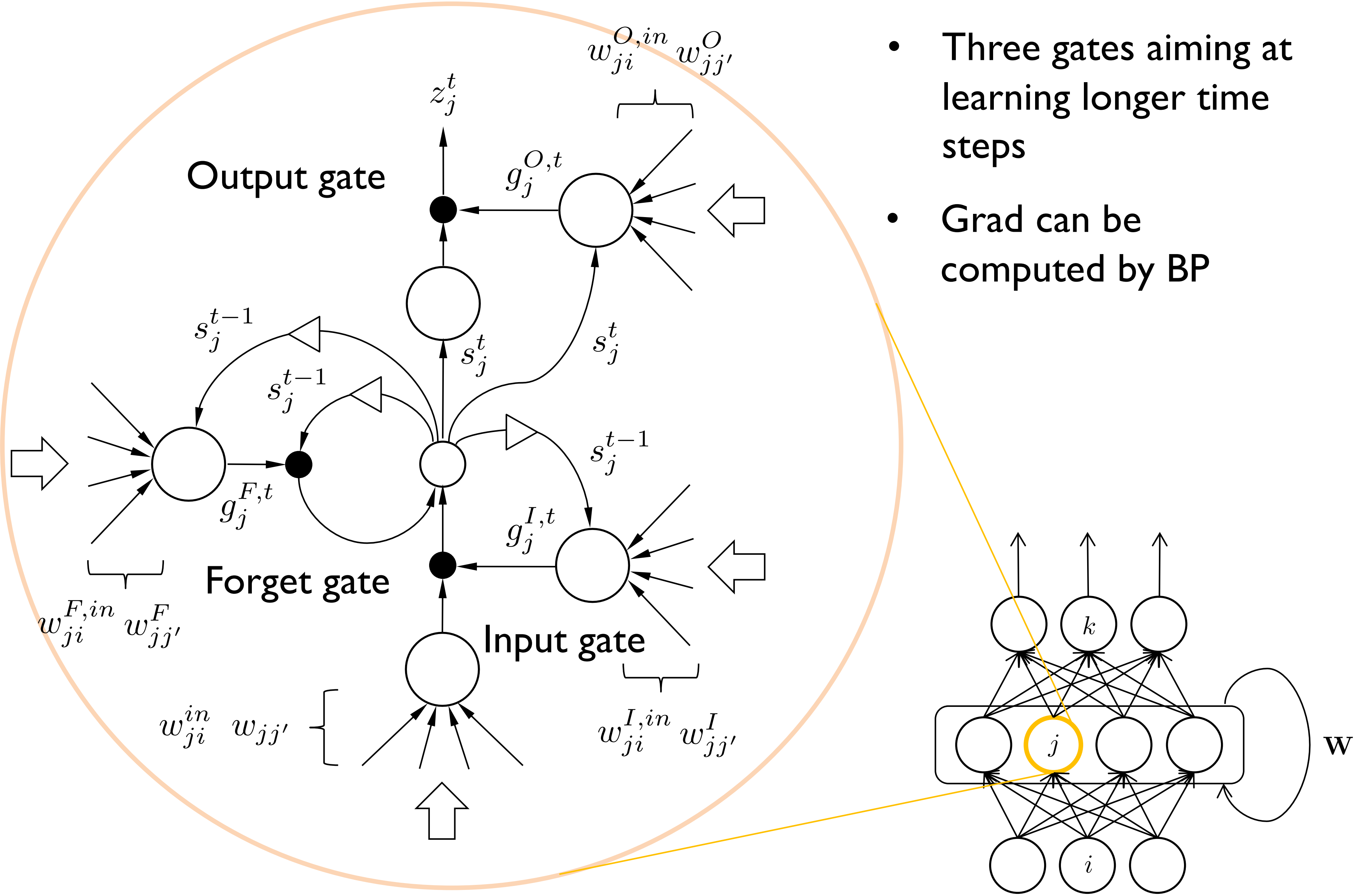
RNNs and gradient vanishing problem

- RNNs are developed in 80-90's;
- They are inherently deep nets!
  - Researchers already faced the problem in those days
  - Maximum number of layers such that training is manageable = the length of sequence that can be learned effectively
  - It is empirically known to be at most 10 steps



How far δ's can survive without vanishing?
= Num of steps that can affect the latest output

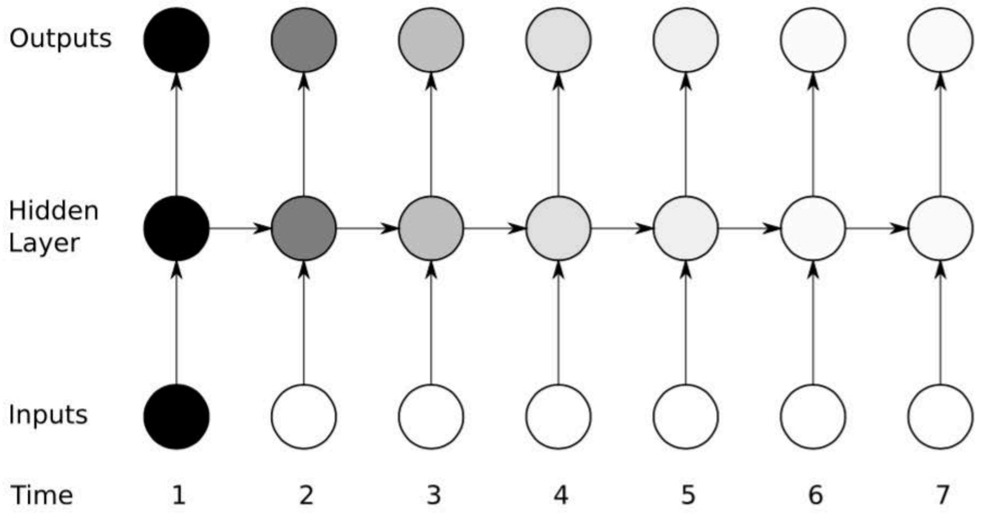
LSTM：Long Short-Term Memory



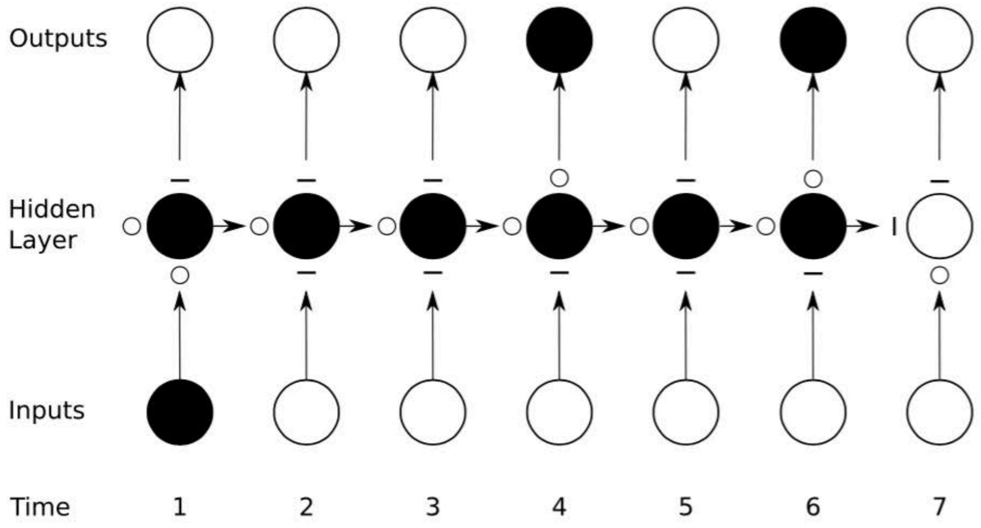
- Three gates aiming at learning longer time steps
- Grad can be computed by BP

LSTM：Long Short-Term Memory

RNN



LSTM



LSTM can learn to use a longer *context*

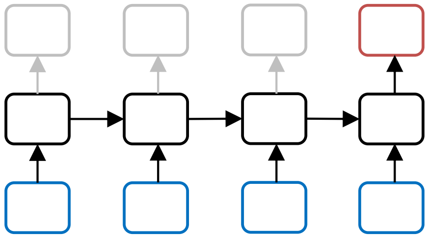
From Graves, Supervised Sequence Labelling with Recurrent Neural Networks

Applications of RNNs

M inputs / 1 output

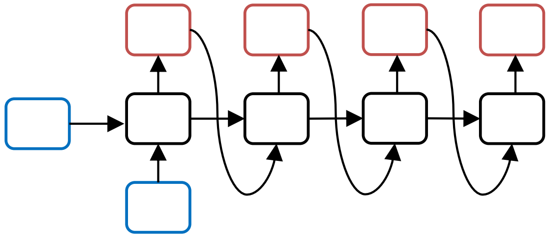
E.g. Sentence classification

In: *"They have the best happy hours, the food is good, and service is even better."*
Out: 4 star



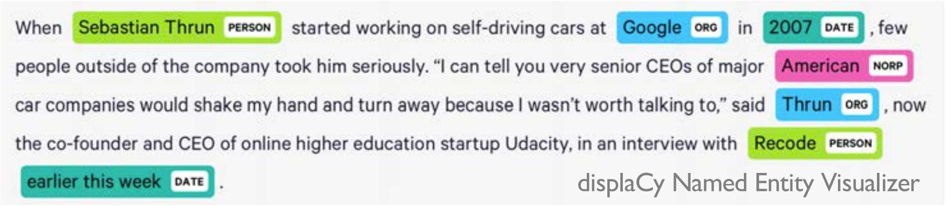
1 inputs / *M* output

E.g. Speech synthesis



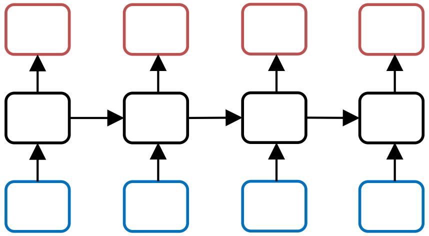
M inputs / *M* output

E.g. Sentence tagging



When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun ORG , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode PERSON earlier this week DATE .

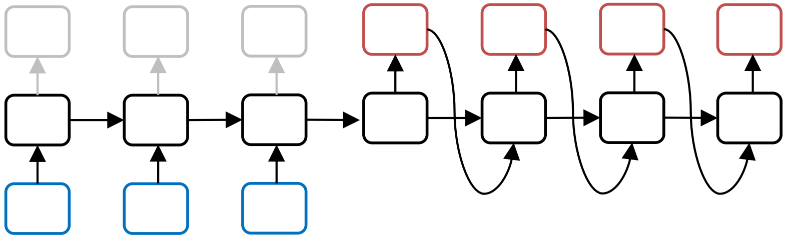
displaCy Named Entity Visualizer



M inputs / *N* output

E.g. Machine translation

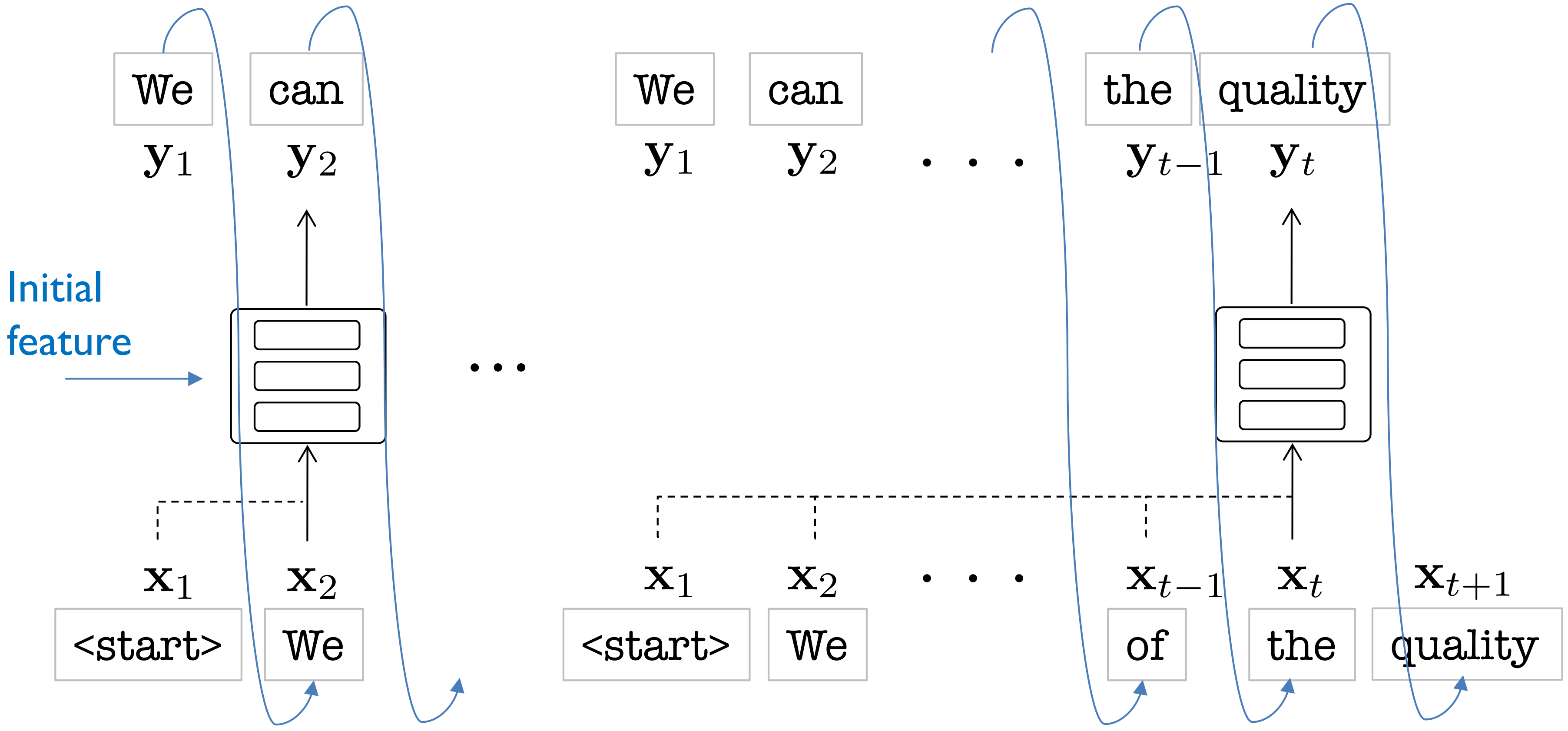
In: *"They have the best happy hours, …"*
Out: "Ils ont les meilleurs happy hours, …"



Autoregressive model

- The output at *t*–1 is used as input at *t*
- An example: Language models
  - The net generates a sentence that matches a given initial condition

Initial feature at the hidden layer or a fist few words etc.

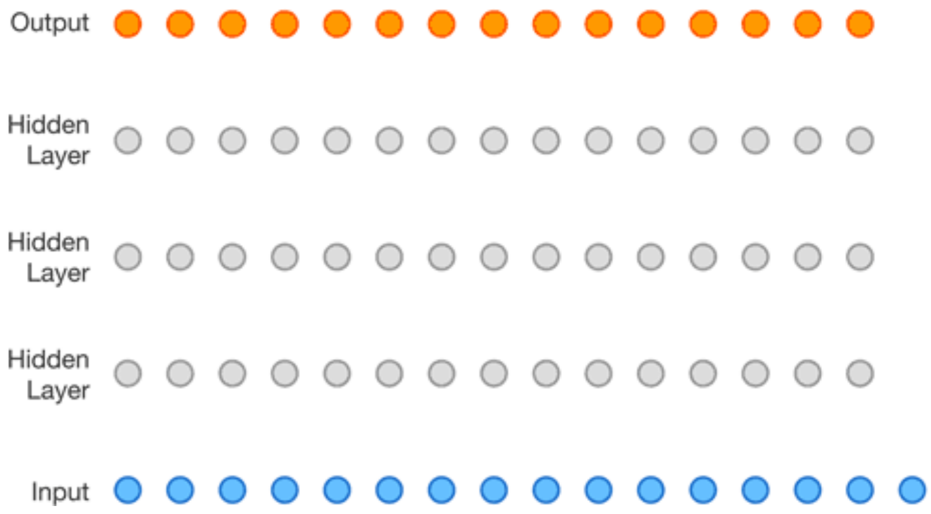


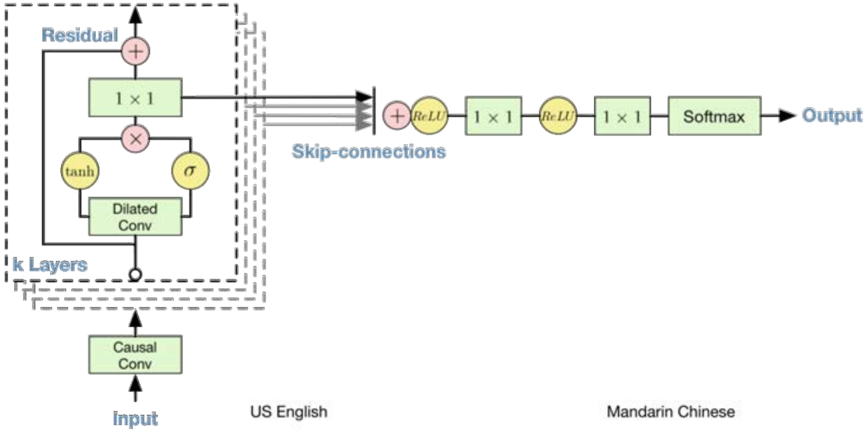
WaveNet

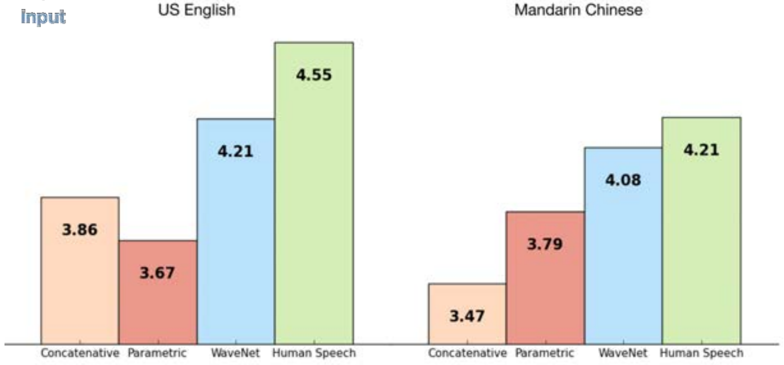
van den Oord+, WAVENET: A GENERATIVE MODEL FOR RAW AUDIO, 2016

- Raw audio signal (sampled/digitized) input to the net
- Dilated conv. / residual connect / gated activation func.
- What and how to speak is controlled by additional input **h**
  - linguistic feature: phone identities, syllable stress, # of syllables

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h})$$



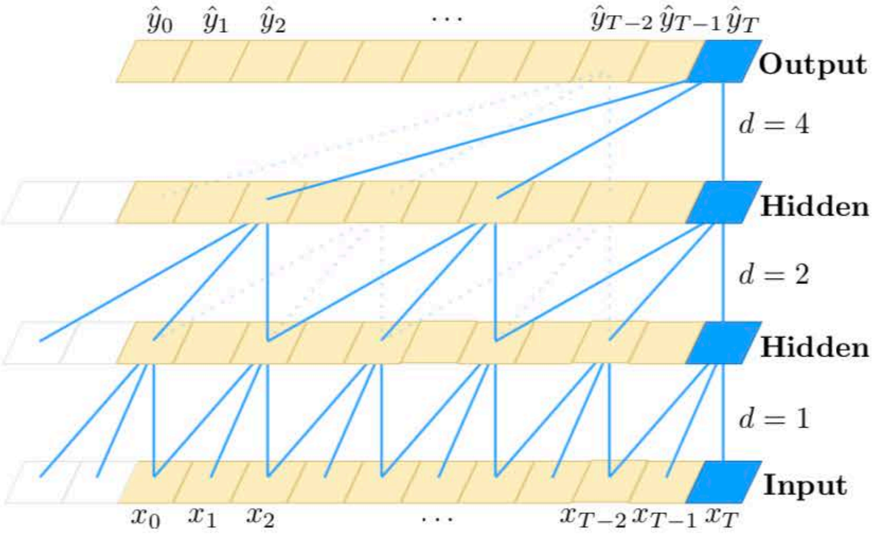


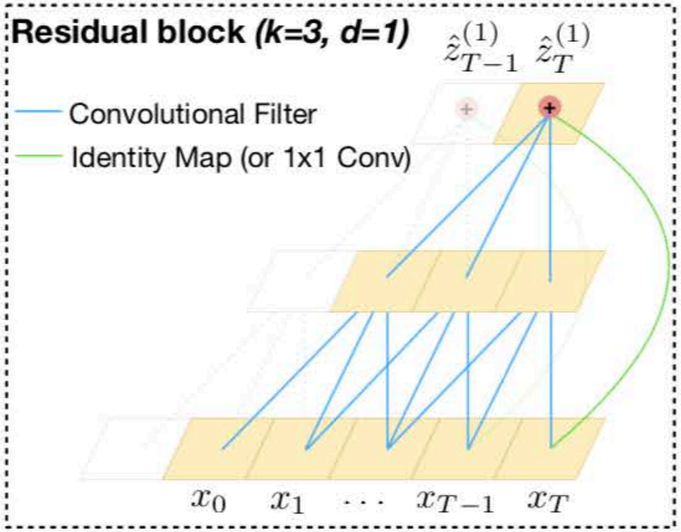


https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

Temporal Convolutional Networks (TCNs)

- Standard 1D CNN w/ modern components performs better than RNNs
  - Dilated convolution
  - Residual connection
- Difference from RNN
  - TCNs can deal with only a finite length of input history
  - Can be used in parallel fashion for training and inference

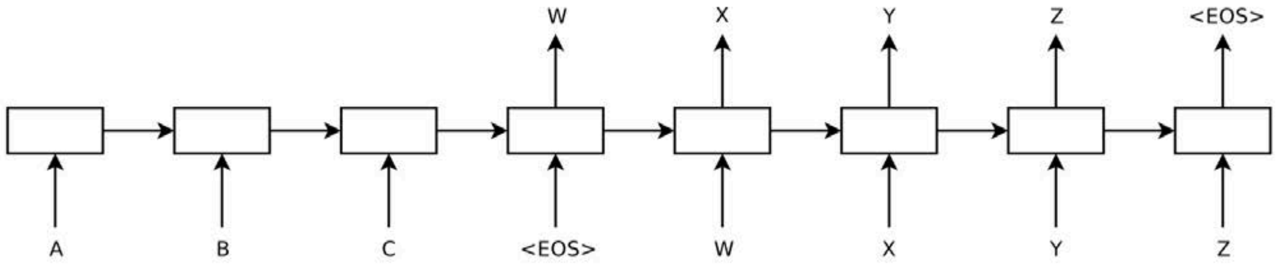

$\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ … $\hat{y}_{T-2}$ $\hat{y}_{T-1}$ $\hat{y}_T$
Output
$d = 4$
Hidden
$d = 2$
Hidden
$d = 1$
Input
x_0 x_1 x_2 … x_{T-2} x_{T-1} x_T


Residual block (k=3, d=1)
$\hat{z}^{(1)}_{T-1}$ $\hat{z}^{(1)}_{T}$
Convolutional Filter
Identity Map (or 1x1 Conv)
x_0 x_1 … x_{T-1} x_T

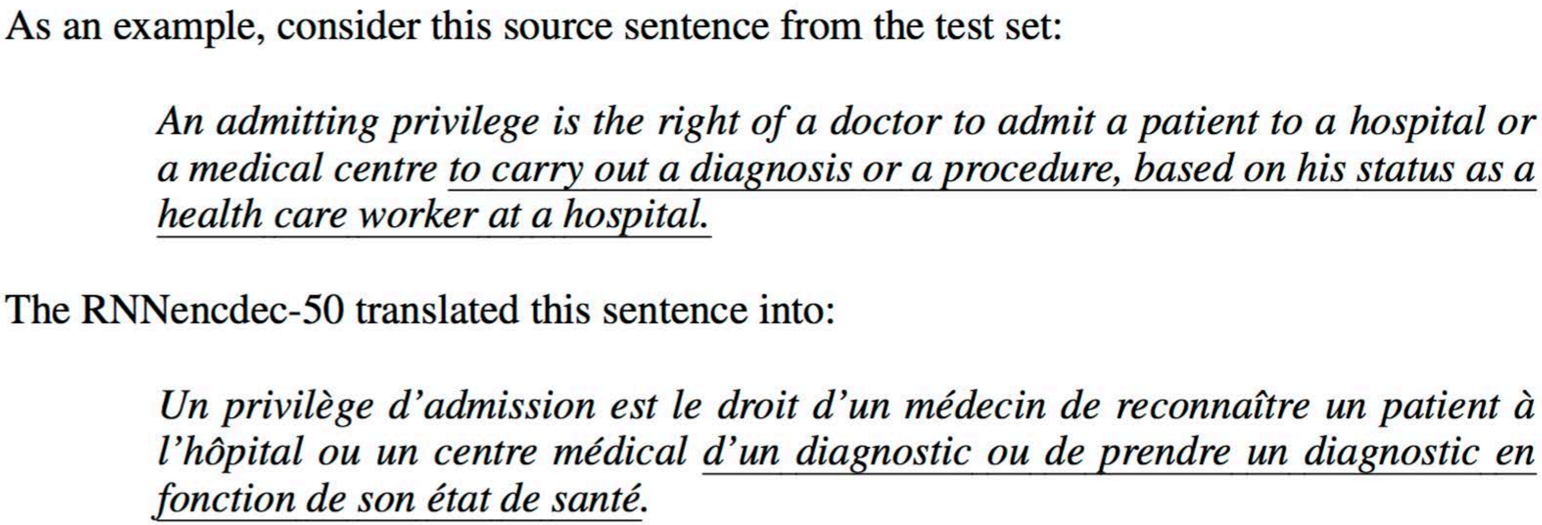
Bai-Kolter-Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, arXiv2018

Neural machine translation (NMT), 1st generation

- *Sequence-to-sequence* (Seq2seq) model
- Generate a target sentence in an autoregressive way



RNN with a hidden layer having 1000 units learns to translate 50 words



As an example, consider this source sentence from the test set:

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

The RNNencdec-50 translated this sentence into:

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

Sutskever+, Sequence to Sequence Learning with Neural Networks, 2014

Attention: General idea

- Weighting a set of entities depending on their importance
  - E.g., Words in a sentence

We can get an idea of the quality of the learned feature vectors by displaying them in a 2-D map.

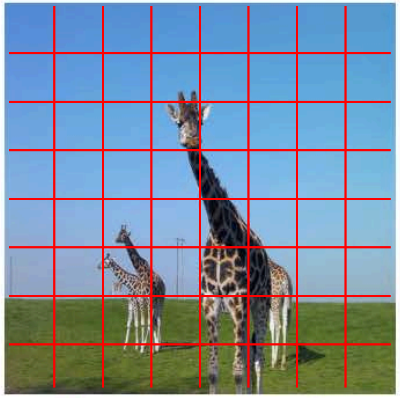
Query: What can we get?

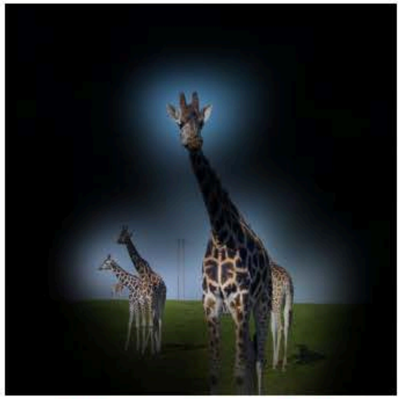
We can get an idea of the quality of the learned feature vectors by displaying them in a 2-D map.

Query: How do we get?

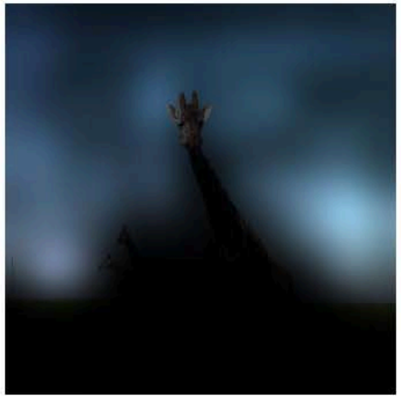
We can get an idea of the quality of the learned feature vectors by displaying them in a 2-D map.

  - E.g., Regions in an image





Query: What is this animal?



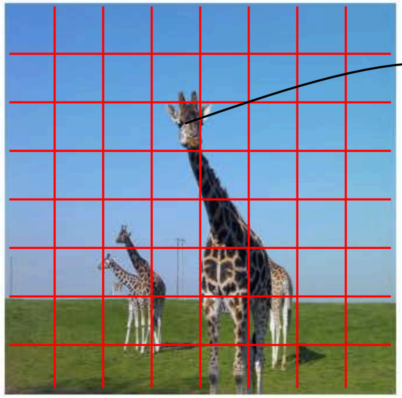
Is it cloudy?

Attention: Computation

- Relevance between query (target) feature & source feature
- Weighted average of source features = attended feature

input | feature | attention weights | attended feature

'source'

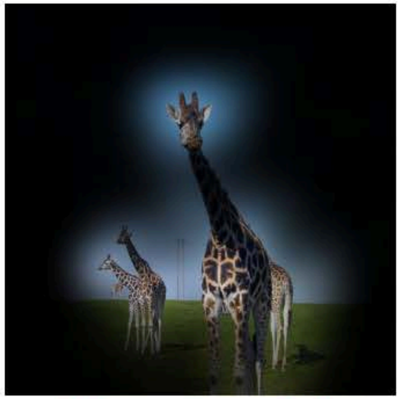


$\mathbf{y}_i$
$(i=1,\ldots,N)$

$\textit{relevance}(\mathbf{q}, \mathbf{y}_i)$

'query' (*'target'*)

What is this animal? → $\mathbf{q}$



$$a_i = \frac{\textit{relevance}(\mathbf{q}, \mathbf{y}_i)}{\Sigma_i\, \textit{relevance}(\mathbf{q}, \mathbf{y}_i)}$$

$\mathbf{q}^T\mathbf{y}_i$ (inner product) is the most popular

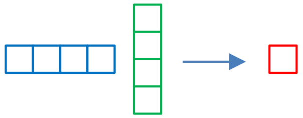
$\mathbf{y} = \Sigma_i\, a_i\, \mathbf{y}_i$

compact representation of the *query-related contents* of the image

Attention: Standard implementation

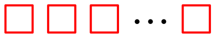
- Use inner product for *relevance* and softmax for the normalization

$$relevance(\mathbf{q}, \mathbf{y}_i) \equiv \mathbf{q}^\top \mathbf{y}_i$$



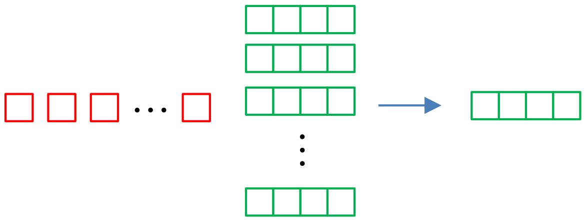
  - Normalize weights with softmax

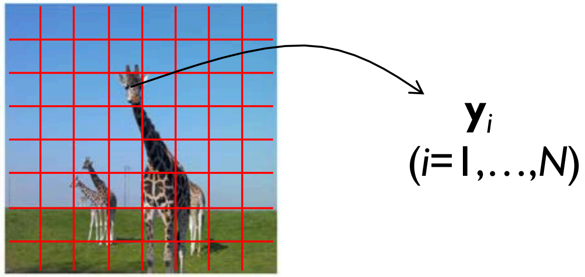
$$a_i \equiv \mathrm{softmax}_i(\mathbf{q}^\top [\mathbf{y}_1, \cdots, \mathbf{y}_N]) = \frac{\exp(\mathbf{q}^\top \mathbf{y}_i)}{\sum_{i=1}^{N} \exp(\mathbf{q}^\top \mathbf{y}_i)}$$



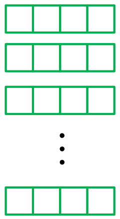
- Attended feature is written as

$$\mathbf{y} \equiv \sum_{i=1}^{N} a_i \mathbf{y}_i \overset{\top}{\rightarrow} \mathrm{softmax}\left(\mathbf{q}^\top \mathbf{Y}^\top\right) \mathbf{Y}$$





$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_N^\top \end{bmatrix}$$



Transformer

Vaswani+, Attention Is All You Need, 2017

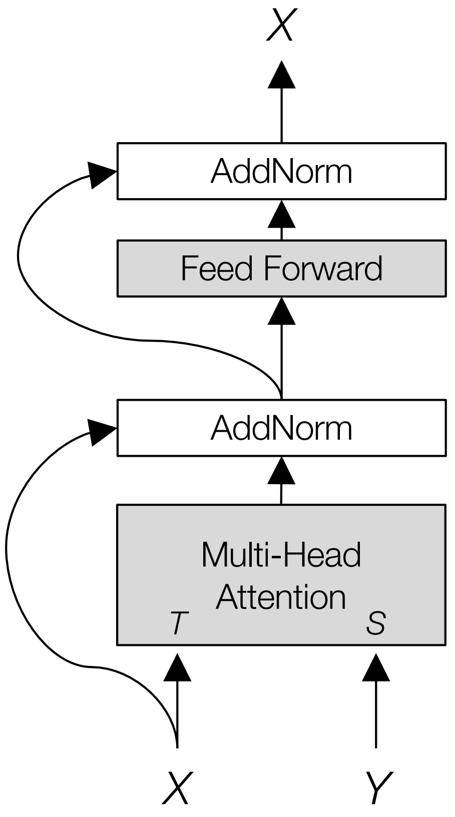
- Insert 3 *d*×*d* weight matrix

$$\mathrm{softmax}\left(\mathbf{q}^\top\mathbf{Y}^\top\right)\mathbf{Y} \rightarrow \mathrm{softmax}\left((\mathbf{q}\mathbf{W}_1)^\top(\mathbf{Y}\mathbf{W}_2)^\top\right)\mathbf{Y}\mathbf{W}_3$$

- 'Multi-head' attention
  - Use multiple (~ 10) sets of the above three matrices
  - To deal with multiple attention maps at the same time

- First applied to NMT and then to many NLP tasks
  - Self-attention: X = Y

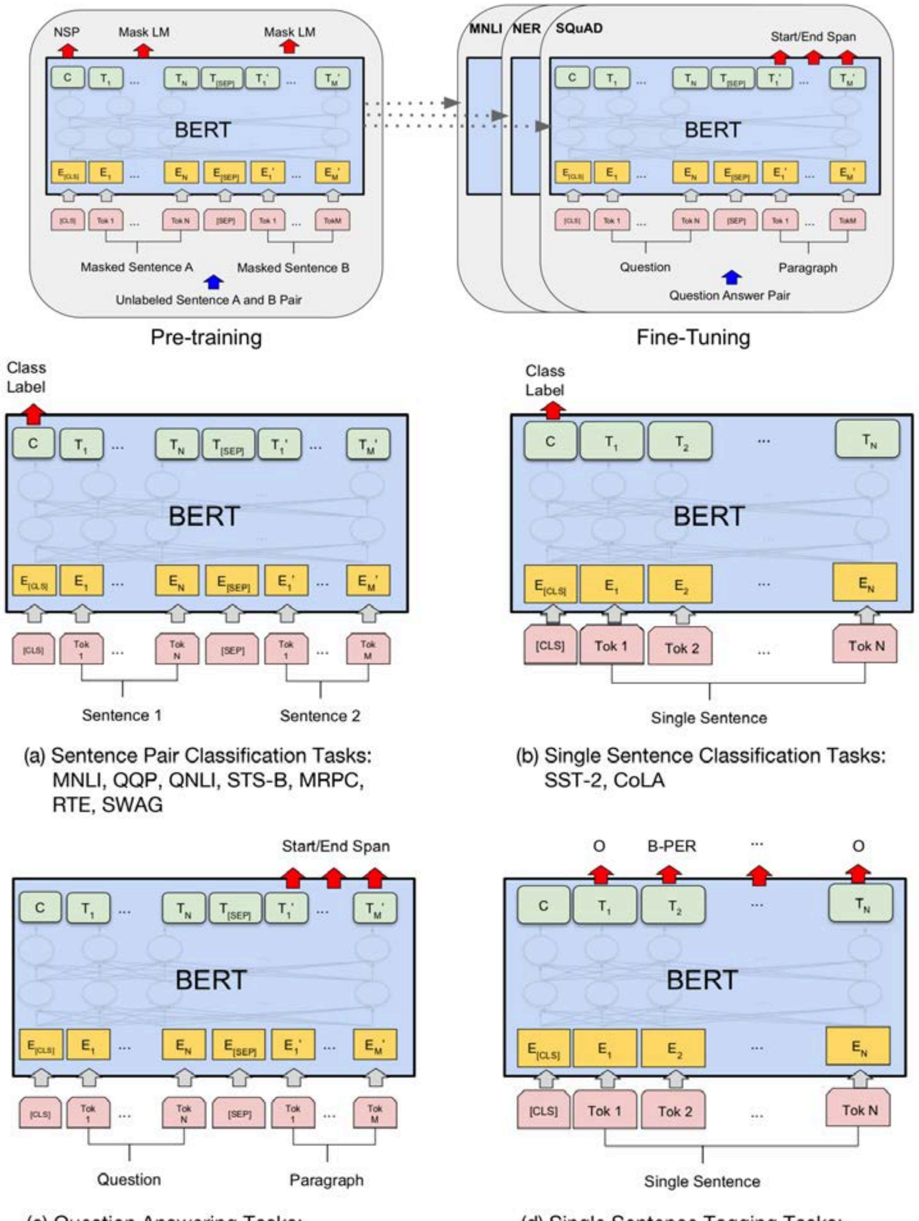
$$\mathbf{X} = \begin{bmatrix} \mathbf{q}_1^\top \\ \vdots \\ \mathbf{q}_N^\top \end{bmatrix}$$



Transformer extended to bi-modal tasks

BERT: Self-supervised learning of transformers

Devlin+, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018



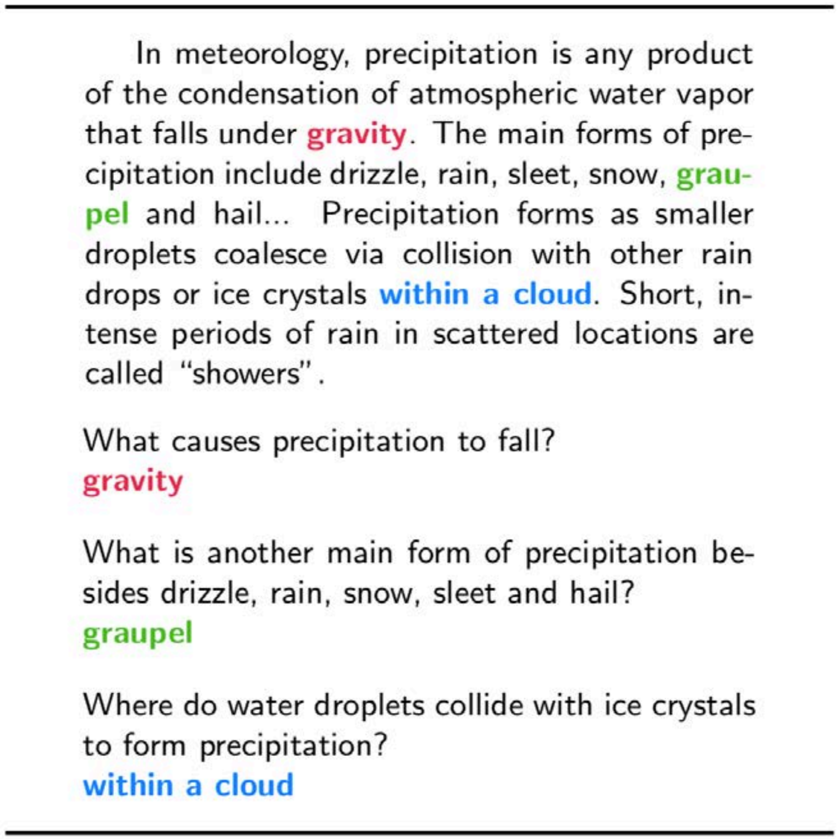
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Self-supervised learning works greatly for NLP tasks
(pre-training on proxy-tasks → fine-tuning on target tasks)



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

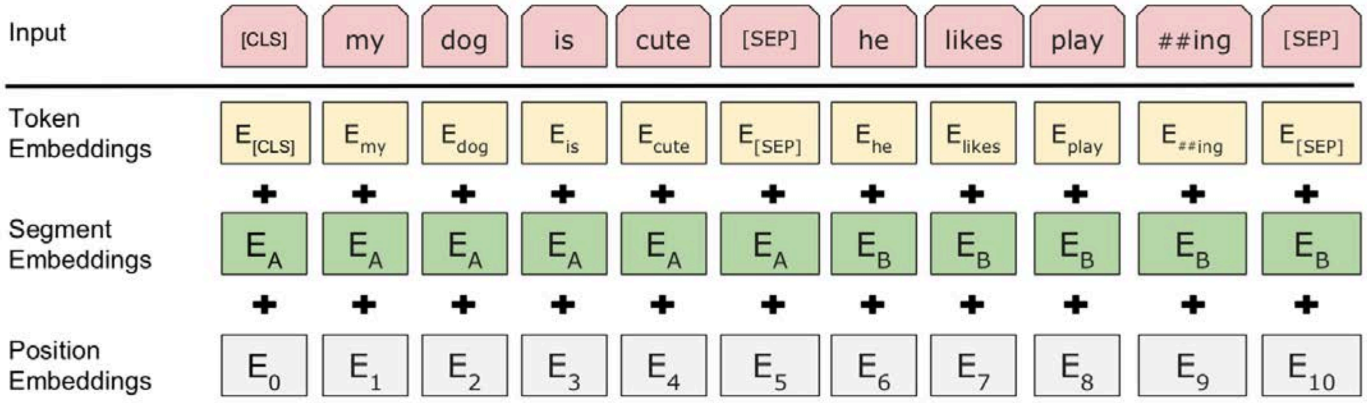
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Rajpurkar+, SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016

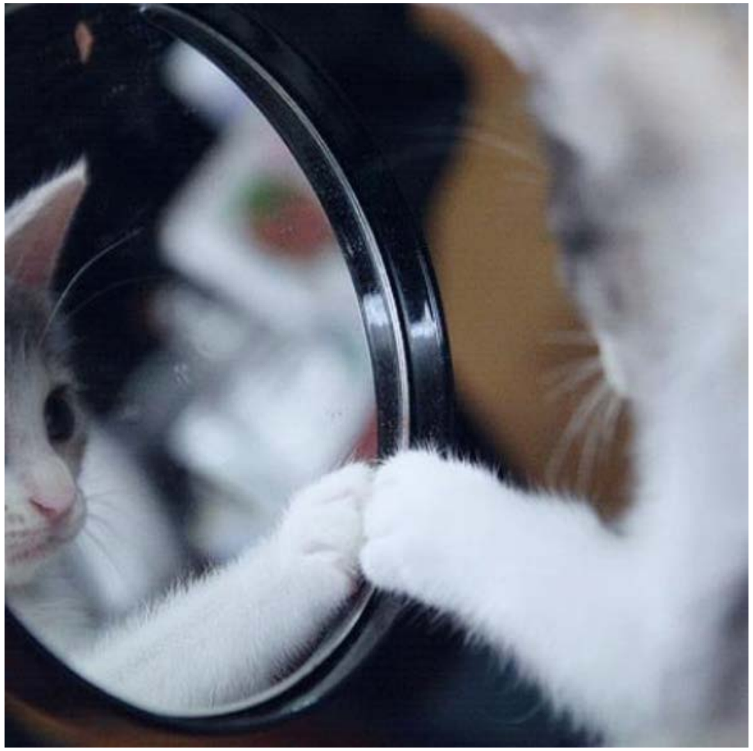
From *set* to *sequence*: Positional encoding

- When applying Transformer (self-attention) to sequential data, the order of inputs does not matter
  - If you change the order of words in a sentence, the output will not change
  - Thus, the relative position of each word in a sentence is encoded and added to its feature

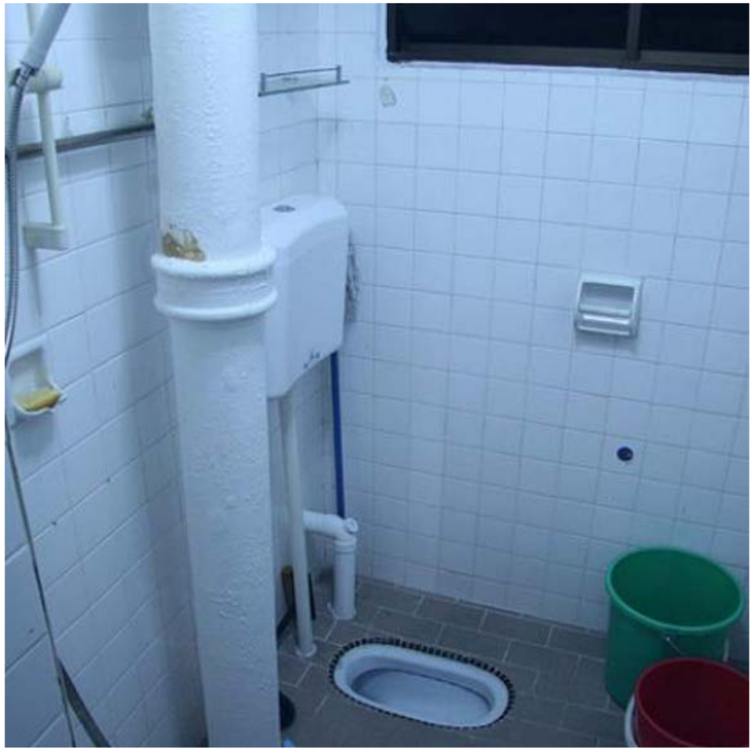


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Visual Question Answering



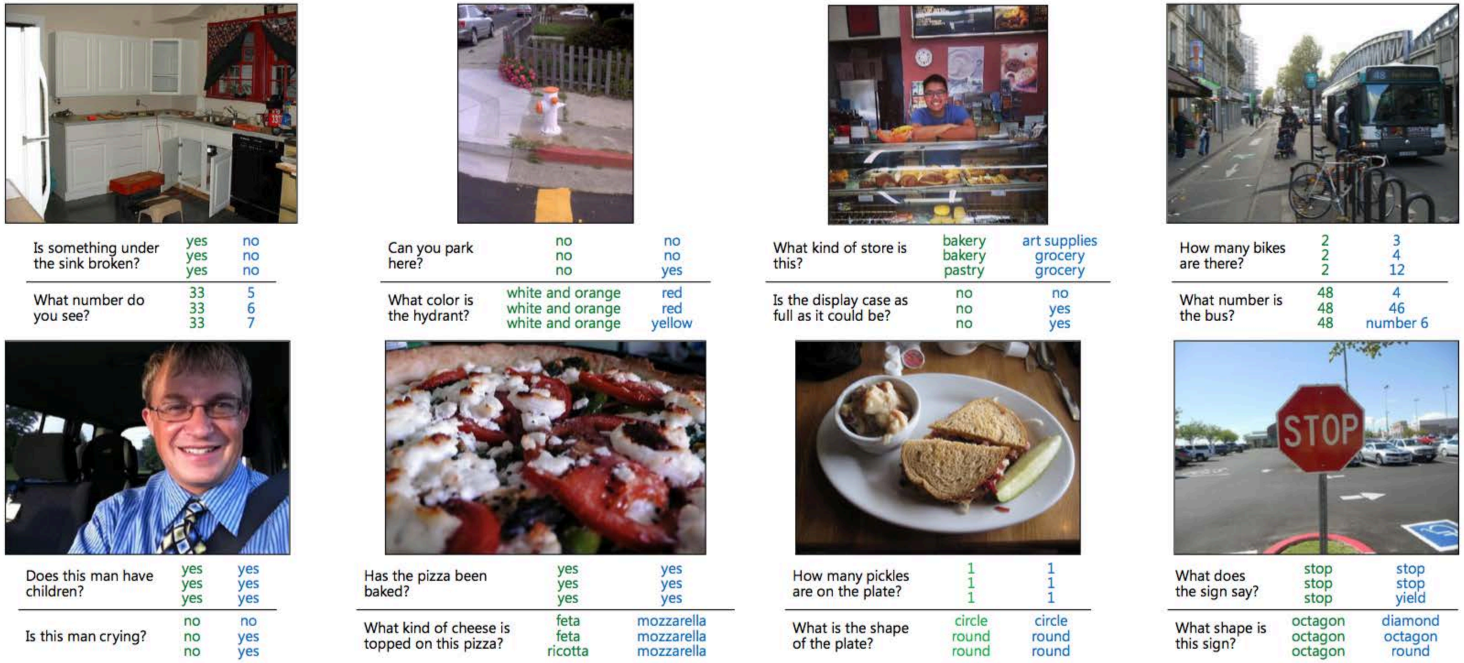
Q: What is reflected in the mirror?
A: Cat



Q: What room is this?
A: Bathroom

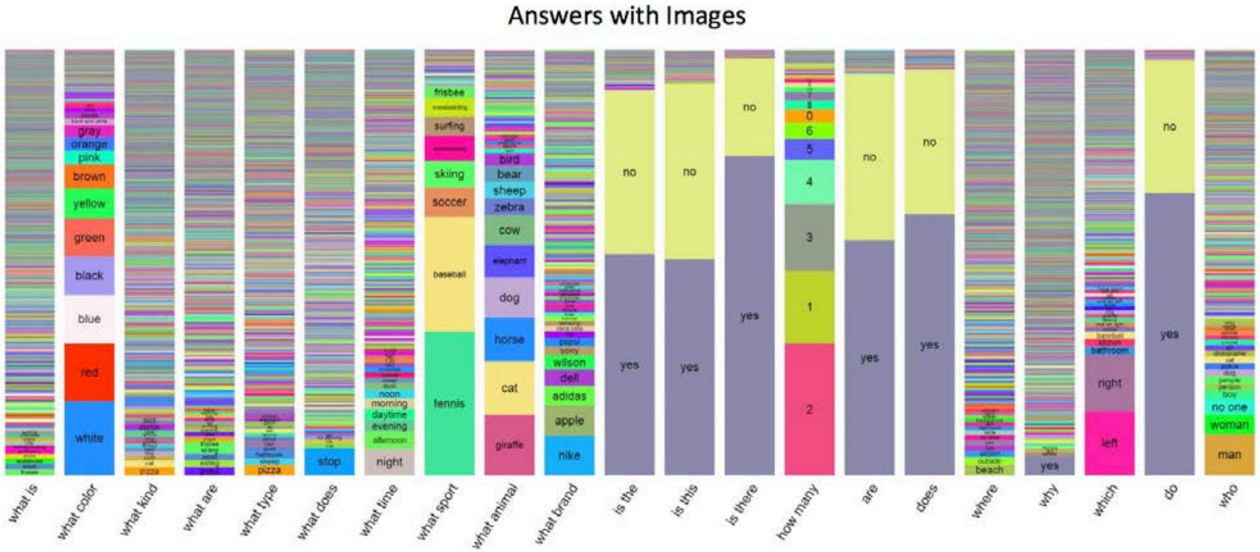
Training data (VQA-1.0/2.0)

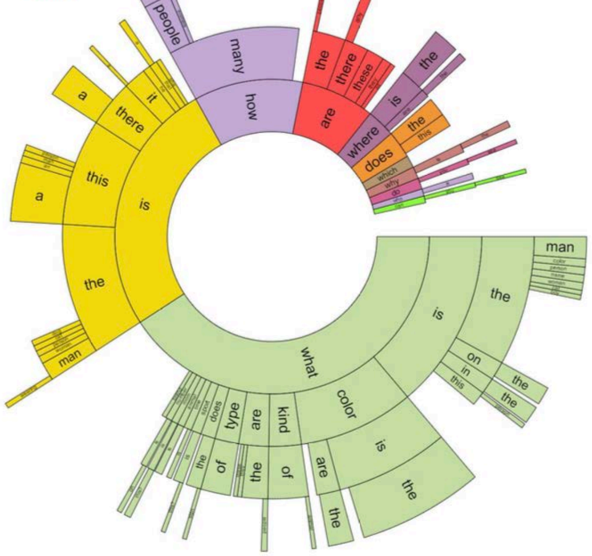
Agrawal+, VQA: Visual Question Answering, ICCV2015



Is something under the sink broken?	yes yes yes	no no no
What number do you see?	33 33 33	5 6 7
Can you park here?	no no no	no no yes
What color is the hydrant?	white and orange white and orange white and orange	red red yellow
What kind of store is this?	bakery bakery pastry	art supplies grocery grocery
Is the display case as full as it could be?	no no no	no yes yes
How many bikes are there?	2 2 2	3 4 12
What number is the bus?	48 48 48	4 46 number 6
Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes
Has the pizza been baked?	yes yes yes	yes yes yes
What kind of cheese is topped on this pizza?	feta feta ricotta	mozzarella mozzarella mozzarella
How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round
What does the sign say?	stop stop stop	stop stop yield
What shape is this sign?	octagon octagon octagon	diamond octagon round

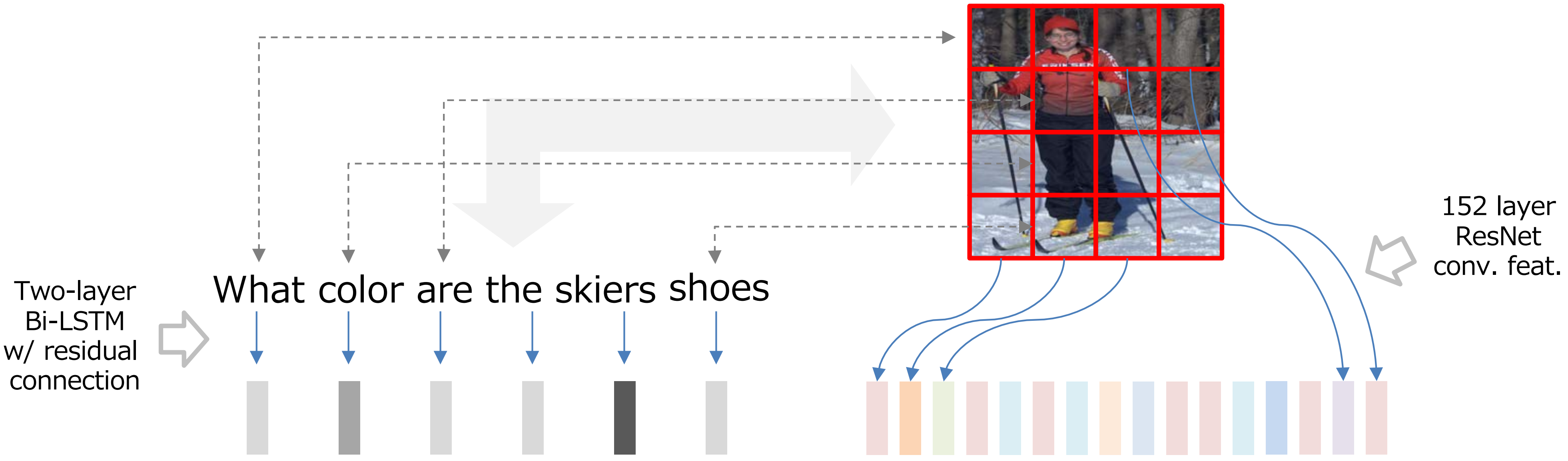
0.2 mil. Images
0.6 mil. Qs
6.1mil As





Attention for vision-language representation

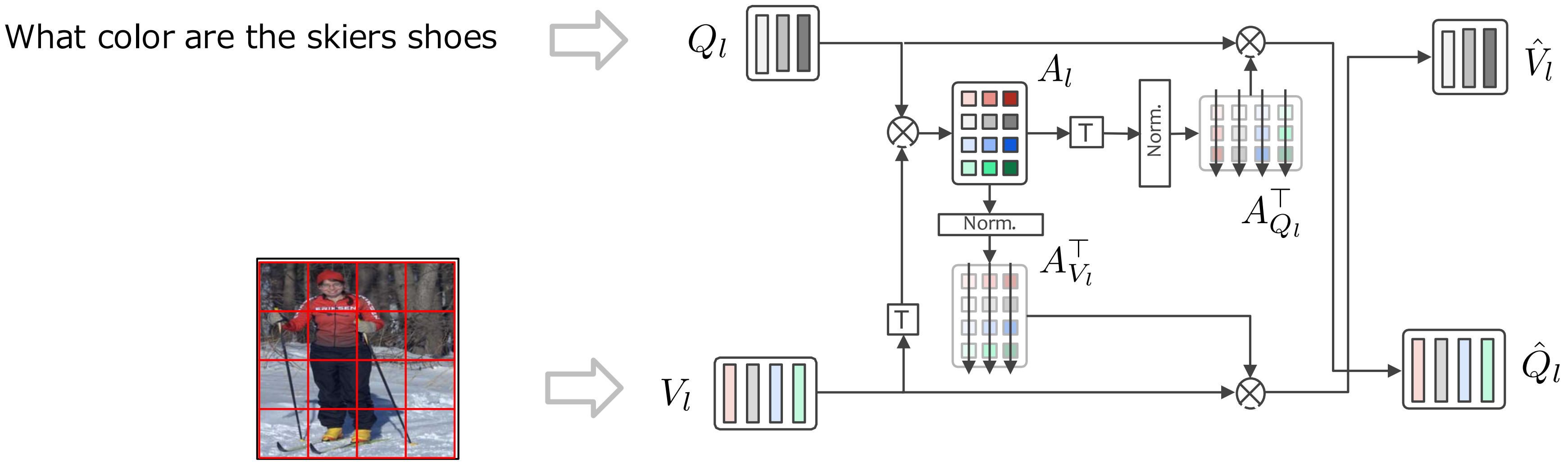
Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



Two-layer Bi-LSTM w/ residual connection

What color are the skiers shoes

152 layer ResNet conv. feat.



What color are the skiers shoes

Q_l

A_l

T

Norm.

$A_{Q_l}^\top$

Norm.

$A_{V_l}^\top$

T

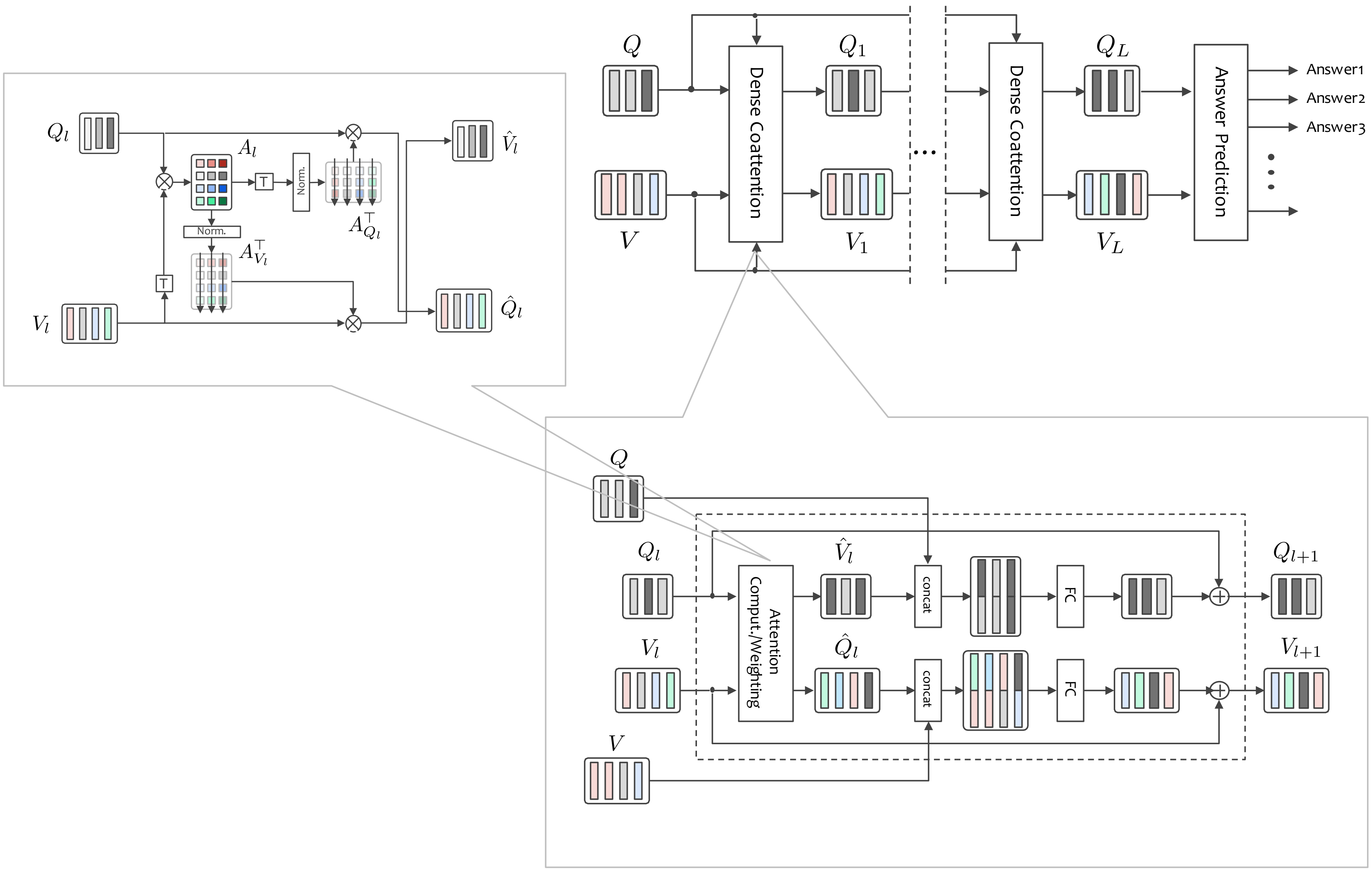
V_l

$\hat{V}_l$

$\hat{Q}_l$

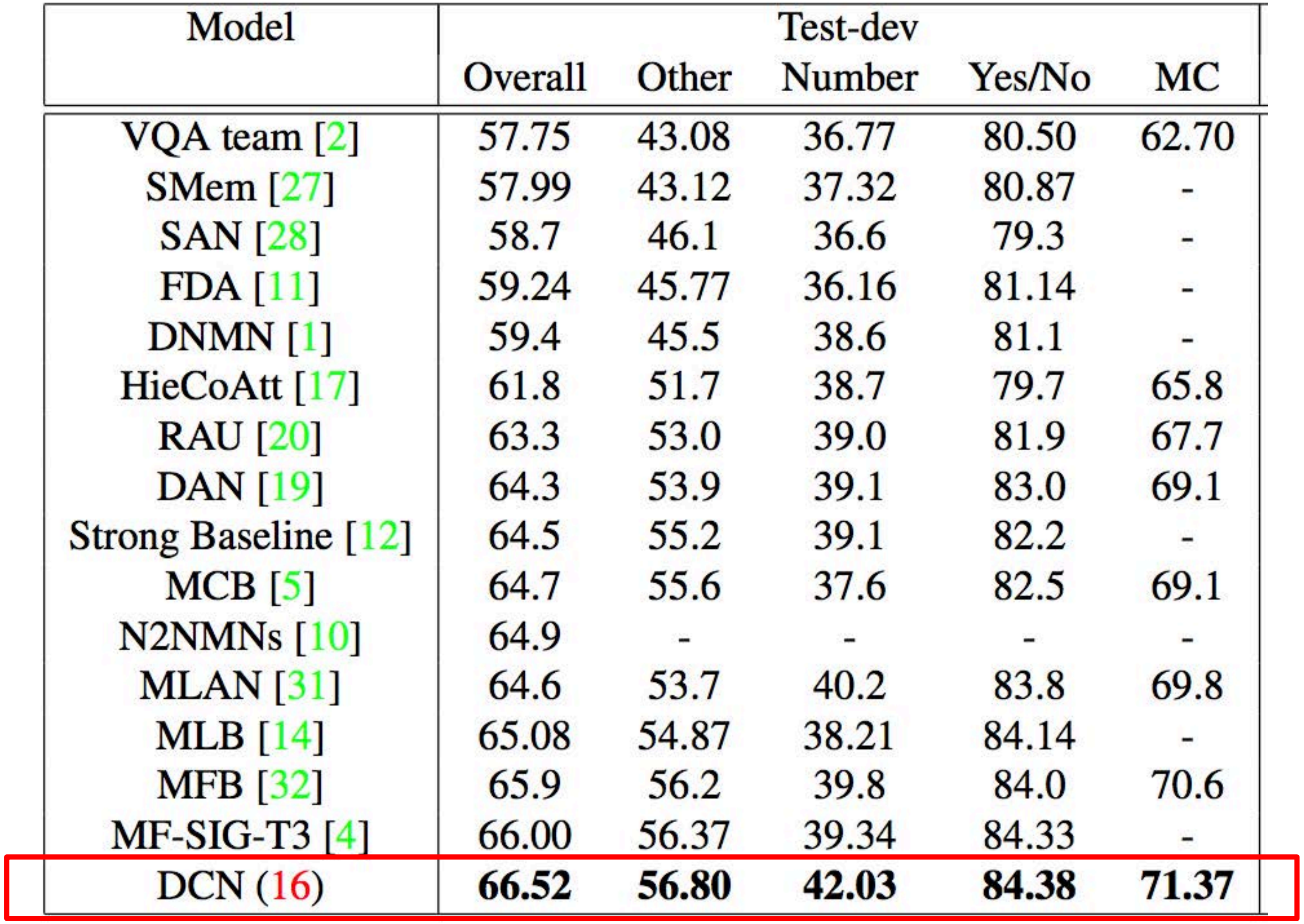
Dense Co-attention Networks

Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



Benchmark results

Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18

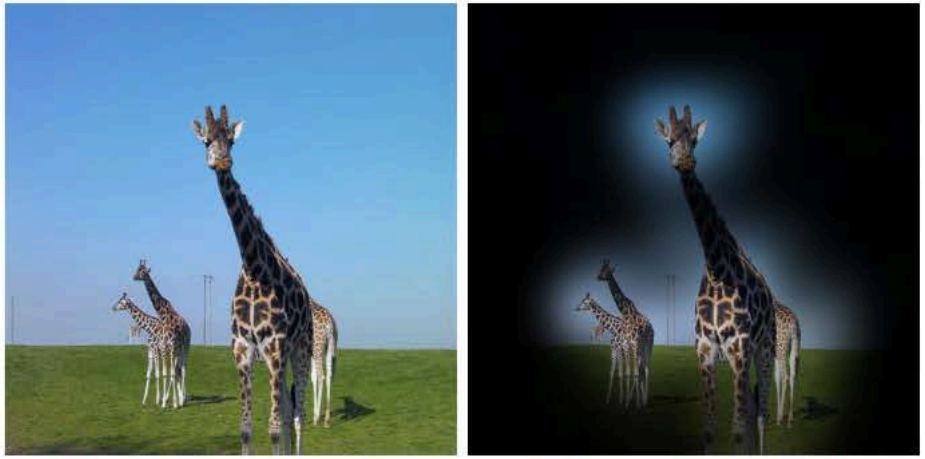


Model	Test-dev				
	Overall	Other	Number	Yes/No	MC
VQA team [2]	57.75	43.08	36.77	80.50	62.70
SMem [27]	57.99	43.12	37.32	80.87	-
SAN [28]	58.7	46.1	36.6	79.3	-
FDA [11]	59.24	45.77	36.16	81.14	-
DNMN [1]	59.4	45.5	38.6	81.1	-
HieCoAtt [17]	61.8	51.7	38.7	79.7	65.8
RAU [20]	63.3	53.0	39.0	81.9	67.7
DAN [19]	64.3	53.9	39.1	83.0	69.1
Strong Baseline [12]	64.5	55.2	39.1	82.2	-
MCB [5]	64.7	55.6	37.6	82.5	69.1
N2NMNs [10]	64.9	-	-	-	-
MLAN [31]	64.6	53.7	40.2	83.8	69.8
MLB [14]	65.08	54.87	38.21	84.14	-
MFB [32]	65.9	56.2	39.8	84.0	70.6
MF-SIG-T3 [4]	66.00	56.37	39.34	84.33	-
DCN (16)	**66.52**	**56.80**	**42.03**	**84.38**	**71.37**

Ref.）Humans **83.30** **72.67** **83.39** **95.77** **-**

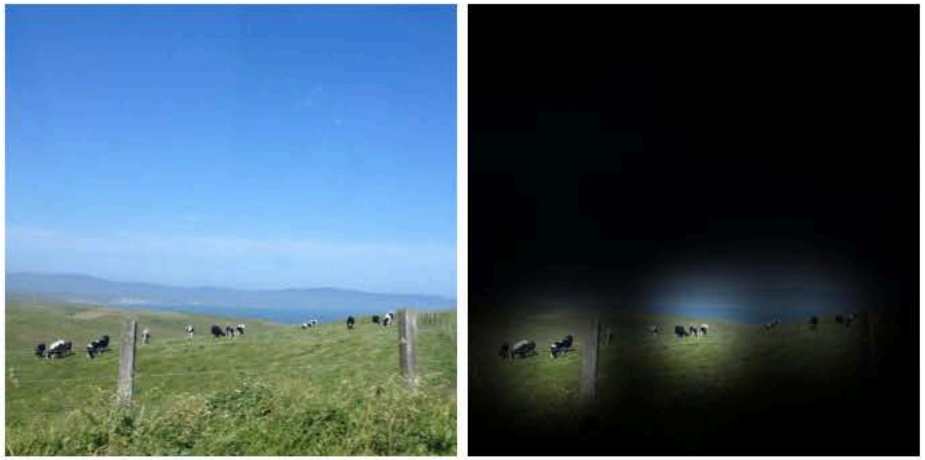
Examples: Correct answers

Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



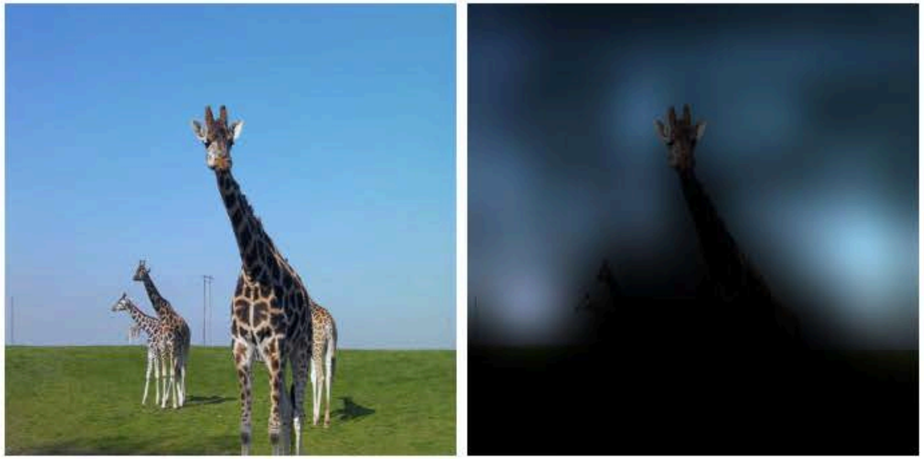
What are these animals   What are these animals

Pred: Giraffes, Ans: Giraffes



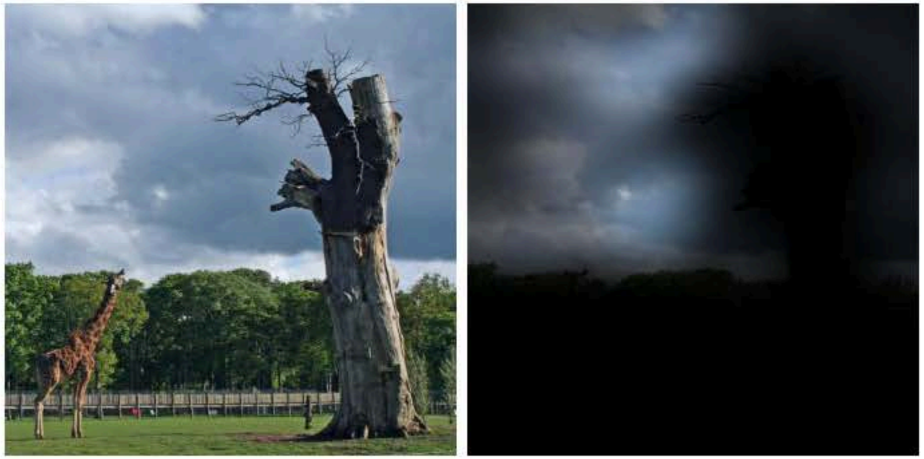
What are these animals   What are these animals

Pred: Cows, Ans: Cows



Is it cloudy   Is it cloudy

Pred: No, Ans: No

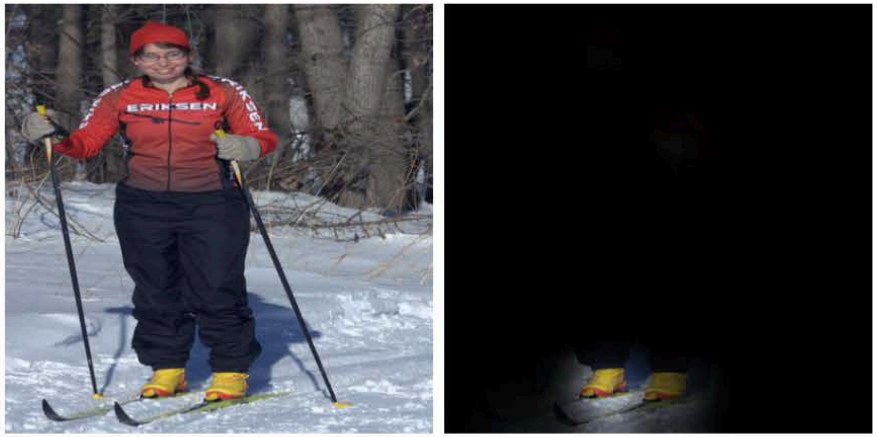


Is it cloudy   Is it cloudy

Pred: Yes, Ans: Yes

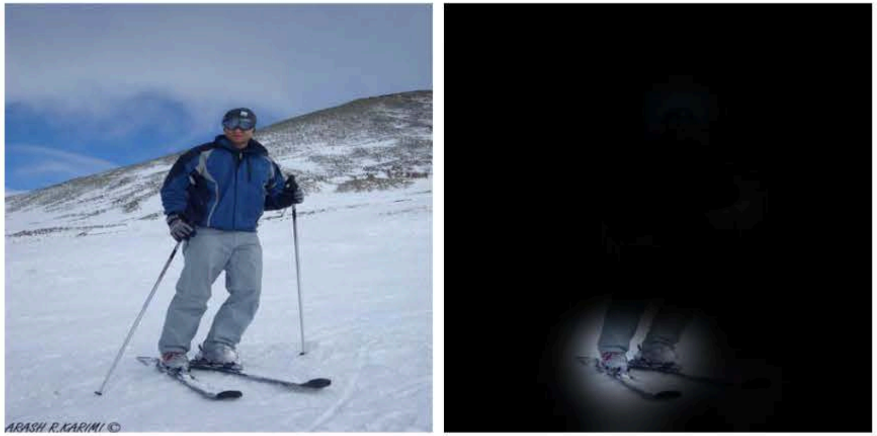
Examples: Correct answers

Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



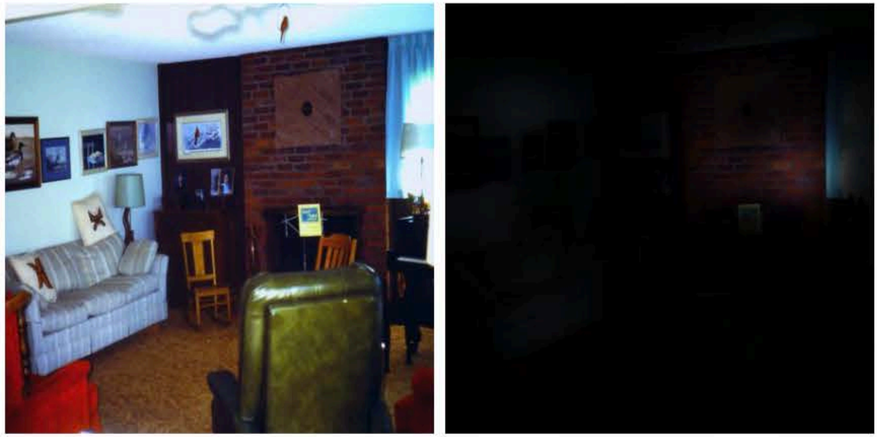
What color are the skiers shoes

Pred: Yellow, Ans: Yellow



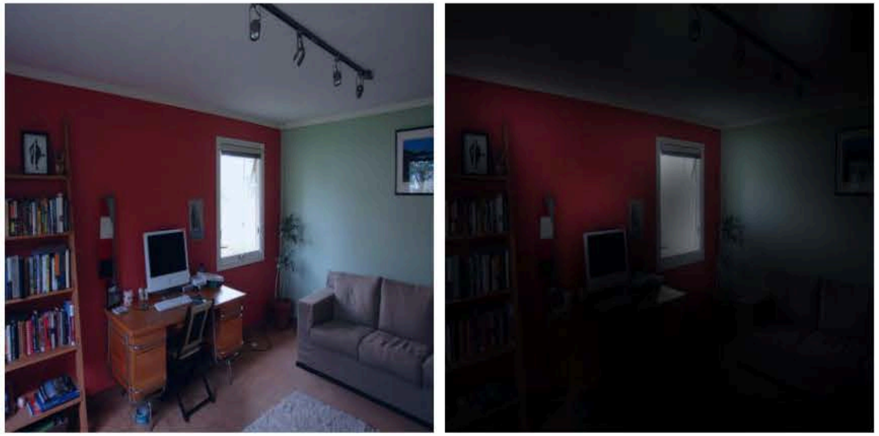
What color are the skiers shoes

Pred: White, Ans: White



What is the darker wall made of

Pred: Brick, Ans: Brick

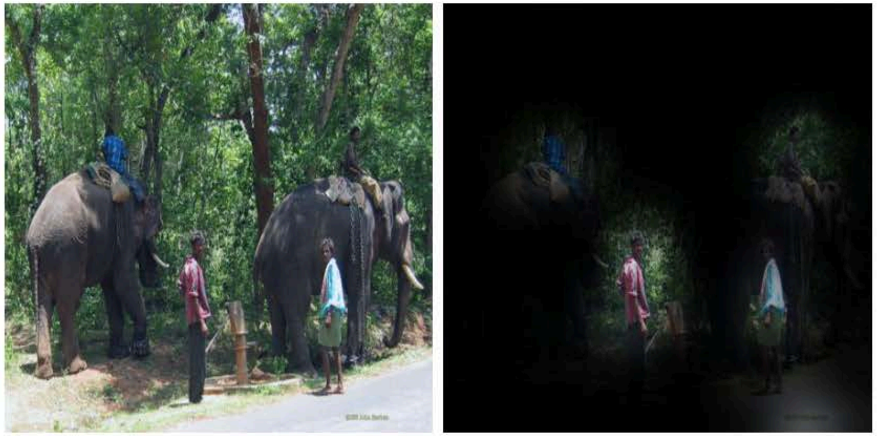


What is the darker wall made of

Pred: Drywall, Ans: Drywall

Examples: Correct answers

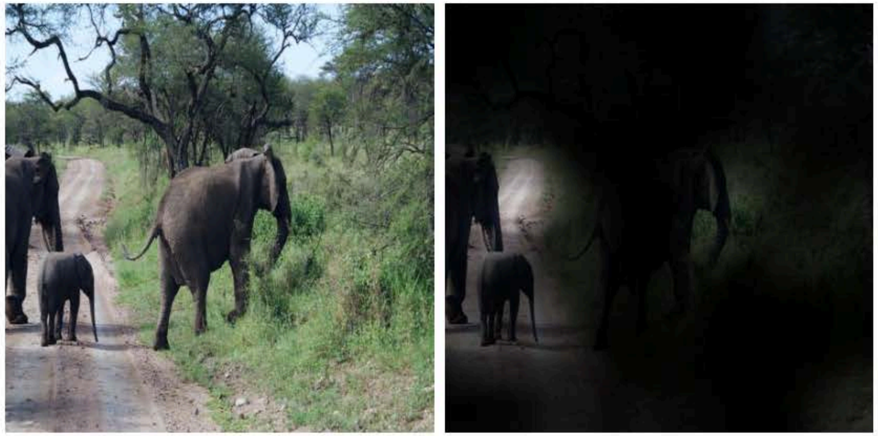
Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



Is there a person standing on the road

Is there a person standing on the road

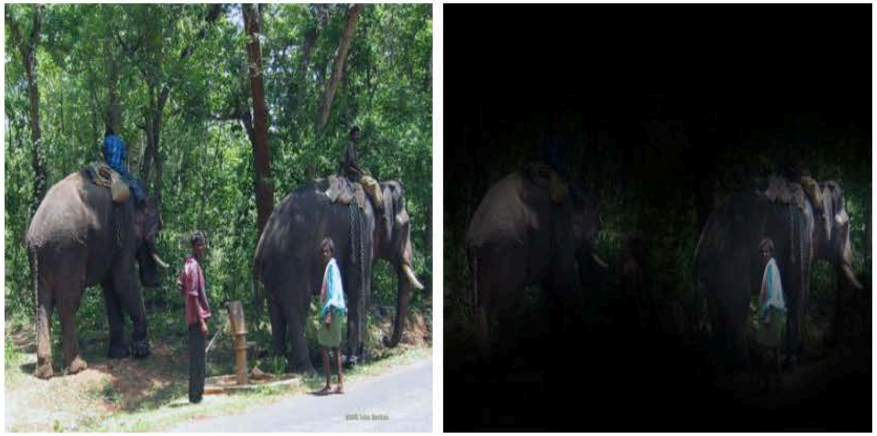
Pred: Yes, Ans: Yes



Is there a person standing on the road

Is there a person standing on the road

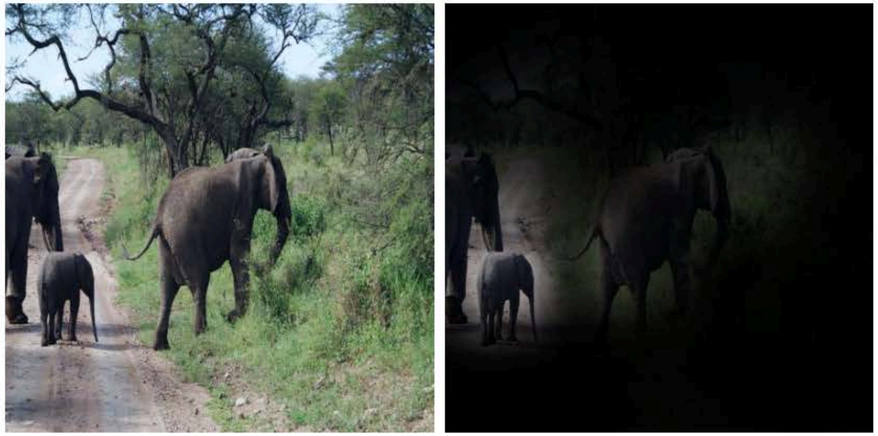
Pred: No, Ans: No



How many elephants

How many elephants

Pred: 2, Ans: 2



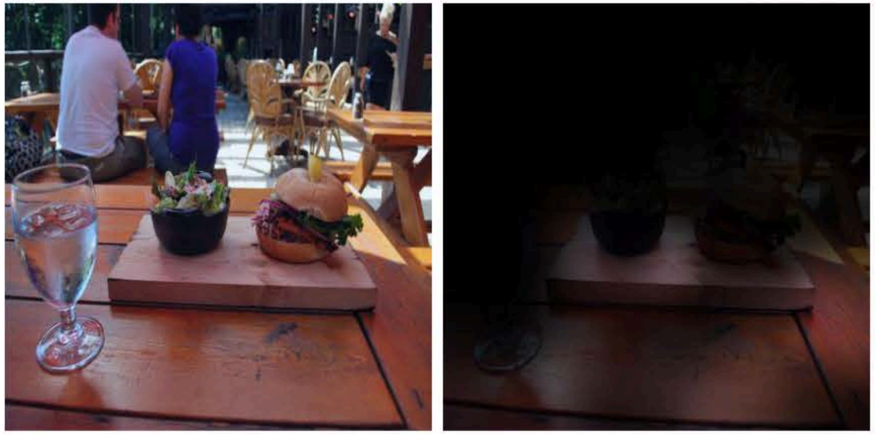
How many elephants

How many elephants

Pred: 3, Ans: 3

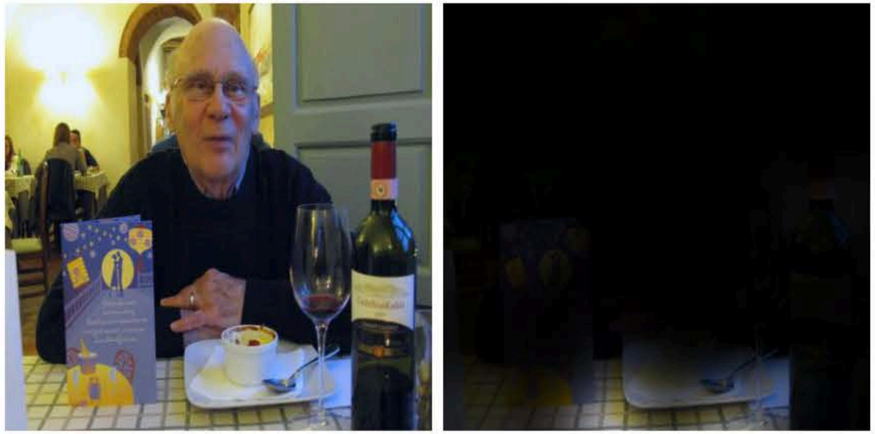
Examples: Wrong answers

Nguyen, Okatani, Improved Fusion of Visual and Language Features by Dense Symmetric Co-attention For VQA, CVPR18



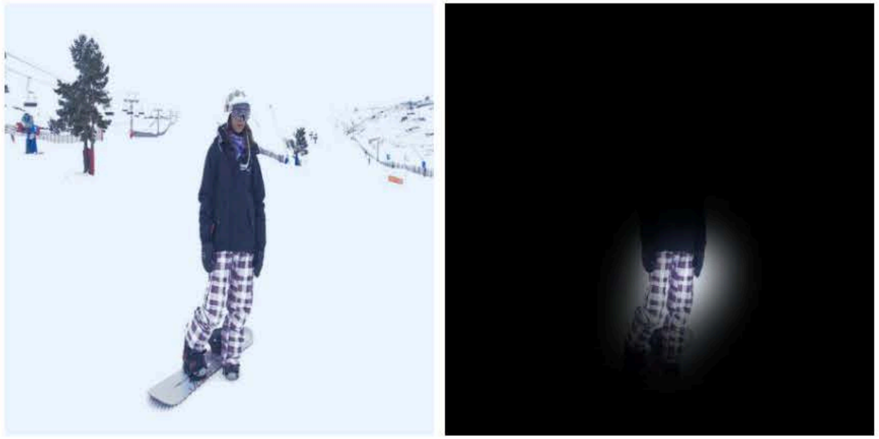
What material is the table made of

Pred: Wood, Ans: Wood



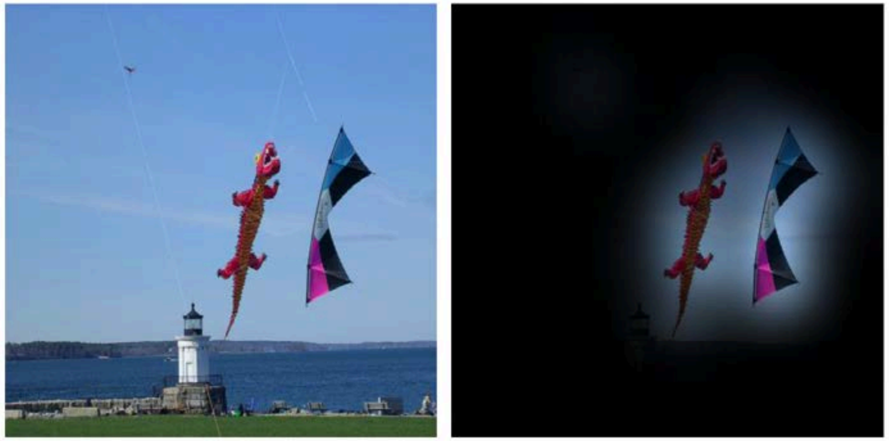
What material is the table made of

Pred: Metal, Ans: Tile *(Error type: 1)*



What is the color of pants the woman is wearing

Pred: Plaid, Ans: Red and White *(Error type: 4)*



What flag is that

Pred: American, Ans: Dragon *(Error type: 2)*