VQA: Visual Question Answering

Cedric Bernard

Cat or dog?

Cat Hypertension | High Blood Pressure in Cats | Vets4Pets

Is this dog happy?

25 Pictures Of Super Happy Dogs That Will Make Your Day Super Better (theawesomedaily.com)

Does this person have 20/20 vision?

1505.00468.pdf (arxiv.org) [1]

1505.00468.pdf (arxiv.org) [1]

AI Complete/Strong AI

Engineers Create Robot Muscles Than Are 1000 Times Stronger (wonderfulengineering.com [5]

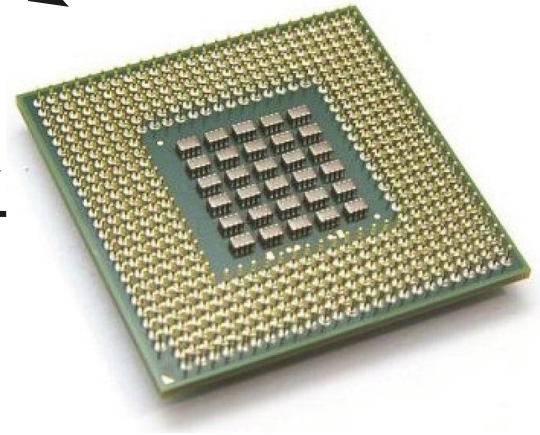
Visual Q/A

1505.00468.pdf (arxiv.org) [1]

7 Reasons why you should lose the fear of frequently asking questions to your collaborators - Team Insights

Question

Answer



(PDF) An Assessment of Direct Chip Cooling Enhancement Using Pin Fins (researchgate.net)

Related Work

- Text Based Q/A
- Describing visual content (captioning)
- Vision and language interaction

Young woman waving hand with windy hair next to inspirational text.

82 Profile Picture Captions for Instagram and Facebook - Healthy Tips (kindyou.com)

VQA Related Work

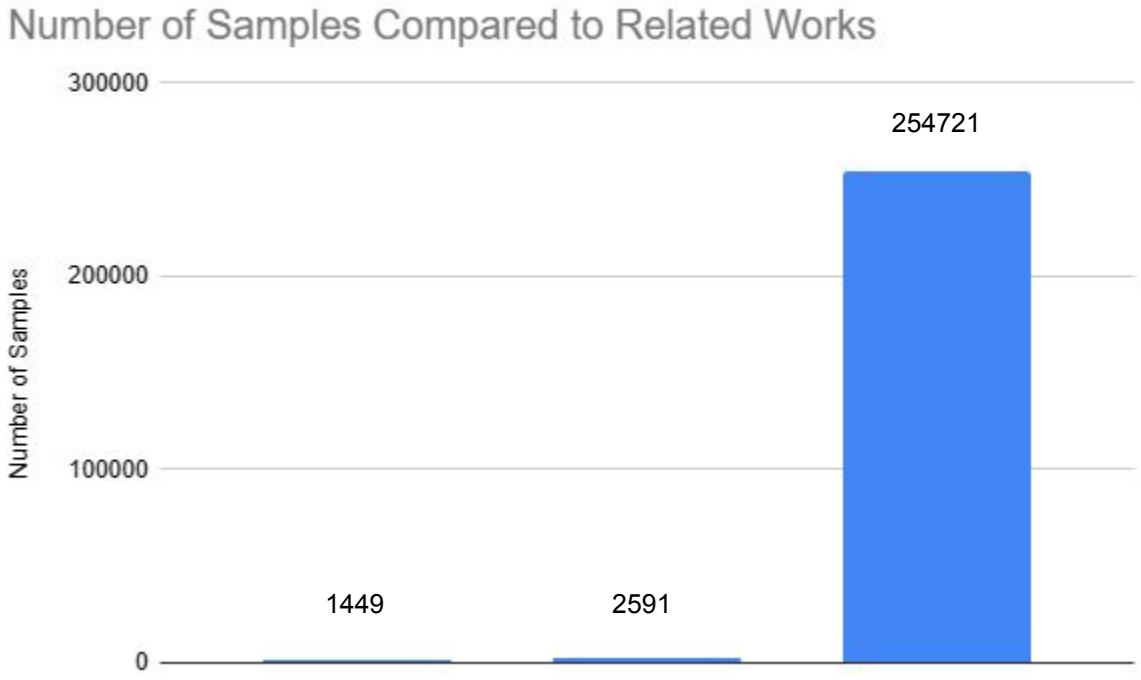
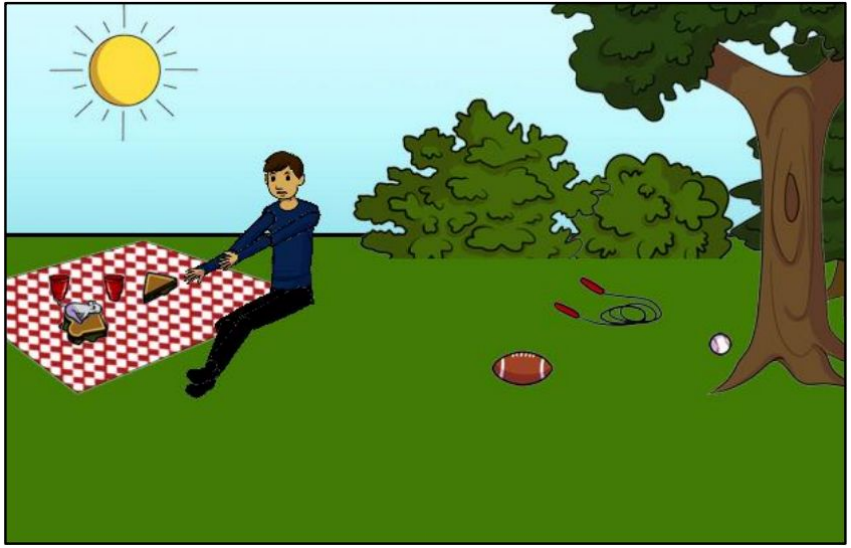


Image Data Collection



Abstract Image Dataset

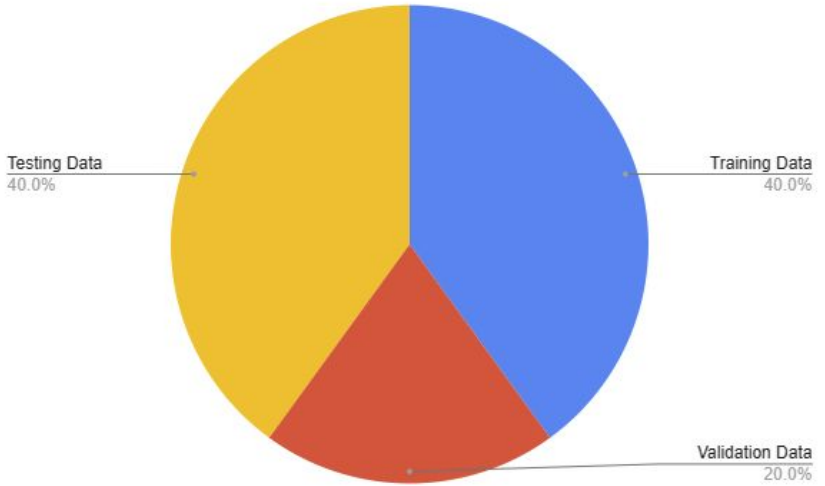
1505.00468.pdf (arxiv.org) [1]

MS COCO

1505.00468.pdf (arxiv.org) [1]

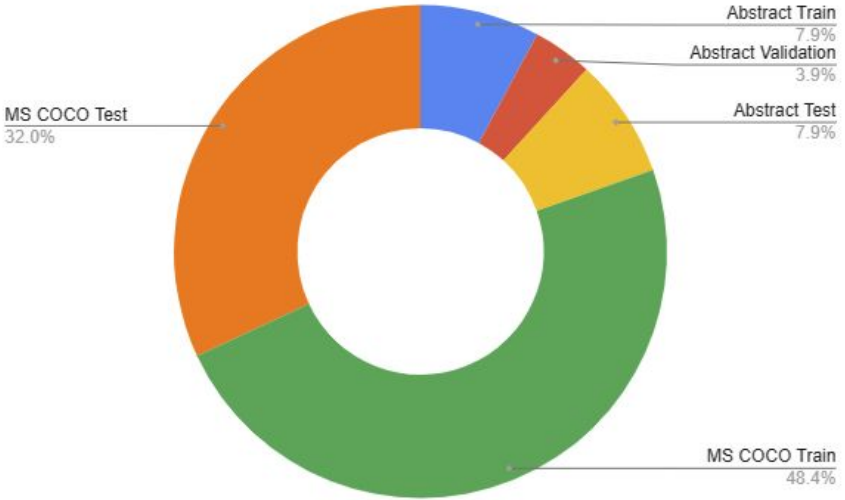
Abstract Image Dataset

- 50,000 images
- 20 Poseable people models
- 31 animals
- 100 objects

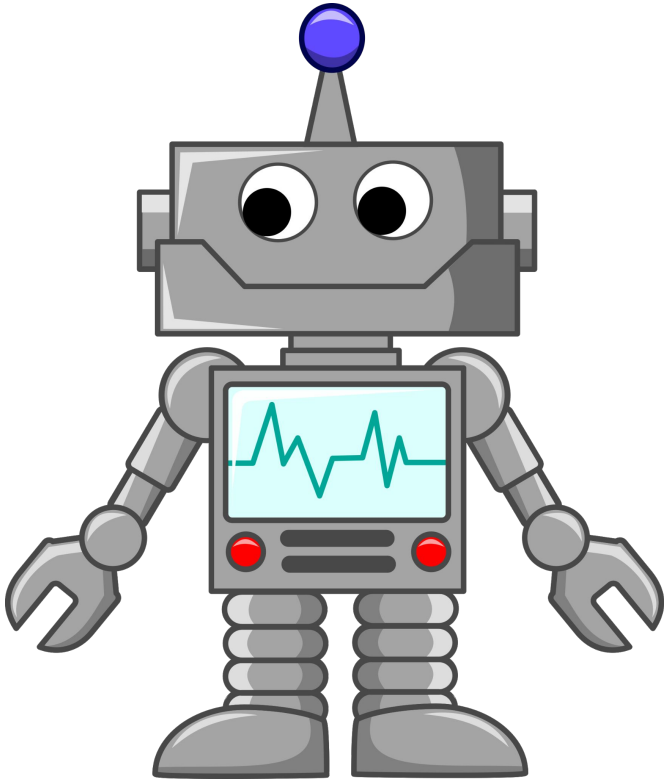


MS COCO

- MicroSoft Common Objects in COntext
- 204672 images with caption



Question Data Collection

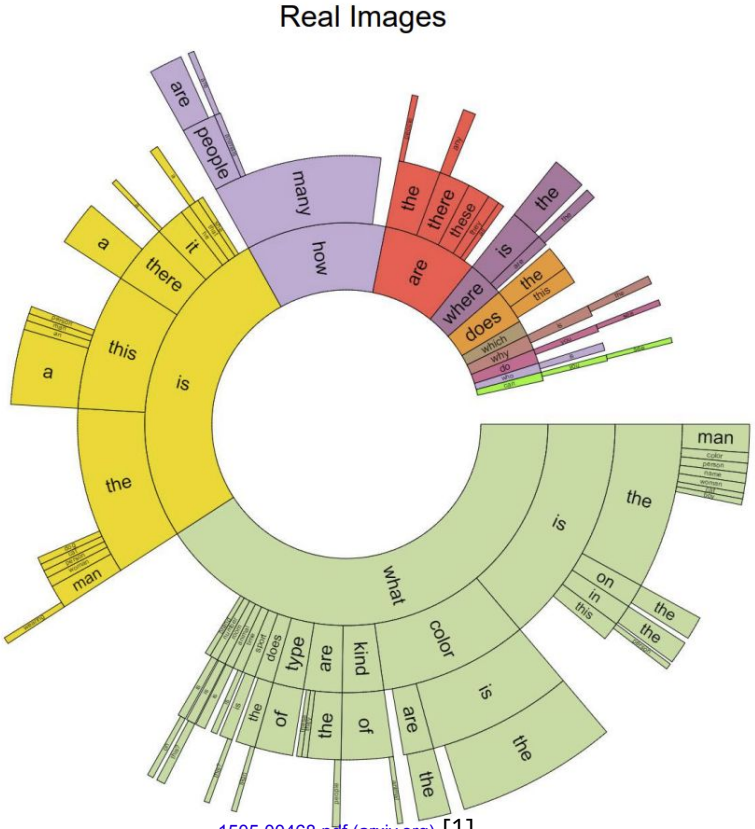


cartoon robot - Openclipart

Cartoon Aliens Pictures - ClipArt Best

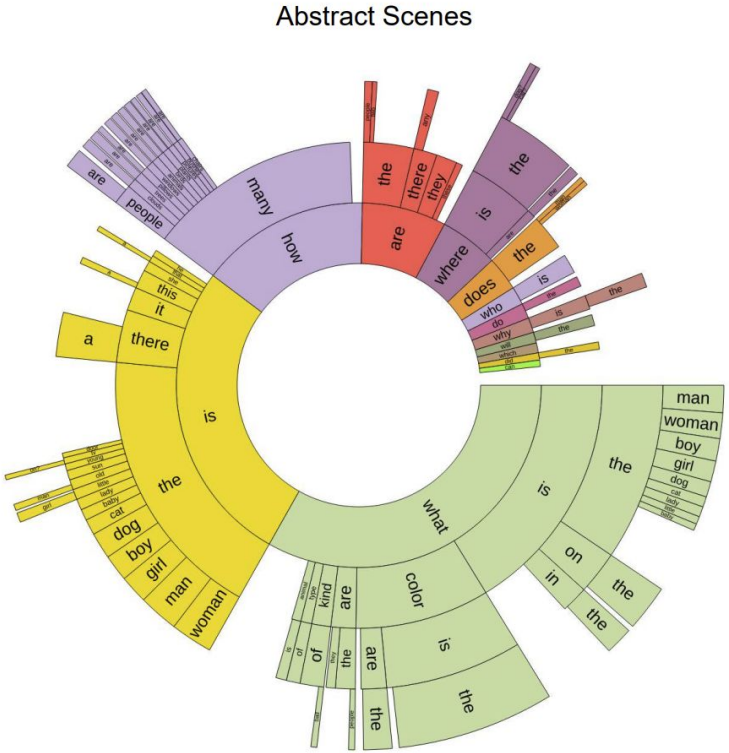
Pin by Rizwan Rasool on Cartoon Character Inspirations | Black love art, Baby art, Baby cartoon characters (pinterest.com)

Real Image Questions



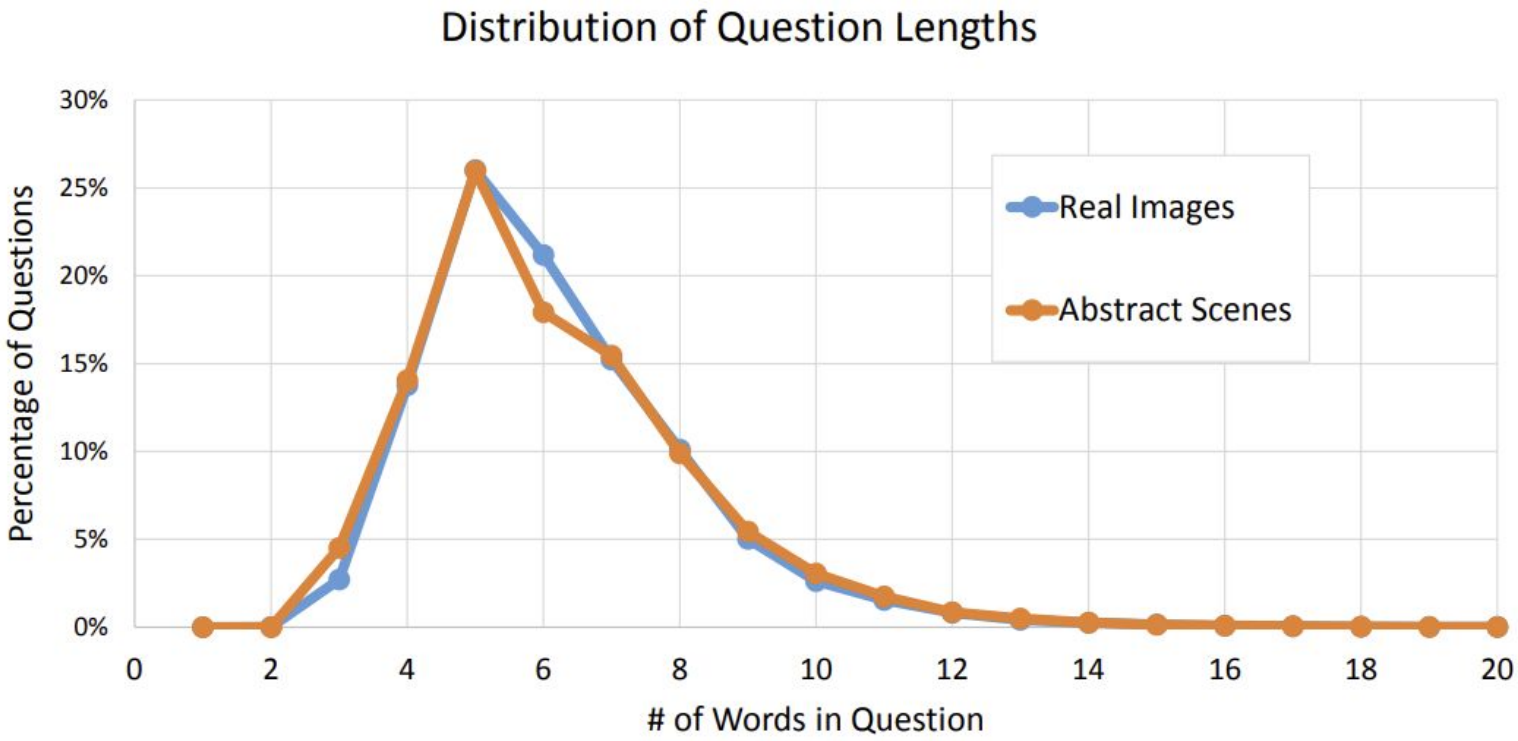
1505.00468.pdf (arxiv.org) [1]

Abstract Questions



1505.00468.pdf (arxiv.org) [1]

Questions Lengths



1505.00468.pdf (arxiv.org) [1]

Answer Collection

- 10 Answers per question
- Single word to brief phrase

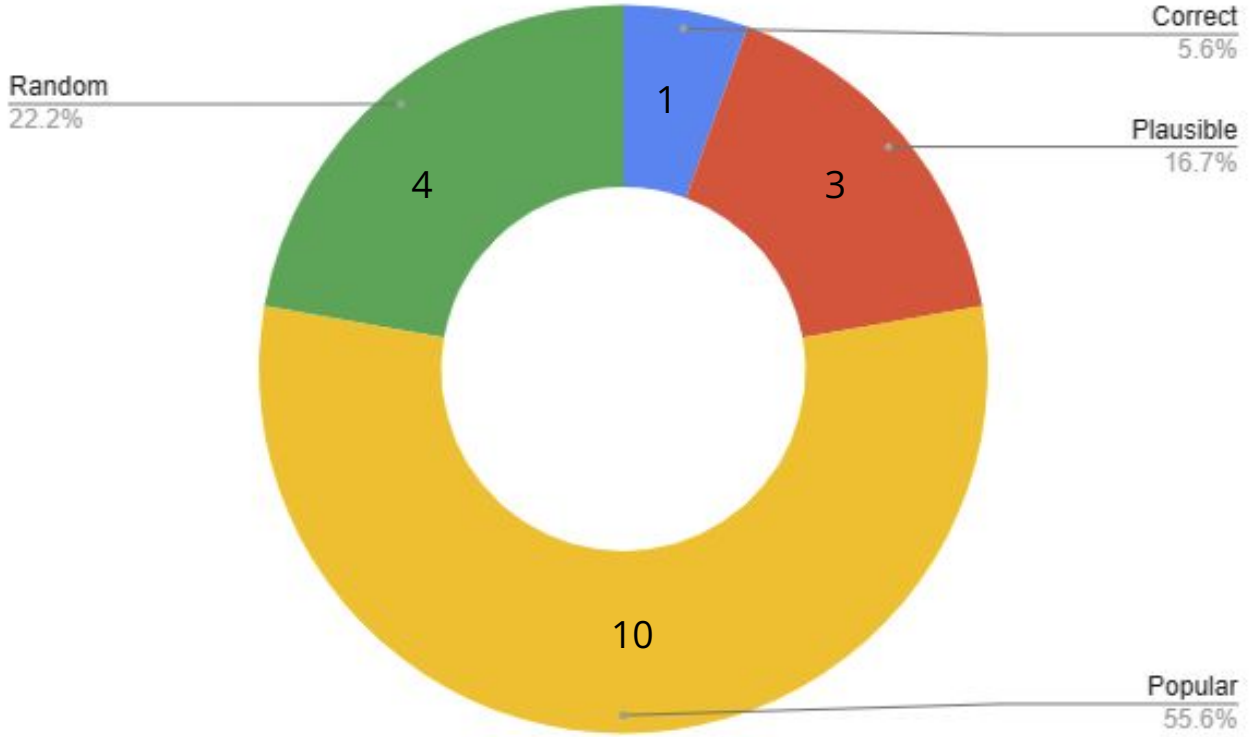
To Force or not to Force (an answer): It's a complicated question | Corona Insights

Accuracy Metric

$$acc = min(h/3, 1)$$

H = number of human subjects providing that exact answer

Multiple Choice Answers



Model Architecture

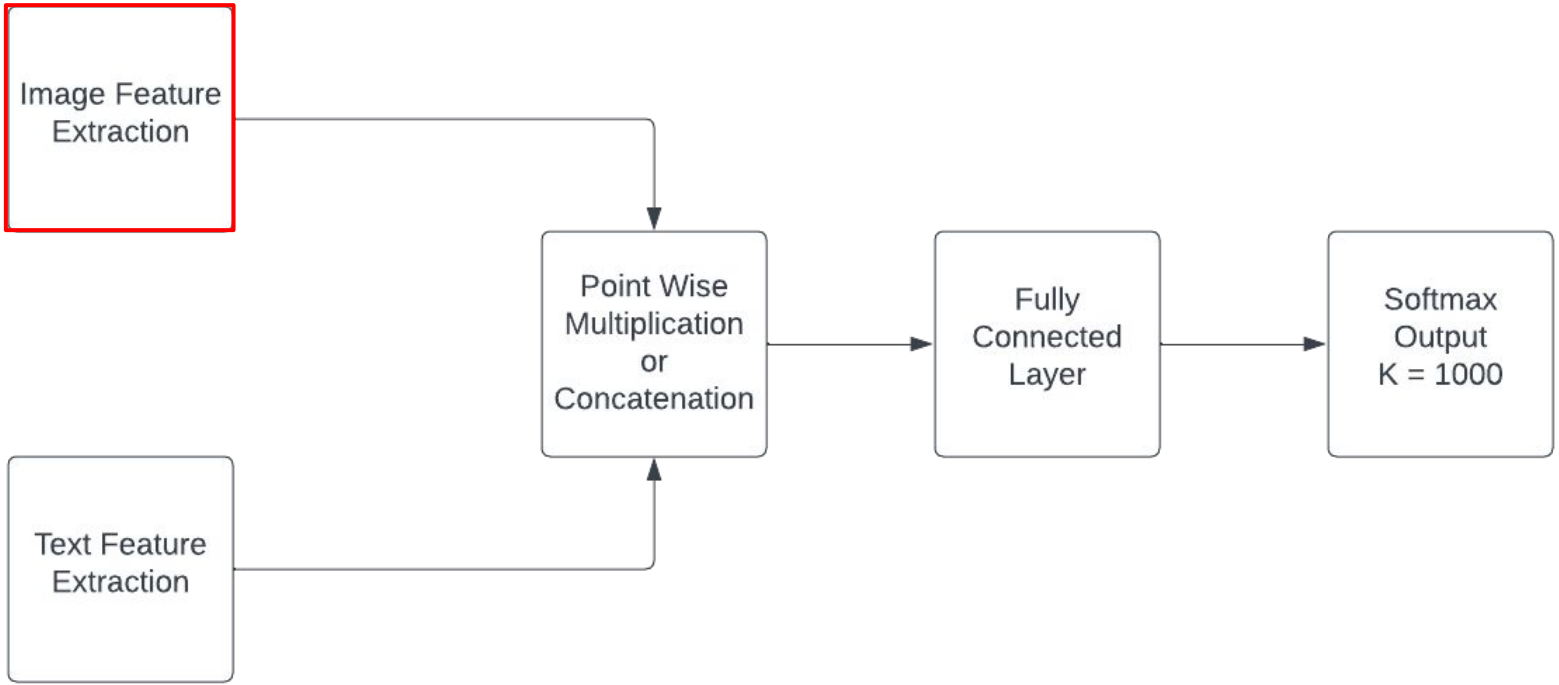
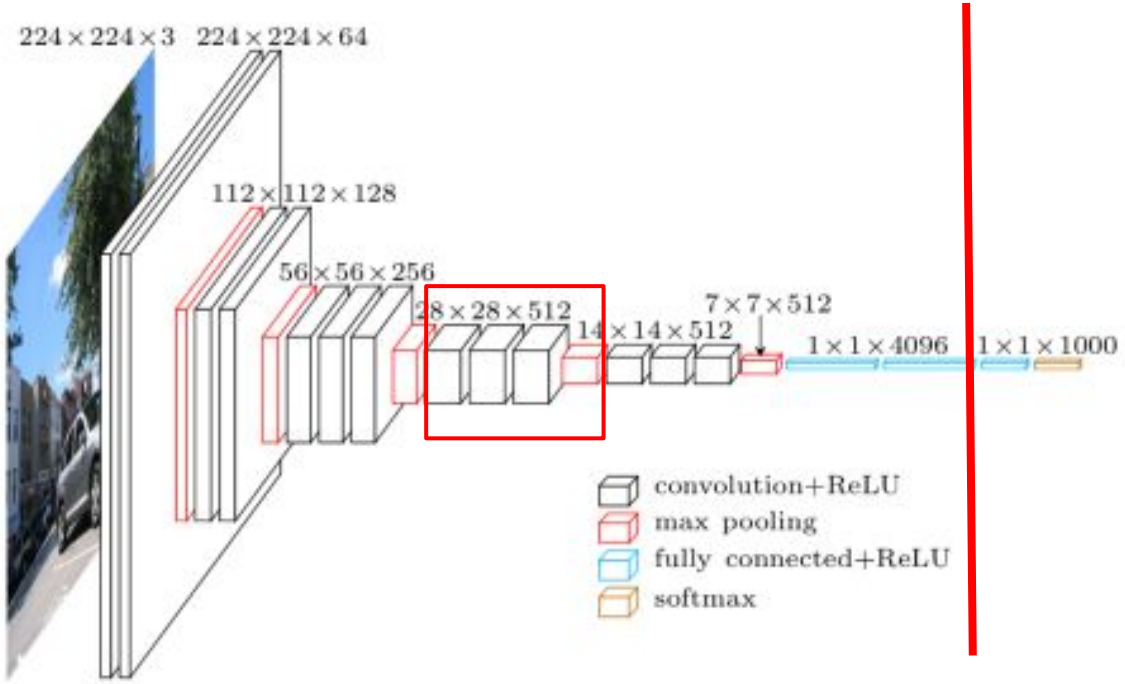
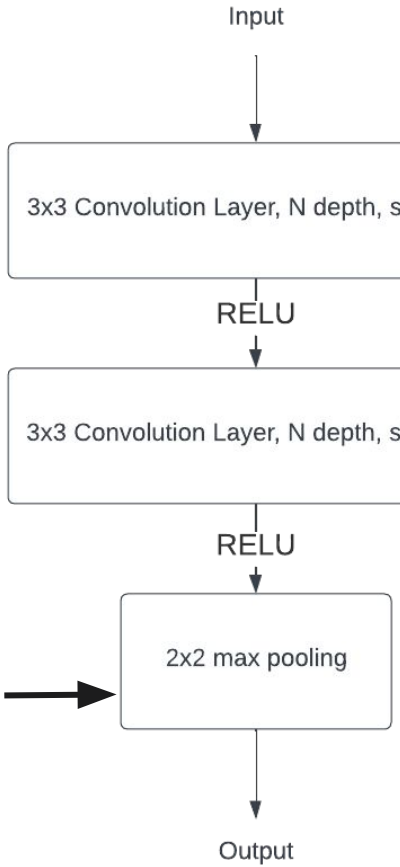


Image Feature Extraction

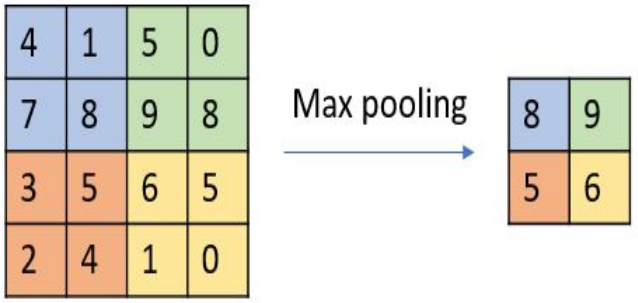


VGG in TensorFlow · Davi Frossard (toronto.edu)

VGGNet

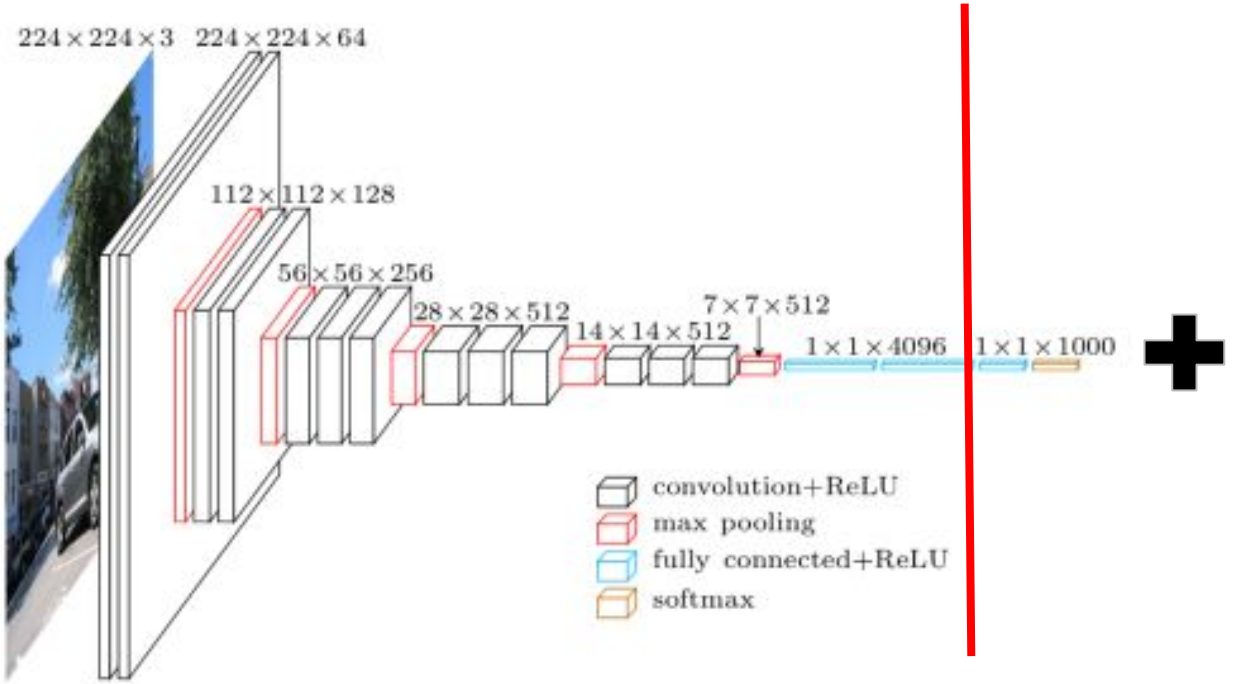


X1	X2	X3
X4	X5	X6
X7	X8	X9



Schematic illustration of a max pooling operation with 2 × 2 filter and... | Download Scientific Diagram (researchgate.net)

Normalized VGGNet



L2 Normalization

VGG in TensorFlow · Davi Frossard (toronto.edu)

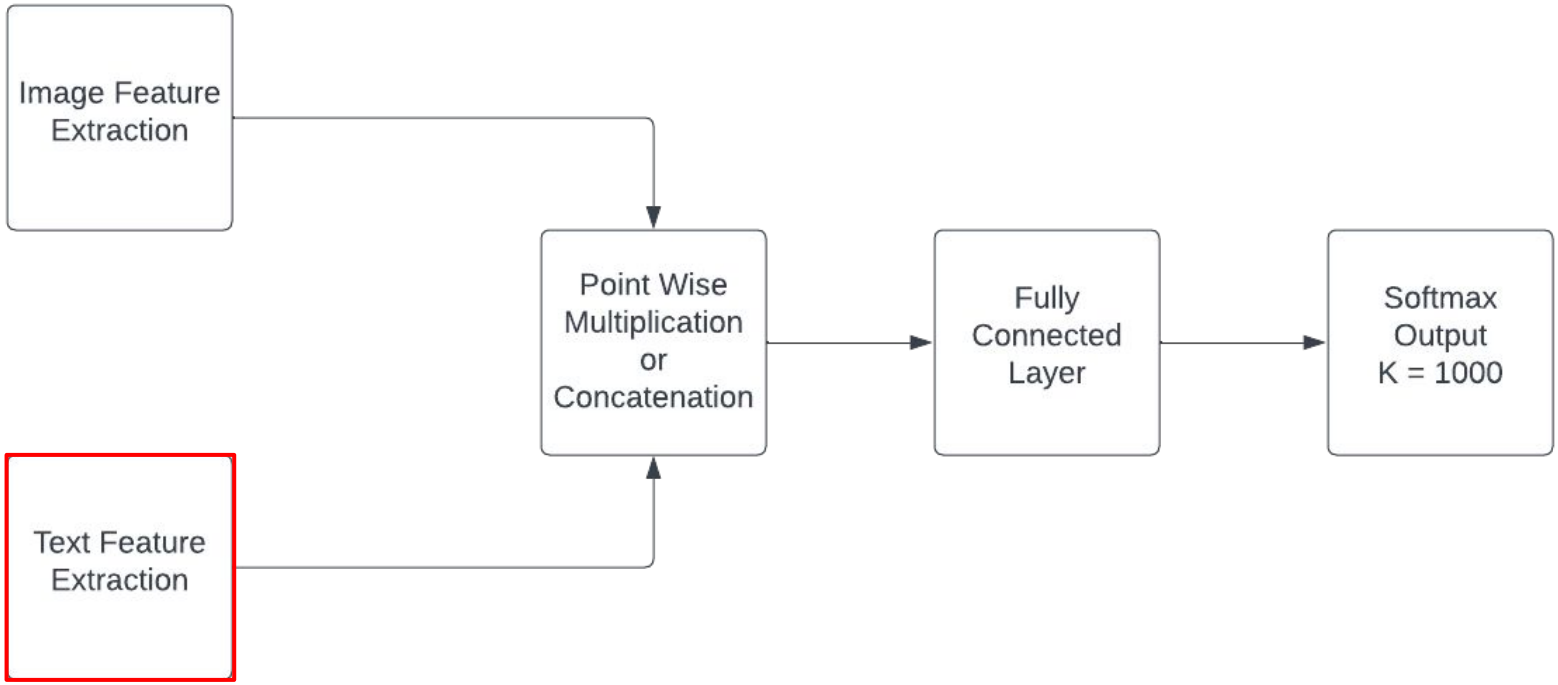
L2 Normalization

$$\mathbf{W} = (\frac{1}{L_2})\mathbf{W}$$

$$L_2 = ||W||^2 = w_1^2 + w_2^2 + \ldots + w_n^2$$

Regularization for Simplicity: L₂ Regularization | Machine Learning | Google Developers [3]

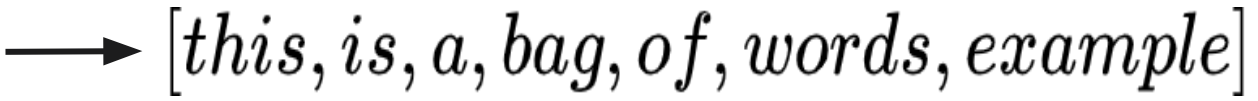
Model Architecture



Bag of Words Representation

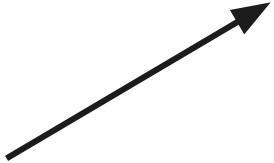
This is a bag of words example.

Maybe not all the words in the corpus are used in the bag of words model

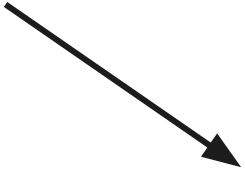


BoW Text to Quantitative Vector

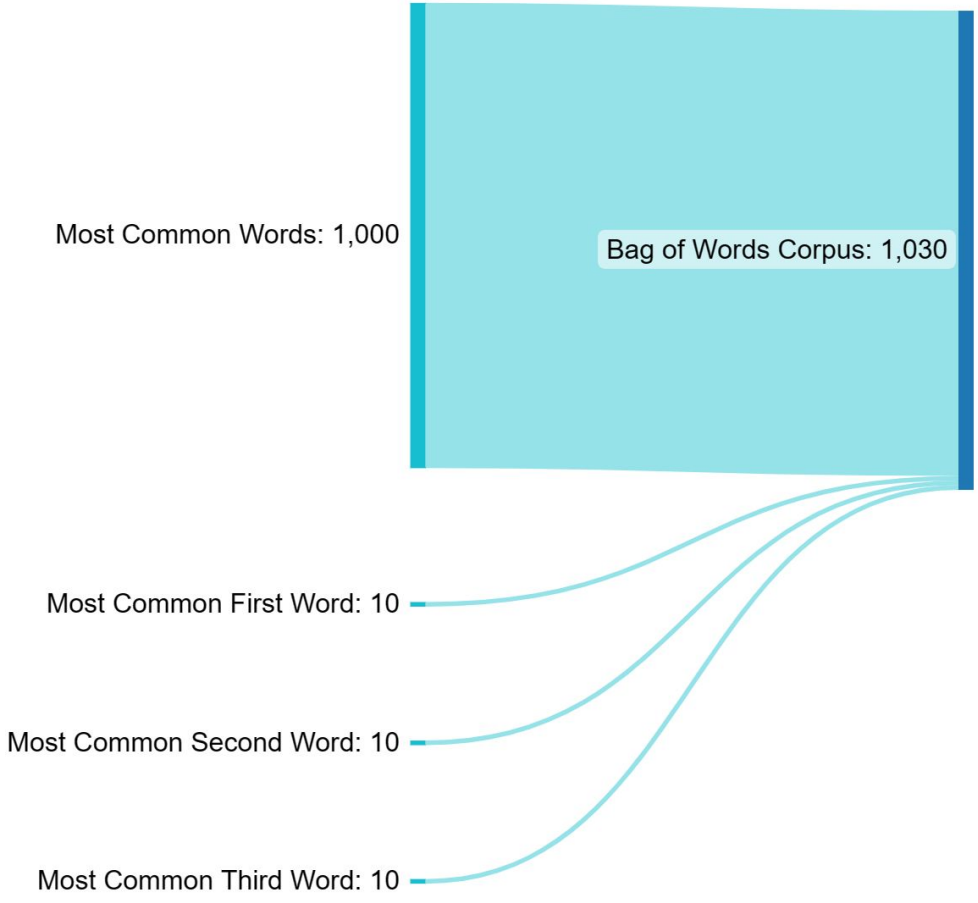
$$[this, is, a, bag, of, words, example]$$



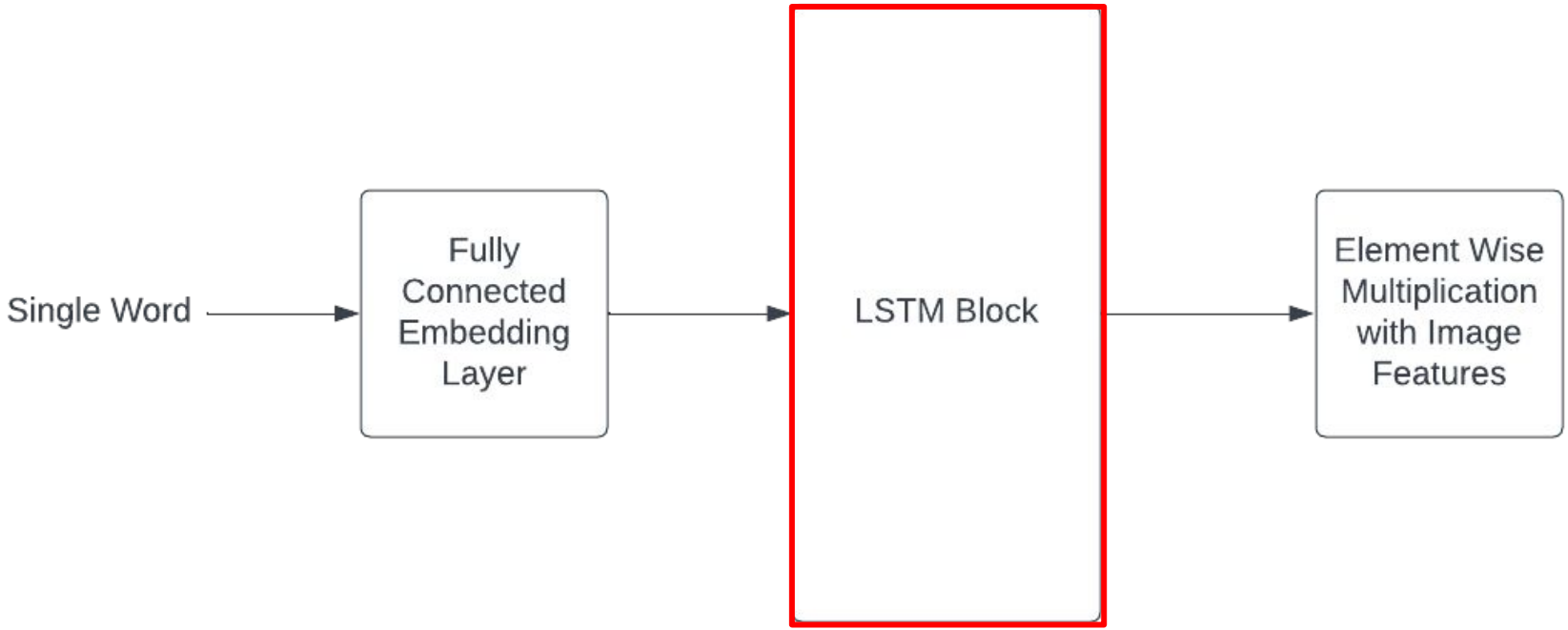
This is new input. This will show bag of words use case.



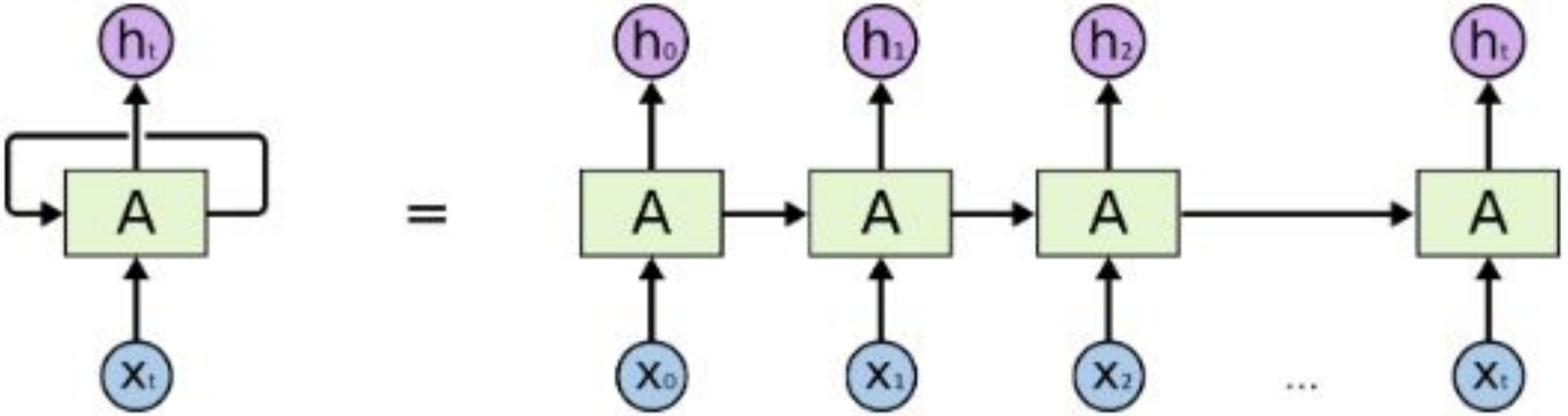
[2, 1, 0, 1, 1, 1, 0]


Most Common Words: 1,000
Bag of Words Corpus: 1,030
Most Common First Word: 10
Most Common Second Word: 10
Most Common Third Word: 10
Made with SankeyMATIC

LSTM Q


Single Word
Fully Connected Embedding Layer
LSTM Block
Element Wise Multiplication with Image Features

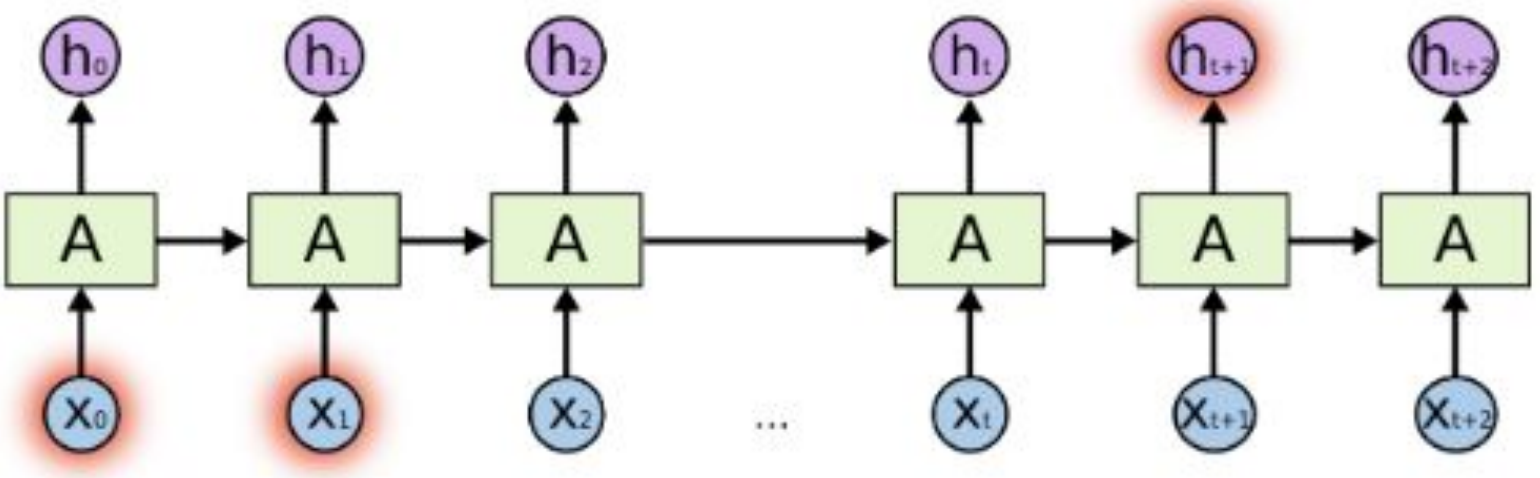
LSTM Structure – Starting with RNN



An unrolled recurrent neural network.

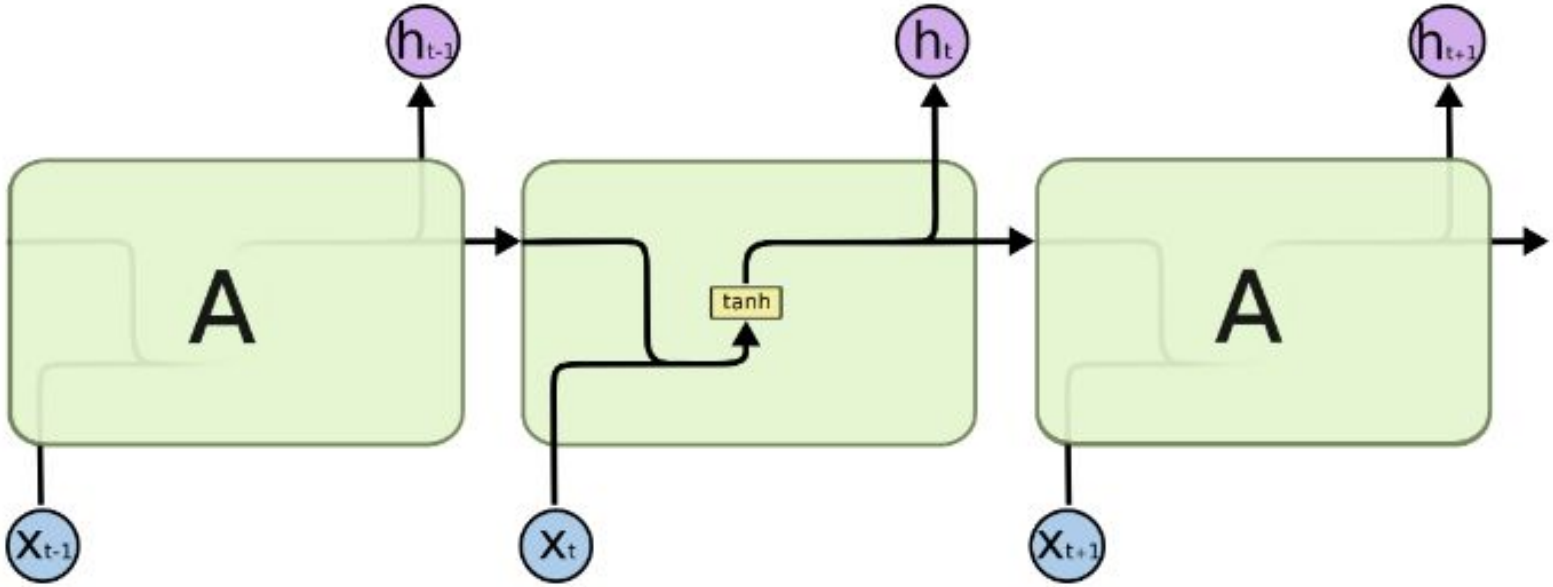
Understanding LSTM Networks -- colah's blog [4]

Long-Term Dependencies



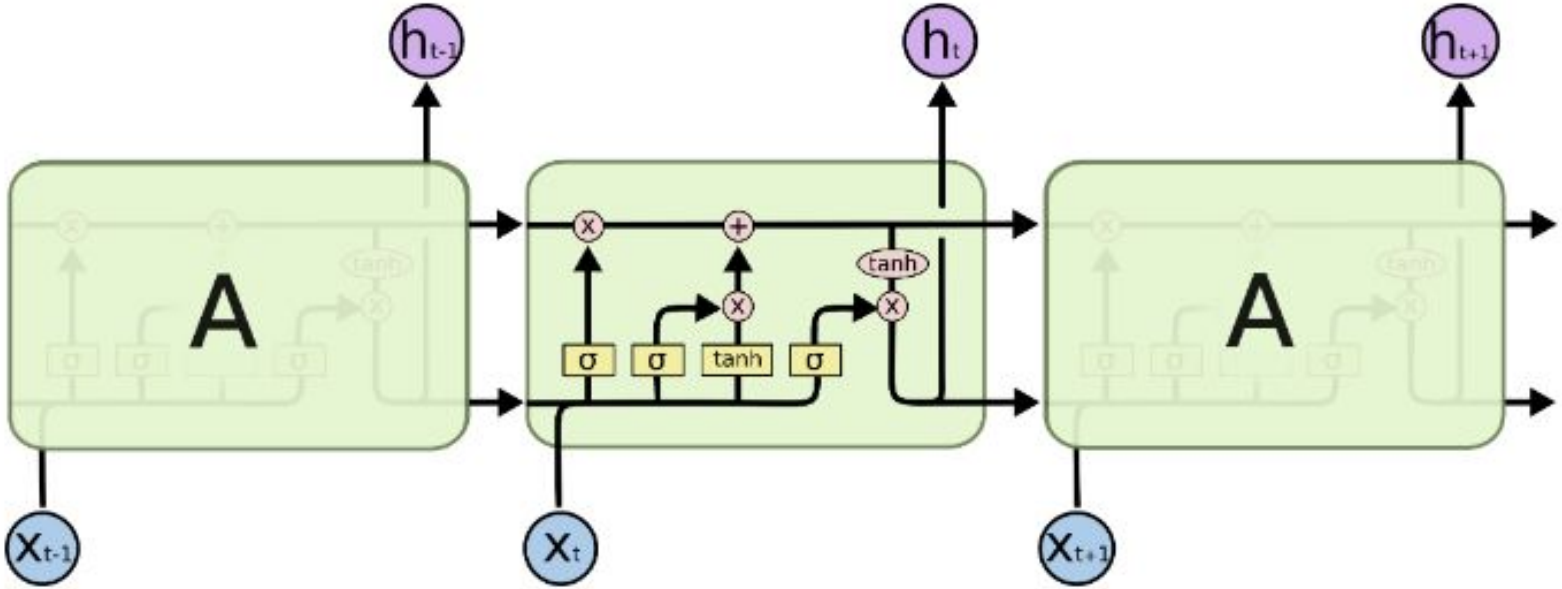
Understanding LSTM Networks -- colah's blog [4]

RNN Cell

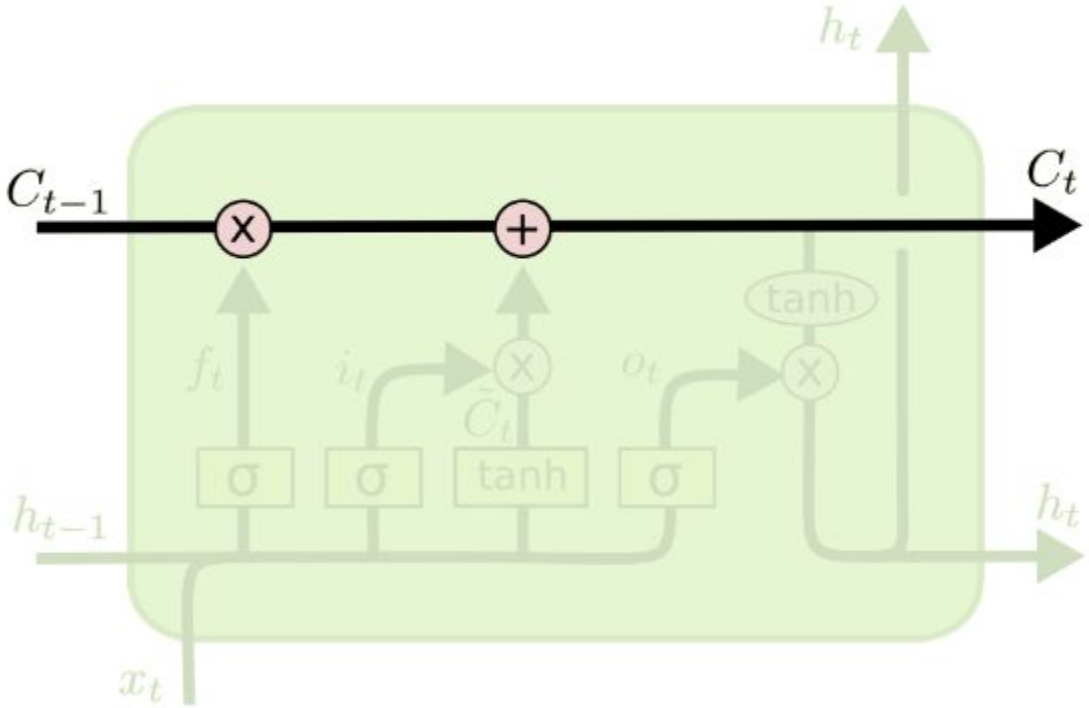


Understanding LSTM Networks -- colah's blog [4]

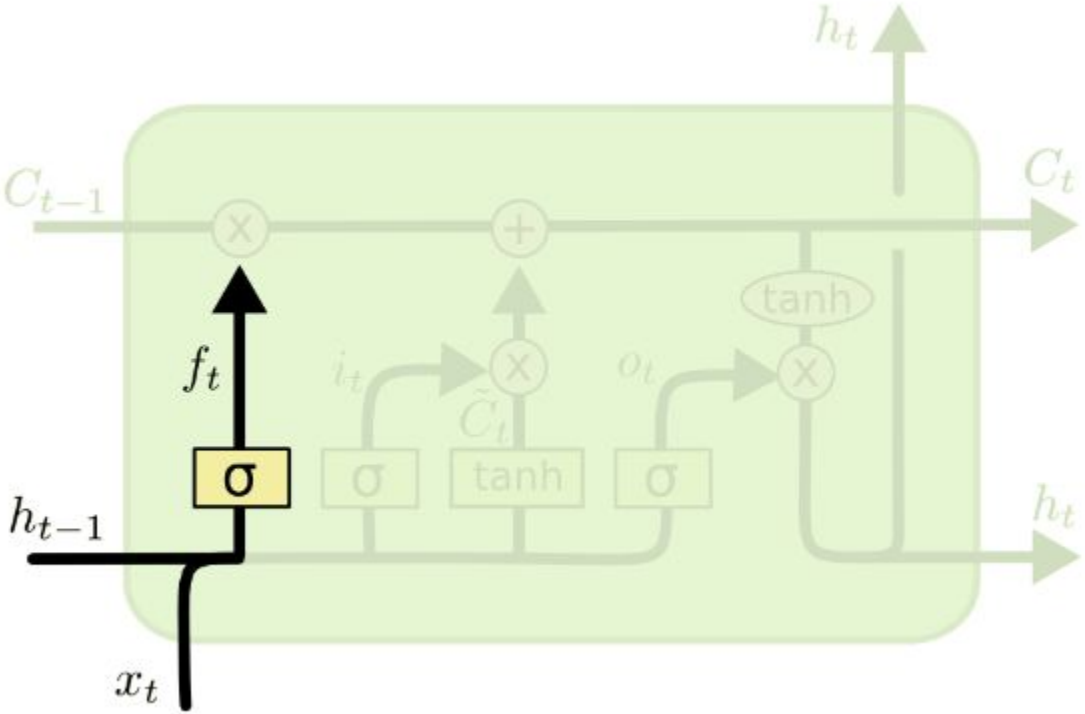
LSTM Cell



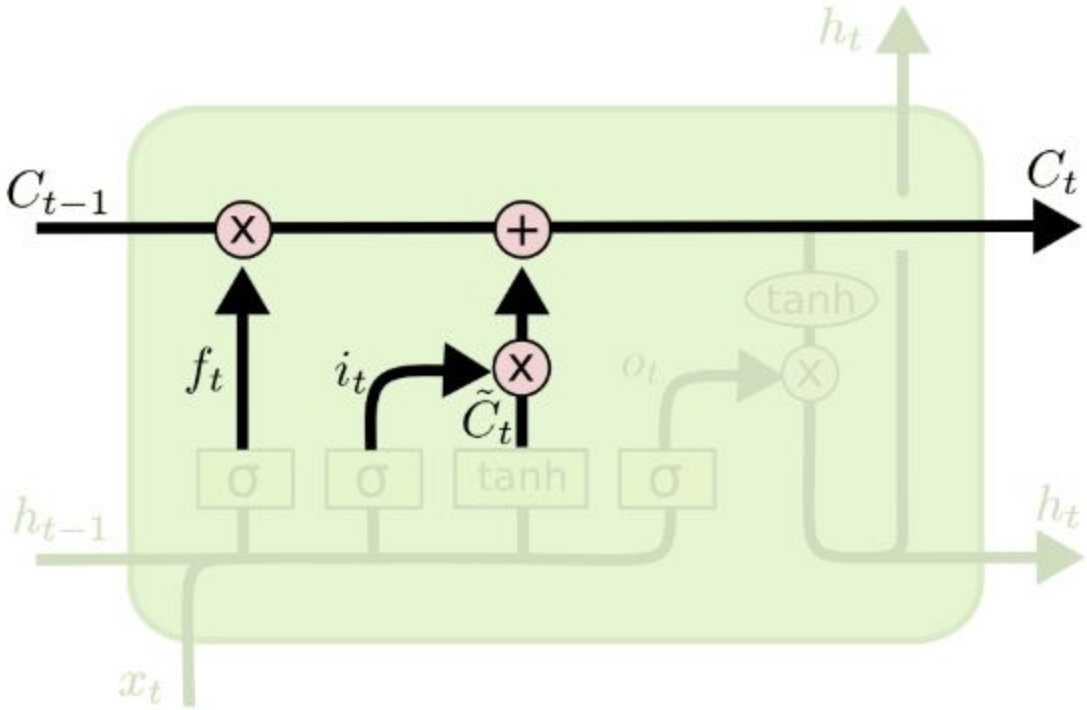
Understanding LSTM Networks -- colah's blog [4]



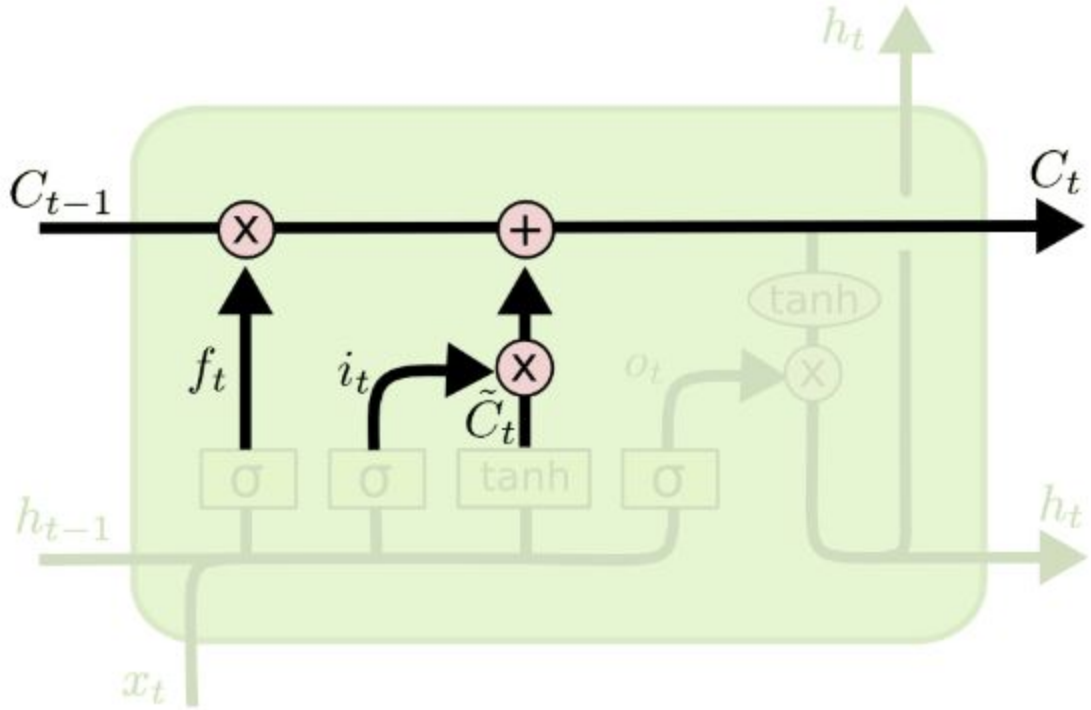
Understanding LSTM Networks -- colah's blog [4]



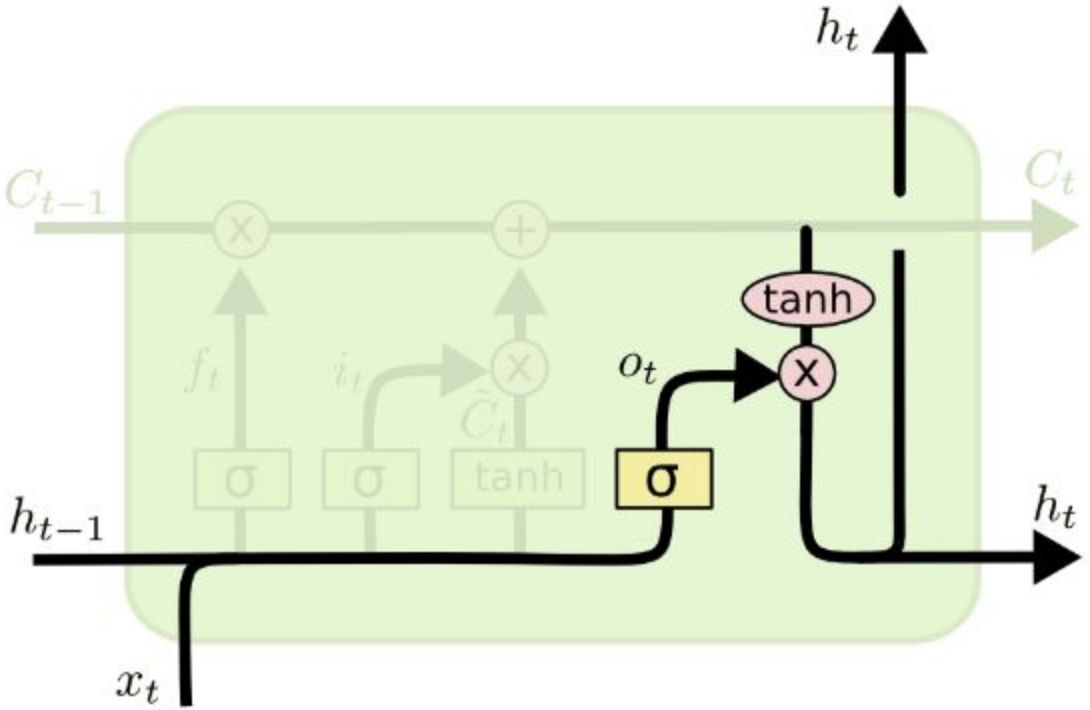
Understanding LSTM Networks -- colah's blog [4]



Understanding LSTM Networks -- colah's blog [4]

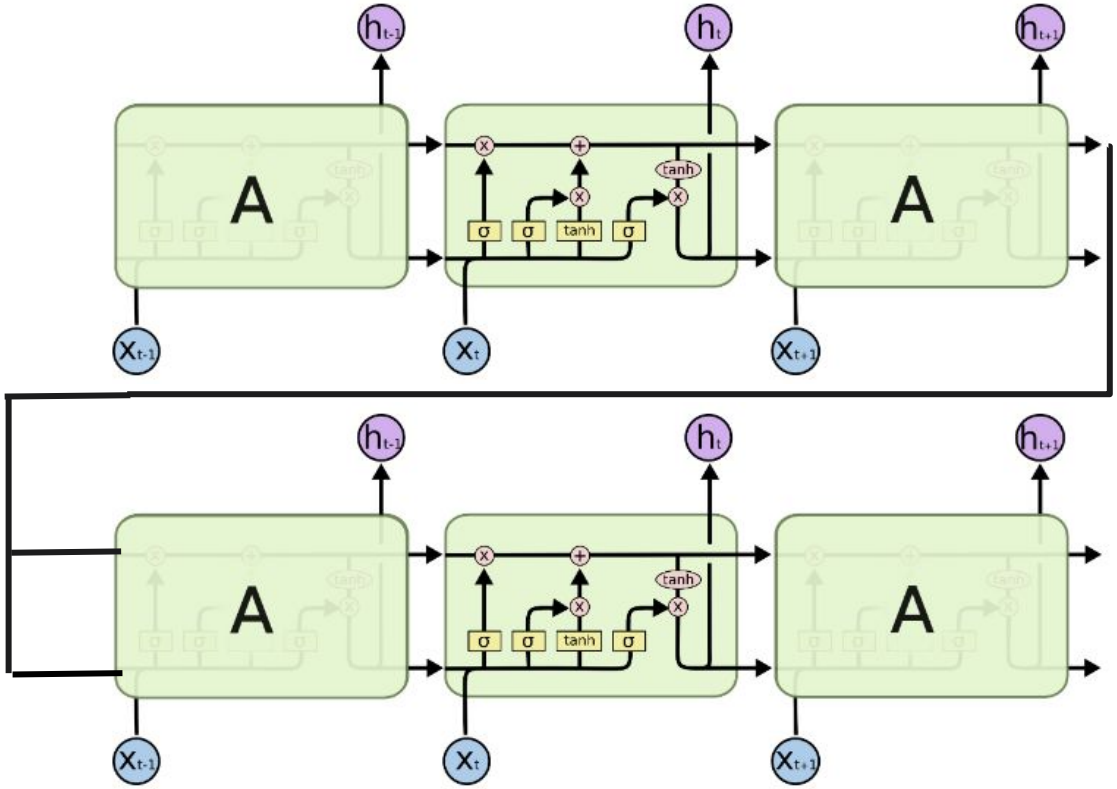


Understanding LSTM Networks -- colah's blog [4]



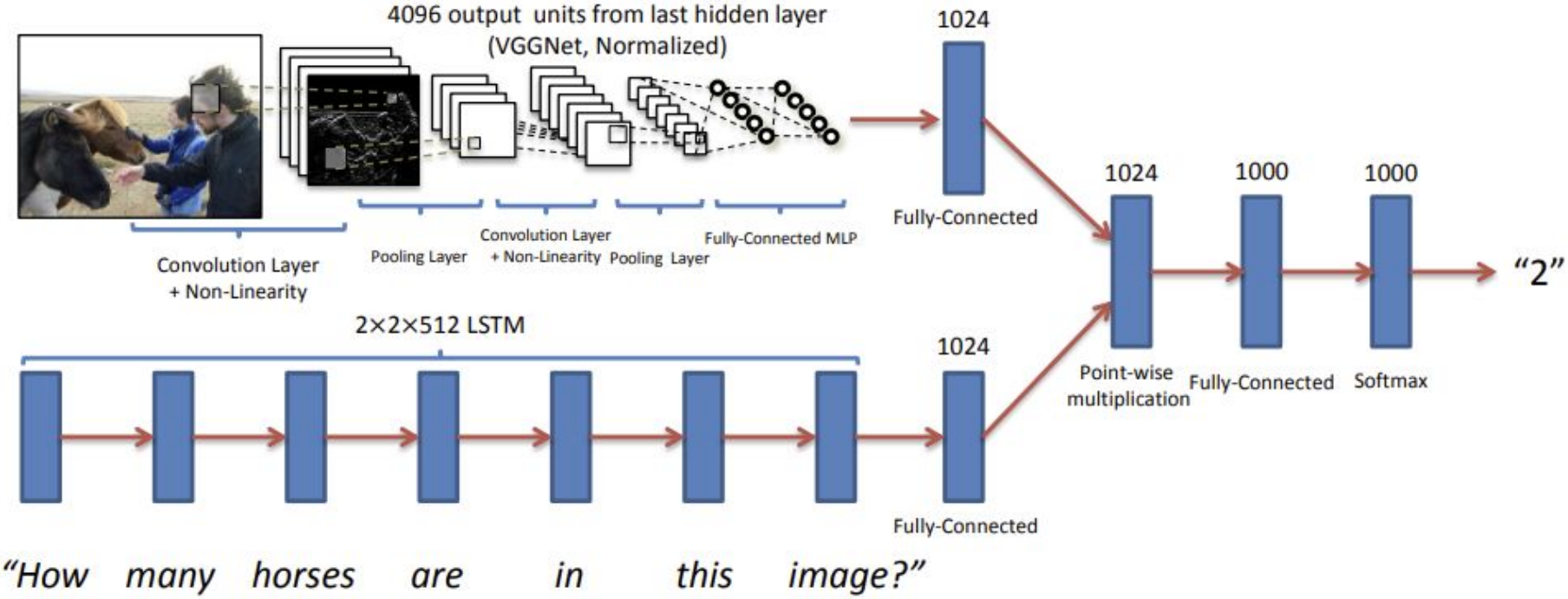
Understanding LSTM Networks -- colah's blog [4]

Deeper LSTM Q



Understanding LSTM Networks -- colah's blog [4]

Model Architecture

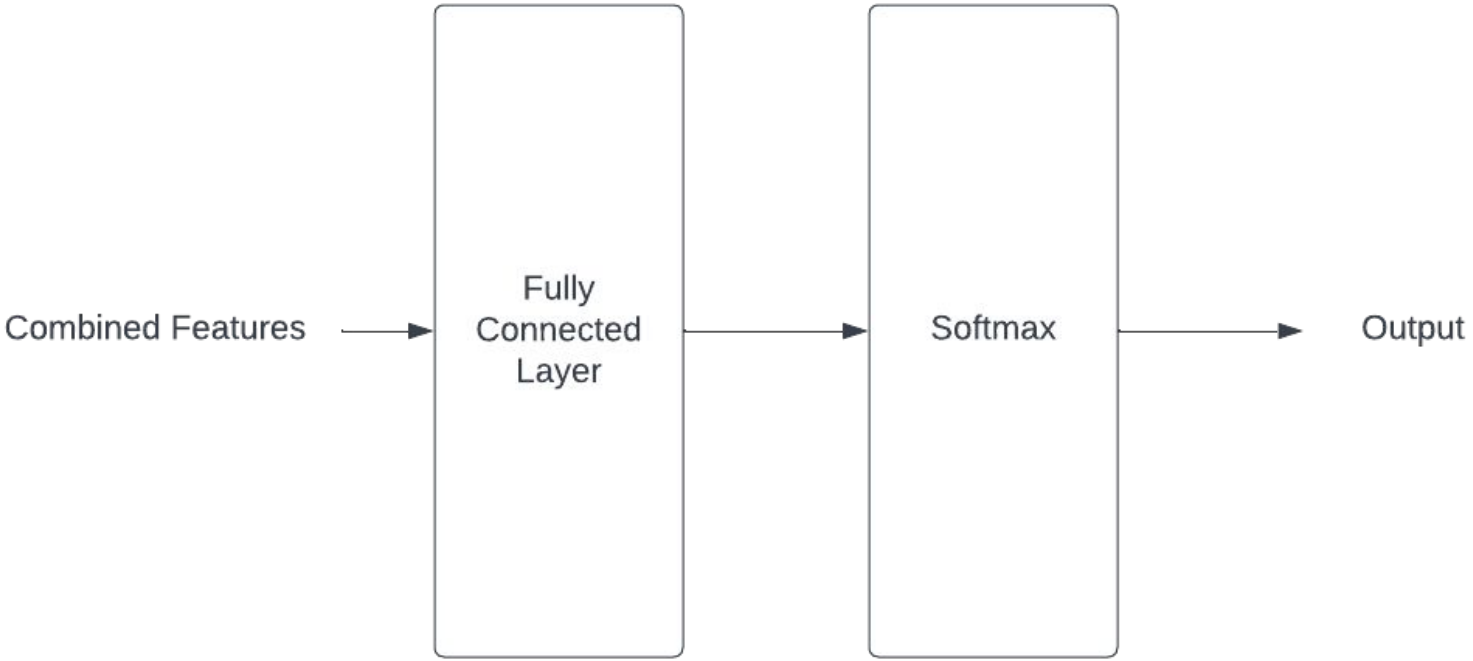


1505.00468.pdf (arxiv.org) [1]

Feature Combination and MLP

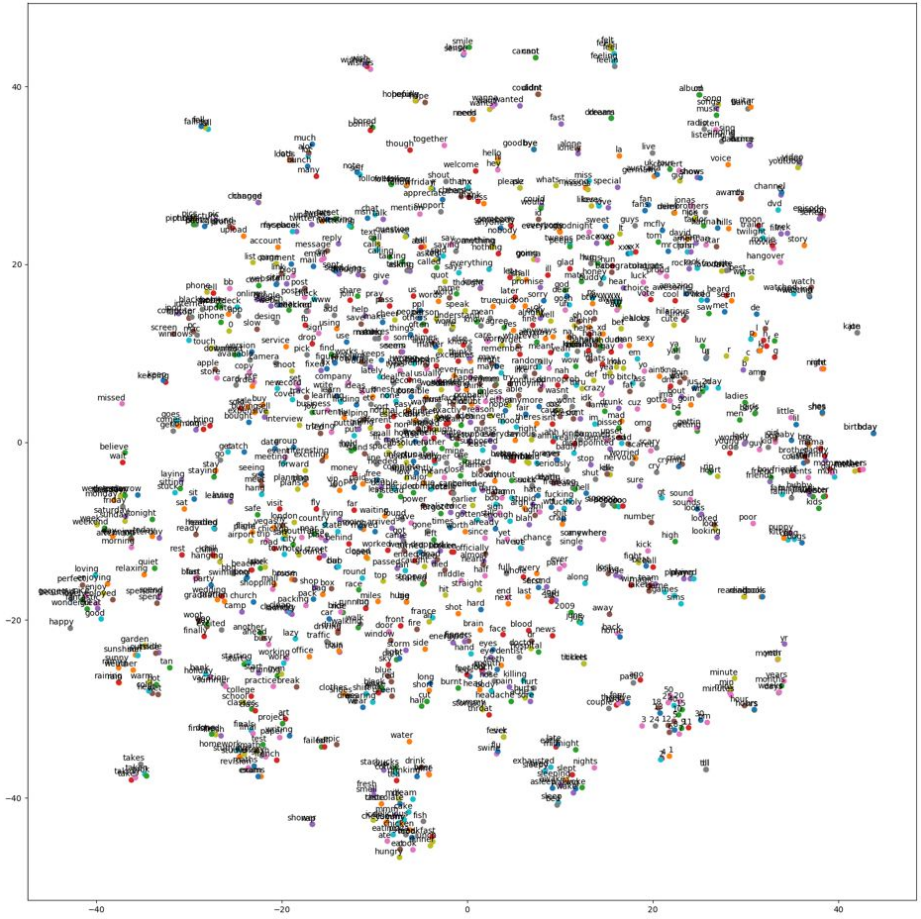
Model Structure	Combination Approach
BoW Q + Image (1030, 1) + (1024, 1)	Concatenation (2054, 1)
LSTM Q + Image (1024, 1) X (1024, 1)	Element-wise multiplication (1024, 1)

Output Perceptron



Model Baselines

- Random
- Prior(yes)
- Per Q-type prior
- Nearest neighbor



vector - word mapping for 2D word embedding - Stack Overflow

Cosine Similarity

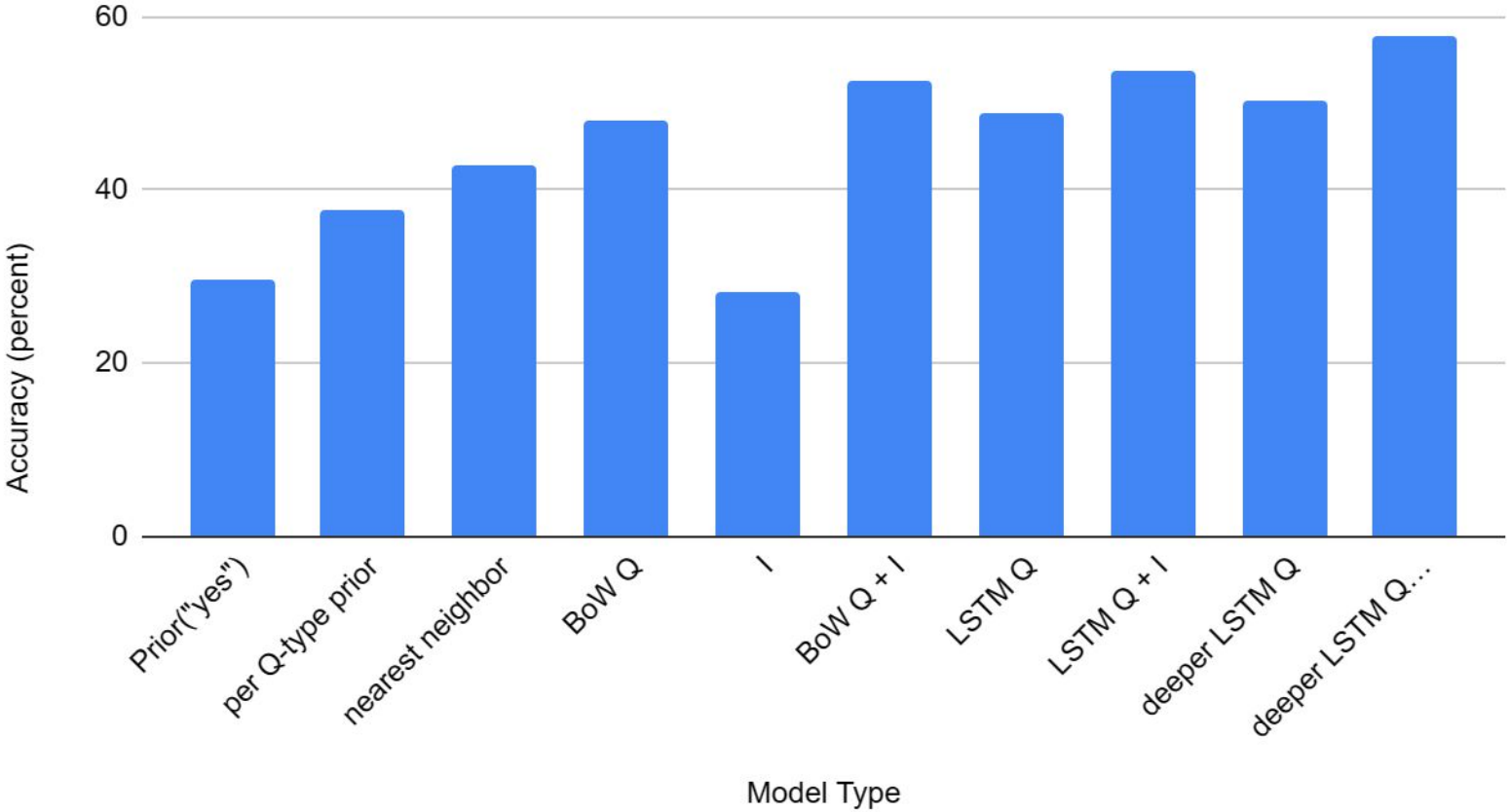
$$similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Cosine Similarity - Understanding the math and how it works? (with python) (machinelearningplus.com) [2]

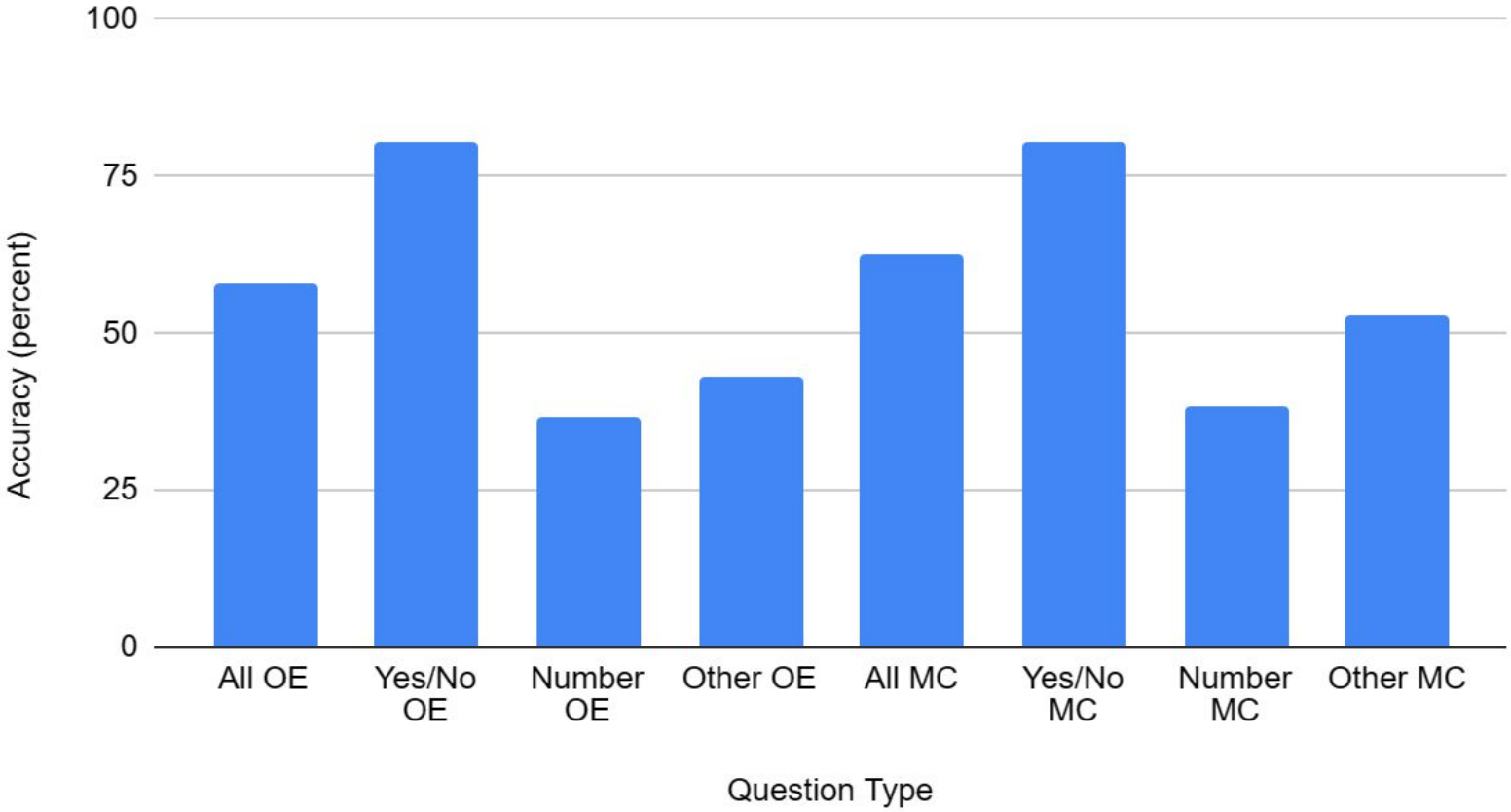
	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
prior ("yes")	29.66	70.81	00.39	01.15	29.66	70.81	00.39	01.15
per Q-type prior	37.54	71.03	35.77	09.38	39.45	71.02	35.86	13.34
nearest neighbor	42.70	71.89	24.36	21.94	48.49	71.94	26.00	33.56
BoW Q	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
I	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BoW Q + I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q + I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
deeper LSTM Q	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**	**62.70**	**80.52**	**38.22**	**53.01**
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
BoW Q + C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

1505.00468.pdf (arxiv.org) [1]

Accuracy by Model Type



Accuracy of deeper LSTM Q + norm I across Q type

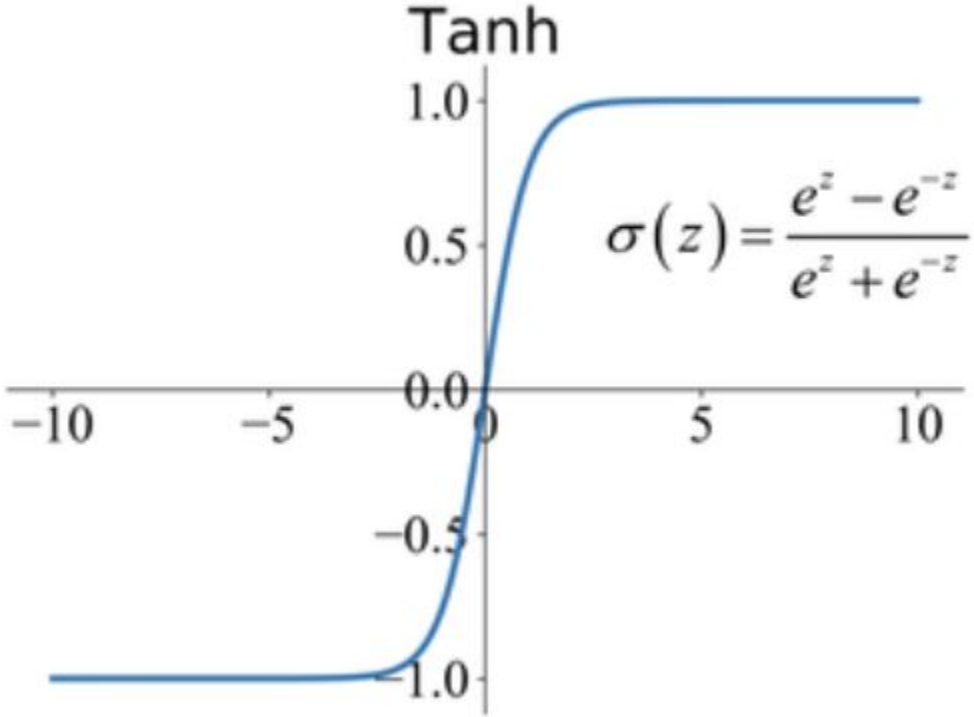


Question Type	Open-Ended K = 1000 Q	Open-Ended K = 1000 Q + I	Open-Ended K = 1000 Q + C	Open-Ended Human Q	Open-Ended Human Q + I	Human Age To Be Able To Answer	Commonsense To Be Able To Answer (%)
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07	27.52
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60	13.22
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55	40.34
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03	28.72
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04	38.92
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51	30.30
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13	45.32
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67	15.93
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65	30.63
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29	38.97
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54	36.51
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25	19.88
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18	73.56
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27	30.00
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23	37.68
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02	33.27
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81	31.83
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49	43.82
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07	31.87
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75	18.04
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50	41.33

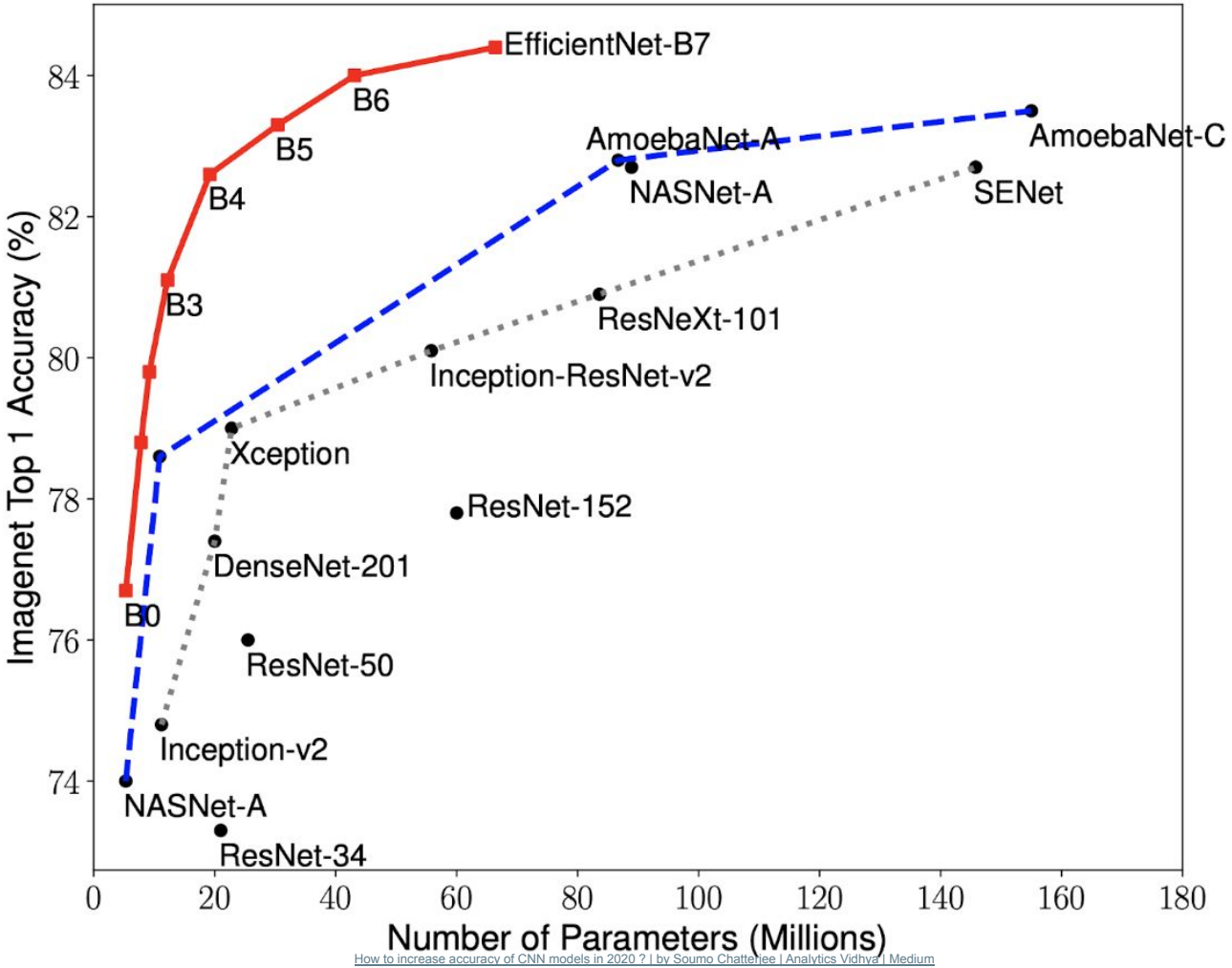
Paper Strengths

- Establishing similarity between abstract image dataset and MS COCO
- Large dataset and no fixed classes/features
- Broad coverage of contexts

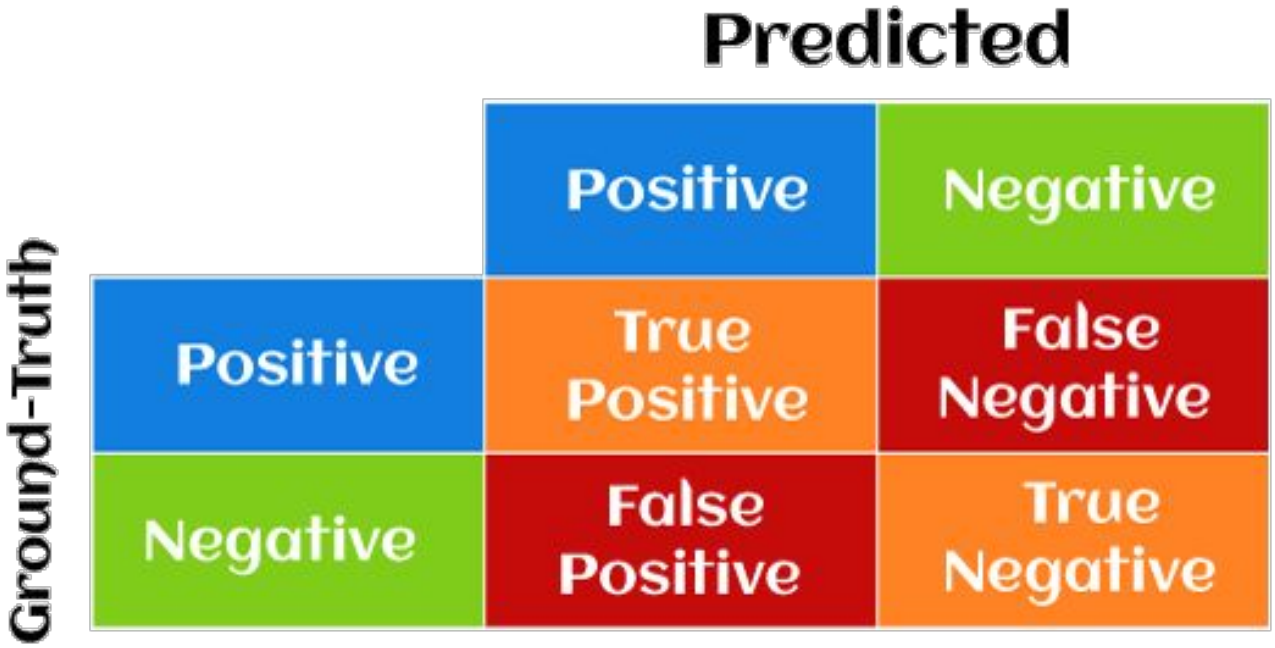
Paper Weaknesses



Tanh Activation Explained | Papers With Code



How to increase accuracy of CNN models in 2020 ? | by Soumo Chatterjee | Analytics Vidhya | Medium



#66: Precision and Recall in Machine Learning - Machine Learning with Python - Tutorial Learning Code (devsfix.com)

Future Work/Open Research Questions

- How does modern techniques effect VQA in respect to AI completeness?
- Does more given context (longer captions, more detailed questions) improve model results?
- Which objects/contexts are most difficult for the model to process
- Does improvements to accuracy on multimodal tasks, correspond to improvements on related tasks

References

[1] Agrawal, Aishwarya, et al. “VQA: Visual Question Answering.” *ArXiv.org*, 27 Oct. 2016, https://arxiv.org/abs/1505.00468.

[2] Prabhakaran, Selva. “Cosine Similarity - Understanding the Math and How It Works? (with Python).” *Machine Learning Plus*, 20 Apr. 2022, https://www.machinelearningplus.com/nlp/cosine-similarity/.

[3] “Regularization for Simplicity: L_2 Regularization | Machine Learning | Google Developers.” *Google*, Google, https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/l2-regularization.

[4] “Understanding LSTM Networks.” *Understanding LSTM Networks -- Colah's Blog*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[5] “What Is Ai-Complete?: Ai Terms Explained - AI for Anyone.” *What Is AI-Complete?: AI Terms Explained - AI For Anyone*, https://www.aiforanyone.org/glossary/ai-complete.