DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day-16

Q1.What is Statistics Learning?

Answer:

Statistical learning: It is the framework for understanding data based on the statistics, which can be classified as the supervised or unsupervised. Supervised statistical learning involves building the statistical model for predicting, or estimating, an output based on one or more inputs, while in unsupervised statistical learning, there are inputs but no supervising output, but we can learn relationships and structure from such data.

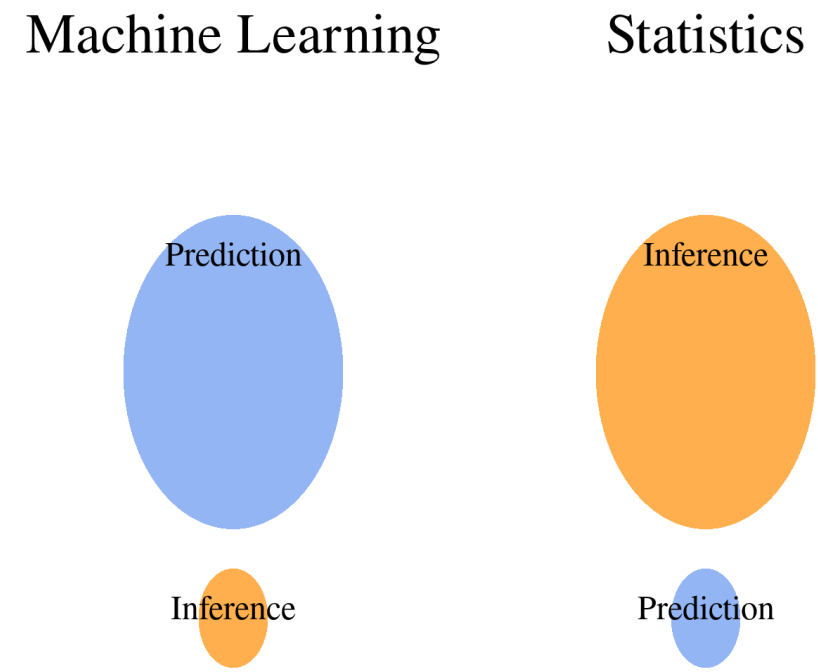
$Y = f(X) + \varepsilon$,$X = (X1, X2, \ldots, Xp)$,

f : It is an unknown function & ε is random error (reducible & irreducible).

Prediction & Inference:

In the situations , where the set of inputs X are readily available, but the output Y is not known, we often treat f as the black box (not concerned with the exact form of "f"), as long as it yields the accurate predictions for Y. This is the *prediction*.

There are the situations where we are interested in understanding the way that Y is affected as X change. In this type of situation, we wish to estimate f, but our goal is not necessarily to make the predictions for Y. Here we are more interested in understanding the relationship between the X and Y. Now f cannot be treated as the black box, because we need to know it's exact form. This is *inference*.

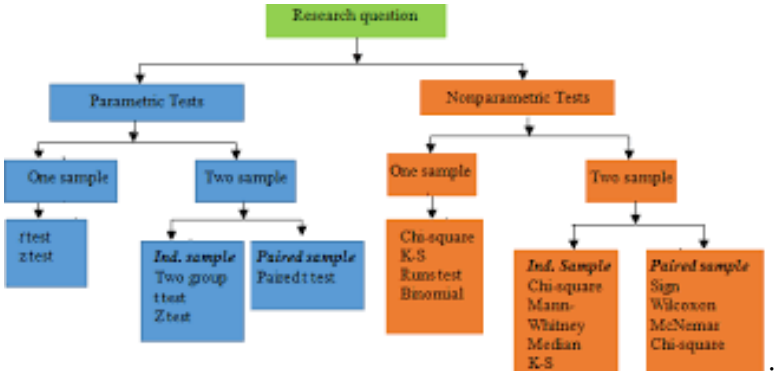


Parametric & Non-parametric methods

Parametric statistics: This statistical tests based on underlying the assumptions about data's distribution. In other words, It is based on the parameters of the normal curve. Because parametric statistics are based on the normal curve, data must meet certain assumptions, or parametric statistics cannot be calculated. Before running any parametric statistics, you should always be sure to test the assumptions for the tests that you are planning to run.

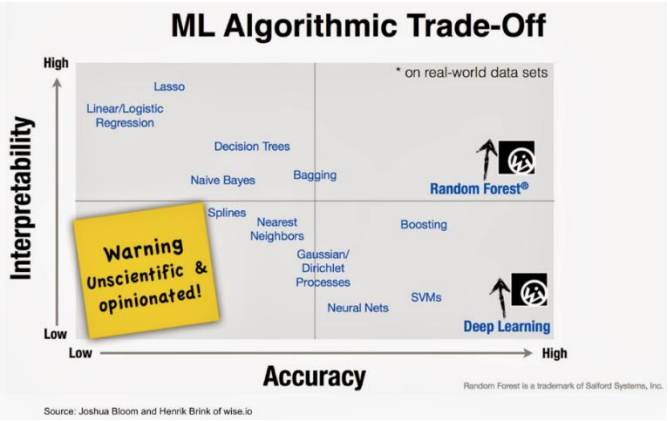
$$f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

As by the name, nonparametric statistics are not based on parameters of the normal curve. Therefore, if our data violate the assumptions of a usual parametric and nonparametric statistics might better define the data, try running the nonparametric equivalent of the parametric test. We should also consider using nonparametric equivalent tests when we have limited sample sizes (e.g., n < 30). Though the nonparametric statistical tests have more flexibility than do parametric statistical tests, nonparametric tests are not as robust; therefore, most statisticians recommend that when appropriate, parametric statistics are preferred.



Prediction Accuracy and Model Interpretability:

Out of many methods that we use for the statistical learning, some are less flexible and more restrictive . When inference is the goal, then there are clear advantages of using the simple and relatively inflexible statistical learning methods. When we are only interested in the prediction, we use flexible models available.



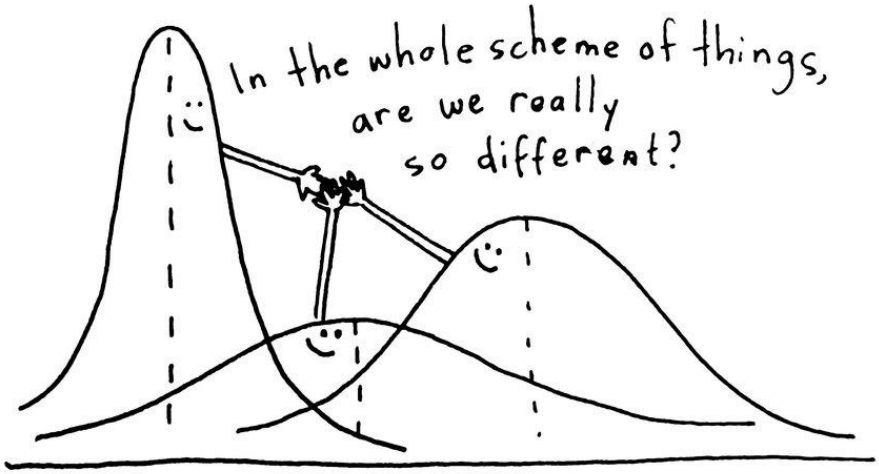
Q2. What is ANOVA?

Answer:

ANOVA: it stands for " Analysis of Variance " is an extremely important tool for analysis of data (both One Way and Two Way ANOVA is used). It is a statistical method to compare the population means of two or more groups by analyzing variance. The variance would differ only when the means are significantly different.

ANOVA test is the way to find out if survey or experiment results are significant. In other words, It helps us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis. We are testing groups to see if there's a difference between them. Examples of when we might want to test different groups:

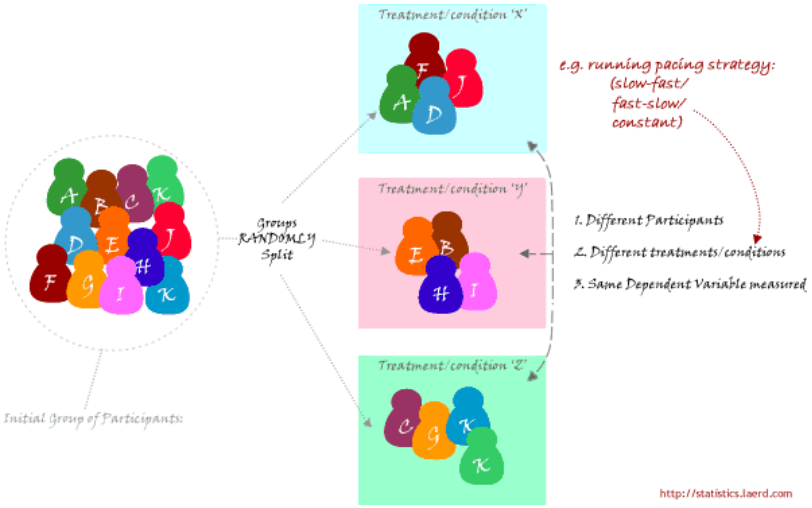
- The group of psychiatric patients are trying three different therapies: counseling, medication, and biofeedback. We want to see if one therapy is better than the others.
- The manufacturer has two different processes to make light bulbs if they want to know which one is better.
- Students from the different colleges take the same exam. We want to see if one college outperforms the other.



Types of ANOVA:

- One-way ANOVA
- Two-way ANOVA

One-way ANOVA is the hypothesis test in which only one categorical variable or the single factor is taken into consideration. With the help of F-distribution, it enables us to compare means of three or more samples. The Null hypothesis (H0) is the equity in all population means while an Alternative hypothesis is the difference in at least one mean.



There are two-ways ANOVA examines the effect of two independent factors on a dependent variable. It also studies the inter-relationship between independent variables influencing the values of the dependent variable, if any.

Q3. What is ANCOVA?

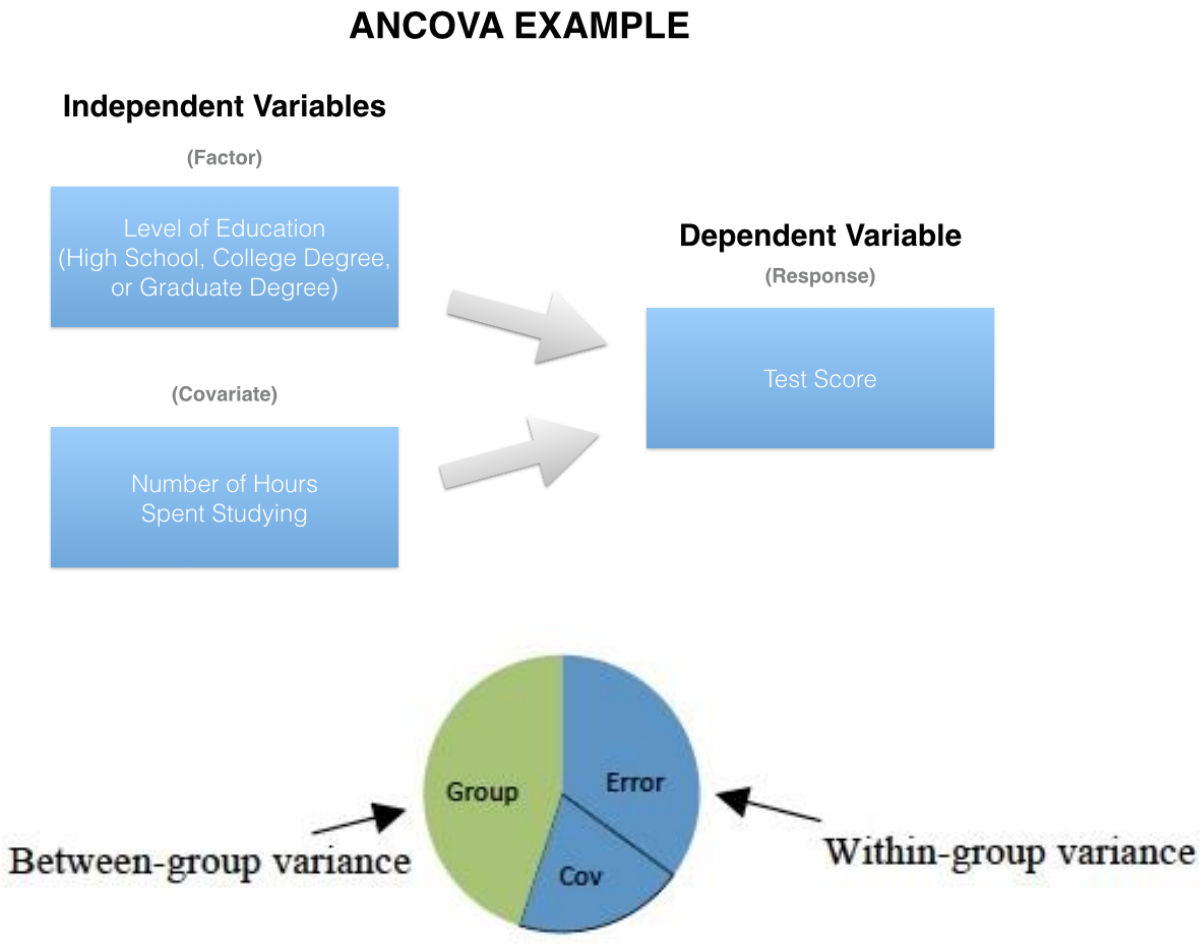
Answer:

Analysis of Covariance (ANCOVA): It is the inclusion of the continuous variable in addition to the variables of interest (the dependent and independent variable) as means for the control. Because the ANCOVA is the extension of the ANOVA, the researcher can still assess main effects and the interactions to answer their research hypotheses. The difference between ANCOVA and an ANOVA is that an ANCOVA model includes the "covariate" that is correlated with dependent variable and means on dependent variable are adjusted due to effects the covariate has on it. Covariates can also

be used in many ANOVA based designs: such as between-subjects, within-subjects (repeated measures), mixed (between – and within – designs), etc. Thus, this technique answers the question

In simple terms, The difference between ANOVA and the ANCOVA is the letter "C", which stands for 'covariance'. Like ANOVA, "Analysis of Covariance" (ANCOVA) has the single continuous response variable. Unlike ANOVA, ANCOVA compares the response variable by both the factor and a continuous independent variable (example comparing test score by both 'level of education' and the 'number of hours spent in studying'). The terms for the continuous independent variable (IV) used in the ANCOVA is "covariate".

Example of ANCOVA

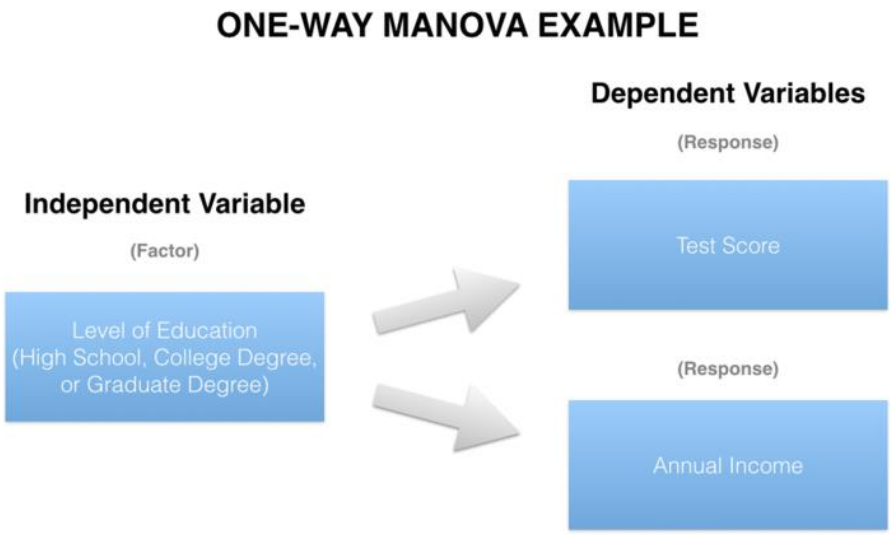


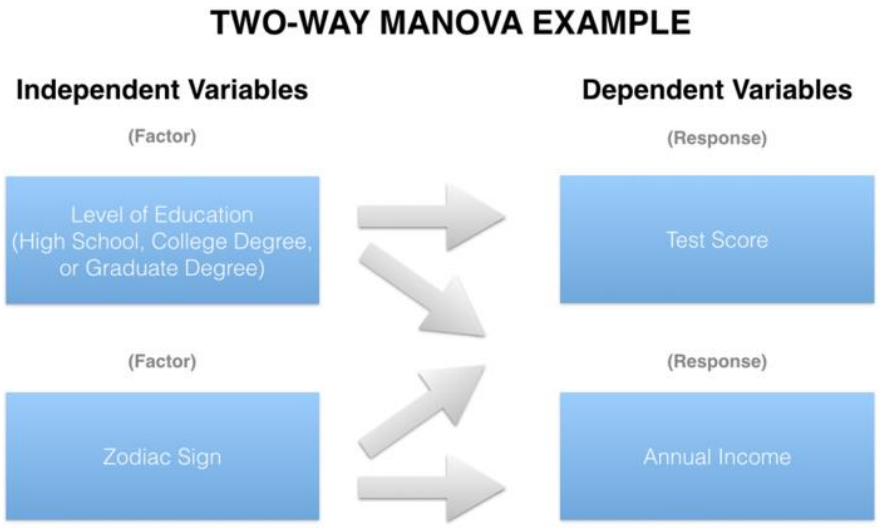
Q4. What is MANOVA?

Answer:

MANOVA (multivariate analysis of variance): It is a type of multivariate analysis used to analyze data that involves more than one dependent variable at a time. MANOVA allows us to test hypotheses regarding the effect of one or more independent variables on two or more dependent variables.

The obvious difference between ANOVA and the "Multivariate Analysis of Variance" (MANOVA) is the “M”, which stands for multivariate. In basic terms, MANOVA is an ANOVA with two or more continuous response variables. Like ANOVA, MANOVA has both the one-way flavor and a two-way flavor. The number of factor variables involved distinguish the one-way MANOVA from a two-way MANOVA.





When comparing the two or more continuous response variables by the single factor, a one-way MANOVA is appropriate (e.g. comparing ‘test score’ and ‘annual income’ together by ‘level of

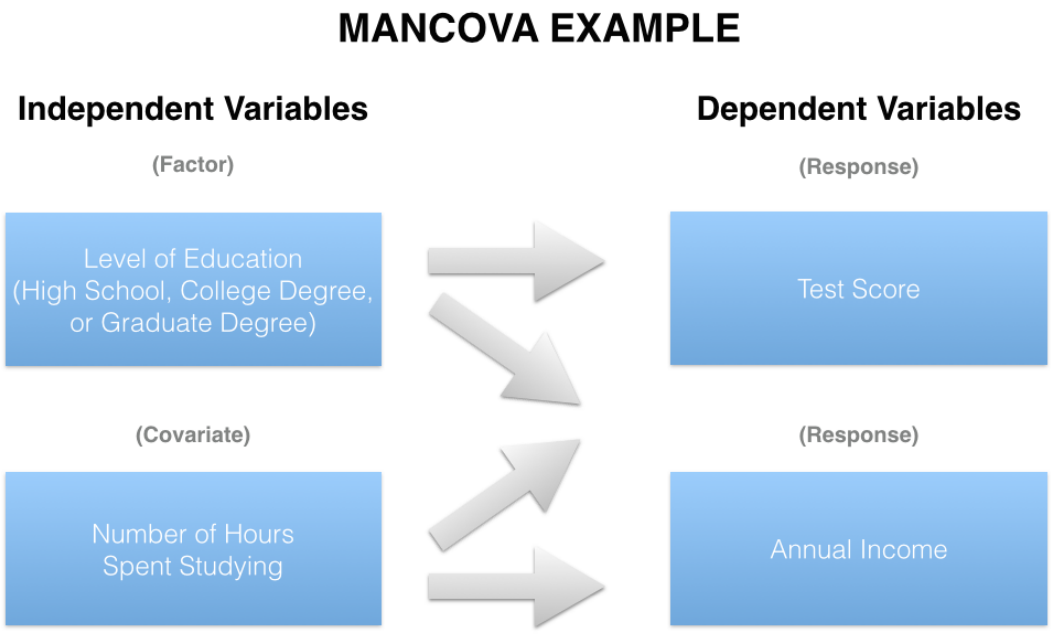
education'). The two-way MANOVA also entails two or more continuous response variables, but compares them by at least two factors (e.g. comparing 'test score' and 'annual income' together by both 'level of education' and 'zodiac sign').

Q5. What is MANCOVA?

Answer:

Multivariate analysis of covariance (MANCOVA): It is a statistical technique that is the extension of analysis of covariance (ANCOVA). It is the multivariate analysis of variance (MANOVA) with a covariate(s).). In MANCOVA, we assess for statistical differences on multiple continuous dependent variables by an independent grouping variable, while controlling for a third variable called the covariate; multiple covariates can be used, depending on the sample size. Covariates are added so that it can reduce error terms and so that the analysis eliminates the covariates' effect on the relationship between the independent grouping variable and the continuous dependent variables.

ANOVA and ANCOVA, the main difference between the MANOVA and MANCOVA, is the "C," which again stands for the "covariance." Both the MANOVA and MANCOVA feature two or more response variables, but the key difference between the two is the nature of the IVs. While the MANOVA can include only factors, an analysis evolves from MANOVA to MANCOVA when one or more covariates are added to the mix.



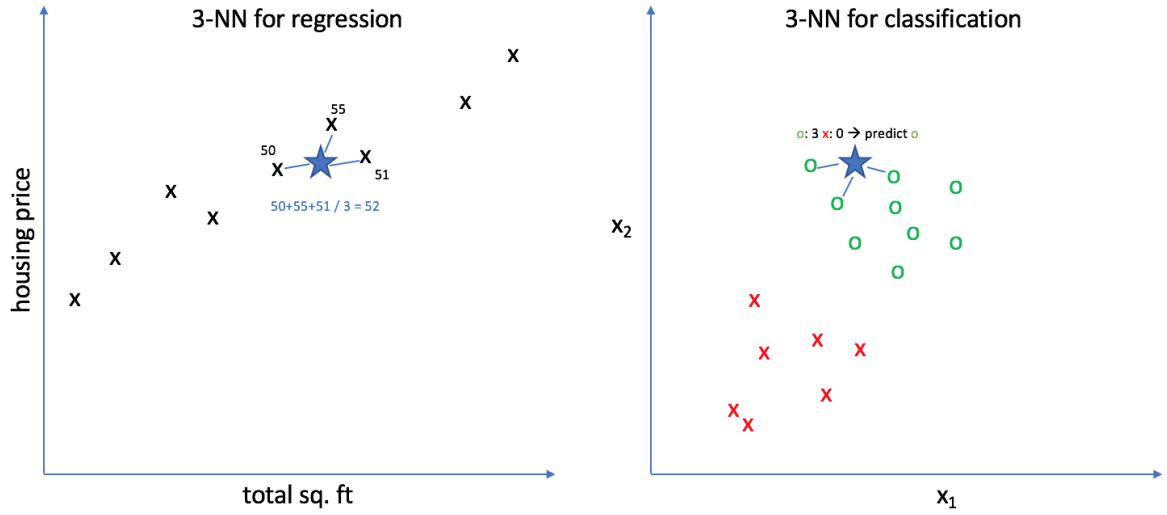
Q6. Explain the differences between KNN classifier and KNN regression methods.

Answer:

They are quite similar. Given a value for KK and a prediction point x0x0, KNN regression first identifies the KK training observations that are closes to x0x0, represented by N0. It then estimates f(x0) using the average of all the training responses in N0. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

So the main difference is the fact that for the classifier approach, the algorithm assumes the outcome as the class of more presence, and on the regression approach, the response is the average value of the nearest neighbors.



Q7. What is t-test?

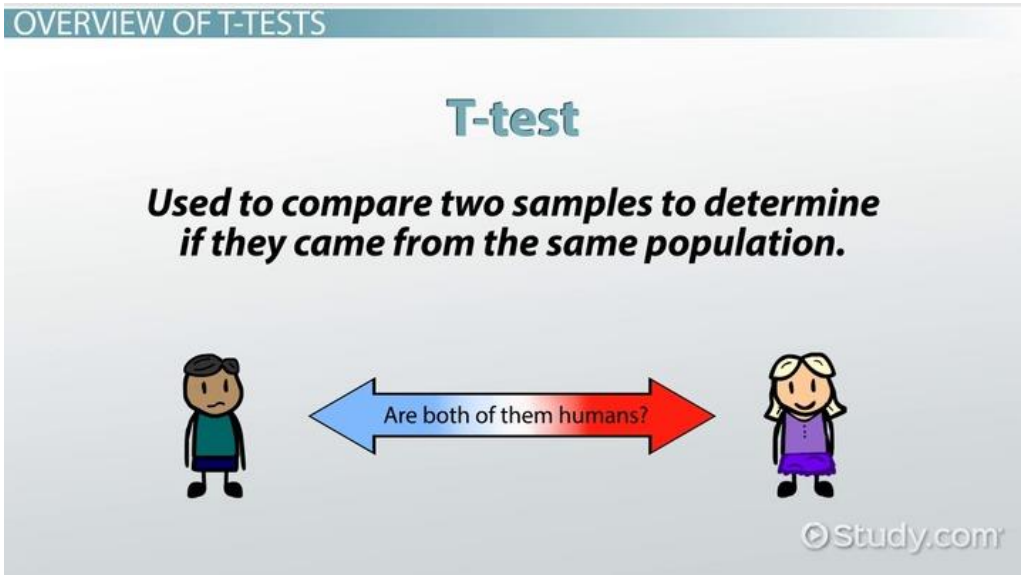
Answer:

To understand T-Test Distribution, Consider the situation, you want to compare the performance of two workers of your company by checking the average sales done by each of them, or to compare the performance of a worker by comparing the average sales done by him with the standard value. In such situations of daily life, t distribution is applicable.

A t-test is the type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

Understand t-test with Example: Let's say you have a cold, and you try a naturopathic remedy. Your cold lasts a couple of days. The next time when you have a cold, you buy an over-the-counter pharmaceutical, and the cold lasts a week. You survey your friends, and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you want to know is, are these results repeatable? A t-test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.

Type	T-statistic	Degrees of freedom
One-sample t-test	$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$	df = n – 1
Paired t-test	$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$	df = n – 1



Q8. What is Z-test?

Answer:

z-test: It is a statistical test used to determine whether the two population means are different when the variances are known, and the sample size is large. The test statistic is assumed to have the normal distribution, and nuisance parameters such as standard deviation should be known for an accurate z-test to be performed.

Another definition of Z-test: A Z-test is a type of hypothesis test. Hypothesis testing is just the way for you to figure out if results from a test are valid or repeatable. Example, if someone said they had found the new drug that cures cancer, you would want to be sure it was probably true. Hypothesis test will tell you if it's probably true or probably not true. A Z test is used when your data is approximately normally distributed.

Z-Tests Working :

Tests that can be conducted as the z-tests include one-sample location test, a two-sample location test, a paired difference test, and a maximum likelihood estimate. Z-tests are related to t-tests, but t-tests are best performed when an experiment has the small sample size. Also, T-tests assumes the standard deviation is unknown, while z-tests assumes that it is known. If the standard deviation of the population is unknown, then the assumption of the sample variance equaling the population variance is made.

When we can run the Z-test :

Different types of tests are used in the statistics (i.e., f test, chi-square test, t-test). You would use a Z test if:

- Your sample size is greater than 30. Otherwise, use a t-test.
- Data points should be independent from each other. Some other words, one data point is not related or doesn't affect another data point.
- Your data should be normally distributed. However, for large sample sizes (over 30), this doesn't always matter.
- Your data should be randomly selected from a population, where each item has an equal chance of being selected.
- Sample sizes should be equal, if at all possible.

Z-TEST

- Formula to find the value of Z (z-test) Is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\bar{x}$ = mean of sample
- μ_0 = mean of population
- σ = standard deviation of population
- n = no. of observations

Q9. What is Chi-Square test?

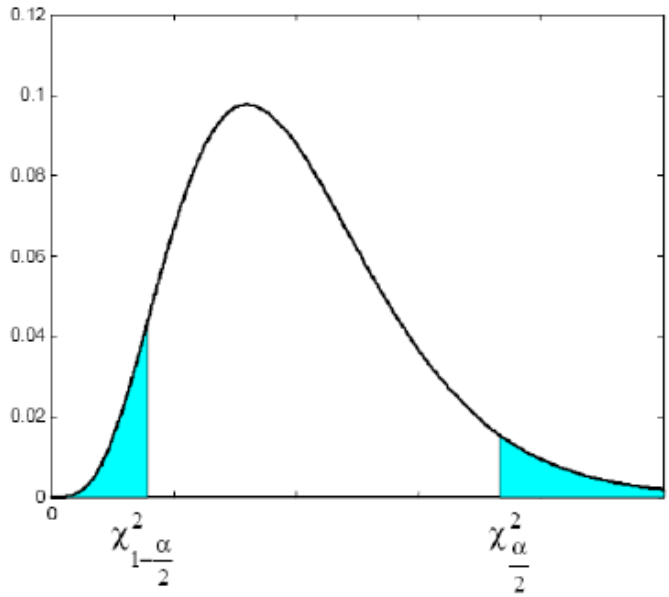
Answer:

Chi-square (χ^2) statistic: It is a test that measures how expectations compare to actual observed data (or model results). The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a coin 100 times meet these criteria.

Chi-square test is intended to test how it is that an observed distribution is due to chance. It is also called the **"goodness of fit"** statistic because it measures how well the observed distribution of the data fits with the distribution that is expected if the variables are independent.

Chi-square test is designed to analyze the **categorical** data. That means that the data has been counted and divided into categories. It will not work with parametric or continuous data (such as height in inches). For example, if you want to test whether attending class influences how students perform on an exam, using test scores (from 0-100) as data would not be appropriate for a Chi-square test. However, arranging students into the categories "Pass" and "Fail" would. Additionally, the data in a Chi-square grid should not be in the form of percentages, or anything other than frequency (count) data.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$



Q10. What is correlation and the covariance in the statistics?

Answer:

The Covariance and Correlation are two mathematical concepts; these two approaches are widely used in the statistics. Both Correlation and the Covariance establish the relationship and also measures the dependency between the two random variables, the work is similar between these two, in the mathematical terms, they are different from each other.

Correlation: It is the statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

Correlation Formula

$$\rho_{xy} = \frac{Con(r_x, r_y)}{\sigma_x \sigma_y}$$

Covariance: It measures the directional relationship between the returns on two assets. The positive covariance means that asset returns move together while a negative covariance means they move inversely. Covariance is calculated by analyzing at-return surprises (standard deviations from the expected return) or by multiplying the correlation between the two variables by the standard deviation of each variable.

Covariance Formula

For Population

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

DATA SCIENCE INTERVIEW PREPARATION

(30 Days of Interview Preparation)

DAY 17

Q1. What is ERM (Empirical Risk Minimization)?

Answer:

Empirical risk minimization (ERM): It is a principle in statistical learning theory which defines a family of learning algorithms and is used to give theoretical bounds on their performance. The idea is that we don't know exactly how well an algorithm will work in practice (the true "risk") because we don't know the true distribution of data that the algorithm will work on, but as an alternative we can measure its performance on a known set of training data.

We assumed that our samples come from this distribution and use our dataset as an approximation. If we compute the loss using the data points in our dataset, it's called empirical risk. It is "empirical" and not "true" because we are using a dataset that's a subset of the whole population.

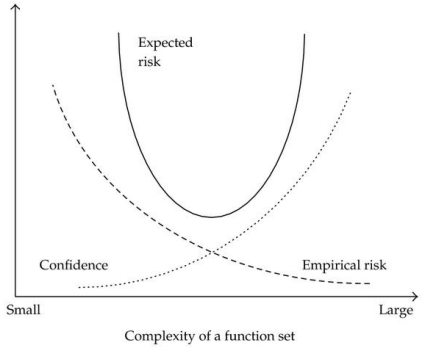
When our learning model is built, we have to pick a function that minimizes the empirical risk that is the delta between predicted output and actual output for data points in the dataset. This process of finding this function is called empirical risk minimization (ERM). We want to minimize the true risk. We don't have information that allows us to achieve that, so we hope that this empirical risk will almost be the same as the true empirical risk.

Let's get a better understanding by Example

We would want to build a model that can differentiate between a male and a female based on specific features. If we select 150 random people where women are really short, and men are really tall, then the model might incorrectly assume that height is the differentiating feature. For building a truly accurate model, we have to gather all the women and men in the world to extract differentiating features. Unfortunately, that is not possible! So we select a small number of people and hope that this sample is representative of the whole population.

$$R_{emp}[f] = \sum_{x \in X} \sum_{j=1}^{2} c(x, y_j, f(x)) p_{emp}(y_j, x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_i, f(x_i))$$



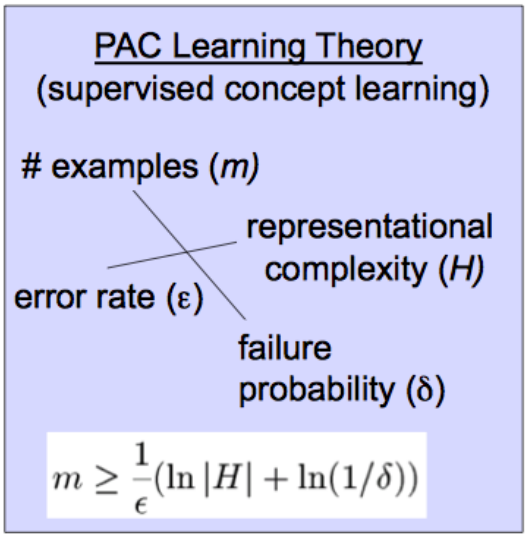
Q2. What is PAC (Probably Approximately Correct)?

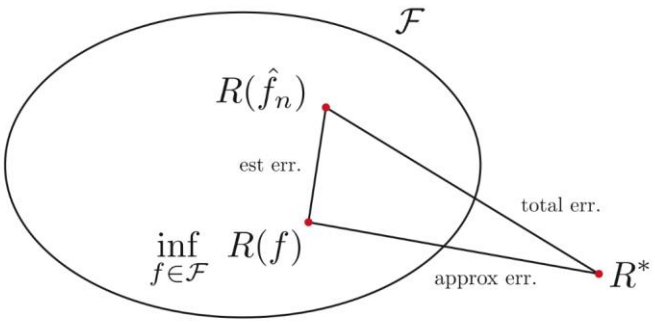
Answer:

PAC: In computational learning theory, probably approximately correct (PAC) learning is a framework for mathematical analysis of machine learning.

The learner receives samples and must have to pick a generalization function (called the *hypothesis*) from a specific class of possible functions. Our goal is that, with high probability, the selected function will have low generalization error. The learner must be able to learn the concept given any arbitrary approximation ratio, probability of success, or distribution of the samples.

Hypothesis class is PAC(Probably Approximately Correct) learnable if there exists a function **m_H** and algorithm that for any labeling function **f,** distribution **D** over the domain of inputs **X,** **delta** and **epsilon** that with **m ≥ m_H** produces a hypothesis **h** like that with probability 1-**delta** it returns a **true error** lower than **epsilon.** Labeling function is nothing other than saying that we have a specific function **f** that labels the data in the domain.





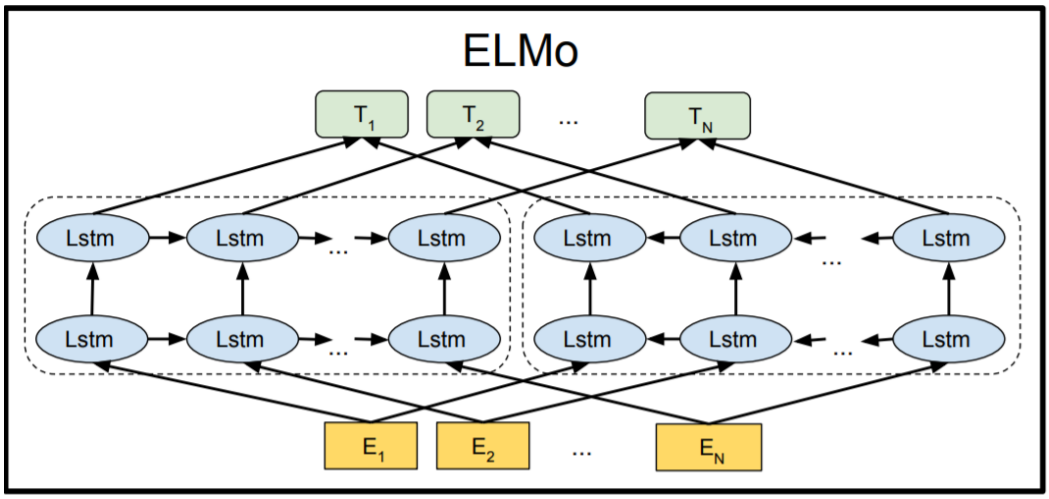
Q3. What is ELMo?

Answer:

ELMo is a novel way to represent words in vectors or embeddings. These word embeddings help achieve state-of-the-art (SOTA) results in several NLP tasks:

Task	Previous SOTA		ELMo + Baseline
SQuAD	SAN	84.4	85.8
SNLI	Chen et al (2017)	88.6	88.7 +/- 0.17
SRL	He et al (2017)	81.7	84.6
Coref	Lee et al (2017)	67.2	70.4
NER	Peters et al (2017)	91.93 +/- 0.19	92.22 +/- 0.10
Sentiment (5-class)	McCann et al (2017)	53.7	54.7 +/- 0.5

It is a deep contextualized word representation that models both complex characteristics of word use (e.g., syntax and semantics), and how these uses vary across linguistic contexts. These word vectors are learned functions of internal states of a deep biLM(bidirectional language model), which is pre-trained on large text corpus. They could be easily added to existing models and significantly improve state of the art across a broad range of challenging NLP problems, including question answering, textual entailment and sentiment analysis.



Q4. What is Pragmatic Analysis in NLP?

Answer:

Pragmatic Analysis(PA): It deals with outside word knowledge, which means understanding i.e external to documents and queries. PA that focuses on what was described is reinterpreted by what it actually meant, deriving the various aspects of language that require real-world knowledge.

It deals with overall communicative and social content and its effect on interpretation. It means abstracting the meaningful use of language in situations. In this analysis, the main focus always on what was said in reinterpreted on what is intended.

It helps users to discover this intended effect by applying a set of rules that characterize cooperative dialogues.

E.g., "close the window?" should be interpreted as a request instead of an order.

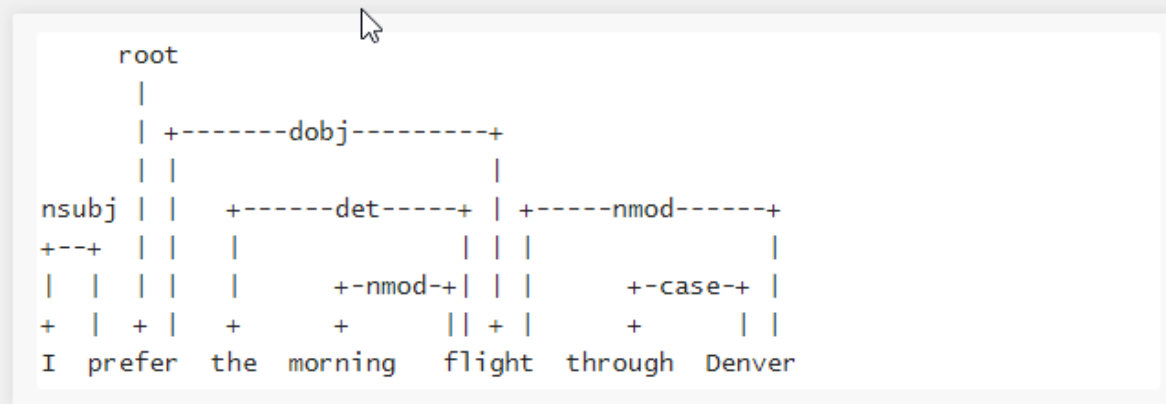
Personal	Organizational devices (Calls/Vocatives)
Request	Questions
Action	Answers/Responses
Permission	Repetition/Imitation
Offering/showing	Elicited identification
Descriptions	Routines
Statements (Internal reports)	Exclamations
Acknowledgements	Unclassified
Performatives	Double coded

Q5. What is Syntactic Parsing?

Answer:

Syntactic Parsing or **Dependency Parsing:** It is a task of recognizing a sentence and assigning a syntactic structure to it. Most Widely we used syntactic structure is the parse tree which can be generated using some parsing algorithms. These parse trees are useful in various applications like grammar checking or more importantly, it plays a critical role in the semantic analysis stage. For example to answer the question "*Who is the point guard for the LA Laker in the next game ?*" we need to figure out its subject, objects, attributes to help us figure out that the user wants the point guard of the LA Lakers specifically for the next game.

Example:



Q6. What is ULMFit?

Answer:

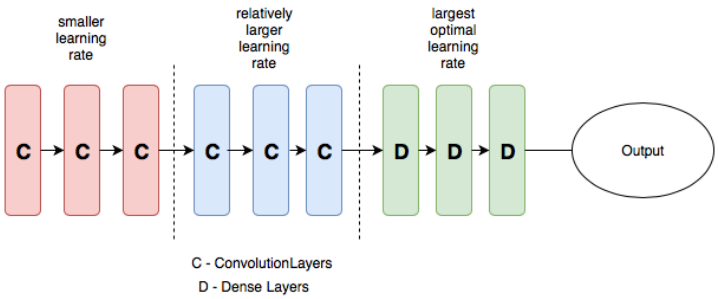
Transfer Learning in **NLP(Natural language Processing)** is an area that had not been explored with great success. But, in May 2018, **Jeremy Howard** and **Sebastian Ruder** came up with the paper – **Universal Language Model Fine-tuning for Text Classification(ULMFit)** which explores the benefits of using a pre trained model on text classification. It proposes **ULMFiT(Universal Language Model Fine-tuning for Text Classification),** a transfer learning method that could be applied to any task in NLP. In this method outperforms the state-of-the-art on six text classification tasks.

ULMFiT uses a **regular LSTM**which is the state-of-the-art language model architecture **(AWD-LSTM)**. The LSTM network has three layers. Single architecture is used throughout – for pre-training as well as for fine-tuning.

ULMFiT achieves the state-of-the-art result using novel techniques like:

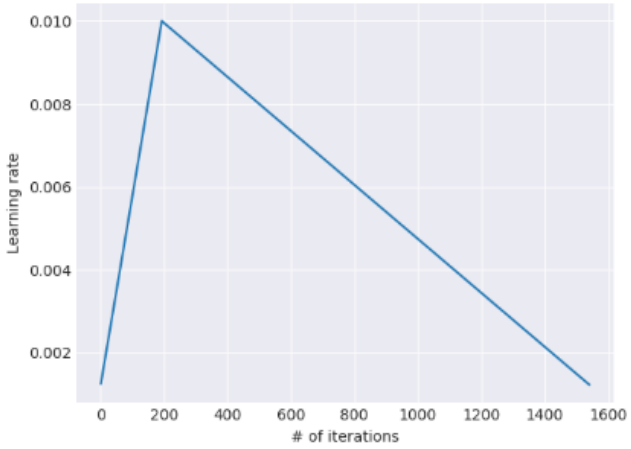
- Discriminative fine-tuning
- Slanted triangular learning rates
- Gradual unfreezing

Discriminative Fine-Tuning



Different layers of a neural network capture different types of information so they should be fine-tuned to varying extents. Instead of using the same learning rates for all layers of the model, discriminative fine-tuning allows us to tune each layer with different learning rates.

Slanted triangular learning



The model should quickly converge to a suitable region of the parameter space in the beginning of training and then later refine its parameters. Using a constant learning rate throughout training is not the best way to achieve this behaviour. Instead Slanted Triangular Learning Rates (STLR) linearly increases the learning rate at first and then linearly decays it.

Gradual Unfreezing

Gradual unfreezing is the concept of unfreezing the layers gradually, which avoids the catastrophic loss of knowledge possessed by the model. It first unfreezes the top layer and fine-tunes all the unfrozen layers for 1 epoch. It then unfreezes the next lower frozen layer and repeats until all the layers have been fine-tuned until convergence at the last iteration.

Q7. What is BERT?

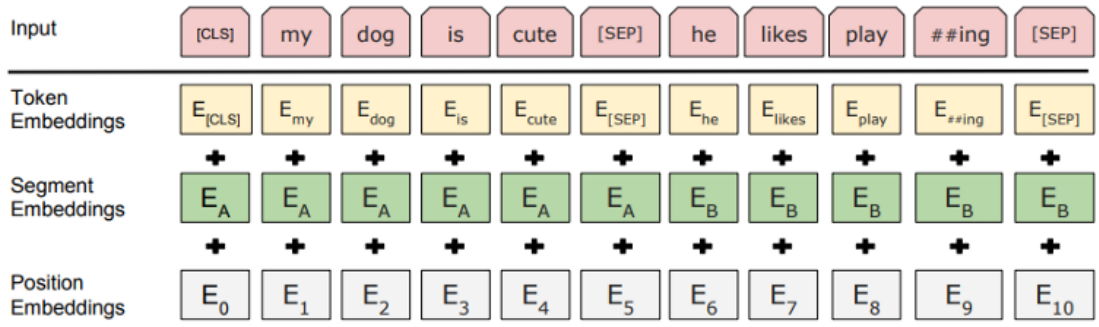
Answer:

BERT (Bidirectional Encoder Representations from Transformers) is an open-sourced NLP pre-training model developed by researchers at Google in 2018. A direct descendant to GPT (Generalized Language Models), BERT has outperformed several models in NLP and provided top results in Question Answering, Natural Language Inference (MNLI), and other frameworks.

What makes it's unique from the rest of the model is that it is the first deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. Since it's open-sourced, anyone with machine learning knowledge can easily build an NLP model without the need for sourcing massive datasets for training the model, thus saving time, energy, knowledge and resources.

How does it work?

Traditional context-free models (like word2vec or GloVe) generate a single word embedding representation for each word in the vocabulary which means the word **"right"** would have the same context-free representation in "I'm sure I'm right" and "Take a right turn." However, BERT would represent based on both previous and next context, making it bidirectional. While the concept of bidirectional was around for a long time, BERT was first on its kind to successfully pre-train bidirectional in a deep neural network.

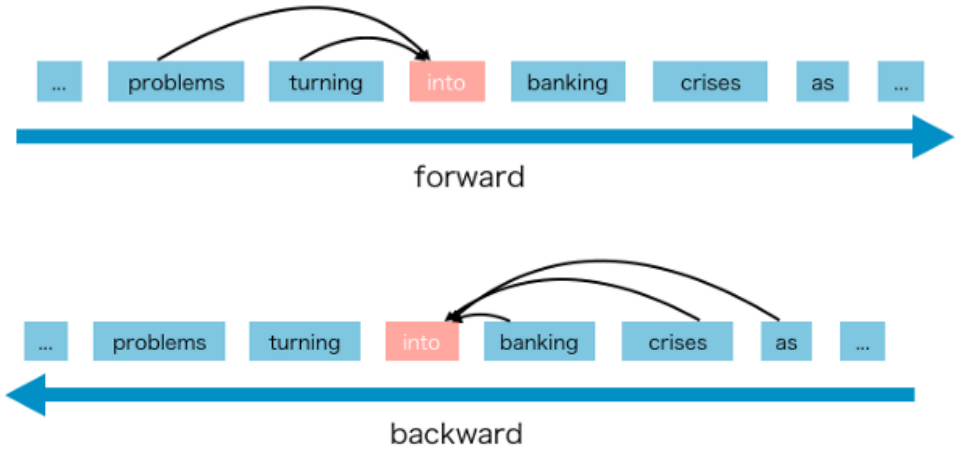


Q8.What is XLNet?

Answer:

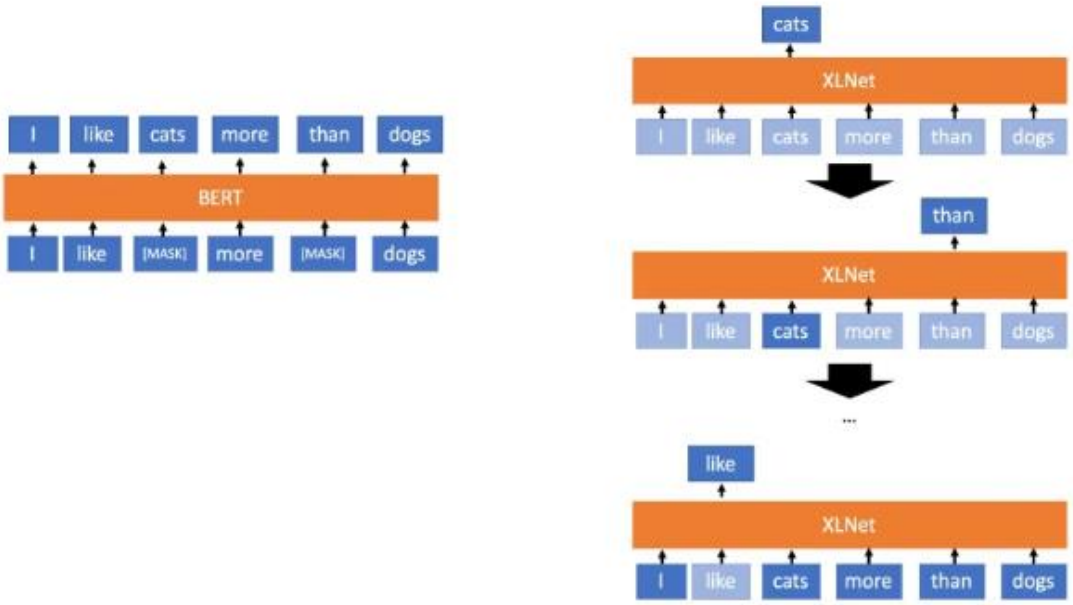
XLNet is a BERT-like model instead of a totally different one. But it is an auspicious and potential one. In one word, **XLNet is a generalized autoregressive pretraining method.**

Autoregressive (AR) language model: It is a kind of model that using the context word to predict the next word. But here the context word is constrained to two directions, either forward or backwards.



The advantages of AR language model are good at generative Natural language Process(NLP) tasks. Because when generating context, usually is the forward direction. AR language model naturally works well on such NLP tasks.

But Autoregressive language model has some disadvantages, and it only can use forward context or backward context, which means it can't use forward and backward context at the same time.



The conceptual difference between BERT and XLNet. Transparent words are masked out so the model cannot rely on them. XLNet learns to predict the words in an arbitrary order but in an autoregressive, sequential manner (not necessarily left-to-right). BERT predicts all masked words simultaneously.

Q9. What is the transformer?

Answer:

Transformer: It is a deep machine learning model introduced in 2017, used primarily in the field of natural language processing (NLP). Like recurrent neural networks(RNN), It is designed to handle ordered sequences of data, such as natural language, for various tasks like machine translation and text summarization. However, Unlike recurrent neural networks(RNN), Transformers do not require that the sequence be processed in the order. So, if the data in question is a natural language, the Transformer does not need to process the beginning of a sentence before it processes the end. Due to this feature, the Transformer allows for much more parallelization than RNNs during training.

Transformers are developed to solve the problem of sequence transduction current neural networks. It means any task that transforms an input sequence to an output sequence. This includes speech recognition, text-to-speech transformation, etc.

For models to perform a sequence transduction, it is necessary to have some sort of memory. example, let us say that we are translating the following sentence to another language (French):

"The Transformers" is a Japanese band. That band was formed in 1968, during the height of the Japanese music history."

In the above example, the word "the band" in the second sentence refers to the band "The Transformers" introduced in the first sentence. When you read about the band in the second sentence, you know that it is referencing to the "The Transformers" band. That may be important for translation.

For translating other sentences like that, a model needs to figure out these sort of dependencies and connections. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have been used to deal with this problem because of their properties.

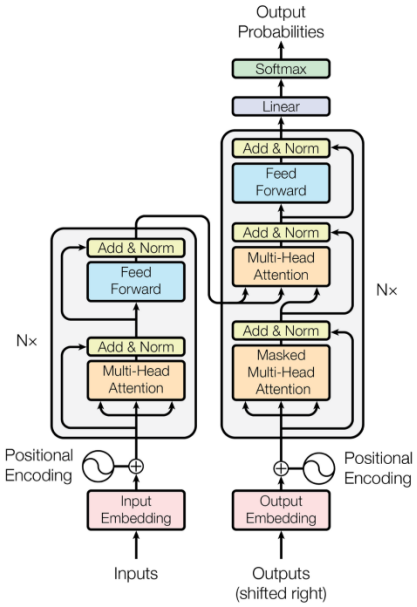


Figure 1: The Transformer - model architecture.

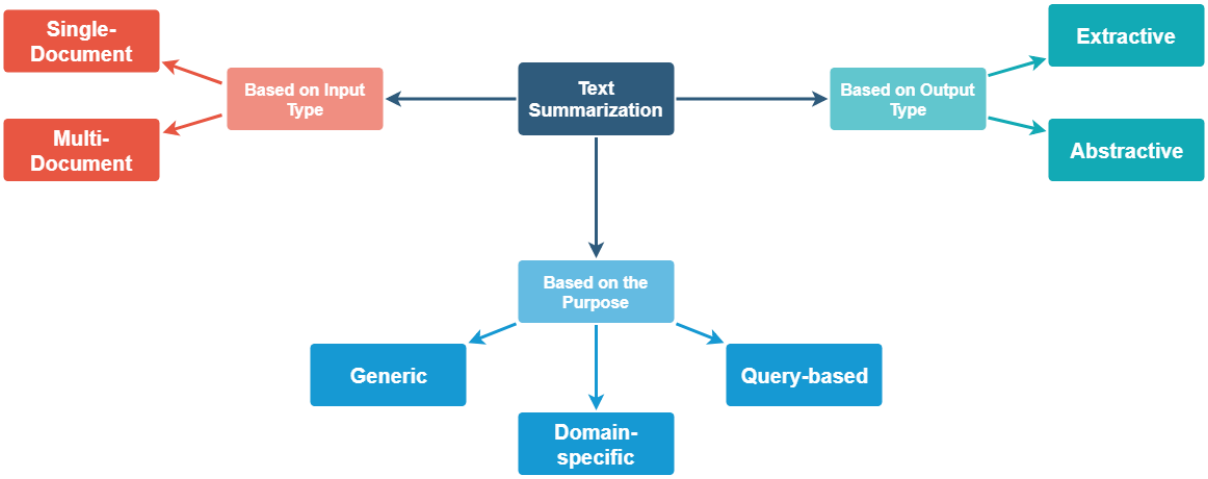
Q10. What is Text summarization?

Answer:

Text summarization: It is the process of shortening a text document, to create a summary of the significant points of the original document.

Types of Text Summarization Methods :

Text summarization methods can be classified into different types.



Example:

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day-18

Q1. What is Levenshtein Algorithm?

Answer:

Levenshtein distance is a string metric for measuring the difference between two sequences. The Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

By Mathematically, the Levenshtein distance between the two strings a, b (of length |a| and |b| respectively) is given by the leva, b(|a| , |b|) where :

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Where, 1 (ai≠bi): This is the indicator function equal to zero when ai≠bi and equal to 1 otherwise, and leva, b(i,j) is the distance between the first i characters of a and the first j characters of b.

Example:

The Levenshtein distance between "HONDA" and "HYUNDAI" is 3, since the following three edits change one into the other, and there is no way to do it with fewer than three edits:

insertion
substitution
deletion

H		O	N	D	A	
H	Y	U	N	D	A	I

H	O		N	D	A	
H	Y	U	N	D	A	I

Q2. What is Soundex?

Answer:

Soundex attempts to find similar names or homophones using phonetic notation. The program retains letters according to detailed equations, to match individual titles for purposes of ample volume research.

Soundex phonetic algorithm: Its indexes strings depend on their English pronunciation. The algorithm is used to describe homophones, words that are pronounced the same, but spelt differently.

Suppose we have the following sourceDF.

```
+-----+-----+
|word1|word2|
+-----+-----+
|   to|  two|
|brake|break|
| here| hear|
| tree| free|
+-----+-----+
```

Let's run below code and see how the soundex algorithm encodes the above words.

```
val actualDF = sourceDF.withColumn(
  "w1_soundex",
  soundex(col("word1"))
).withColumn(
  "w2_soundex",
  soundex(col("word2"))
)

actualDF.show()
```

```
+-----+-----+----------+----------+
|word1|word2|w1_soundex|w2_soundex|
+-----+-----+----------+----------+
|   to|  two|      T000|      T000|
|brake|break|      B620|      B620|
| here| hear|      H600|      H600|
| tree| free|      T600|      F600|
+-----+-----+----------+----------+
```

Let's summarize the above results:

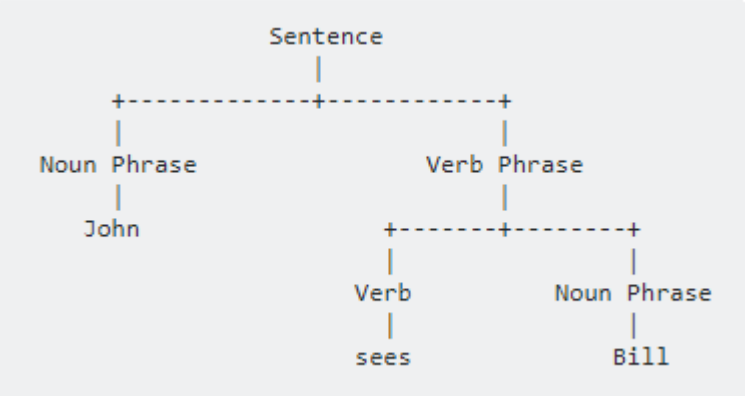
- "two" and "to" both are encoded as T000
- "break" and "brake" both are encoded as B620
- "hear" and "here" both are encoded as H600
- "free" is encoded as F600 and "tree" is encoded as T600: Encodings are similar, but word is different

The Soundex algorithm was often used to compare first names that were spelt differently.

Q3. What is Constituency parse?

Answer:

A constituency parse tree breaks a text into sub-phrases. Non-terminals in the tree are types of phrases, the terminals are the words in the sentence, and the edges are unlabeled. For a simple sentence, "John sees Bill", a constituency parse would be:



Above approaches convert the parse tree into a sequence following a depth-first traversal to be able to apply sequence-to-sequence models to it. The linearized version of the above parse tree looks as follows: (S (N) (VP V N)).

Q4. What is LDA(Latent Dirichlet Allocation)?

Answer:

LDA: It is used to classify text in the document to a specific topic. LDA builds a topic per document model and words per topic model, modelled as Dirichlet distributions.

- Each document is modeled as a distribution of topics, and each topic is modelled as multinomial distribution of words.
- LDA assumes that every chunk of text we feed into it will contain words that are somehow related. Therefore choosing the right corpus of data is crucial.

- It also assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.

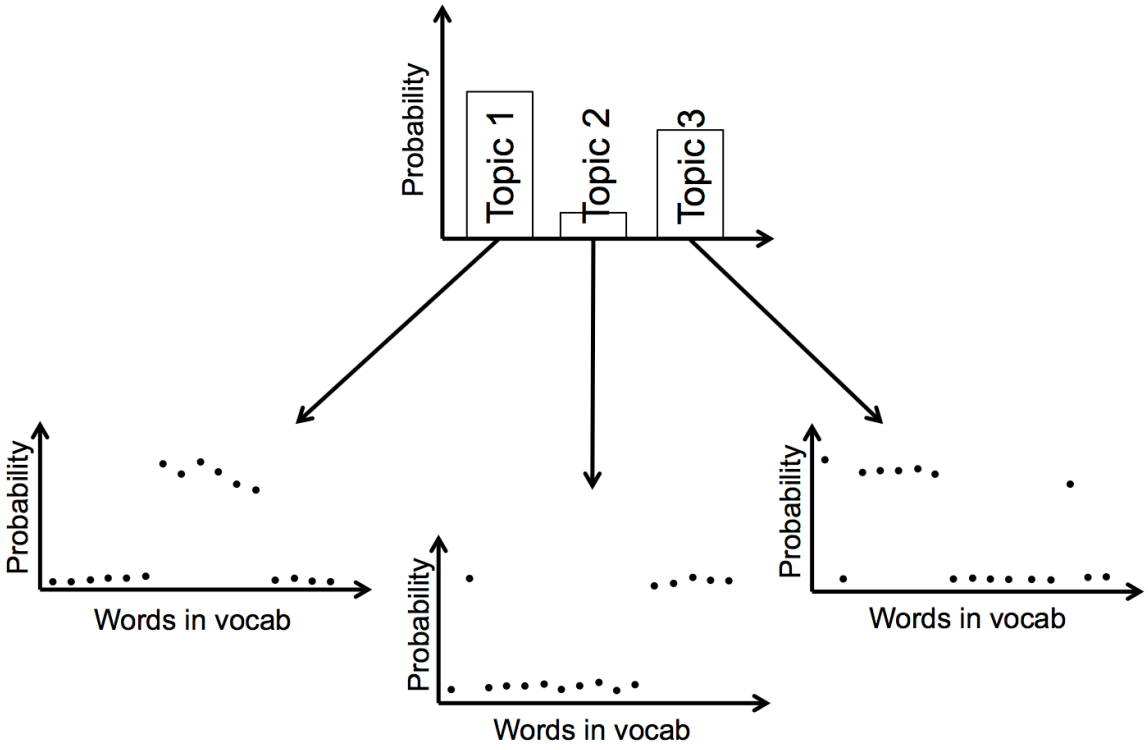
The Bayesian version of PLSA is LDA. It uses Dirichlet priors for the word-topic and document-topic distributions, lending itself to better generalization.

What LDA give us?

It is a probabilistic method. For every document, the results give us a mixture of topics that make up the document. To be precise, we can get probability distribution over the k topics for every document. Every word in the document is attributed to the particular topic with probability given by distribution.

These topics themselves were defined as probability distributions over vocabulary. Our results are two sets of probability distributions:

- The collection of distributions of topics for each document
- The collection of distributions of words for each topic.



Q5.What is LSA?

Answer:

Latent Semantic Analysis (LSA): It is a theory and the method for extract and represents the contextual usage meaning of words by statistical computation applied to large corpus of texts.

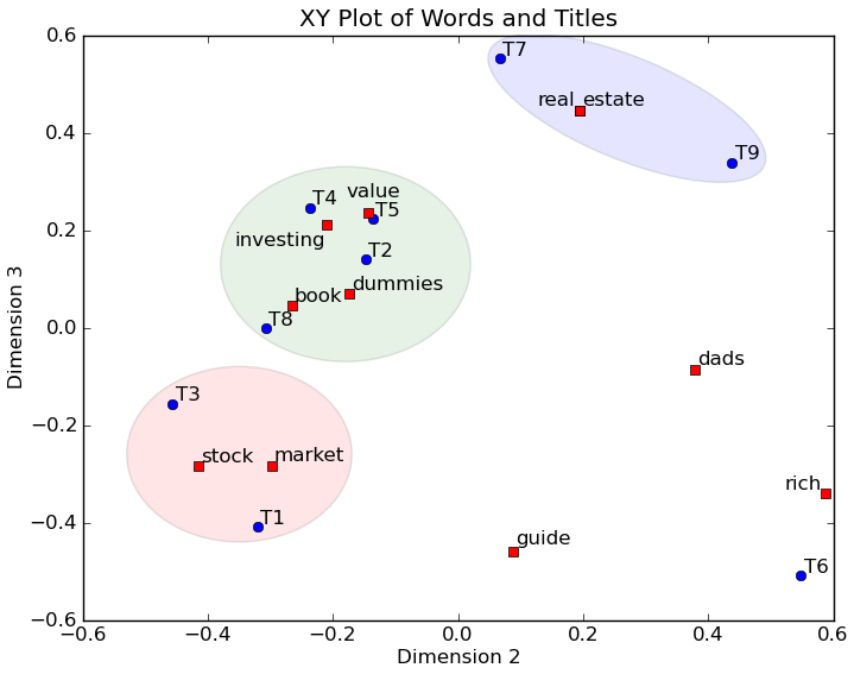
It is an information retrieval technique which analyzes and identifies the pattern in an unstructured collection of text and relationship between them.

Latent Semantic Analysis itself is an unsupervised way of uncovering synonyms in a collection of documents.

Why LSA(Latent Semantic Analysis)?

LSA is a technique for creating vector representation of the document. Having a vector representation of the document gives us a way to compare documents for their similarity by calculating the distance between vectors. In turn, means we can do handy things such as classify documents to find out which of a set knows topics they most likely reside to.

Classification implies we have some known topics that we want to group documents into, and that you have some labelled training data. If you're going to identify natural groupings of the documents without any labelled data, you can use clustering



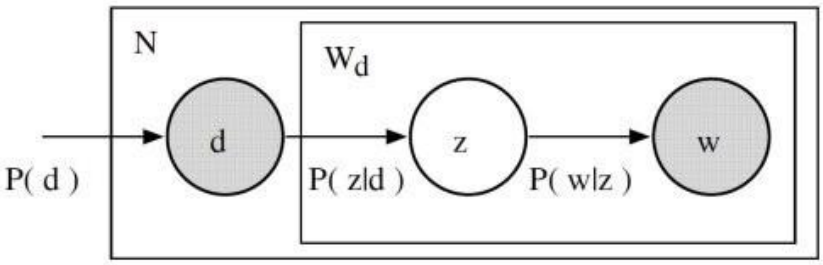
Q6. What is PLSA?

Answer:

PLSA stands for Probabilistic Latent Semantic Analysis, uses a probabilistic method instead of SVD to tackle problem. The main idea is to find the probabilistic model with latent topics that we can *generate* data we observe in our document term matrix. Specifically, we want a model P(D, W) such that for any document d and word w, P(d,w) corresponds to that entry in document-term matrix.

Each document is found in the mixture of topics, and each topic consists of the collection of words. PLSA adds the probabilistic spin to these assumptions:

- Given document d, topic z is available in that document with the probability P(z|d)
- Given the topic z, word w is drawn from z with probability P(w|z)



The joint probability of seeing the given document and word together is:

$$P(D,W) = P(D)\sum_{Z} P(Z|D)P(W|Z)$$

In the above case, P(D), P(Z|D), and P(W|Z) are the parameters of our models. P(D) can be determined directly from corpus. P(Z|D) and the P(W|Z) are modelled as multinomial distributions and can be trained using the expectation-maximisation algorithm (EM).

Q7. What is LDA2Vec?

Answer:

It is inspired by LDA, word2vec model is expanded to simultaneously learn word, document, topic and paragraph topic vectors.

Lda2vec is obtained by modifying the skip-gram word2vec variant. In the original skip-gram method, the model is trained to predict context words based on a pivot word. In lda2vec, the pivot word vector and a document vector are added to obtain a context vector. This context vector is then used to predict context words.

At the document level, we know how to represent the text as mixtures of topics. At the word-level, we typically used something like word2vec to obtain vector representations. It is an extension of word2vec and LDA that jointly learns word, document, and topic vectors.

How does it work?

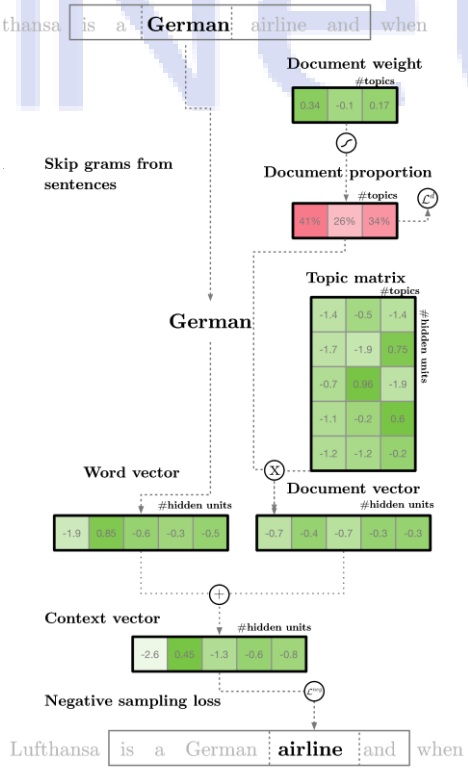
It correctly builds on top of the skip-gram model of word2vec to generate word vectors. Neural net that learns word embedding by trying to use input word to predict enclosing context words.

With Lda2vec, other than using the word vector directly to predict context words, you leverage a context vector to make the predictions. Context vector is created as the sum of two other vectors: the word vector and the document vector.

The same skip-gram word2vec model generates the word vector. The document vector is most impressive. It is a really weighted combination of two other components:

- the document weight vector, representing the "weights" of each topic in a document
- Topic matrix represents each topic and its corresponding vector embedding.

Together, a document vector and word vector generate "context" vectors for each word in a document. lda2vec power lies in the fact that it not only learns word embeddings for words; it simultaneously learns topic representations and document representations as well.



Q8. What is Expectation-Maximization Algorithm(EM)?

Answer:

The Expectation-Maximization Algorithm, in short, EM algorithm, is an approach for maximum likelihood estimation in the presence of latent variables.

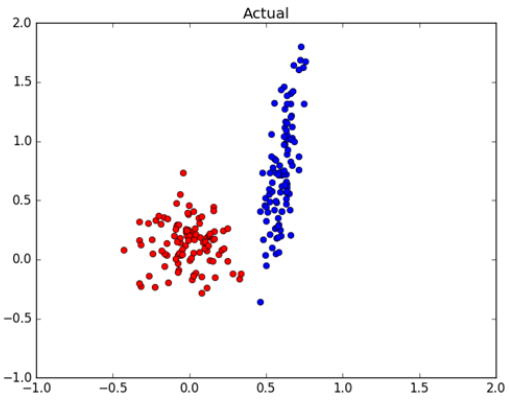
This algorithm is an iterative approach that cycles between two modes. The first mode attempts to predict the missing or latent variables called the estimation-step or E-step. The second mode attempts to optimise the parameters of the model to explain the data best called the maximization-step or M-step.

- **E-Step**. Estimate the missing variables in the dataset.
- **M-Step**. Maximize the parameters of the model in the presence of the data.

The EM algorithm can be applied quite widely, although it is perhaps most well known in machine learning for use in unsupervised learning problems, such as density estimation and clustering.

For detail explanation of EM is, let us first consider this example. Say that we are in a school, and interested to learn the height distribution of female and male students in the school. The most sensible thing to do, as we probably would agree with me, is to randomly take a sample of N students of both genders, collect their height information and estimate the mean and standard deviation for male and female separately by way of maximum likelihood method.

Now say that you are not able to know the gender of student while we collect their height information, and so there are two things you have to guess/estimate: (1) whether the individual sample of height information belongs to a male or a female and (2) the parameters (μ, θ) for each gender which is now unobservable. This is tricky because only with the knowledge of who belongs to which group, can we make reasonable estimates of the group parameters separately. Similarly, only if we know the parameters that define the groups, can we assign a subject properly. How do you break out of this infinite loop? Well, EM algorithm just says to start with initial random guesses.



Q9.What is Text classification in NLP?

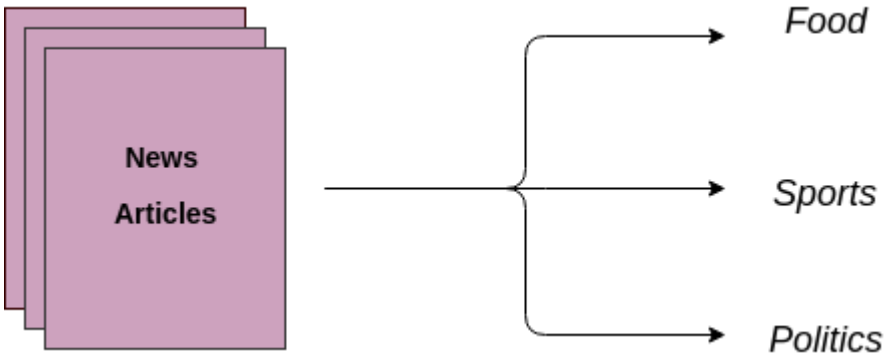
Answer:

Text classification is also known as text tagging or text categorization is a process of categorizing text into organized groups. By using NLP, text classification can automatically analyze text and then assign a set of pre-defined tags or categories based on content.

Unstructured text is everywhere on the internet, such as emails, chat conversations, websites, and the social media but it's hard to extract value from given data unless it's organized in a certain way. Doing so used to be a difficult and expensive process since it required spending time and resources to manually sort the data or creating handcrafted rules that are difficult to maintain. Text classifiers with NLP have proven to be a great alternative to structure textual data in a fast, cost-effective, and scalable way.

Text classification is becoming an increasingly important part of businesses as it allows us to get insights from data and automate business processes quickly. Some of the most common examples and the use cases for automatic text classification include the following:

- **Sentiment Analysis:** It is the process of understanding if a given text is talking positively or negatively about a given subject (e.g. for brand monitoring purposes).
- **Topic Detection:** In this, the task of identifying the theme or topic of a piece of text (e.g. know if a product review is about Ease of Use, Customer Support, or Pricing when analysing customer feedback).
- **Language Detection:** the procedure of detecting the language of a given text (e.g. know if an incoming support ticket is written in English or Spanish for automatically routing tickets to the appropriate team).



Q10. What is Word Sense Disambiguation (WSD)?

Answer:

WSD (Word Sense Disambiguation) is a solution to the ambiguity which arises due to different meaning of words in a different context.

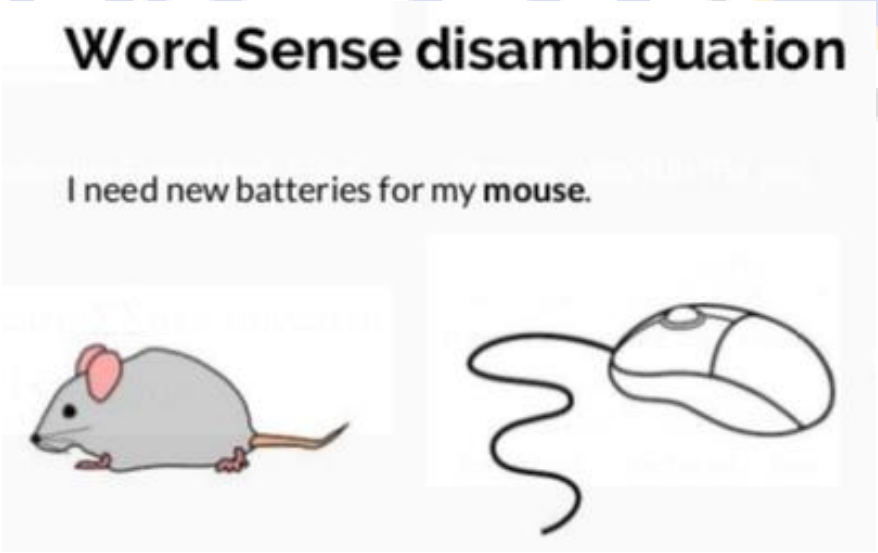
In natural language processing, **word sense disambiguation** (WSD) is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be mostly unconscious in people. WSD is the natural classification problem: Given a word and its possible senses, as defined by the dictionary, classify an occurrence of the word in the context into one or more of its sense classes. The features of the context (such as the neighbouring words) provide the evidence for classification.

For example, consider these two below sentences.

" The **bank** will not be accepting the cash on Saturdays. "

" The river overflowed the **bank** ."

The word " **bank "** in the given sentence refers to commercial (finance) banks, while in the second sentence, it refers to a riverbank. The uncertainty that arises, due to this is tough for the machine to detect and resolve. Detection of change is the first issue and fixing it and displaying the correct output is the second issue.



DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

DAY 19

Q1. What is LSI(Latent Semantic Indexing)?

Answer:

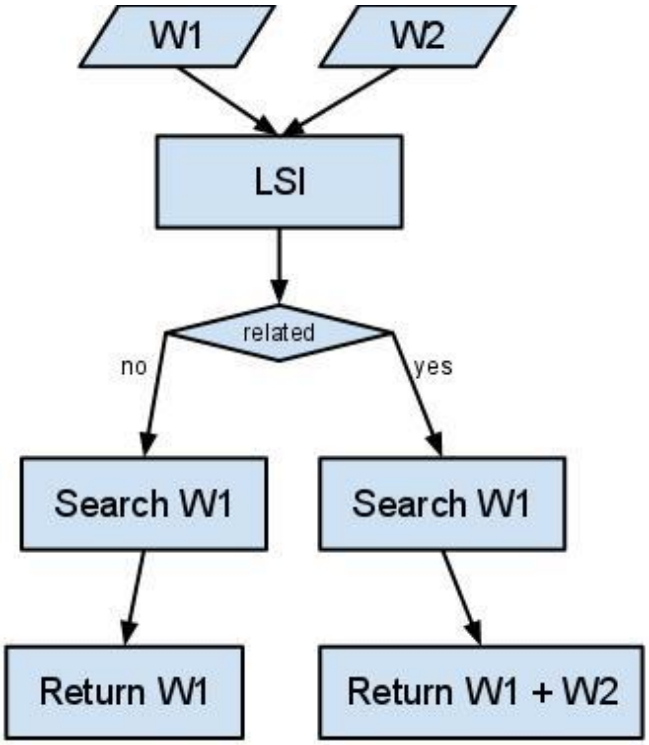
Latent Semantic Indexing (LSI): It is an indexing and retrieval method that uses a mathematical technique called SVD(Singular value decomposition) to find patterns in relationships between terms and concepts contained in an unstructured collection of text. It is based on the principle that words that are used in the same contexts tend to have similar meanings.

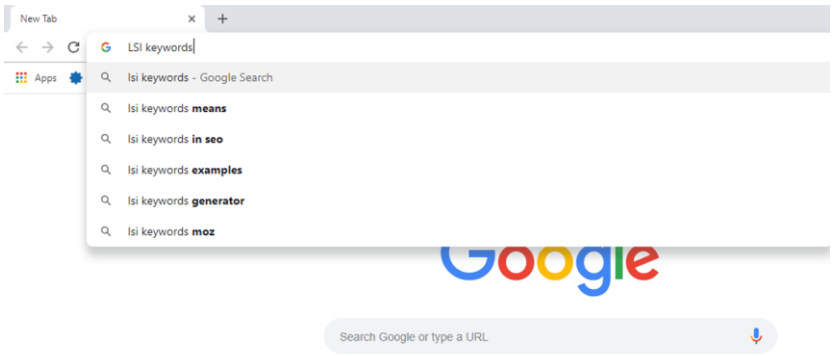
For example, Tiger and Woods are associated with men instead of an animal, and a Wood, Parris, and Hilton are associated with the singer.

Example:

If you use LSI to index a collection of articles and the words “fan” and “regulator” appear together frequently enough, the search algorithm would notice that the two terms are semantically close. A search for “fan” will, therefore, return a set of items containing that phrase, but also items that contain just the word “regulator”. It doesn't understand word distance, but by examining a sufficient number of documents, it only knows the two terms are interrelated. It then uses that information to provide an expanded set of results with better recall than an understandable keyword search.

The diagram below describes the effect between LSI and keyword searches. W stands for a document.





Q2. What is Named Entity Recognition? And tell some use cases of NER?

Answer:

Named-entity recognition (NER): It is also known as entity extraction, and entity identification is a subtask of information extraction that explore to locate and classify atomic elements in text into predefined categories like the names of persons, organizations, places, expressions of times, quantities, monetary values, percentages and more.

In each text document, particular terms represent specific entities that are more informative and have a different context. These entities are called named entities, which more accurately refer to conditions that represent real-world objects like people, places, organizations or institutions, and so on, which are often expressed by proper names. The naive approach could be to find these by having a look at the noun phrases in text documents. It also is known as entity chunking/extraction, which is a popular technique used in information extraction to analyze and segment the named entities and categorize or classify them under various predefined classes.

Named Entity Recognition use-case

- **Classifying content for news providers-**
  NER can automatically scan entire articles and reveal which are the significant people, organizations, and places discussed in them. Knowing the relevant tags for each item helps in automatically categorizing the articles in defined hierarchies and enable smooth content discovery.
- **Customer Support:**
  Let's say we are handling the customer support department of an electronics store with multiple branches worldwide; we go through a number mentions in our customers' feedback. Such as this for instance.

Now, if we pass it through the Named Entity Recognition API, it pulls out the entities Bangalore (location) and Fitbit (Product). This can be then used to categorize the complaint and assign it to the relevant department within the organization that should be handling this.

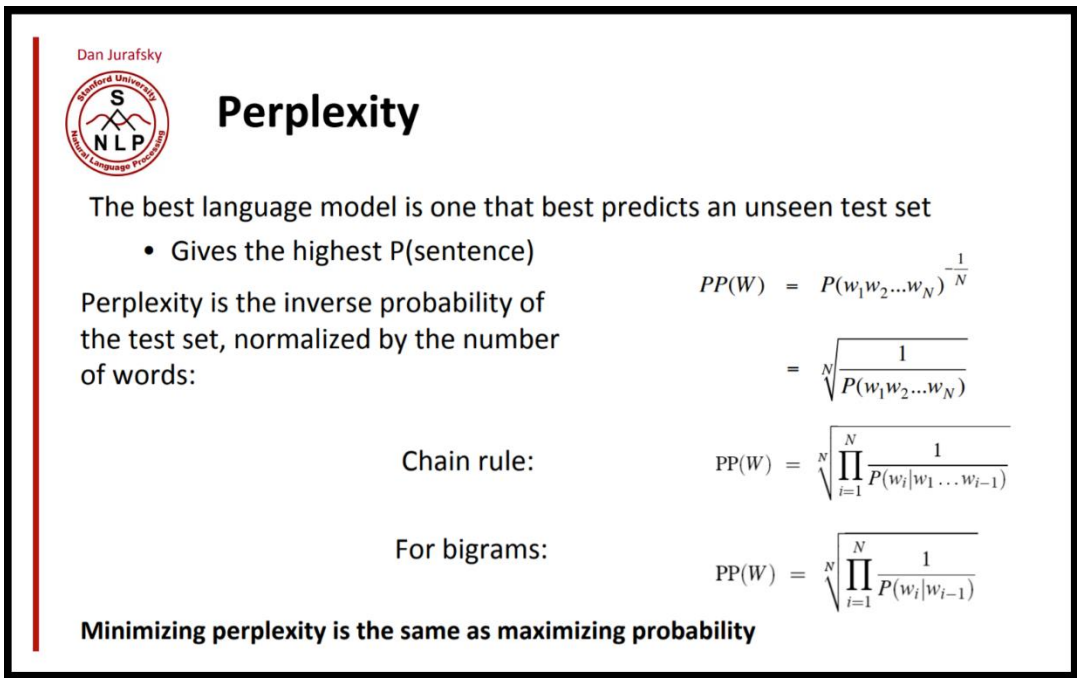
Figure 1: An example of NER application on an example text

Q3. What is perplexity?

Answer:

Perplexity: It is a measurement of how well a probability model predicts a sample. In the context of NLP, perplexity(Confusion) is one way to evaluate language models.

The term perplexity has three closely related meanings. It is a measure of how easy a probability distribution is to predict. It is a measure of how variable a prediction model is. And It is a measure of prediction error. The third meaning of perplexity is calculated slightly differently, but all three have the same fundamental idea.



Q4. What is the language model?

Answer:

Language Modelling (LM): It is one of the essential parts of modern NLP. There are many sorts of applications for Language Modelling, like Machine Translation, Spell Correction Speech Recognition, Summarization, Question Answering, Sentiment analysis, etc. Each of those tasks requires the use of the language model. The language model is needed to represent the text to a form understandable from the machine point of view.

The statistical language model is a probability distribution over a series of words. Given such a series, say of length m, it assigns a probability to the whole series.

It provides context to distinguish between phrases and words that sounds are similar. For example, in American English, the phrases " wreck a nice beach " and "recognize speech" sound alike but mean different things.

Data sparsity is a significant problem in building language models. Most possible word sequences are not noticed in training. One solution is to make the inference that the probability of a word only depends on the previous n words. This is called as an n-gram model or unigram model when $n = 1$. The unigram model is also known as the bag of words model.

How does this Language Model help in NLP Tasks?

The probabilities restoration by a language model is most useful to compare the likelihood that different sentences are "good sentences." This was useful in many practical tasks, for example:

Spell checking: You observe a word that is not identified as a known word as part of a sentence. Using the edit distance algorithm, we find the closest known words to the unknown words. These are the candidate corrections. For example, we observe the word "wurd" in the context of the sentence, "I like to write this wurd." The candidate corrections are ["word", "weird", "wind"]. How can we select among these candidates the most likely correction for the suspected error "weird"?

Automatic Speech Recognition: we receive as input a string of phonemes; a first model predicts for sub-sequences of the stream of phonemes candidate words; the language model helps in ranking the most likely sequence of words compatible with the candidate words produced by the acoustic model.

Machine Translation: each word from the source language is mapped to multiple candidate words in the target language; the language model in the target language can rank the most likely sequence of candidate target words.

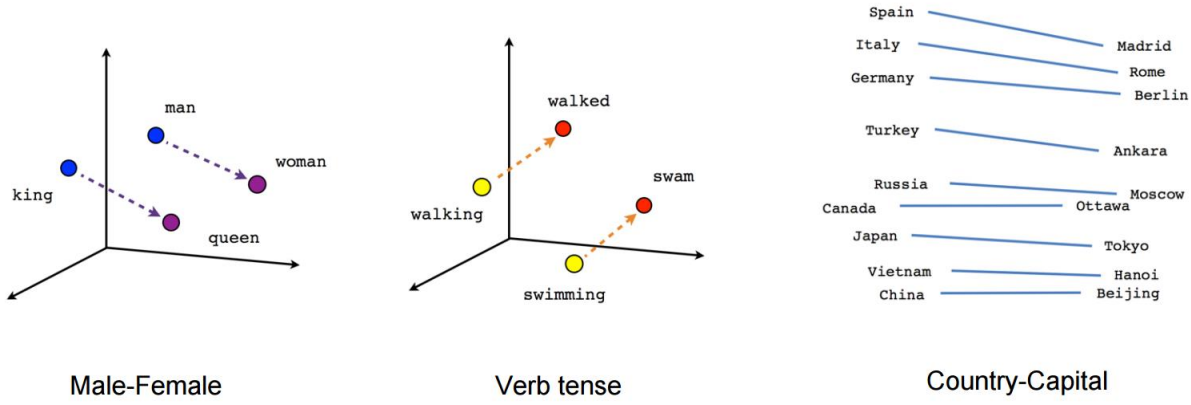
Q5. What is Word Embedding?

Answer:

A word embedding is a learned representation for text where words that have the same meaning have a similar observation.

It is basically a form of word representation that bridges the human understanding of language to that of a machine. Word embeddings divide representations of text in an n-dimensional space. These are essential for solving most NLP problems.

And the other point worth considering is how we obtain word embeddings as no two sets of word embeddings are similar. Word embeddings aren't random; they're developed by training the neural network. A recent powerful word embedding usage comes from Google named Word2Vec, which is trained by predicting several words that appear next to other words in a language. For example, the word "cat", the neural network would predict the words like "kitten" and "feline." This intuition of words comes out "near" each other allows us to place them in vector space.



Male-Female | Verb tense | Country-Capital

Q6. Do you have an idea about fastText?

Answer:

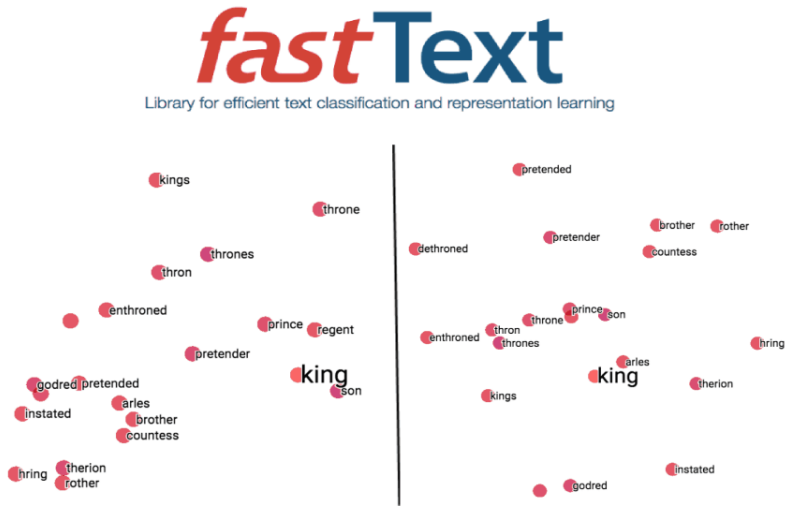
fastText: It is another word embedding method that is an extension of the word2vec model. Alternatively, learning vectors for words directly. It represents each word as an n-gram of characters. So, for example, take the word, “artificial” with n=3, the fastText representation of this word is <ar, art, rti, tif, ifi, fic, ici, ial, al>, where the angular brackets indicate the beginning and end of the word.

This helps to capture the meaning of shorter words and grant the embeddings to understand prefixes and suffixes. Once the word has been showed using character skip-grams, a n-gram model is trained to learn the embeddings. This model is acknowledged to be a bag of words model with a sliding

window over a word because no internal structure of the word is taken into account. As long as the characters are within this window, the order of the n-grams doesn't matter.

fastText works well with rare words. So even if a word wasn't seen during training, it can be broken down into n-grams to get its embeddings.

Word2vec and GloVe both fail to provide any vector representation for words that are not in the model dictionary. This is a huge advantage of this method.



Q7. What is GloVe?

Answer:

GloVe(global vectors) is for word representation. GloVe is an unsupervised learning algorithm developed by Stanford for achieving word embeddings by aggregating a global word-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the word in vector space.

The GloVe model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a new word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

How GloVe find meaning in statistics?

Produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a new word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

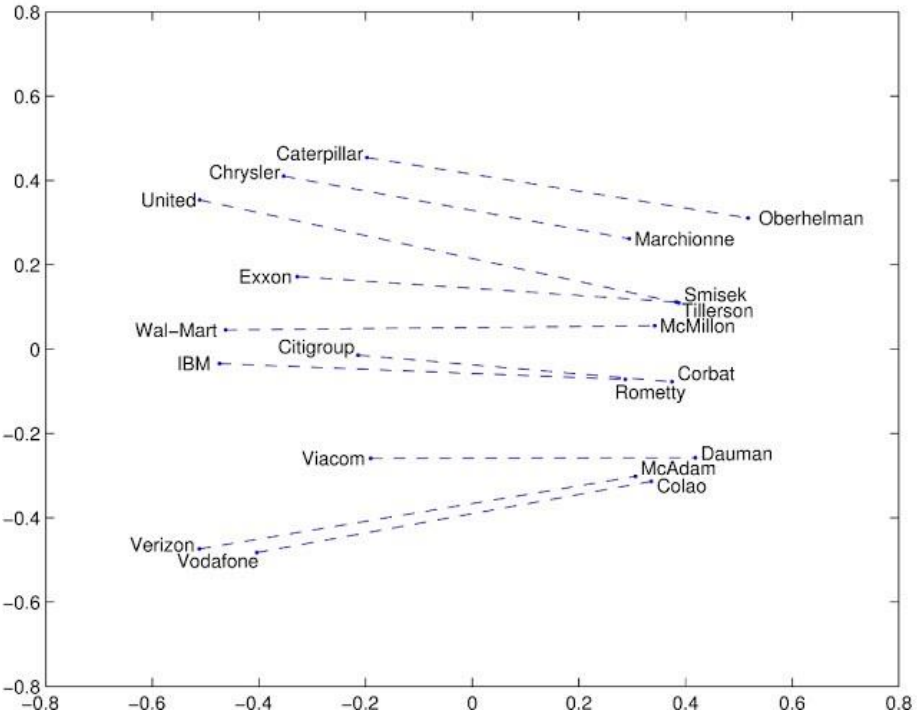
GloVe aims to achieve two goals:

- (1) Create word vectors that **capture meaning in vector space**
- (2) Takes advantage of **global count statistics** instead of only local information

Unlike word2vec – which learns by streaming sentences – GloVe determines based on a **co-occurrence matrix** and trains word vectors, so their differences predict **co-occurrence ratios**

GloVe weights the loss based on word frequency.

Somewhat surprisingly, word2vec and GloVe turn out to be remarkably similar, despite starting off from entirely different starting points.



Q8. Explain Gensim?

Answer:

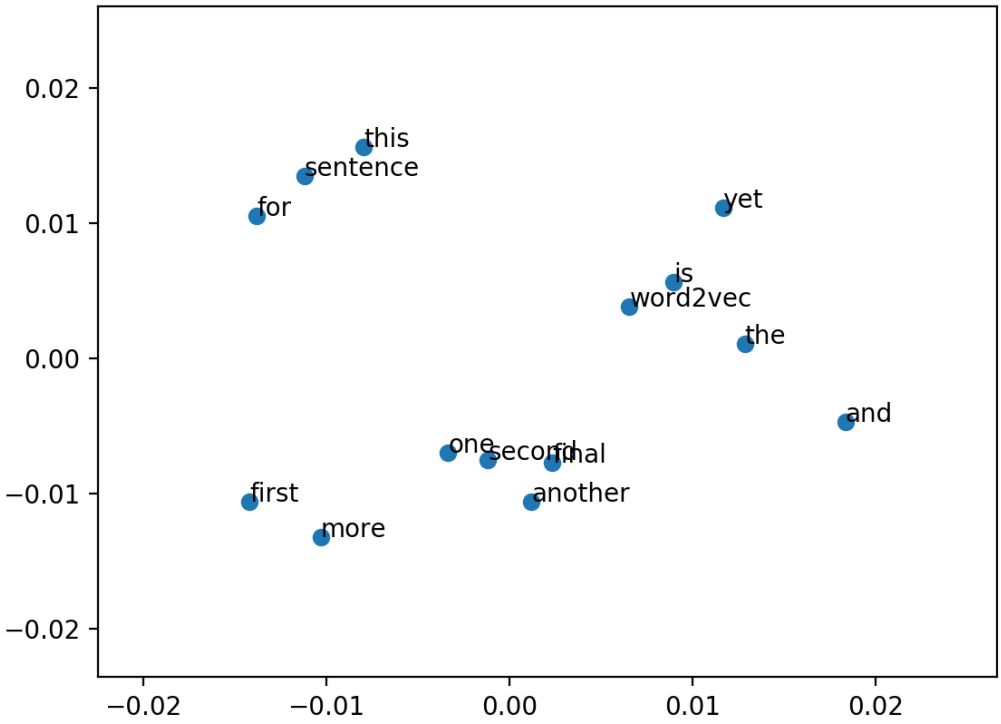
Gensim: It is billed as a Natural Language Processing package that does 'Topic Modeling for Humans'. But its practically much more than that.

If you are unfamiliar with topic modeling, it is a technique to extract the underlying topics from large volumes of text. Gensim provides algorithms like LDA and LSI (which we already seen in previous interview questions) and the necessary sophistication to built high-quality topic models.

It is an excellent library package for processing texts, working with word vector models (such as FastText, Word2Vec, etc) and for building the topic models. Another significant advantage with gensim is: it lets us handle large text files without having to load the entire file in memory.

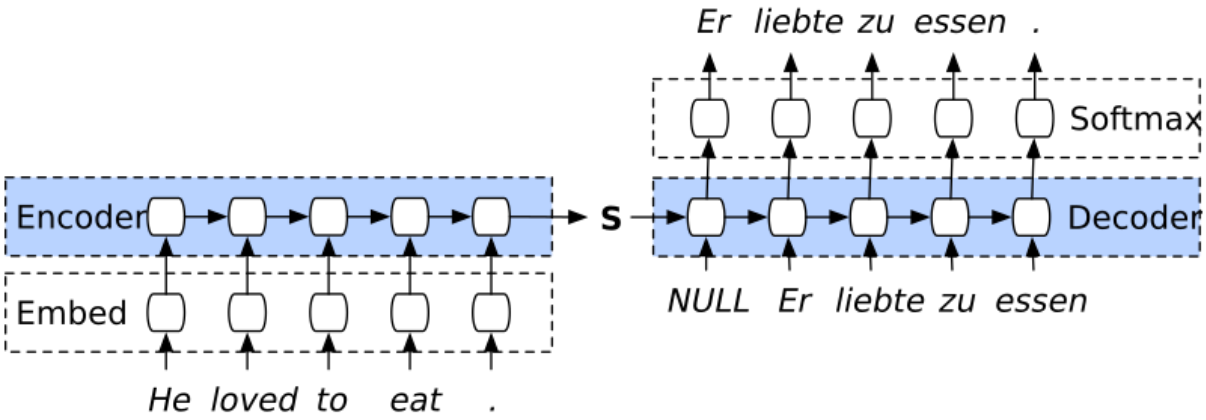
We can also tell as It is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning.

Gensim is implemented in Python and Cython. Gensim is designed to handle extensive text collections using data streaming and incremental online algorithms, which differentiates it from most other machine learning software packages that target only in-memory processing.



Q9. What is Encoder-Decoder Architecture?

Answer:



The encoder-decoder architecture consists of two main parts :

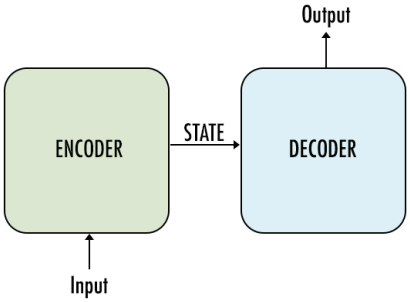
- **Encoder**:
  Encoder simply takes the input data, and trains on it, then it passes the final state of its recurrent layer as an initial state to the first recurrent layer of the decoder part.

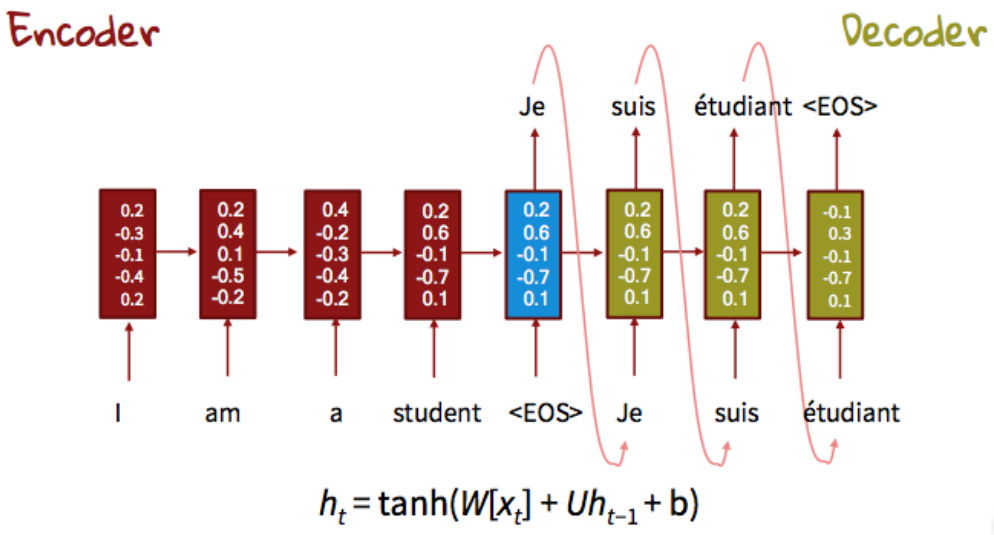
```
Encoder input : English sentences

Encoder initial state : It depends on the initializer we use
```

- **Decoder :**
  The decoder takes the final state of encoder's final recurrent layer and uses it as an initial state to its initial, recurrent layer, the input of the decoder is sequences that we want to get French sentences.

```
Decoder input : French sentences

Decoder initial state : The last state of encoder's last recurrent
layer
```

Some more example for better understanding:



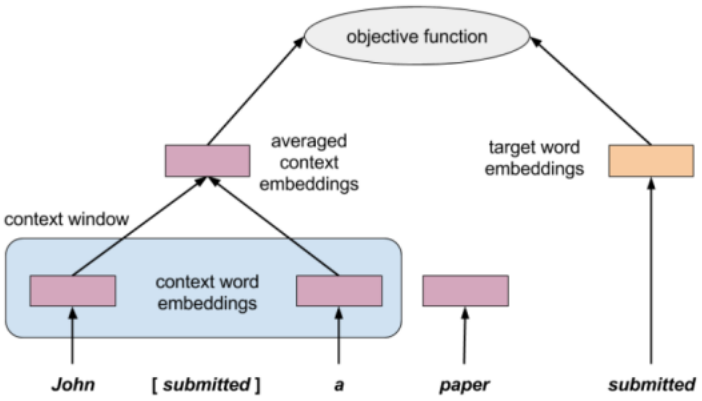


Q10. What is Context2Vec?

Answer:

Assume a case where you have a sentence like. I can't find May. Word May maybe refers to a month's name or a person's name. You use the words surround it (context) to help yourself to determine the best suitable option. Actually, this problem refers to the Word Sense Disambiguation task, on which you investigate the actual semantics of the word based on several semantic and linguistic techniques. The Context2Vec idea is taken from the original CBOW Word2Vec model, but instead of relying on averaging the embedding of the words, it relies on a much more complex parametric model that is based on one layer of Bi-LSTM. Figure1 shows the architecture of the CBOW model.

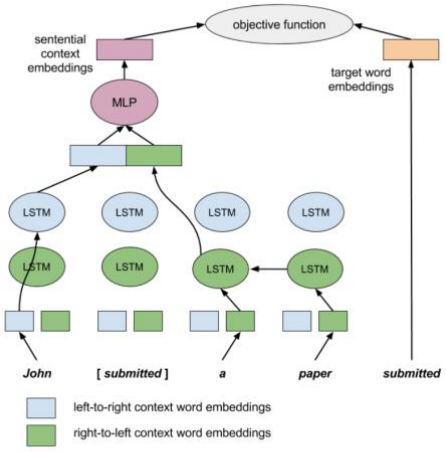
Figure1



(a) word2vec *CBOW*

Context2Vec applied the same concept of windowing, but instead of using a simple average function, it uses 3 stages to learn complex parametric networks.

- A Bi-LSTM layer that takes left-to-right and right-to-left representations
- A feedforward network that takes the concatenated hidden representation and produces a hidden representation through learning the network parameters.
- Finally, we apply the objective function to the network output.



(b) *context2vec*

Figure 1: *word2vec* and *context2vec* architectures.

We used the Word2Vec negative sampling idea to get better performance while calculating the loss value.

The following are some samples of the closest words to a given context.

Sentential Context	Closest target words
This [ ] is due	item, fact-sheet, offer, pack, card
This [ ] is due not just to mere luck	offer, suggestion, announcement, item, prize
This [ ] is due not just to mere luck, but to outstanding work and dedication	award, prize, turnabout, offer, gift
[ ] is due not just to mere luck, but to outstanding work and dedication	it, success, this, victory, prize-money

Table 1: Closest target words to various sentential contexts, illustrating *context2vec*'s sensitivity to long range dependencies, and both sides of the target word.

DATA SCIENCE INTERVIEW PREPARATION

(30 Days of Interview Preparation)

DAY 20

Q1. Do you have any idea about Event2Mind in NLP?

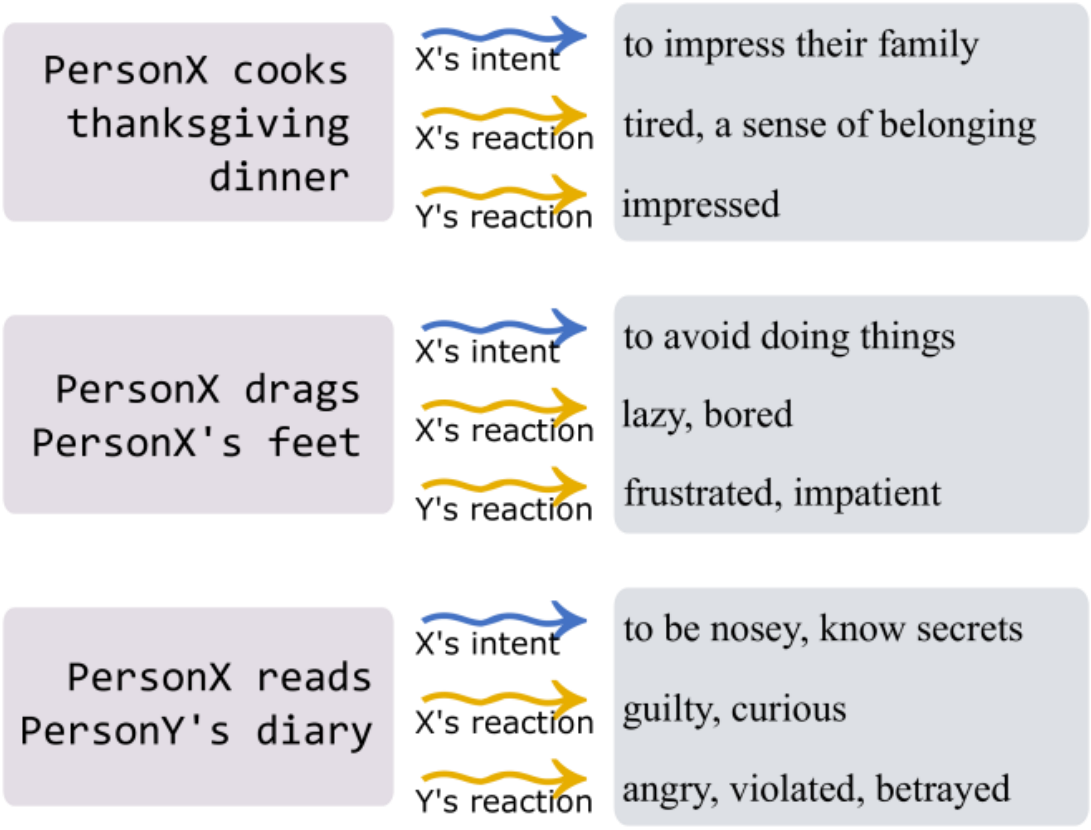
Answer:

Yes, it is based on NLP research paper to understand the common-sense inference from sentences.

Event2Mind: Common-sense Inference on Events, Intents, and Reactions

The study of "Commonsense Reasoning" in NLP deals with teaching computers how to gain and employ common sense knowledge. NLP systems require common sense to adapt quickly and understand humans as we talk to each other in a natural environment.

This paper proposes a new task to teach systems commonsense reasoning: given an event described in a short "event phrase" (e.g. "PersonX drinks coffee in the morning"), the researchers teach a system to reason about the likely intents ("PersonX wants to stay awake") and reactions ("PersonX feels alert") of the event's participants.



Understanding a narrative requires common-sense reasoning about the mental states of people in relation to events. For example, if "Robert is dragging his feet at work," pragmatic implications about Robert's *intent* are that "Robert wants to avoid doing things" (Above Fig). You can also infer that Robert's *emotional reaction* might be feeling "bored" or "lazy." Furthermore, while not explicitly mentioned, you can assume that people other than Robert are affected by the situation, and these people are likely to feel "impatient" or "frustrated."

This type of pragmatic inference can likely be useful for a wide range of NLP applications that require accurate anticipation of people's intents and emotional reactions, even when they are not expressly mentioned. For example, an ideal dialogue system should react in empathetic ways by reasoning about the human user's mental state based on the events the user has experienced, without the user explicitly stating how they are feeling. Furthermore, advertisement systems on social media should be able to reason about the emotional reactions of people after events such as mass shootings and remove ads for guns, which might increase social distress. Also, the pragmatic inference is a necessary step toward automatic narrative understanding and generation. However, this type of commonsense social reasoning goes far beyond the widely studied entailment tasks and thus falls outside the scope of existing benchmarks.

Q2. What is SWAG in NLP?

Answer:

SWAG stands for **S**ituations **w**ith **A**dversarial **G**enerations is a dataset consisting of 113k multiple-choice questions about a rich spectrum of grounded situations.

Swag: A Large Scale Adversarial Dataset for Grounded Commonsense Inference

According to NLP research paper on SWAG is "Given a partial description like "he opened the hood of the car," humans can reason about the situation and anticipate what might come next ("then, he examined the engine"). In this paper, you introduce the task of grounded commonsense inference, unifying natural language inference(NLI), and common-sense reasoning.

We present SWAG, a dataset with 113k multiple-choice questions about the rich spectrum of grounded positions. To address recurring challenges of annotation artifacts and human biases found in many existing datasets, we propose AF(Adversarial Filtering), a novel procedure that constructs a de-biased dataset by iteratively training an ensemble of stylistic classifiers, and using them to filter the data. To account for the aggressive adversarial filtering, we use state-of-the-art language models to oversample a diverse set of potential counterfactuals massively. Empirical results present that while humans can solve the resulting inference problems with high accuracy (88%), various competitive models make an effort on our task. We provide a comprehensive analysis that indicates significant opportunities for future research.

When we read a tale, we bring to it a large body of implied knowledge about the physical world. For instance, given the context "on stage, a man takes a seat at the piano," we can easily infer what the situation might look like: a man is giving a piano performance, with a crowd watching him. We can furthermore infer him likely next action: he will most likely set his fingers on the piano key and start playing.

This type of natural language inference(NLI) requires common-sense reasoning, substantially broadening the scope of prior work that focused primarily on linguistic entailment. Whereas the

dominant entailment paradigm asks if 2 natural language sentences (the 'premise' and the 'hypothesis') describe the same set of possible worlds, here we focus on whether a (multiple-choice) ending represents a possible (*future*) world that can a from the situation described in the premise, even when it is not strictly entailed. Making such inference necessitates a rich understanding of everyday physical conditions, including object affordances and frame semantics.

On stage, a woman takes a seat at the piano. She	a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. **d) nervously sets her fingers on the keys.**
A girl is going across a set of monkey bars. She	a) jumps up across the monkey bars. b) struggles onto the monkey bars to grab her head. **c) gets to the end and stands on a wooden plank.** d) jumps up and does a back flip.
The woman is now blow drying the dog. The dog	**a) is placed in the kennel next to a woman's feet.** b) washes her face with the shampoo. c) walks into frame and walks towards the dog. d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from Swag; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

Q3. What is the Pix2Pix network?

Answer:

Pix2Pix network: It is a Conditional GANs (cGAN) that learn the mapping from an input image to output an image.

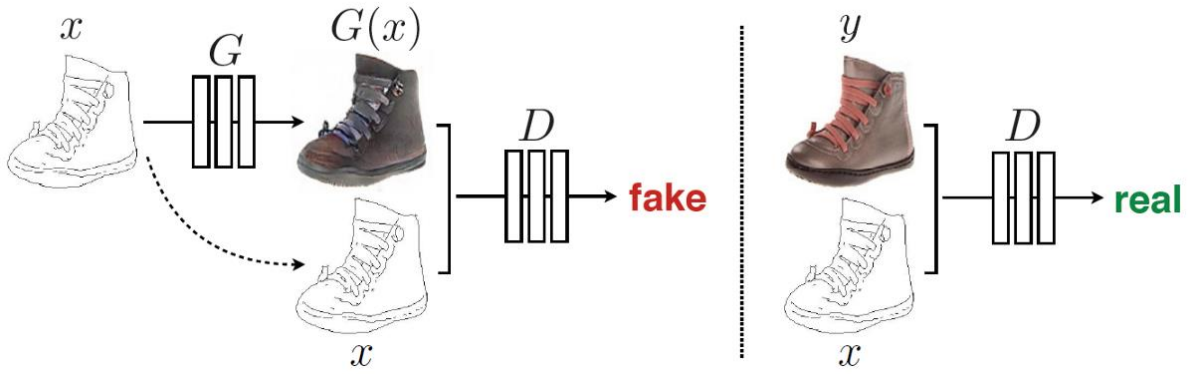
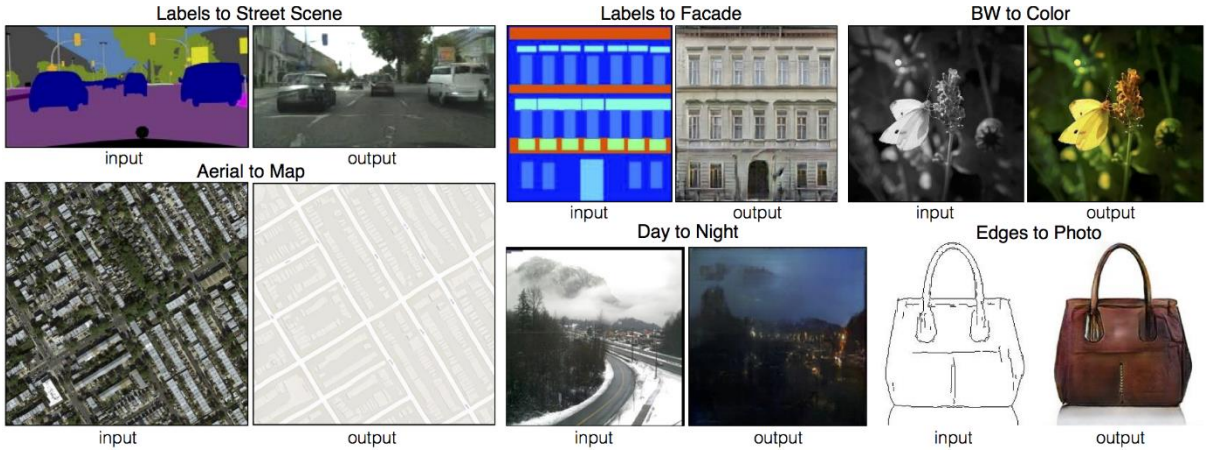


Image-To-Image Translation is the process for translating one representation of the image into another representation.

The image-to-image translation is another example of a task that GANs (Generative Adversarial Networks) are ideally suited for. These are tasks in which it is nearly impossible to hard-code a loss function. Studies on GANs are concerned with novel image synthesis, translating from a random vector z into an image. Image-to-Image translation converts one image to another like the edges of the bag below to the photo image. Another exciting example of this is shown below:



In Pix2Pix Dual Objective Function with an Adversarial and L1 Loss

A naive way to do Image-to-Image translation would be to discard the adversarial framework altogether. A source image would just be passed through a parametric function, and the difference in the resulting image and the ground truth output would be used to update the weights of the network. However, designing this loss function with standard distance measures such as L1 and L2 will fail to capture many of the essential distinctive characteristics between these images. However,

authors do find some value to the L1 loss function as a weighted sidekick to the adversarial loss function.

The Conditional-Adversarial Loss (Generator versus Discriminator) is very popularly formatted as follows:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))],$$

The L1 loss function previously mentioned is shown below:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Combining these functions results in:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

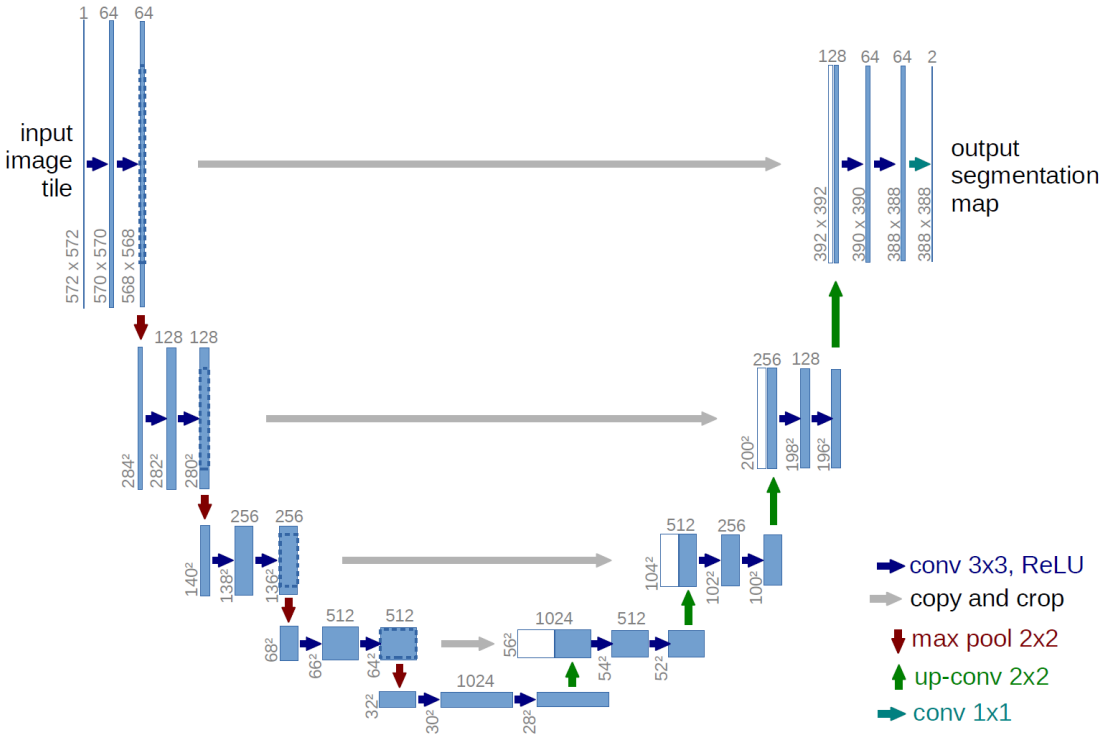
In the experiments, the authors report that they found the most success with the lambda parameter equal to 100.

Q4. Explain UNet Architecture?

Answer:

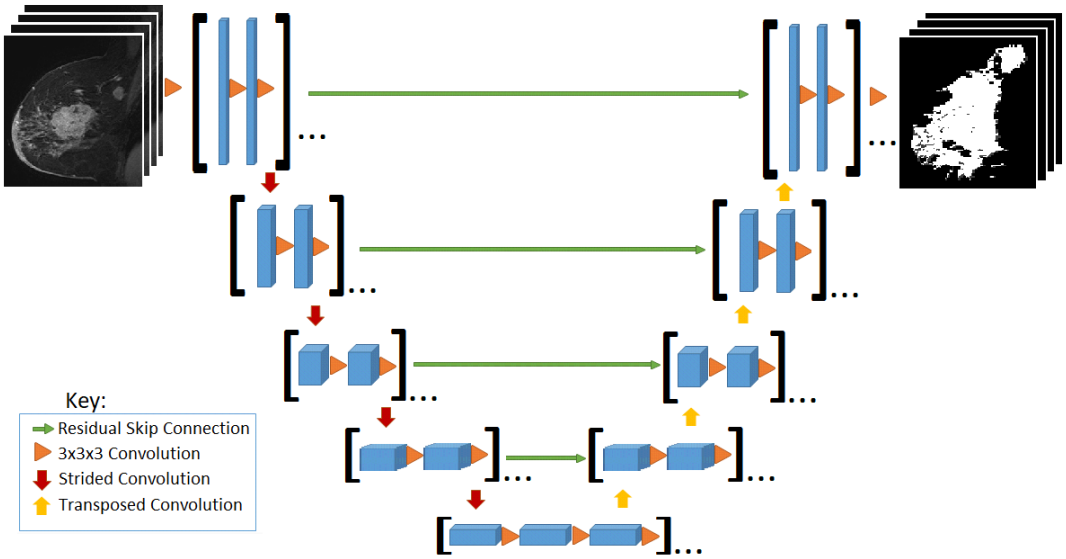
U-Net architecture: It is built upon the Fully Convolutional Network and modified in a way that it yields better segmentation in medical imaging. Compared to FCN-8, the two main differences are (a) U-net is symmetric and (b) the skip connections between the downsampling path and upsampling path apply a concatenation operator instead of a sum. These skip connections intend to provide local information to the global information while upsampling. Because of its symmetry, the network has a large number of feature maps in the upsampling path, which allows transferring information. By comparison, the underlying FCN architecture only had the *number of classes* feature maps in its upsampling way.

How does it work?



The UNet architecture looks like a 'U,' which justifies its name. This UNet architecture consists of 3 sections: The contraction, the bottleneck, and the expansion section. The contraction section is made of many contraction blocks. Each block takes an input that applies two 3X3 convolution layers, followed by a 2X2 max pooling. The number of features or kernel maps after each block doubles so that UNet architecture can learn complex structures. Bottommost layer mediates between the contraction layer and the expansion layer. It uses two 3X3 CNN layers followed by 2X2 up convolution layer.

But the heart of this architecture lies in the expansion section. Similar to the contraction layer, it also has several expansion blocks. Each block passes input to two 3X3 CNN layers, followed by a 2X2 upsampling layer. After each block number of feature maps used by the convolutional layer, get half to maintain symmetry. However, every time input is also get appended by feature maps of the corresponding contraction layer. This action would ensure that features that are learned while contracting the image will be used to reconstruct it. The number of expansion blocks is as same as the number of contraction blocks. After that, the resultant mapping passes through another 3X3 CNN layer, with the number of feature maps equal to the number of segments desired.



Q5. What is pair2vec?

Answer:

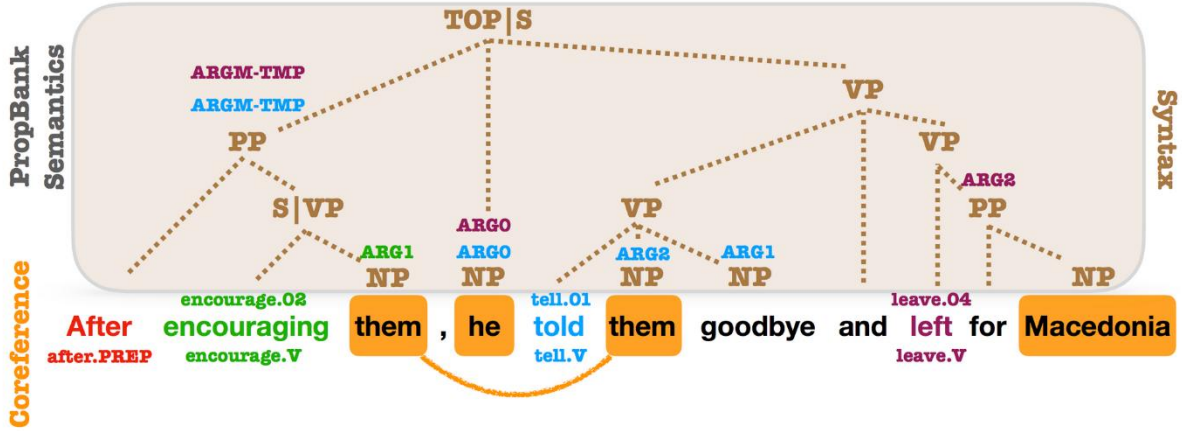
This paper pre trains *word pair representations* by maximizing pointwise mutual information of pairs of words with their context. This encourages a model to learn more meaningful representations of word pairs than with more general objectives, like modeling. The pre-trained representations are useful in tasks like SQuAD and MultiNLI that require cross-sentence inference. You can expect to see more pretraining tasks that capture properties particularly suited to specific downstream tasks and are complementary to more general-purpose tasks like language modeling.

Reasoning about implied relationships between pairs of words is crucial for cross sentences inference problems like question answering (QA) and natural language inference (NLI). In NLI, e.g., given a premise such as "*golf is prohibitively expensive*," inferring that the hypothesis "*golf is a cheap pastime*" is a contradiction requires one to know that *expensive* and *cheap* are antonyms. Recent work has shown that current models, which rely heavily on unsupervised single-word embeddings, struggle to grasp such relationships. In this pair2vec paper, we show that they can be learned with word pair2vec(pair vector), which are trained, unsupervised, at a huge scale, and which significantly improve performance when added to existing cross-sentence attention mechanisms.

X	Y	Contexts
		with X and Y baths
hot	*cold*	too X or too Y
		neither X nor Y
		in X, Y
Portland	*Oregon*	the X metropolitan area in Y
		X International Airport in Y
		food X are maize, Y, etc
crop	*wheat*	dry X, such as Y,
		more X circles appeared in Y fields
		X OS comes with Y play
Android	*Google*	the X team at Y
		X is developed by Y

Table 1: Example word pairs (italicized) and their contexts (Wikipedia).

Unlike single word representations, which are typically trained by modeling the co-occurrence of a target word x with its context c, our word-pair representations are learned by modeling the three-way co-occurrence between two words (x,y) and the context c that ties them together, as illustrated in above Table. While similar training signal has been used to learn models for ontology construction and knowledge base completion, this paper shows, for the first time, that considerable scale learning of pairwise embeddings can be used to improve the performance of neural cross-sentence inference models directly.



Q6. What is Meta-Learning?

Answer:

Meta-learning: It is an exciting area of research that tackles the problem of learning to learn. The goal is to design models that can learn new skills or fastly to adapt to new environments with minimum training examples. Not only does this dramatically speed up and improve the design of ML(Machine Learning) pipelines or neural architectures, but it also allows us to replace hand-engineered algorithms with novel approaches learned in a data-driven way.

The goal of meta-learning is to train the model on a variety of learning tasks, such that it can solve new learning tasks with only a small number of training samples. It tends to focus on finding **model agnostic** solutions, whereas multi-task learning remains deeply tied to model architecture.

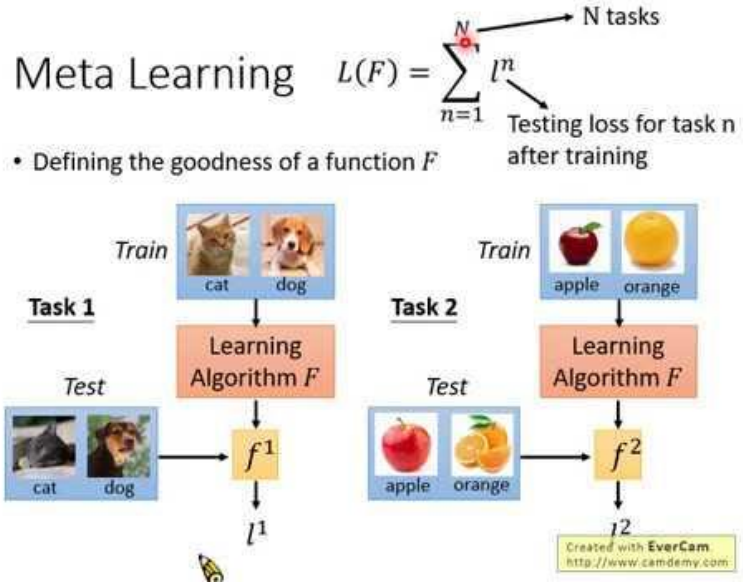
Thus, meta-level AI algorithms make AI systems:

· Learn faster

· Generalizable to many tasks

· Adaptable to environmental changes like in Reinforcement Learning

One can solve any problem with a single model, but meta-learning should not be confused with one-shot learning.



Q7. What is ALiPy(Active Learning in Python)?

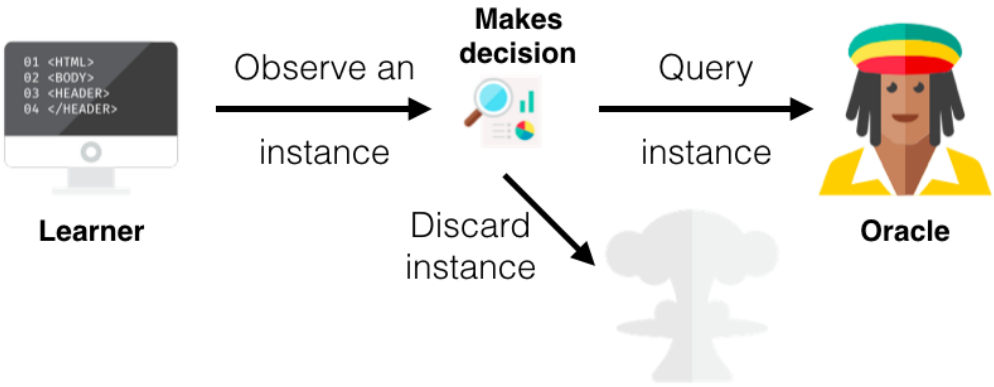
Answer:

Supervised ML methods usually require a large set of labeled examples for model training. However, in many real applications, there are ample unlabeled data but limited labeled data; and acquisition of labels is costly. Active learning (AL) reduces labeling costs by iteratively selecting the most valuable data to query their labels from the annotator.

Active learning is the leading approach to learning with limited labeled data. It tries to reduce human efforts on data annotation by actively querying the most prominent examples.

ALiPy is a Python toolbox for active learning(AL), which is suitable for various users. On the one hand, the entire process of active learning has been well implemented. Users can efficiently perform experiments by many lines of codes to finish the entire process from data pre-processes to

result in visualization. More than 20 commonly used active learning(AL) methods have been implemented in the toolbox, providing users many choices.



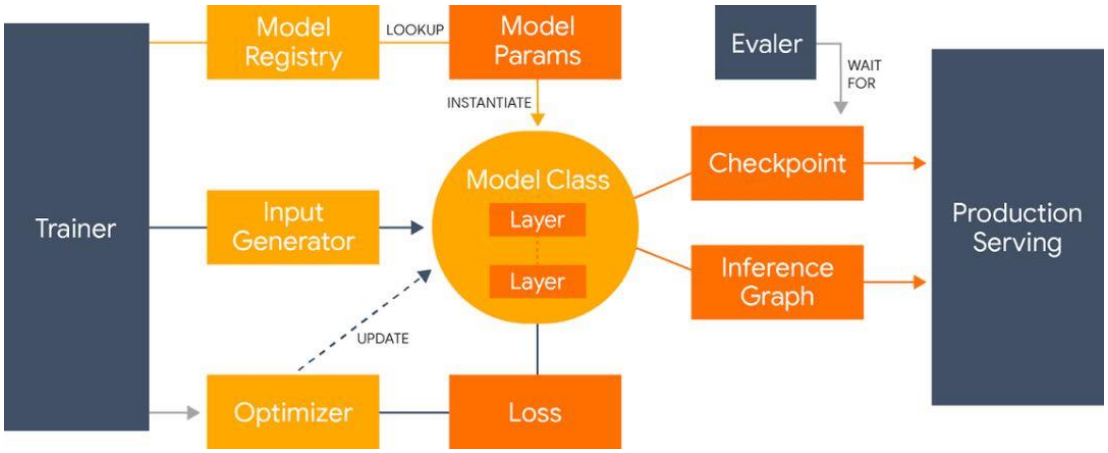
Q8.What is the Lingvo model?

Answer:

Lingvo: It is a Tensorflow framework offering a complete solution for collaborative deep learning research, with a particular focus towards sequence-to-sequence models. These models are composed of modular building blocks that are flexible and easily extensible, and experiment configurations are centralized and highly customizable. Distributed training and quantized inference are supported directly within a framework, and it contains existing implementations of an ample number of utilities, helper functions, and newest research ideas. This model has been used in collaboration by dozens of researchers in more than 20 papers over the last two years.

Why does this Lingvo research matter?

The process of establishing a new deep learning(DL) system is quite complicated. It involves exploring an ample space of design choices involving training data, data processing logic, the size, and type of model components, the optimization procedures, and the path to deployment. This complexity requires the framework that quickly facilitates the production of new combinations and the modifications from existing documents and experiments and shares these new results. It is a workspace ready to be used by deep learning researchers or developers. Nguyen Says: “We have researchers working on state-of-the-art(SOTA) products and research algorithms, basing their research off of the same codebase. This ensures that code is battle-tested. Our collective experience is encoded in means of good defaults and primitives that we have found useful over these tasks.”



Q9. What is Dropout Neural Networks?

Answer:

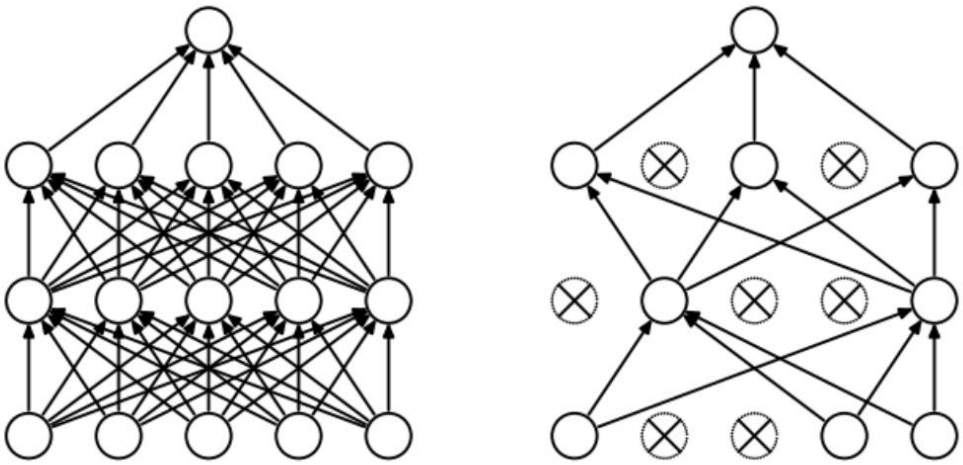
The term "dropout" refers to dropping out units (both hidden and visible) in a neural network.

At each training stage, individual nodes are either dropped out of the net with probability 1-p or kept with probability p, so that a reduced network is left; incoming and outgoing edges to a dropped-out node are also removed.

Why do we need Dropout?

The answer to these questions is "to prevent over-fitting."

A fully connected layer occupies most of the parameters, and hence, neurons develop co-dependency amongst each other during training, which curbs the individual power of each neuron leading to over-fitting of training data.



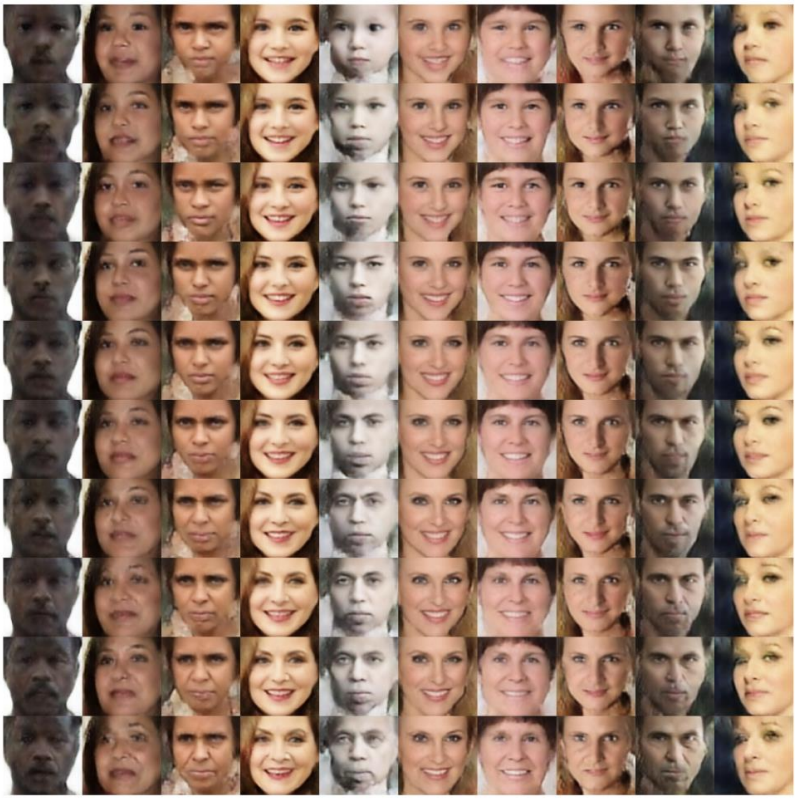
(a) Standard Neural Net (b) After applying dropout.

Q10. What is GAN?

Answer:

A **generative adversarial network** (**GAN**): It is a class of machine learning systems invented by Ian Goodfellow and his colleagues in 2014. Two neural networks are contesting with each other in a game (in the idea of game theory, often but not always in the form of a zero-sum game). Given a training set, this technique learns to generate new data with the same statistics as the training set. E.g., a GAN trained on photographs can produce original pictures that look at least superficially authentic to human observers, having many realistic characteristics. Though initially proposed as a form of a generative model for unsupervised learning, GANs have also proven useful for semi-supervised learning,[2] fully supervised learning, and reinforcement learning.

Example of GAN



- Given an image of a face, the network can construct an image that represents how that person could look when they are old.

Generative Adversarial Networks takes up a game-theoretic approach, unlike a conventional neural network. The network learns to generate from a training distribution through a 2-player game. The two entities are Generator and Discriminator. These two adversaries are in constant battle throughout the training process.

DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day21

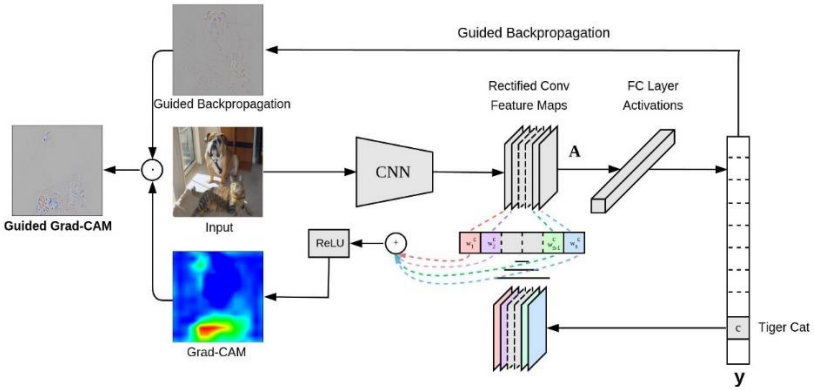
Q1. Explain Grad-CAM architecture?

Answer:

According to the research paper, "We propose a technique for making Convolutional Neural Network (CNN)-based models more transparent by visualizing input regions that are 'important' for predictions – producing *visual explanations*. Our approach is called Gradient-weighted Class Activation Mapping (Grad-CAM), which uses class-specific gradient information to localize the crucial regions. These localizations are combined with the existing pixel-space visualizations to create a new high-resolution, and class-discriminative display called the Guided Grad-CAM. These methods help better to understand CNN-based models, including image captioning and the apparent question answering (VQA) models. We evaluate our visual explanations by measuring the ability to discriminate between the classes and to inspire trust in humans, and their correlation with the occlusion maps. Grad-CAM provides a new way to understand the CNN-based models."

A technique for making CNN(Convolutional Neural Network)-based models more transparent by visualizing the regions of input that are "important" for predictions from these models — or visual explanations.

This visualization is both high-resolution (when the class of interest is 'tiger cat,' it identifies crucial 'tiger cat' features like stripes, pointy ears and eyes) and class-discriminative (it shows the 'tiger cat' but not the 'boxer (dog)').



Q2.Explain squeeze-net architecture?

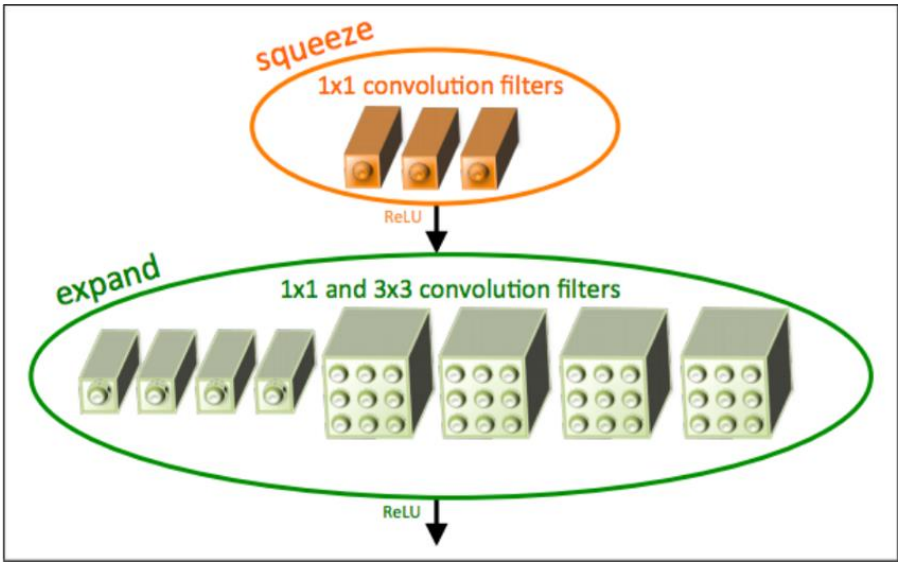
Answer:

Nowadays, technology is at its peak. Self-driving cars and IoT is going to be household talks in the next few years to come. Therefore, everything is controlled remotely, say, e.g., in self-driving cars, we will need our system to communicate with the servers regularly. So accordingly, if we have a model that has a small size, then we can quickly deploy it in the cloud. So that's why we needed an architecture that is less in size and also achieves the same level of accuracy that other architecture achieves.

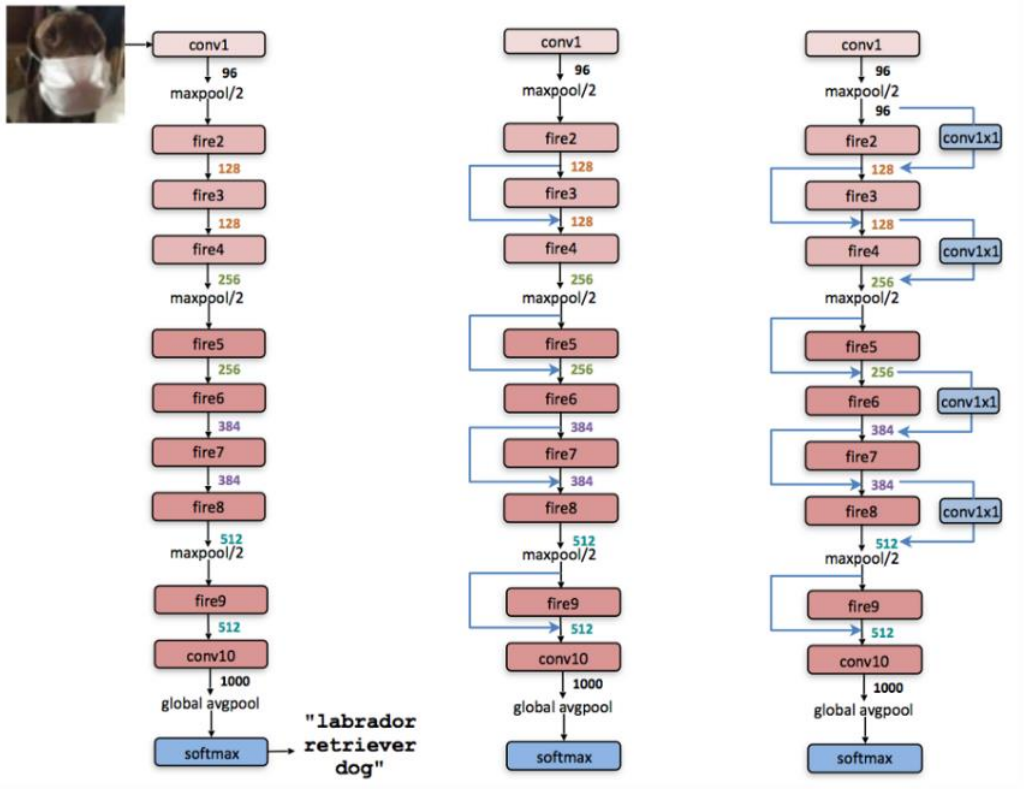
It's Architecture

- **Replace 3x3 filters with 1x1 filter-** We plan to use the maximum number of 1x1 filters as using a 1X1 filter rather than a 3X3 filter can reduce the number of parameters by 9X. We may think that replacing 3X3 filters with 1X1 filters may perform badly as it has less information to work on. But this is not a case. Typically 3X3 filter may capture the spatial information of pixels close to each other while the 1X1 filter zeros in on pixel and captures features amongst its channels.

- **Decrease number of input channels to 3x3 filters-** to maintain a small total number of parameters in a CNN, and it is crucial not only to decrease the number of 3x3 filters, but also to decrease the number of input channels to 3x3 filters. We reduce the number of input channels to 3x3 filters using *squeeze layers.* The author of this paper has used a term called the "*fire module,"* in which there is a squeeze layer and an expanded layer. In the squeeze layer, we are using 1X1 filters, while in the expanded layer, we are using a combo of 3X3 filters and 1X1 filters. The author is trying to limit the number of inputs to 3X3 filters to reduce the number of parameters in the layer.



- **Downsample late in a network so that convolution layers have a large activation map-** Having got an intuition about contracting the sheer number of parameters we are working with, how the model is getting most out of the remaining set of parameters. The author in this paper has downsampled the feature map in later layers, and this increases the accuracy. But this is an excellent contrast to networks like VGG where a large feature map is taken, and then it gets smaller as network approach towards the end. This different approach is too interesting, and they cite the paper by K. He and H. Sun that similarly applies delayed downsampling that leads to higher classification accuracy.
  This architecture consists of the fire module, which enables it to bring down the number of parameters.



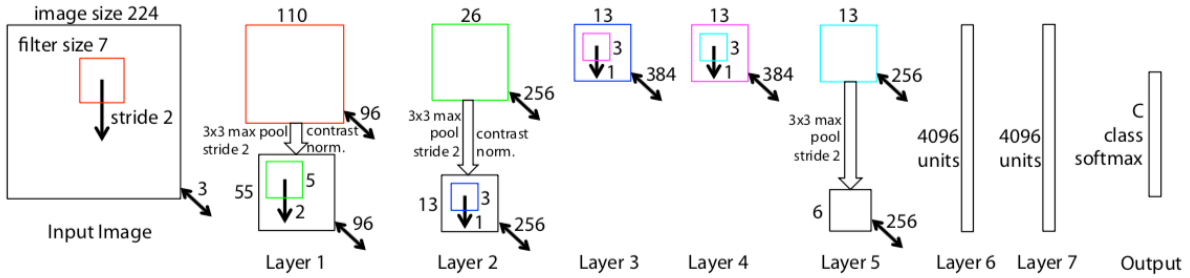
And other thing that surprises me is the lack of fully connected layers or dense layers at the end, which one will see in a typical CNN architecture. The dense layers, in the end, learn all the relationships between the high-level features and the classes it is trying to identify. The fully connected layers are designed to learn that noses and ears make up a face, and wheels and lights indicate cars. However, in this architecture, that extra learning step seems to be embedded within the transformations between various “fire modules.”

CNN architecture	Compression Approach	Data Type	Original → Compressed Model Size	Reduction in Model Size vs. AlexNet	Top-1 ImageNet Accuracy	Top-5 ImageNet Accuracy
AlexNet	None (baseline)	32 bit	240MB	1x	57.2%	80.3%
AlexNet	SVD [5]	32 bit	240MB → 48MB	5x	56.0%	79.4%
AlexNet	Network Pruning [11]	32 bit	240MB → 27MB	9x	57.2%	80.3%
AlexNet	Deep Compression [10]	5-8 bit	240MB → 6.9MB	35x	57.2%	80.3%
SqueezeNet (ours)	None	32 bit	4.8MB	**50x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	8 bit	4.8MB → 0.66MB	**363x**	57.5%	80.3%
SqueezeNet (ours)	Deep Compression	6 bit	4.8MB → 0.47MB	**510x**	57.5%	80.3%

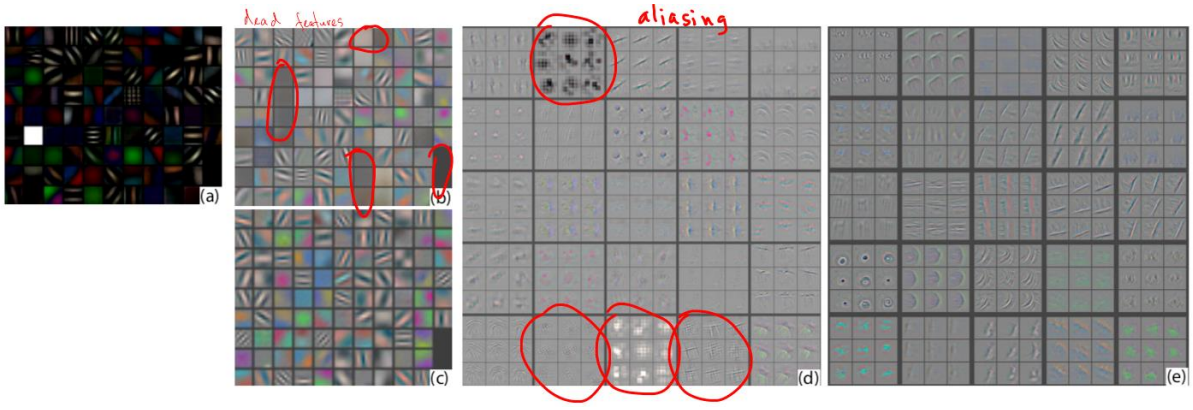
The squeeze-net can accomplish an accuracy nearly equal to AlexNet with 50X less number of parameters. The most impressive part is that if we apply Deep compression to the already smaller model, then it can reduce the size of the squeeze-net model to 510x times that of AlexNet.

Q3.ZFNet architecture

Answer:



The architecture of the network is an optimized version of the last year's winner - AlexNet. The authors spent some time to find out the bottlenecks of AlexNet and removing them, achieving superior performance.



(a): First layer ZFNET features without feature scale clipping. (b): the First layer features from AlexNet. Note that there are lot of dead features - ones where the network did not learn any patterns. (c): the First layer features for ZFNet. Note that there are only a few dead features. (d): Second layer features from AlexNet. The grid-like patterns are so-called aliasing artifacts. They appear when

receptive fields of convolutional neurons overlap, and neighboring neurons learn similar structures. (e): 2nd layer features for ZFNet. Note that there are no aliasing artifacts. Source: original paper.

In particular, they reduced the filter size in the 1st convolutional layer from 11x11 to 7x7, which resulted in fewer dead features learned in the first layer (see the image below for an example of that). A dead feature is a situation where a convolutional kernel fails to learn any significant representation. Visually it looks like a monotonic single-color image, where all the values are close to each other.

In addition to changing the filter size, the authors of FZNet have doubled the number of filters in all convolutional layers and the number of neurons in the fully connected layers as compared to the AlexNet. In the AlexNet, there were 48-128-192-192-128-2048-2048 kernels/neurons, and in the ZFNet, all these doubled to 96-256-384-384-256-4096-4096. This modification allowed the network to increase the complexity of internal representations and as a result, decrease the error rate from 15.4% for last year's winner, to 14.8% to become the winner in 2013.

Q4. What is NAS (Neural Architecture Search)?

Answer:

Developing the neural network models often requires significant architecture engineering. We can sometimes get by with transfer learning, but if we want the best possible performance, it's usually best to design your network. This requires specialized skills and is challenging in general; we may not even know the limits of the current state-of-the-art(SOTA) techniques. Its a lot of trial and error, and experimentation itself is time-consuming and expensive.

This is the NAS(Neural Architecture Search) comes in. NAS(Neural Architecture Search) is an algorithm that searches for the best neural network architecture. Most of the algorithms work in the following way. Start off by defining the set of "building blocks" that can be used for our network. E.g., the state-of-the-art(SOTA) NASNet paper proposes these commonly used blocks for an image recognition network-

- identity
- 1x7 then 7x1 convolution
- 3x3 average pooling
- 5x5 max pooling
- 1x1 convolution
- 3x3 depthwise-separable conv
- 7x7 depthwise-separable conv
- 1x3 then 3x1 convolution
- 3x3 dilated convolution
- 3x3 max pooling
- 7x7 max pooling
- 3x3 convolution
- 5x5 depthwise-seperable conv

In the NAS algorithm, the controller Recurrent Neural Network (RNN) samples the building blocks, putting them together to create some end to end architecture. Architecture generally combines the same style as state-of-the-art(SOTA) networks, such as DenseNets or ResNets, but uses a much different combination and the configuration of blocks.

This new network architecture is then trained to convergence to obtain the least accuracy on the held-out validation set. The resulting efficiencies are used to update the controller so that the controller will generate better architectures over time, perhaps by selecting better blocks or making better connections. The controller weights are updated with a policy gradient. The whole end-to-end setup is shown below.

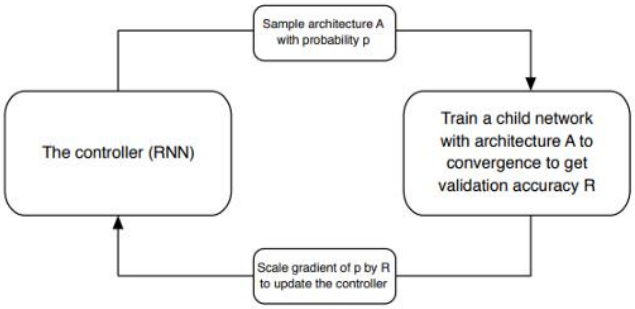
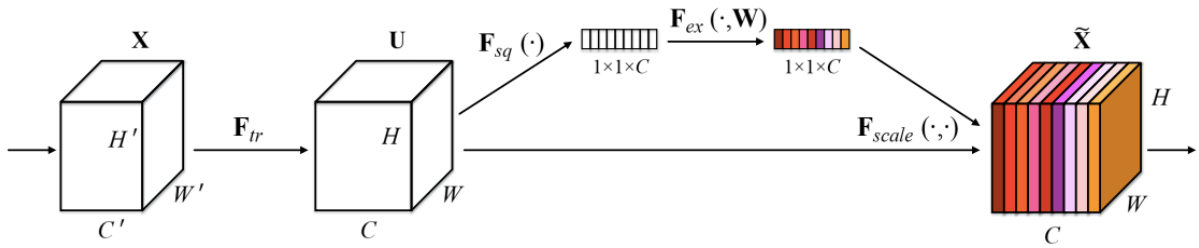


Figure 1. Overview of Neural Architecture Search [71]. A controller RNN predicts architecture A from a search space with probability p. A child network with architecture A is trained to convergence achieving accuracy R. Scale the gradients of p by R to update the RNN controller.

It's a reasonably intuitive approach! In simple means: have an algorithm grab different blocks and put those blocks together to make the network. Train and test out that network. Based on our results, adjust the blocks we used to make the network and how you put them together!

Q5. What is SENets?

Answer:



SENets stands for Squeeze-and-Excitation Networks introduces a building block for CNNs that improves channel interdependencies at almost no computational cost. They have used in the 2017 ImageNet competition and helped to improve the result from last year by 25%. Besides this large performance boost, they can be easily added to existing architectures. The idea is this:

Let's add parameters to each channel of the convolutional block so that the network can adaptively adjust the weighting of each feature map.

As simple as may it sound, this is it. So, let's take a closer look at why this works so well.

Why it works too well?

CNN's uses its convolutional filters to extract hierarchal information from the images. Lower layers find little pieces of context like high frequencies or edges, while upper layers can detect faces, text, or other complex geometrical shapes. They extract whatever is necessary to solve the task precisely.

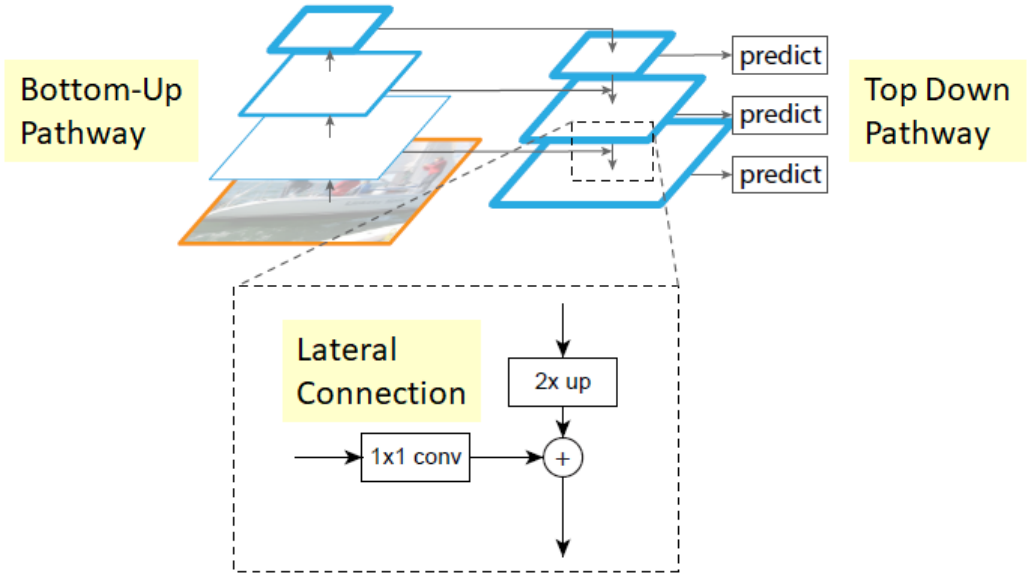
All of this works by fusing spatial and channel information of an image. The different filters will first find the spatial features in each input channel before adding the information across all available output channels.

All we need to understand for now is that the network weights each of its channels equally when creating output feature maps. It is all about changing this by adding a content-aware mechanism to weight each channel adaptively. In its too basic form, this could mean adding a single parameter to each channel and giving it linear scalar how relevant each one is.

However, the authors push it a little further. First, they get the global understanding of each channel by squeezing feature maps to a single numeric value. This results in the vector of size *n*, where *n* is equal to the number of convolutional channels. Afterward, it is fed through a two-layer neural network, which outputs a vector of the same size. These *n* values can now be used as weights on the original features maps, scaling each channel based on its importance.

Q6. Feature Pyramid Network (FPN)

Answer:



The Bottom-Up Pathway

The bottom-up pathway is feedforward computation of backbone ConvNet. It is known as one pyramid level is for each stage. The output of last layer of each step will be used as the reference set of feature maps for enriching the top-down pathway by lateral connection.

Top-Down Pathway and Lateral Connection

- The higher resolution features are upsampled spatially coarser, but semantically stronger, feature maps from higher pyramid levels. More particularly, the spatial resolution is upsampled by a factor of 2 using nearest neighbor for simplicity.
- Each lateral connection adds feature maps of the same spatial size from the bottom-up pathway and top-down pathway.
- Specifically, **the feature maps from the bottom-up pathway undergo 1×1 convolutions to reduce channel dimensions.**
- And **feature maps from the bottom-up pathway and top-down pathway are merged by element-wise addition.**

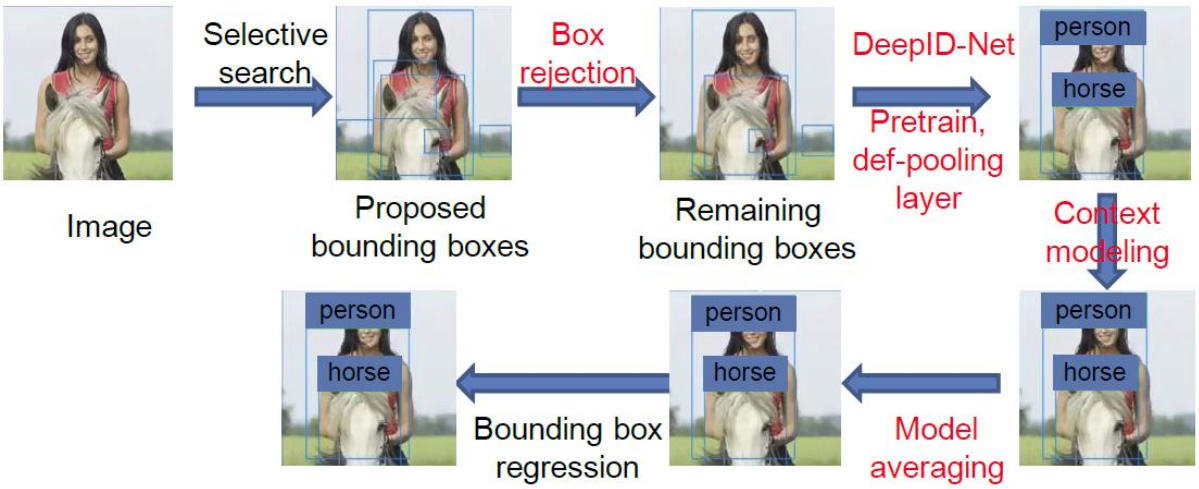
Prediction in FPN

- Finally, **the 3×3 convolution is appended on each merged map to generate a final feature map, which is to reduce the aliasing effect of upsampling.** This last set of feature maps is called {P2, P3, P4, P5}, corresponding to {C2, C3, C4, C5} that are respectively of same spatial sizes.
- Because all levels of pyramid use shared classifiers/regressors as in a traditional featured image pyramid, feature dimension at output d is fixed with $d = 256$. Thus, all extra convolutional layers have 256 channel outputs.

Q7. DeepID-Net(Def-Pooling Layer)

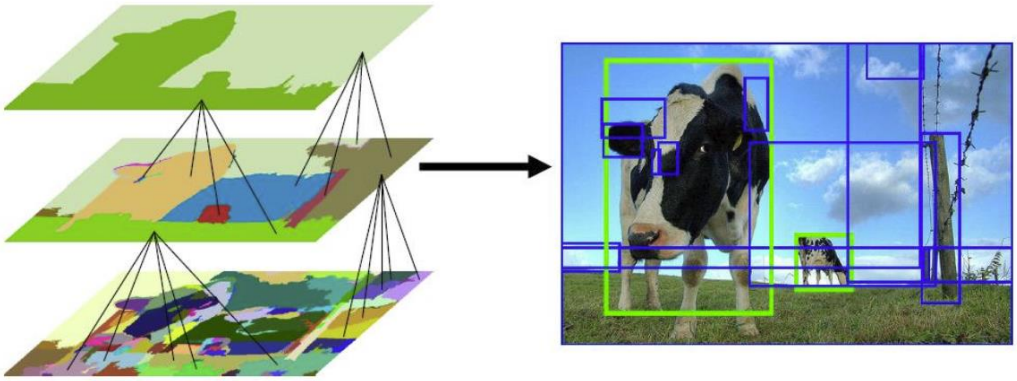
Answer:

A new def-pooling (deformable constrained pooling) layer is used to model the deformation of the object parts with geometric constraints and penalties. That means, except detecting the whole object directly, it is also important to identify object parts, which can then assist in detecting the whole object.



The steps in **black** color are the **old stuff** that **existed in R-CNN**. The stages in **red** color do not appear **in R-CNN**.

1. **Selective Search**

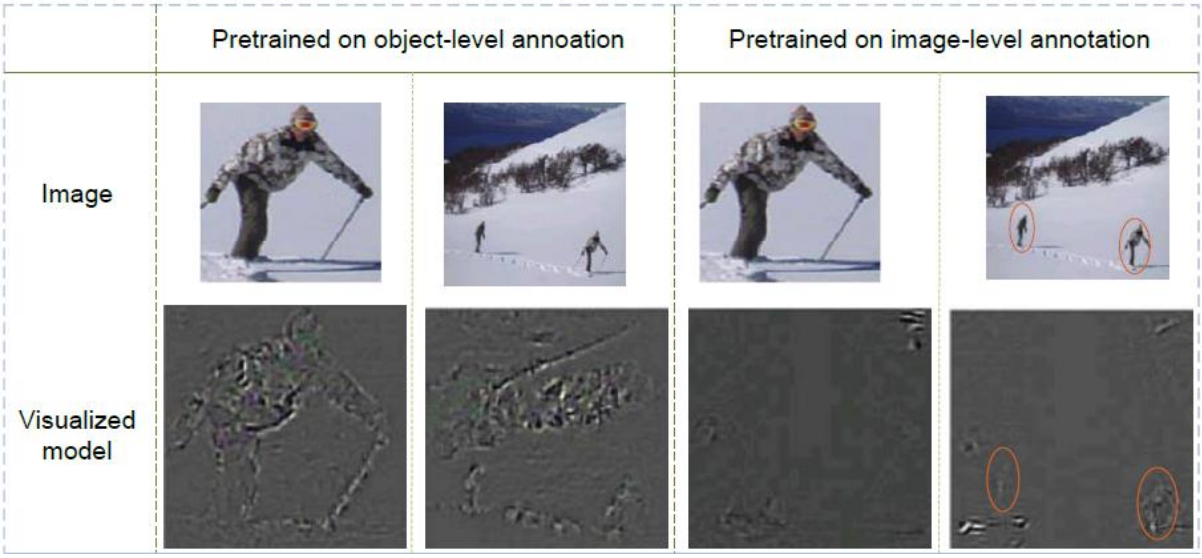


- First, color similarities, texture similarities, regions size, and region filling are used as **non-object-based segmentation**. Therefore you obtain **many small segmented areas** as shown at the bottom left of the image above.
- Then, the bottom-up approach is used that **small segmented areas are merged to form the larger segment areas.**
- Thus, **about 2K regions, proposals (bounding box candidates) are generated,** as shown in the above image.

2. Box Rejection

R-CNN is used to reject bounding boxes that are most likely to be the background.

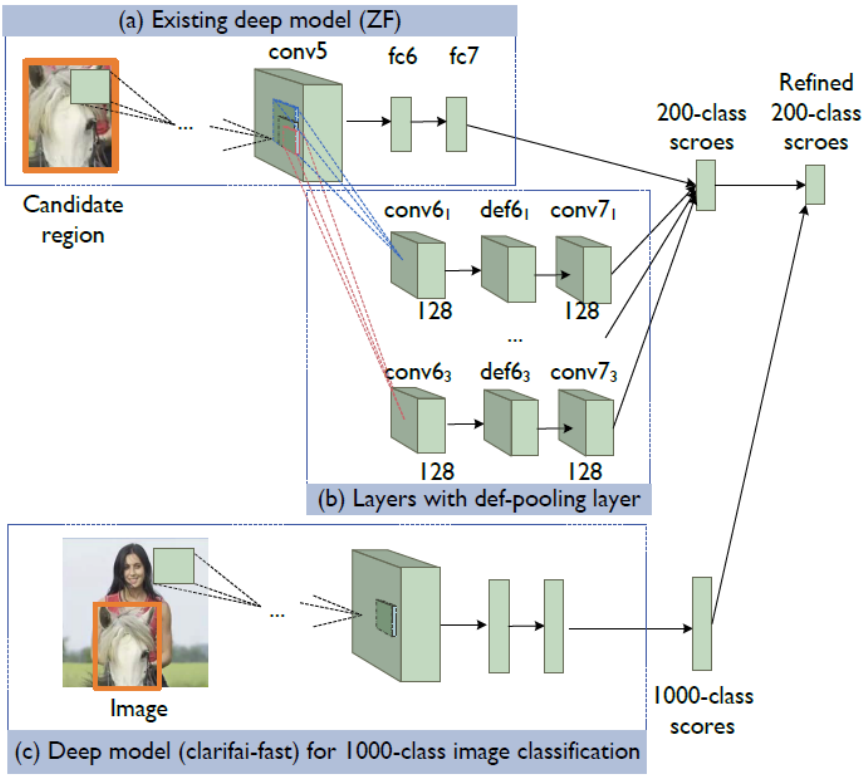
3. Pre train Using Object-Level Annotations

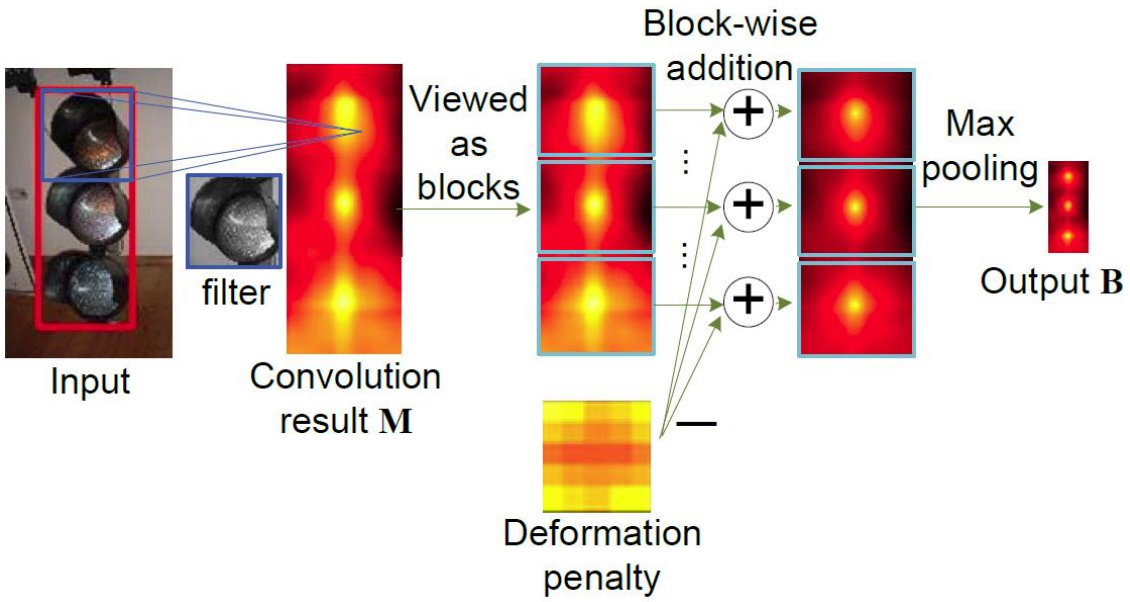


Usually, pretraining is on **image-level annotation**. It is **not good when an object is too small within the image** because the object should occupy a large area within the bounding box created by the selective search.

Thus, **pretraining is on object-level annotation**. And **the deep learning(DL) model can be any models** such as ZFNet, VGGNet, and GoogLeNet.

4. Def-Pooling Layer





$$b_c^{(x,y)} = \max_{\delta_x,\delta_y \in \{-R,\cdots,R\}} \{m_c^{\mathbf{z}_{\delta_x,\delta_y}} - \sum_{n=1}^{N} a_{c,n} d_{c,n}^{\delta_x,\delta_y}\}$$

$$\text{where } \mathbf{z}_{\delta_x,\delta_y} = (s_x \cdot x + \delta_x, s_y \cdot y + \delta_y).$$

For the def-pooling path, output from conv5, goes through the Conv layer, then goes through the def-pooling layer, and then has a max-pooling layer.

In simple terms, the summation of ac multiplied by dc,n, is the 5×5 deformation penalty in the figure above. The penalty of placing object part from assumed the central position.

By training the DeepID-Net, object parts of the object to be detected will give a high activation value after the def-pooling layer if they are closed to their anchor places. And this output will connect to 200-class scores for improvement.

5. Context Modeling

In object detection tasks in ILSVRC, there are 200 classes. And there is also the classification competition task in ILSVRC for classifying and localizing 1000-class objects. The contents are more diverse compared with the object detection task. Hence, **1000-class scores, obtained by classification network, are used to refine 200-class scores.**

6. The Model Averaging-

Multiple models are used to increase the accuracy, and **the results from all models are averaged**. This technique has been used since AlexNet, LeNet, and so on.

7. Bounding Box Regression

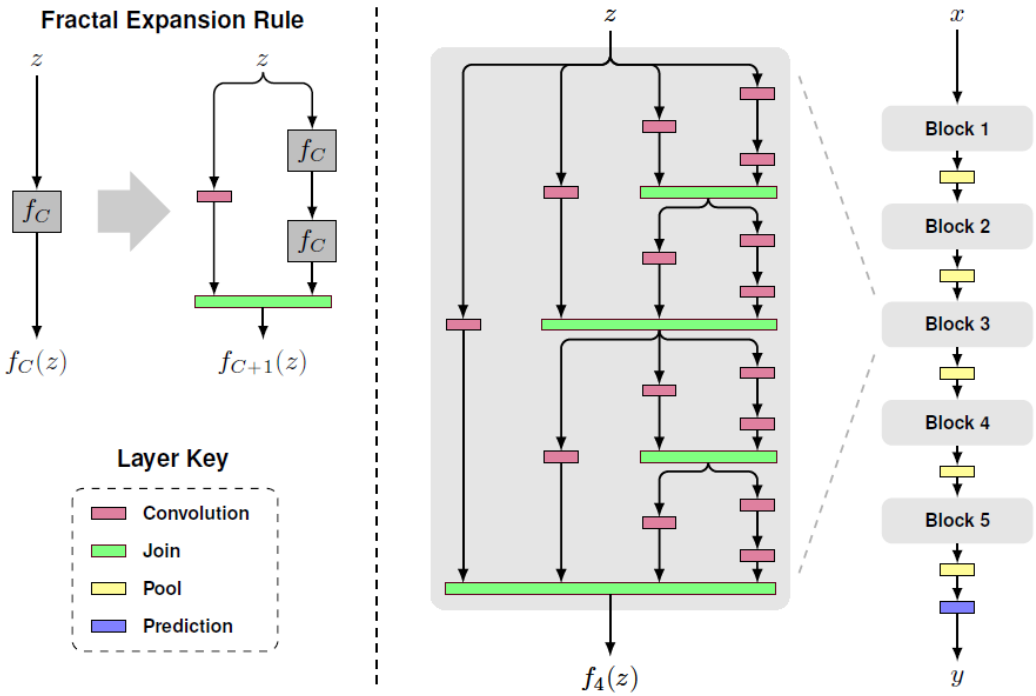
Bounding box regression is to **fine-tune the bounding box location**, which has been used in R-CNN.

Q8. What is FractalNet Architecture?

Answer:

In2015, after the invention of ResNet, with numerous champion won, there are plenty of researchers working on how to improve the ResNet, such as Pre-Activation ResNet, RiR, RoR, Stochastic Depth, and WRN. In this story, conversely, a non-residual-network approach, FractalNet, is shortly reviewed. When VGGNet is starting to degrade when it goes from 16 layers (VGG-16) to 19 layers (VGG-19), FractalNet can go up to 40 layers or even 80 layers.

Architecture



In the above picture: A Simple Fractal Expansion (on Left), Recursively Stacking of Fractal Expansion as One Block (in the Middle), 5 Blocks Cascaded as FractalNet (on the Right)

For the base case, f1(z) is the convolutional layer:

$$f_1(z) = \text{conv}(z)$$

After that, recursive fractals are:

$$f_{C+1}(z) = [(f_C \circ f_C)(z)] \oplus [\text{conv}(z)]$$

Where C is a number of columns as in the middle of the above figure. The number of the convolutional layers at the deepest path within the block will have 2^(C-1). In this case, **C=4**, thereby, a number of convolutional layers are 2^3=8 layers.

For the **join layer** (green), the **element-wise mean** is computed. It is not concatenation or addition.

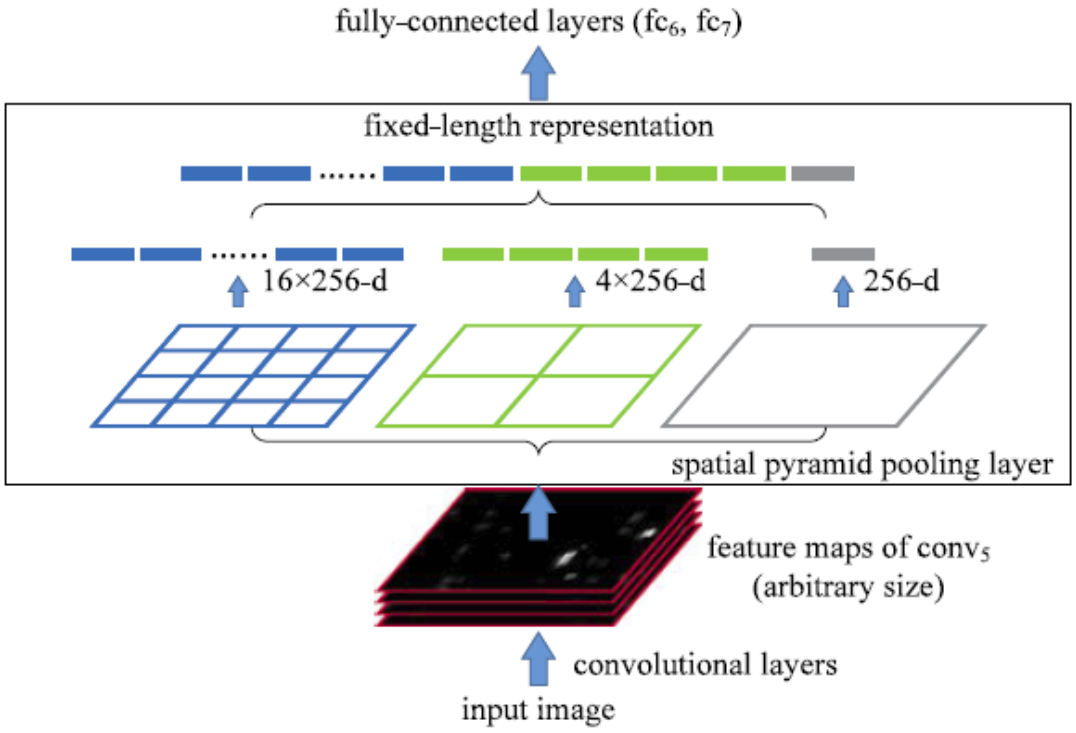
With five blocks (**B=5**) cascaded as FractalNet at the right of the figure, then the number of convolutional layers at the most profound path within the whole network is B×2^(C-1), i.e., 5×2^3=**40 layers**.

In between 2 blocks, 2×2 max pooling is done to reduce the size of feature maps. Batch Norm and ReLU are used after each convolution.

Q9. What is the SSPNet architecture?

Answer:

SPPNet has introduced the new technique in CNN called **Spatial Pyramid Pooling (SPP)** at the transition of the convolutional layer and fully connected layer. This is a work from **Microsoft**.



Conventionally, at the transformation of the Conv layer and FC layer, there is one single pooling layer or even no pooling layer. In SPPNet, it suggests having **multiple pooling layers with different scales**.

In the figure, **3-level SPP** is used. Suppose conv5 layer has 256 feature maps. Then at the SPP layer,

1. first, each feature map is **pooled to become one value (which is grey)**. Thus **256-d vector is formed**.
2. Then, each feature map is **pooled to have four values (which is green)**, and form the **4×256-d vector**.
3. Similarly, each feature map is **pooled to have 16 values (in blue)**, and form the **16×256-d vector**.
4. The **above three vectors are concatenated to form a 1-d vector**.
5. Finally, this **1-d vector is going into FC layers** as usual.

With SPP, you don't need to crop the image to a fixed size, like AlexNet, before going into CNN. **Any image sizes can be inputted.**

DATA SCIENCE INTERVIEW PREPARATION

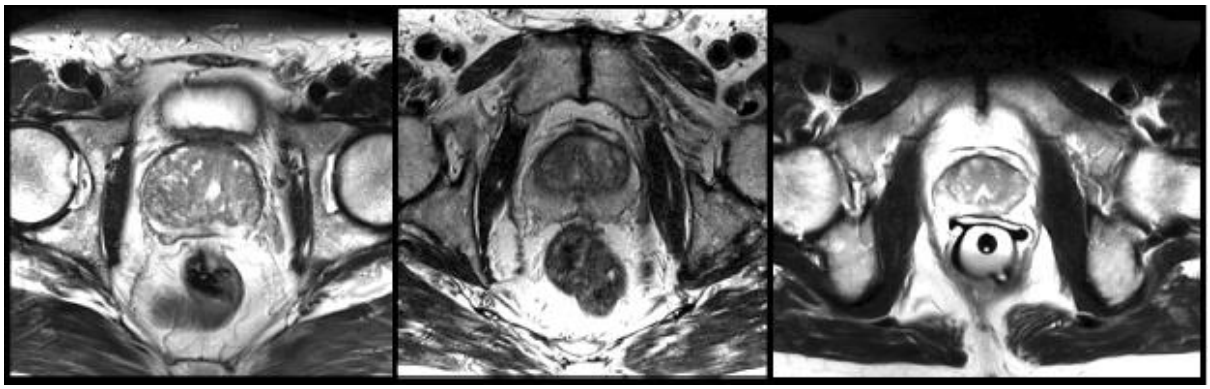
(30 Days of Interview Preparation)

Day22

Q1. Explain V-Net (Volumetric Convolution) Architecture with related to Biomedical Image Segmentation?

Answer:

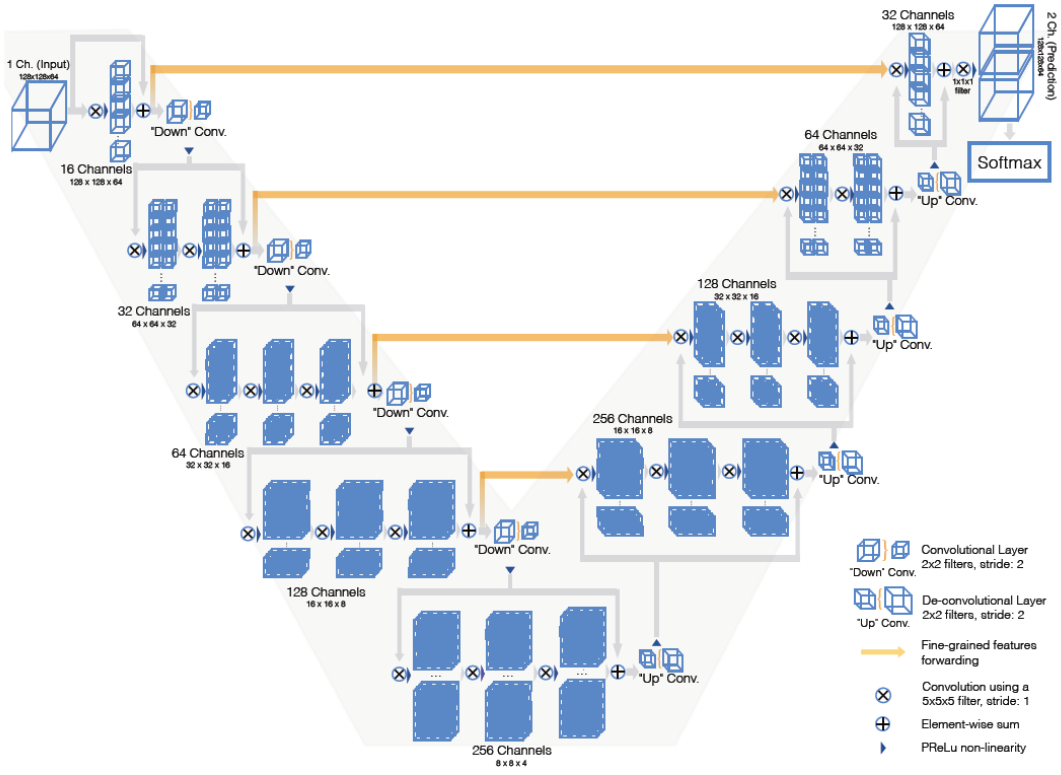
There were several medical data used in clinical practice consists of 3D volumes, such as MRI volumes illustrate prostate, while most approaches are only able to process 2D images. A 3D image segmentation based on a volumetric, fully convolutional neural network is proposed in this work.



Slices from MRI volumes depicting prostate

Prostate segmentation nevertheless is the crucial task having clinical relevance both during diagnosis, where the volume of the prostate needs to be assessed and during treatment planning, where the estimate of the anatomical boundary needs to be accurate.

Architecture



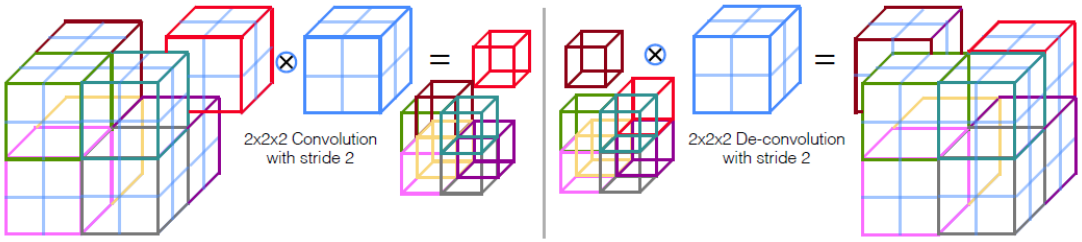
- V-Net, justifies by its name, it is shown as V-shape. The left part of the network consists of a compression path, while on the right part decompresses signal until its original size is reached.
- This is the same as U-Net, but with some difference.

On Left

- The left side of the network is divided into different stages that operate at various resolutions. Each stage comprises one to 3 convolutional layers.
- **At each stage, a residual function is learned.** The input of each stage is used in convolutional layers and processed through non-linearities and added to the output of the last convolutional layer of that stage to enable learning a residual function. This V-net architecture ensures convergence compared with non-residual learning networks such as U-Net.
- The **convolutions** performed in each stage use **volumetric kernels** having the size of **5×5×5 voxels**. (A voxel represents a value on a regular grid in 3D-space. The term voxel is commonly used in 3D much 3D space, just like voxelization in a point cloud.)
- Along the compression path, the **resolution is reduced by convolution with 2×2×2 voxels full kernels applied with stride 2**. Thus, the size of the resulting feature maps is halved, with a **similar purpose as pooling layers**. And **number of feature channels doubles at each stage** of the compression path of V-Net.
- Replacing pooling operations with convolutional ones helps to have a smaller memory footprint during training because no switches mapping the output of pooling layers back to their inputs are needed for back-propagation.
- Downsampling helps to increase the receptive field.
- **PReLU** is used as a non-linearity activation function.

On Right Part

- The network extracts features and expands spatial support of the lower resolution feature maps to gather and assemble the necessary information to output a two-channel volumetric segmentation.



- At each stage, a **deconvolution** operation is employed to **increase the size of the inputs followed by one to three convolutional layers,** involving **half the number of 5×5×5 kernels** applied in the previous layer.
- **The residual function** is learned, similar to left part of the network.
- The 2 features maps computed by **a very last convolutional layer**, having **1×1×1 kernel size** and producing **outputs of the same size as input volume.**
- These two output feature maps are **the probabilistic segmentation of the foreground and background regions by applying soft-max voxelwise**.

Q2. Highway Networks- Gating Function to highway

Answer:

It is found that difficulties are optimizing a very deep neural network. However, it's still an open problem with why it is difficult to optimize a deep network. (it is due to gradient vanishing problem.) Inspired by LSTM (Long Short-Term Memory), authors thereby **make use of gating function to adaptively bypass or transform the signal so that the network can go deeper.** The deep network with more than 1000 layers can also be optimized.

Plain Network

Before going into Highway Networks, Let us start with plain network which consists of *L* layers where the *l*-th layer (with omitting the symbol for the layer):

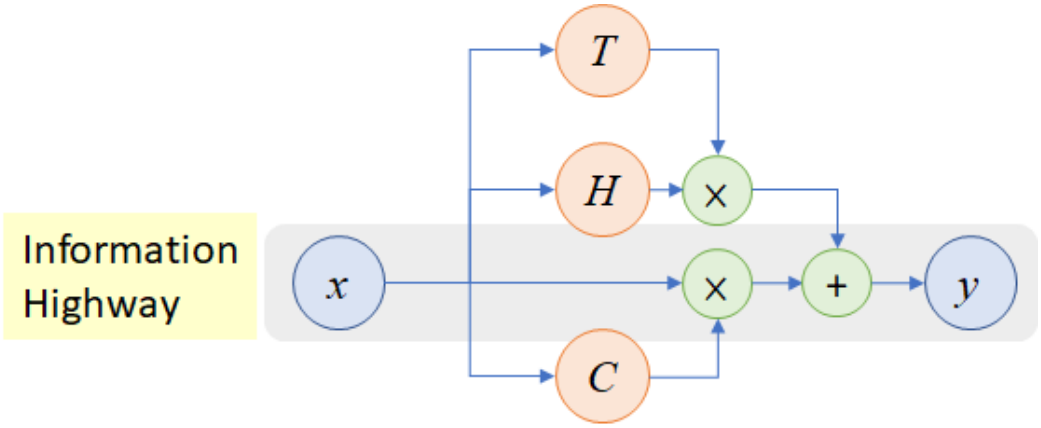
$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}).$$

Where *x* is input, *WH* is the weight, *H* is the transform function followed by an activation function, and *y* is the output. And for *i*-th unit:

$$y_i = H_i(\mathbf{x})$$

We compute the *yi* and pass it to the next layer.

Highway Network



In a highway network, 2 non-linear transforms ***T*** and ***C*** are introduced:

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot C(\mathbf{x}, \mathbf{W_C}).$$

where ***T* is Transform Gate**, and **C is the Carry Gate**.

In particular, ***C* = 1 - *T***:

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot (1 - T(\mathbf{x}, \mathbf{W_T})).$$

We can have below conditions for specific *T* values:

$$\mathbf{y} = \begin{cases} \mathbf{x}, & \text{if } T(\mathbf{x}, \mathbf{W_T}) = \mathbf{0}, \\ H(\mathbf{x}, \mathbf{W_H}), & \text{if } T(\mathbf{x}, \mathbf{W_T}) = \mathbf{1}. \end{cases}$$

When *T*=0, we pass input as output directly, which creates an information highway. That's why it is called the Highway Network.

When *T*=1, we use non-linear activated transformed input as output.

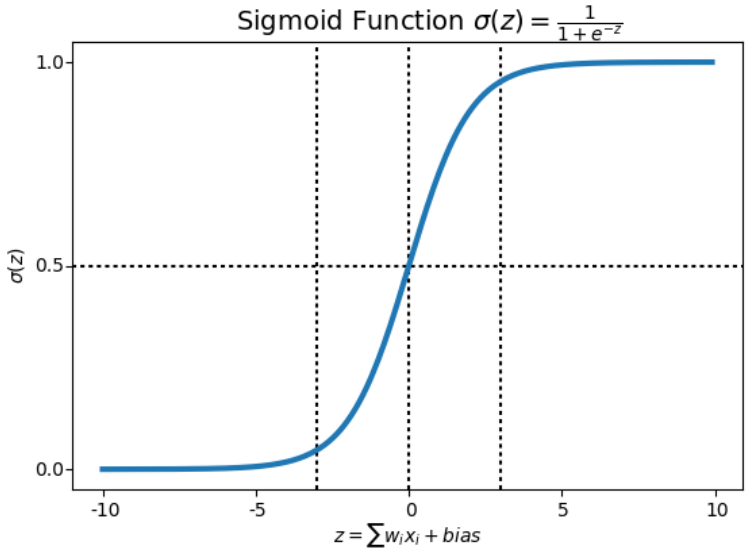
Here, in contrast to the *i*-th unit in plain network, the authors introduce the ***block*** concept. For ***i*-th block**, there is a **block state *Hi*(*x*)**, and **transform gate output *Ti*(*x*)**. And the corresponding **block output *yi***:

$$y_i = H_i(\mathbf{x}) * T_i(\mathbf{x}) + x_i * (1 - T_i(\mathbf{x}))$$

which is connected to the next layer.

- Formally, **T(*x*) is the sigmoid function**:

$$f_{C+1}(z) = [(f_C \circ f_C)(z)] \oplus [\text{conv}(z)]$$



Sigmoid function caps the output between 0 to 1. When the input has a too-small value, it becomes 0. When the input has a too-large amount, it becomes 1. **Therefore, by learning *WT* and *bT*, a network can adaptively pass *H*(*x*) or pass *x* to the next layer.**

And the author claims that this helps to have the simple initialization scheme for *WT* which is independent of nature of *H*.

bT can be initialized with the negative value (e.g., -1, -3, etc.) such that the network is initially biased towards carrying behaviour.

LSTM inspires the above idea as the authors mentioned.

And SGD(**Stochastic Gradient Descent) did not stall for networks with more than 1000 layers.** However, the exact results have not been provided.

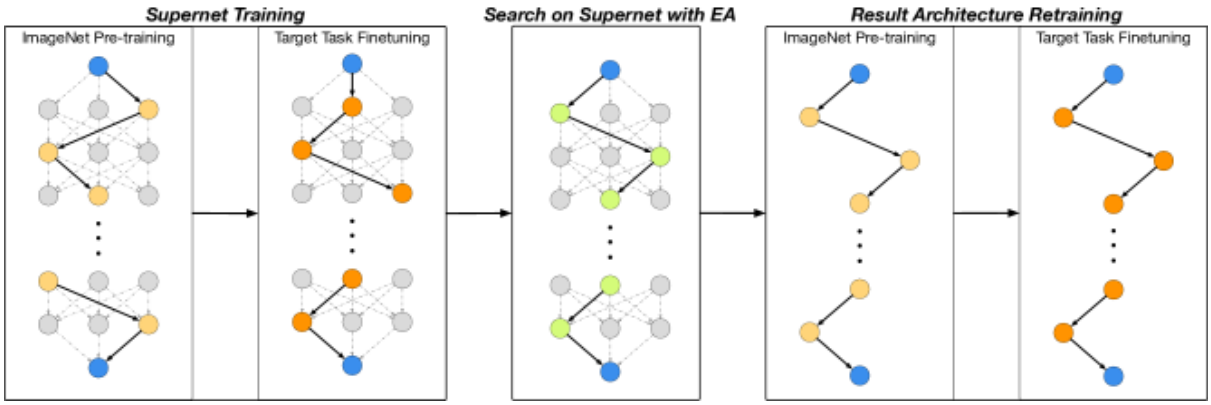
Q3. What is DetNAS: Neural Architecture Search(NAS) on Object Detection?

Answer:

Object detection is one of the most fundamental computer vision(OpenCV) tasks and has been widely used in real-world applications. The performance of object detectors highly relies on features extracted by backbones. However, most works on object detection directly use networks designed for classification as a backbone the feature extractors, e.g., ResNet. The architectures optimized on image classification can not guarantee performance on object detection. It is known that there is an essential gap between these two different tasks. Image classification basically focuses on ”What” main object of the image is, while object detection aims at finding ”Where” and ”What” each object

instance in an image. There have been little works focusing on backbone design for object detector, except the hand-craft network, DetNet.

Neural architecture search (NAS) has achieved significant progress in image classification and semantic segmentation. The networks produced by search have reached or even surpassed the performance of the hand-crafted ones on this task. But object detection has never been supported by NAS before. Some NAS(Neural architecture search) work directly applies architecture searched on CIFAR-10 classification on object detection.



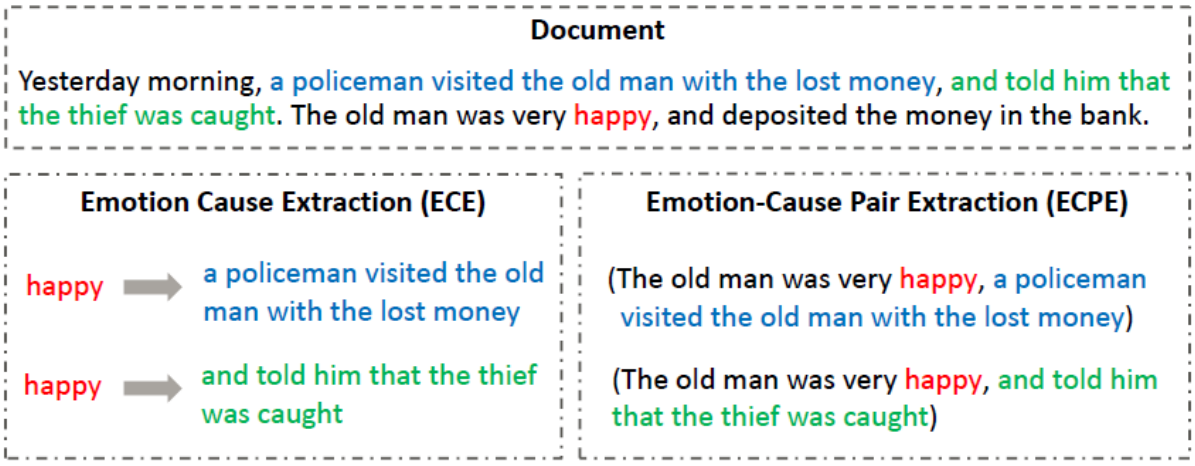
In this work, we present the first effort towards learning a backbone network for object detection tasks. Unlike previous NAS works, our method does not involve any architecture-level transfer. We propose DetNAS to conduct neural architecture search directly on the target tasks. The quests are even performed with precisely the same settings to the target task. Training an objector detector usually needs several days and GPUs, no matter using a pre-train-finetune scheme or training from scratch. Thus, it is not affordable to directly use reinforcement learning (RL) or evolution algorithm (EA) to search the architectures independently. To overcome this obstacle, we formulate this problem into searching the optimal path in the large graph or supernet. In simple terms, DetNAS consists of three steps: (1) training a supernet that includes all sub-networks in search space; (2) searching for the sub-network with the highest performance on the validation set with EA; (3) retraining the resulting network and evaluating it on the test set.

Q4.You have any idea about ECE (Emotion cause extraction).

Answer:

Emotion cause extraction (ECE) aims at extracting potential causes that lead to emotion expressions in the text. The ECE task was first proposed and defined as a word-level sequence labeling problem in Lee et al. To solve the shortcoming of extracting causes at the word level, Gui et al. 2016 released a new corpus which has received much attention in the following study and becomes a benchmark dataset for ECE research.

Below Fig. Displays an example from this corpus, there are five clauses in a document. The emotion "happy" is contained in fourth clause. We denote this clause as an *emotion clause*, which refers to a term that includes emotions. It has two corresponding causes: "a policeman visited the old man with the lost money" in the second clause and, "told him that the thief was caught" in the third clause. We name them as *cause clause*, which refers to a term that contains causes.



In this work, we propose a new task: emotion-cause pair extraction (ECPE), which aims to extract all potential pairs of emotions and corresponding causes in the document. In Above Fig, we show the difference between the traditional ECE task and our new ECPE task. The goal of ECE is to extract the corresponding cause clause of the given emotion. In addition to a document as the input, ECE needs to provide annotated feeling at first before cause extraction.

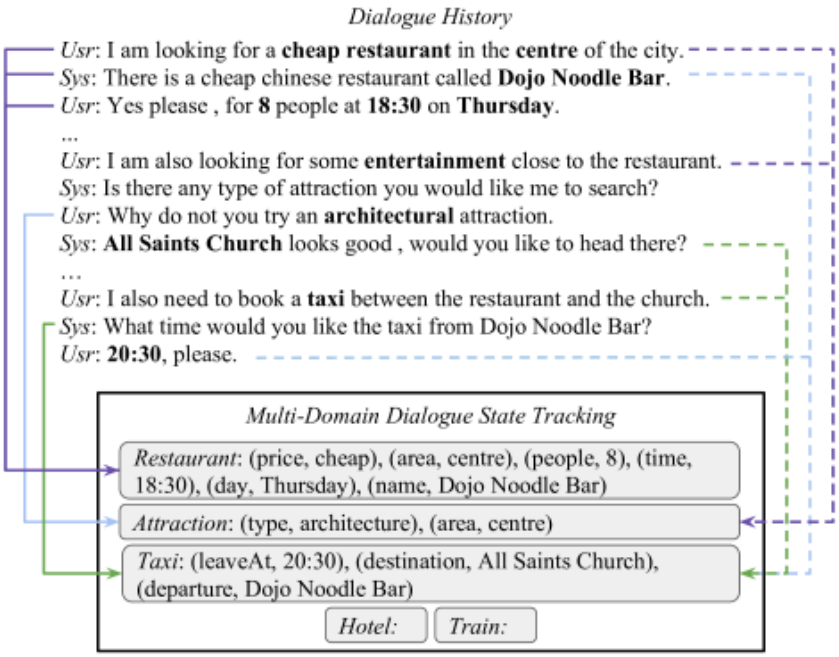
In contrast, the output of our ECPE task is a pair of emotion-cause, without the need of providing emotion annotation in advance. From Above fig., e.g., given the annotation of feeling: "happy," the goal of ECE is to track the two corresponding cause clauses: "a policeman visited the old man with the lost money" and "and told him that the thief was caught." While in the ECPE task, the goal is to directly extract all pairs of emotion clause and cause clause, including ("The old man was delighted", "a policeman visited the old man with the lost money") and ("The old man was pleased", "and told him that the thief was caught"), without providing the emotion annotation "happy".

To address this new ECPE task, we propose a two-step framework. Step 1 converts the emotion-cause pair extraction task to two individual sub-tasks (emotion extraction and cause extraction respectively) via two kinds of multi-task learning networks, intending to extract a set of emotion clauses and a set of cause clauses. Step 2 performs emotion-cause pairing and filtering. We combine all the elements of the two sets into pairs and finally train a filter to eliminate the couples that do not contain a causal relationship.

Q5.What is DST (Dialogue state tracking)?

Answer:

Dialogue state tracking (DST) is a core component in task-oriented dialogue systems, such as restaurant reservations or ticket bookings. The goal of DST is to extract user goals expressed during conversation and to encode them as a compact set of the dialogue states, i.e., a set of slots and their corresponding values. E.g., as shown in below fig., *(slot, value)* pairs such as *(price, cheap)* and *(area, centre)* are extracted from the conversation. Accurate DST performance is important for appropriate dialogue management, where user intention determines the next system action and the content to query from the databases.



State tracking approaches are based on the assumption that ontology is defined in advance, where all slots and their values are known. Having a predefined ontology can simplify DST into a classification problem and improve performance (Henderson et al., 2014b; Mrkšić et al., 2017; Zhong et al., 2018). However, there are two significant drawbacks to this approach: 1) A full ontology is hard to obtain in advance (Xu and Hu, 2018). In the industry, databases are usually exposed through an external API only, which is owned and maintained by others. It is not feasible to gain access to enumerate all the possible values for each slot. 2) Even if a full ontology exists, the number of possible slot values could be significant and variable. For example, a restaurant name or a train departure time can contain a large number of possible values. Therefore, many of the previous works that are based on neural classification models may not be applicable in real scenarios.

Q6.What is NMT(Neural machine translation)?

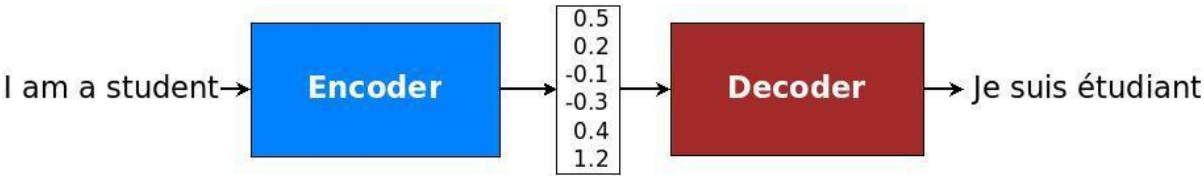
Answer:

NMT stands for Neural machine translation, which is the use of neural network models to learn the statistical model for machine translation.

The key benefit to the approach is that the single system can be trained directly on the source and target text, no longer requiring the pipeline of specialized methods used in statistical (ML)machine learning.

Unlike the traditional phrase-based translation system which consists of many sub-components that are tuned separately, neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation.

As such, neural machine translation(NMT) systems are said to be end-to-end systems as only one model is required for the translation.



In Encoder

The task of the encoder is to provide the representation of a input sentence.The input sentence is a sequence of words, for which we first consult embedding matrix. Then, as in the primary language model described previously, we process these words with a recurrent neural network(RNN). This results in hidden states that encode each word with its left context, i.e., all the preceding words. To also get the right context, we also build a recurrent neural network(RNN) that runs right-to-left, or, from the end of the sentence to beginning. Having two recurrent neural networks(RNN) running in two directions is known as the bidirectional recurrent neural network(RNN).

In Decoder

The decoder is the recurrent neural network(RNN). It takes some representation of input context (more on that in the next section on the attention mechanism) and previous hidden state and the output word prediction, and generates a new hidden decoder state and the new output word prediction.

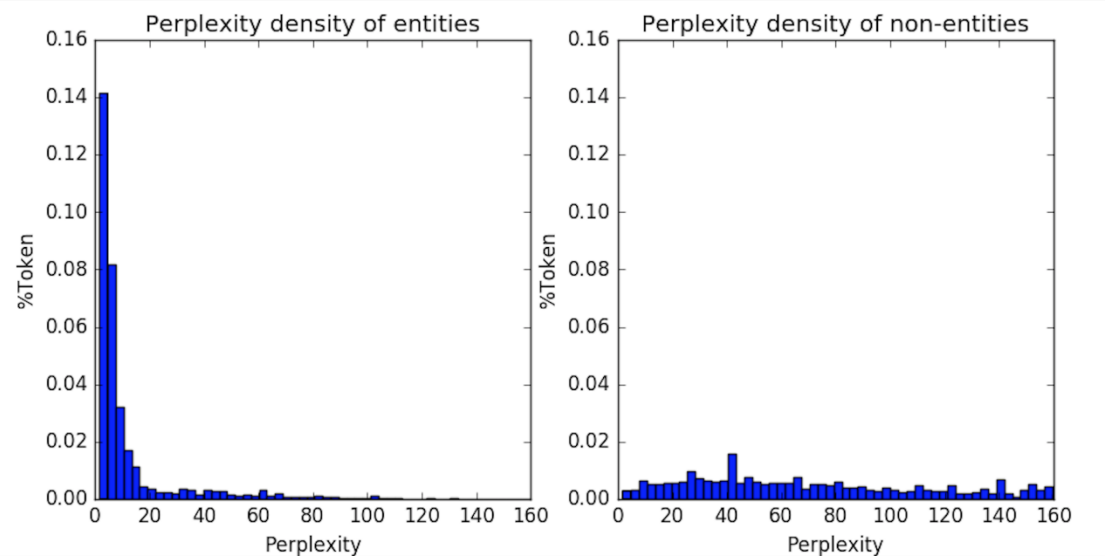
If you use LSTMs for the encoder, then you also use LSTMs for the decoder. From hidden state. You now predict the output word. This prediction takes the form of the probability distribution over entire output vocabulary. If you have a vocabulary of, say, 50,000 words, then the prediction is a 50,000 dimensional vector, each element corresponding to the probability predicted for one word in the vocabulary.

Q7. What is Character-Level models (CLM)?

Answer:

In English, there is strong empirical evidence that the character sequence that create up proper nouns tend to be distinctive. Even divorced of context,human reader can predict that “hoekstenberger” is an entity, but “abstractually” is not. Some NER research explores use of character-level features

including capitalization, prefixes and suffixes Cucerzan and Yarowsky; Ratinov and Roth (2009), and character-level models (CLMs) Klein et al. (2003) to improve the performance of NER, but to date there has been no systematic study isolating utility of CLMs in capturing the distinctions between name and non-name tokens in English or across other languages.



We conduct the experimental assessment of the discriminative power of CLMs for a range of languages: English, Arabic, Amharic, Bengali, Farsi, Hindi, Somali, and Tagalog. These languages use the variety of scripts and orthographic conventions (e.g, only three use capitalization), come from different language families, and vary in their morphological complexity. We represent the effectiveness of CLMs(character-level models) in distinguishing name tokens from non-name tokens, as illustrated by the above Figure, which shows confusion in histograms from a CLM trained on entity tokens. Our models use individual tokens, but perform extremely well in spite of taking no account of the word context.

We then assess the utility of directly adding simple features based on this CLM(character-level model) implementation to an existing NER system, and show that they have the significant positive impact on performance across many of the languages we tried. By adding very simple CLM-based features to the system, our scores approach those of a state-of-the-art(SOTA) NER system Lample et al. (2016) across multiple languages, representing both the unique importance and broad utility of this approach.

Q8.What is LexNLP package?

Answer:

Over the last 2 decades, many high-quality, open-source packages for natural language processing(NLP) and machine learning(ML) have been released. Developers and researchers can quickly write applications in languages such as Python, Java, and R that stand on shoulders of

comprehensive, well-tested libraries such as Stanford NLP (Manning et al. (2014)), OpenNLP (ApacheOpenNLP (2018)), NLTK (Bird et al. (2009)), spaCy (Honnibal and Montani (2017), scikit-learn library (Buitinck et al. (2013), Pedregosa et al. (2011)), and Gensim (Řehůřek and Sojka (2010)). Consequently, for most of the domains, rate of research has increased and cost of the application development has decreased.

For some specialized areas like marketing and medicines, there are focused libraries and organizations like BioMedICUS (Consortium (2018)), RadLex (Langlotz (2006)), and the Open Health Natural Language Processing(NLP) Consortium. Law, however, has received substantially less attention than others, despite its ubiquity, societal importance, and the specialized form. LexNLP is designed to fill this gap by providing both tools and data for developers and researchers to work with real legal and regulatory text, including statutes, regulations, the court opinions, briefs, contracts, and the other legal work products.

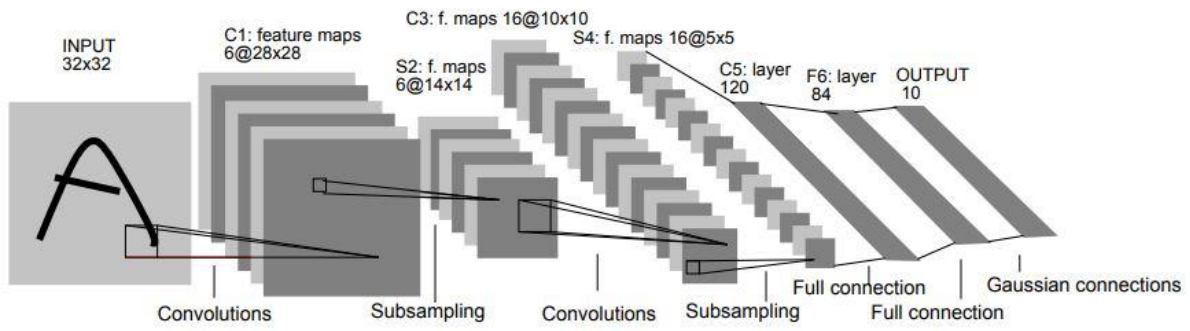
Law is the domain driven by language, logic, and the conceptual relationships, ripe for computation and analysis (Ruhl et al. (2017)). However, in our experience, natural language processing(NLP) and machine learning(ML) have not been applied as fruitfully or widely in legal as one might hope. We believe that the key impediment to academic and commercial application has been lack of tools that allow users to turn the real, unstructured legal document into structured data objects. The Goal of LexNLP is to make this task simple, whether for the analysis of statutes, regulations, court opinions, briefs or the migration of legacy contracts to smart contract or distributed ledger systems.

Q9.Explain The Architecture of LeNet-5.

Answer:

Yann LeCun, Leon Bottou, Yosuha Bengio and Patrick Haffner proposed the neural network architecture for the handwritten and machine-printed character recognition in the 1990's which they called them LeNet-5. The architecture is straightforward and too simple to understand that's why it is mostly used as a first step for teaching (CNN)Convolutional Neural Network.

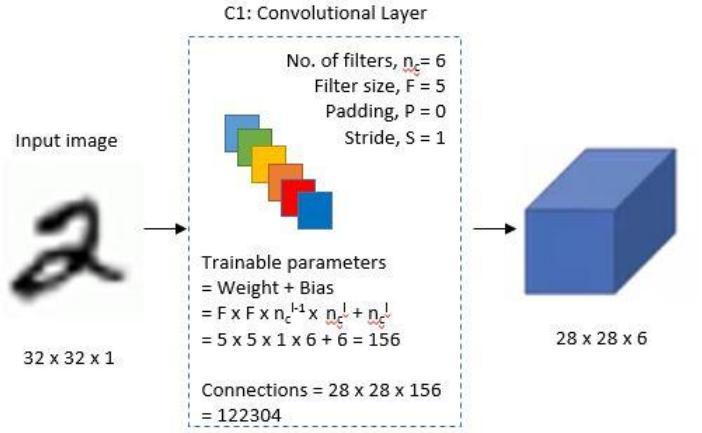
Architecture



This architecture consists of two sets of convolutional and average pooling layers, followed by the flattening convolutional layer, then 2 fully-connected layers and finally the softmax classifier.

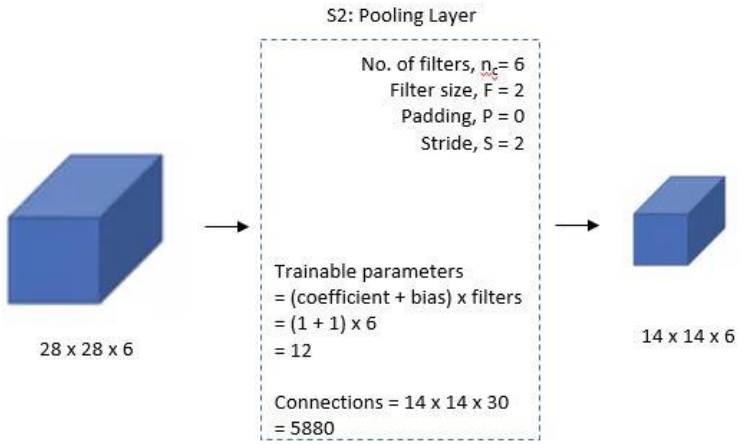
In the First Layer:

The input for LeNet-5 is the 32×32 grayscale image which passes through first convolutional layer with 6 feature maps or filters having size 5×5 and the stride of one. Image dimensions changes from 32x32x1 to 28x28x6.



In Second Layer:

Then it applies average pooling layer or sub-sampling layer with the filter size 2×2 and stride of two. The resulting image dimension will be reduced to 14x14x6.



Third Layer:

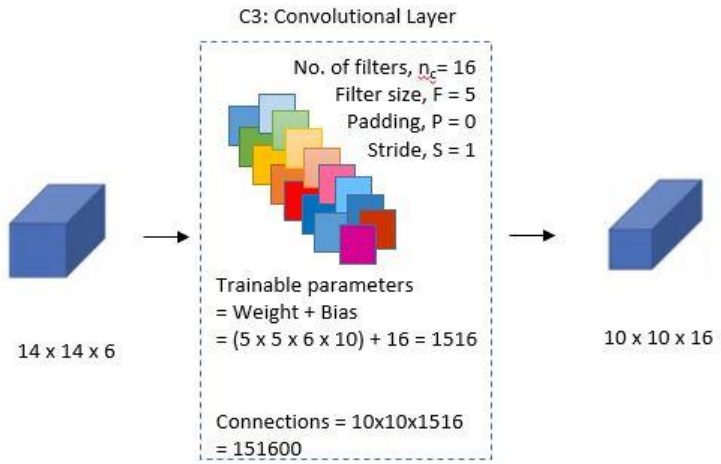
Next, there is the second convolutional layer with 16 feature maps having size 5×5 and the stride of 1. In this layer, only ten out of sixteen feature maps are connected to 6 feature maps of previous layer as shown below.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

TABLE I

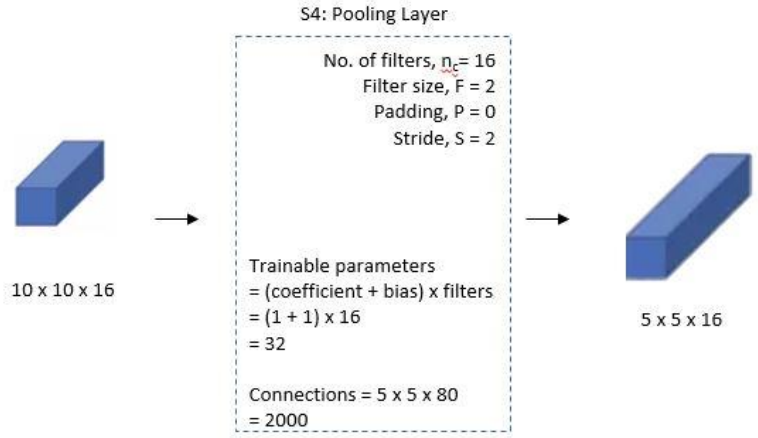
EACH COLUMN INDICATES WHICH FEATURE MAP IN S2 ARE COMBINED BY THE UNITS IN A PARTICULAR FEATURE MAP OF C3.

The main reason is to break symmetry in the network and keeps a number of connections within reasonable bounds. That is why the number of training parameters in this layers are 1516 instead of 2400 and similarly, number of connections are 151600 instead of 240000.



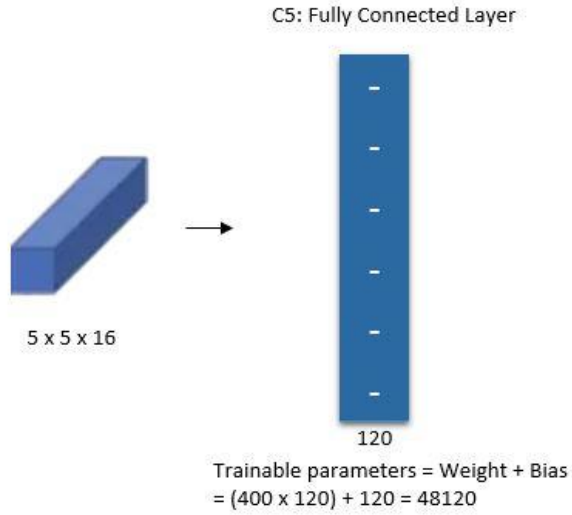
Fourth Layer:

In the fourth layer (S4) is an average pooling layer with filter size 2×2 and stride of 2. This layer is same as second layer (S2) except it has 16 feature maps so output will be reduced to 5x5x16.



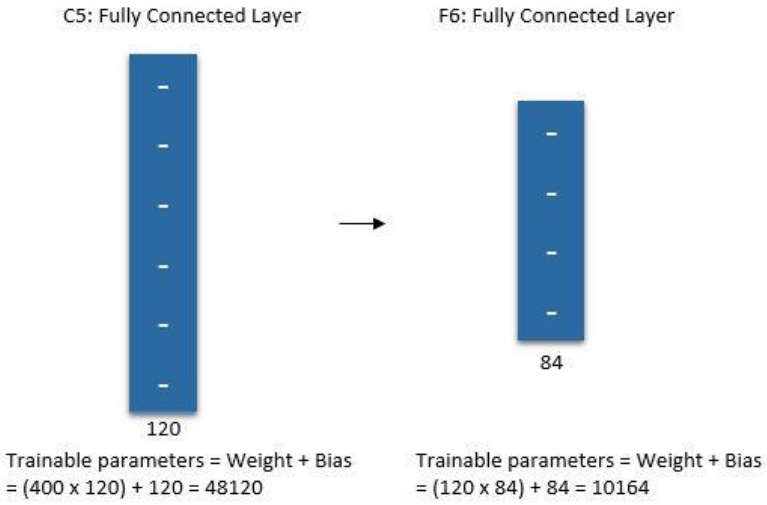
Fifth Layer:

The fifth layer (C5) is the fully connected convolutional layer with 120 feature maps each of the size 1×1. Each of 120 units in C5 is connected to all the 400 nodes (5x5x16) in the fourth layer S4.



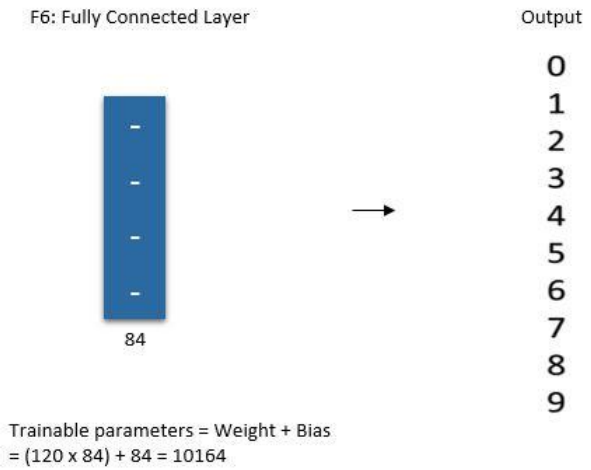
Sixth Layer:

The sixth layer is also fully connected layer (F6) with 84 units.



Output Layer:

Finally, there is fully connected softmax output layer $\hat{y}$ with 10 possible values corresponding to digits from 0 to 9.



DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

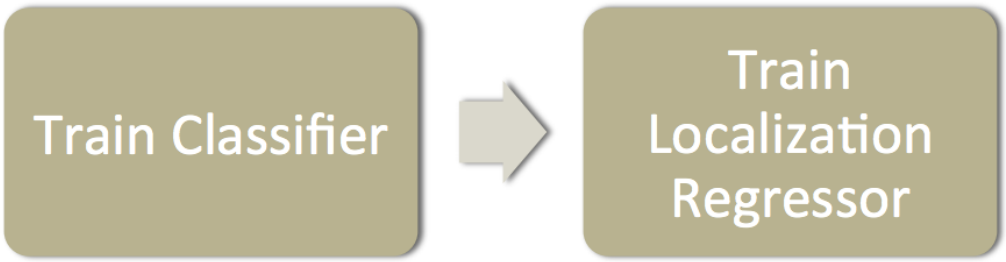
DAY 23

Q1.Explain Overfeat in Object detection.

Answer:

Overfeat: It is a typical model of integrating object detection, localization, and classification tasks whole into one convolutional neural network(CNN). The main idea is to do image classification at different locations on regions of multiple scales of the image in a sliding window fashion, and second, predict bounding box locations with the regressor trained on top of the same convolution layers.

This model architecture is too similar to AlexNet. This model is trained as follows:



1. Train a CNN model (identical to AlexNet) on image classification tasks.
2. Then, we replace top classifier layers by the regression network and trained it to predict object bounding boxes at each spatial location and scale. Regressor is class-specific, each generated for one class image.

- Input: Images with classification and bounding box.
- Output: (xleft,xright,ytop,ybottom)(xleft,xright,ytop,ybottom), 4 values in total, representing the coordinates of the bounding box edges.
- Loss: The regressor is trained to minimize l2 norm between the generated bounding box and truth for each training example.

At the detection time,

1. It Performs classification at each location using the pretrained CNN model.
2. It Predicts object bounding boxes on all classified regions generated by the classifier.
3. Merge bounding boxes with sufficient overlap from localization and sufficient confidence of being the same object from the classifier.

Q2. What is Multipath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction?

Answer:

In this paper, we focus on problem of predicting future agent states, which is the crucial task for robot planning in real-world environments. We are specifically interested in addressing this problem for self-driving vehicles, application with a potentially enormous societal impact. Mainly, predicting the future of other agents in this domain is vital for safe, comfortable, and efficient operation. E.g., it is important to know whether to yield to the vehicle if they are going to cut in front of our robot or when would be the best time to add into traffic. Such future prediction requires an understanding of a static and dynamic world context: road semantics (*like* lane connectivity, stop lines), traffic light informations, and past observations of other agents, as in below Fig.

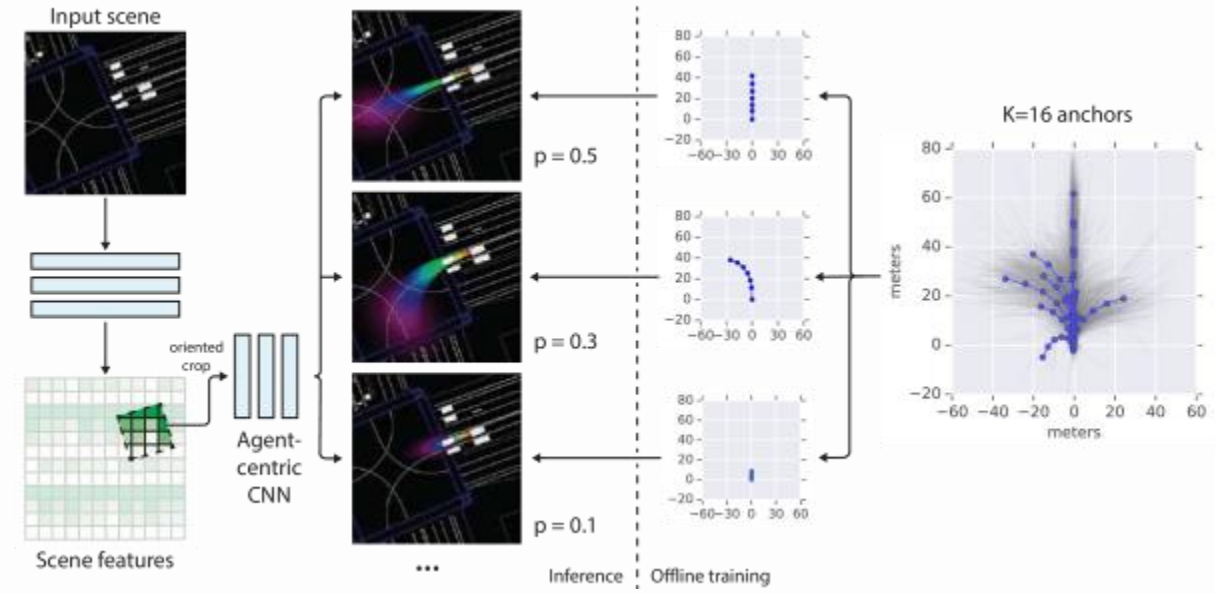
A fundamental aspect of the future state prediction is that it is inherently *stochastic*, as agents can't know each other's motivations. When we are driving, we can never really be sure what other drivers will do next, and it is essential to consider multiple outcomes and their likelihood.

We seek the model of the future that can provide both (i) a weighted, parsimonious set of discrete trajectories that covers space of likely outcomes and (ii) a closed-form evaluation of the likelihood of any trajectory. These two attributes enable efficient reasoning in relevant planning use-cases, e.g., human-like reactions to discrete trajectory hypotheses (*e.g.*, yielding, following), and probabilistic queries such as the expected risk of collision in a space-time region.

This model addresses these issues with critical insight: it employs a fixed set of *trajectory anchors* as the basis of our modeling. This lets us factor stochastic uncertainty hierarchically: First, *intent uncertainty* captures the uncertainty of *what* an agent intends to do and is encoded as a distribution over the set of anchor trajectories. Second, given an intent, *control uncertainty* represents our uncertainty over *how* they might achieve it. We assume control uncertainty is normally distributed at each future time step [Thrun05], parameterized such that the mean corresponds to a context-specific offset from the anchor state, with the associated covariance capturing the unimodal aleatoric uncertainty [Kendall17]. In Fig. Illustrates a typical scenario where there are three likely intents given the scene context, with control mean offset refinements respecting road geometry, and control uncertainty intuitively growing over time.

Our trajectory anchors are modes found in our training data in state-sequence space via unsupervised learning. These anchors provide templates for coarse-granularity futures for an agent and might correspond to semantic concepts like "change lanes," or "slow down" (although to be clear, we don't use any semantic concepts in our modeling).

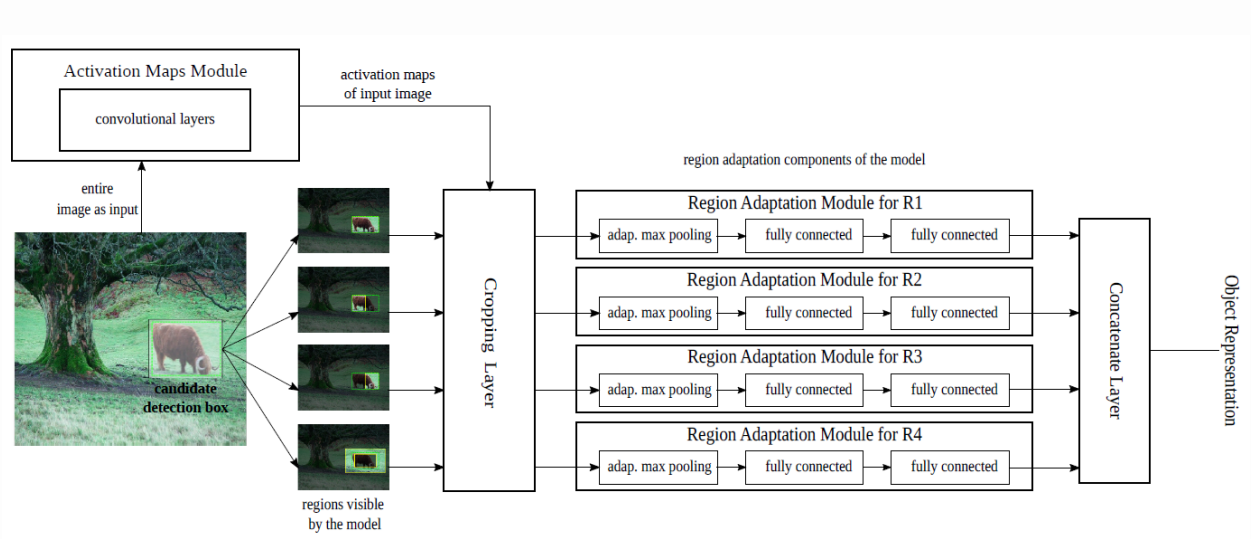
Our complete model predicts a Gaussian mixture model (GMM) at each time step, with the mixture weights (intent distribution) fixed over time. Given such a parametric distribution model, we can directly evaluate the likelihood of any future trajectory and have a simple way to obtain a compact, diverse weighted set of trajectory samples: the MAP sample from each anchor-intent.



Q3. An Object detection approach using MR-CNN

Answer:

Multi-Region CNN (MR-CNN): Object representation using multiple regions to capture several different aspects of one object.

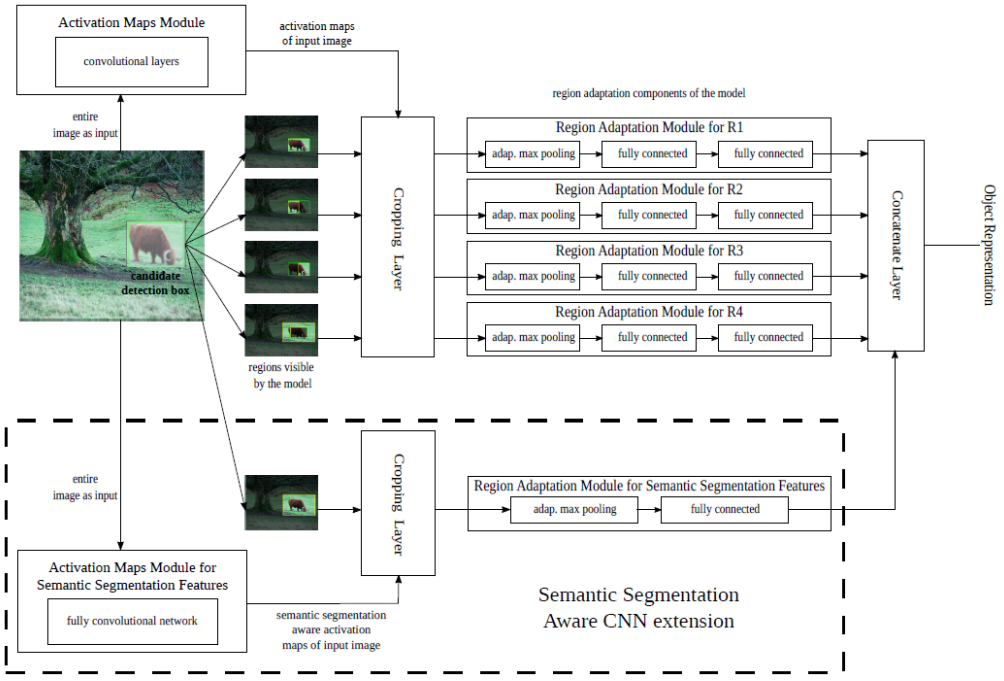


Network Architecture of MR-CNN

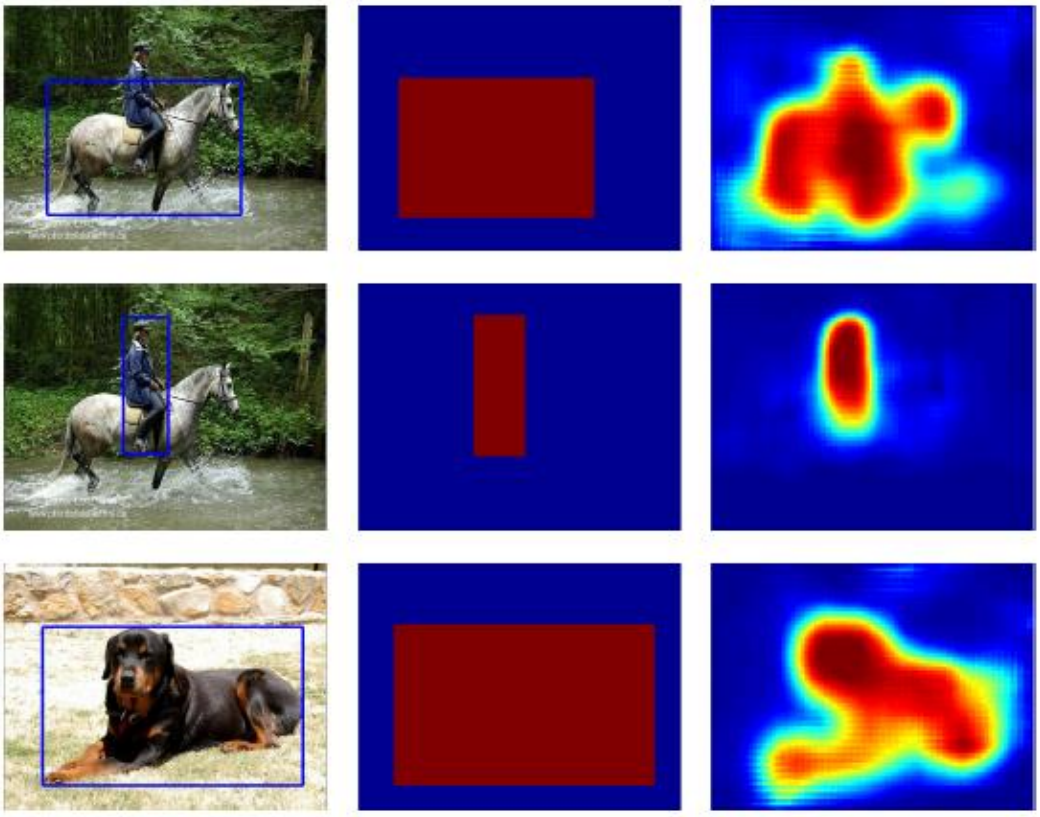
- First, the input image goes through Activation Maps Module, as shown above, and outputs the activation map.
- **Bounding box** or Region proposals candidates are generated using Selective Search.
- For each bounding box candidate *B*, a set of regions {*Ri*}, with *i*=1 to *k,* are generated, that is why it is known as multi-region. More details about the choices of multiple areas are described in next sub-section.
- ROI pooling is performed for each region *Ri*,cropped or pooled area goes through fully connected (FC) layers, at each Region Adaptation Module.
- Finally, the output from all FC layers are added together to form a 1D feature vector, which is an object representation of the bounding box *B*.
- Here, VGG-16 ImageNet pre-trained model is used. The max-pooling layer after the last conv layer is removed.

Q4. Object detection using Segmentation-aware CNN

Answer:



- There are close connections between segmentation and detection. And segmentation related ques are empirically known to help object detection often.
- Two modules are added: **Activation maps module for semantic segmentation-aware features**, and **regions adaptation module for grammarly segmentation-aware feature.**
- There is no additional annotation used for training here.
- FCN is used for an activation map module.
- The last FC7 layer channels number is changed from 4096 to 512.



- **The weakly supervised training** strategy is used. **Artificial foreground class-specific segmentation mask is created using bounding box annotations**.

- More particularly, the ground truth bounding boxes of an image are projected on the spatial domain of the last hidden layer of the FCN, and the ”pixels” that lay inside the projected boxes are labelled as foreground while the rest are labelled as background.

- After training the FCN using the mask, the last classification layer is dropped. Only the rest of FCN is used.

- Though it is weakly supervised training, the foreground probabilities shown as above still carry some information, as shown above.

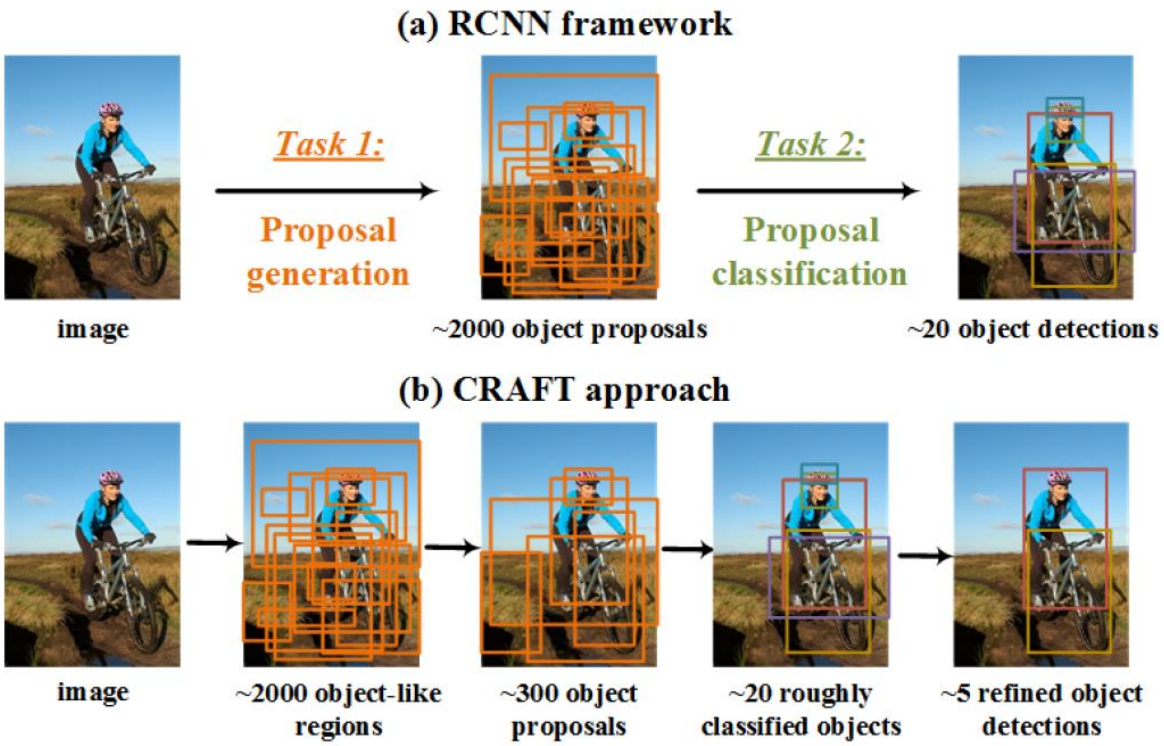
- The bounding box used is 1.5× larger than the original bounding box.

Q5. What is CRAFT (Object detection)?

Answer:

CRAFT stands for Cascade **R**egion-proposal-network **A**nd **F**as**T** R-CNN. It is reviewed by the Chinese Academy of Sciences **and** Tsinghua University. In Faster R-CNN, region proposal network is used to generate proposals. These proposals, after ROI pooling, are going through network for classification. However, CRAFT is found that there is a core problem in Faster R-CNN:

- In proposal generation, there is still a large proportion of background regions. The existence of many background sample causes many false positives.



In CRAFT(**C**ascade **R**egion-proposal-network**),** as shown above, another CNN(Convolutional neural network) is added after RPN to generate fewer proposals (i.e., 300 here)**.** Then, classification is performed on 300 proposals and outputs about 20 first detection results. For each primitive result**,** refined object detection is performed using one-vs-rest classification.

Cascade Proposal Generation

Baseline RPN

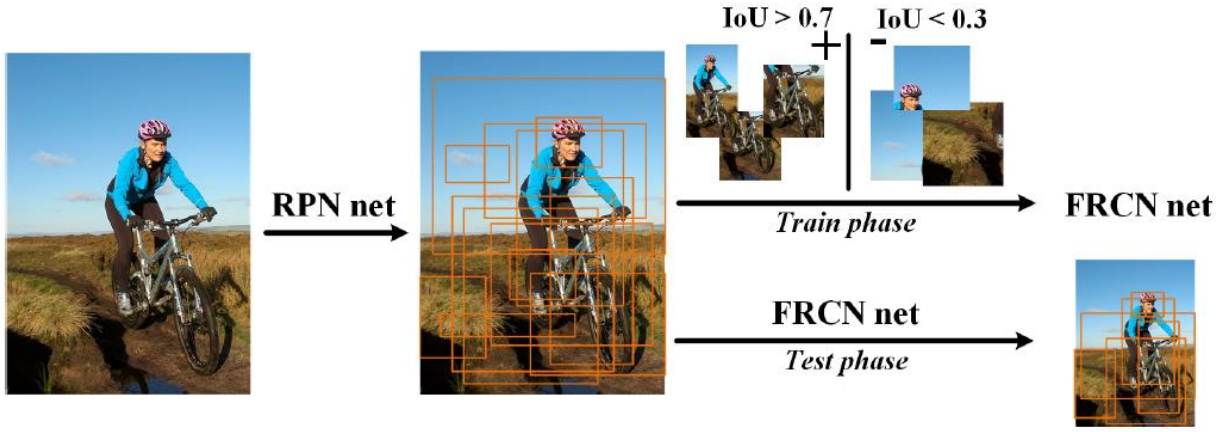
- An ideal proposal generator should generate as few proposal as possible while covering almost all object instances. Due to resolution loss caused by CNN pooling operation and the fixed aspect ratio of the sliding window, RPN is weak at covering objects with extreme shapes or scales.

aero	bike	**bird**	**boat**	**bottle**
95.44	98.81	**93.90**	**92.78**	**80.38**
bus	car	cat	**chair**	cow
98.12	96.00	99.16	**91.80**	99.18
table	dog	horse	mbike	persn
95.15	99.59	97.70	96.31	95.49
plant	sheep	sofa	train	**tv**
86.87	98.76	98.74	97.52	**90.58**

Recall Rates (is in %), Overall is 94.87%, lower than 94.87% is bold in the text.

- The above results are baseline RPN based on VGG_M trained using PASCAL VOC 2007 train+val, and tested on the test set.
- The recall rate on each object category varies a lot. Object with extreme aspect ratio and scale are hard to be detected, such as boat and bottle.

Proposed Cascade Structure



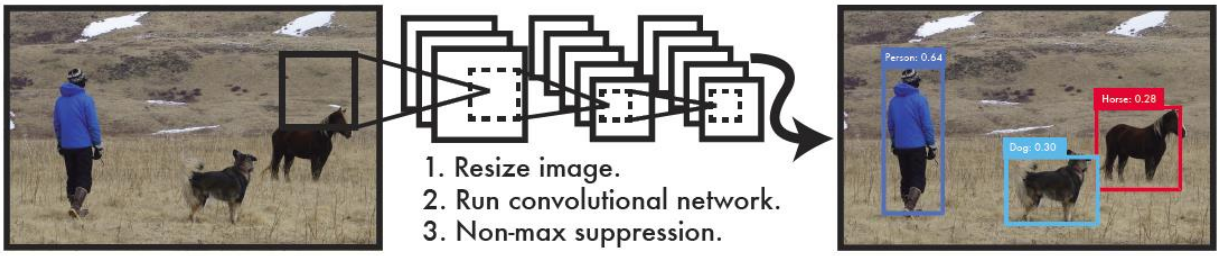
The concatenation classification network after RPN is denoted as FRCN Net here

- An additional classification network that comes after RPN.
- The additional network is the 2- class detection network denoted as FRCN net in above figure. It uses output of RPN as training data.
- After RPN net is trained, the 2000 first proposals of each training image are used as training data for the FRCN net.
- During training, +ve and -ve sampling are based on 0.7 IoU for negatives and below 0.3 IoU for negatives, respectively.
- **There are 2 advantages:**
- 1) First, additional FRCN net further **improves quality of the object proposals** and **shrinks more background regions**, making proposals fit better with task requirement.
- 2) Second, **proposals from multiple sources can be merged** as the input of the FRCN net so that complementary information can be used.

Q6. Explain YOLOv1 for Object Detection.

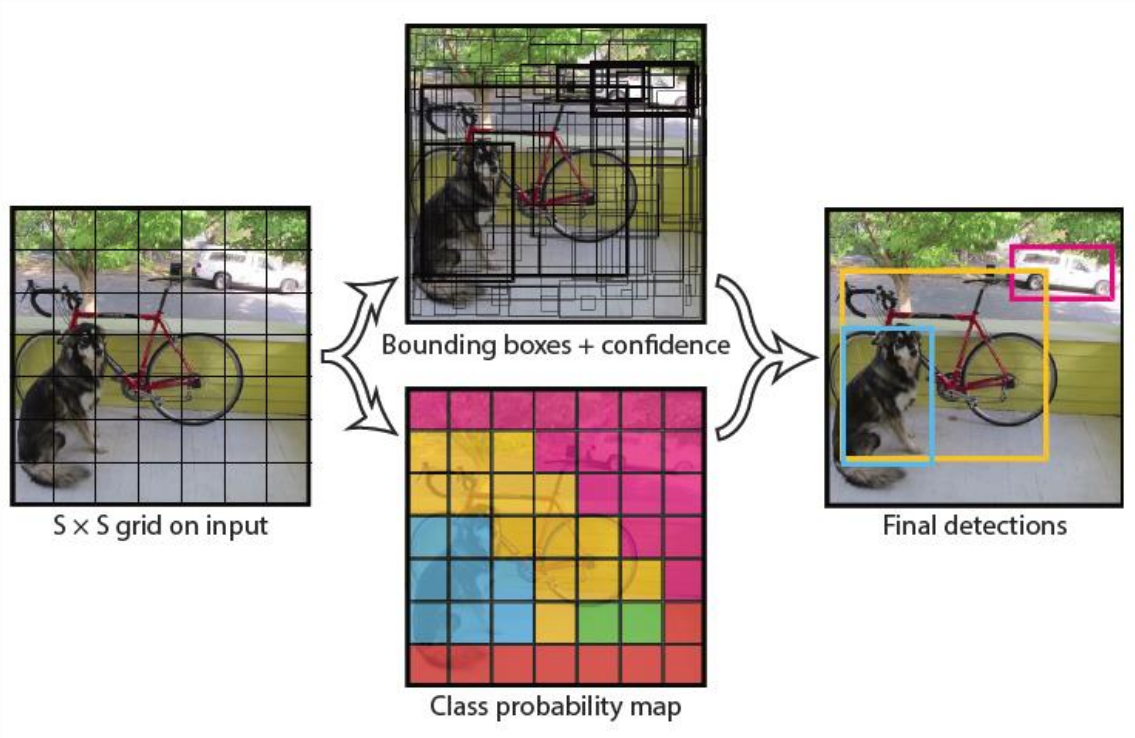
Answer:

YOLOv1 stands for You Look Only Once, it is reviewed by FAIR (Facebook AI Research). The network only looks at the image once to detect multiple objects.



By just looking image once, the detection speed is in real-time (45 fps). Fast YOLOv1 achieves 155 fps.

YOLO suggests having a unified network to perform all at once. Also, an end-to-end training network can be achieved.



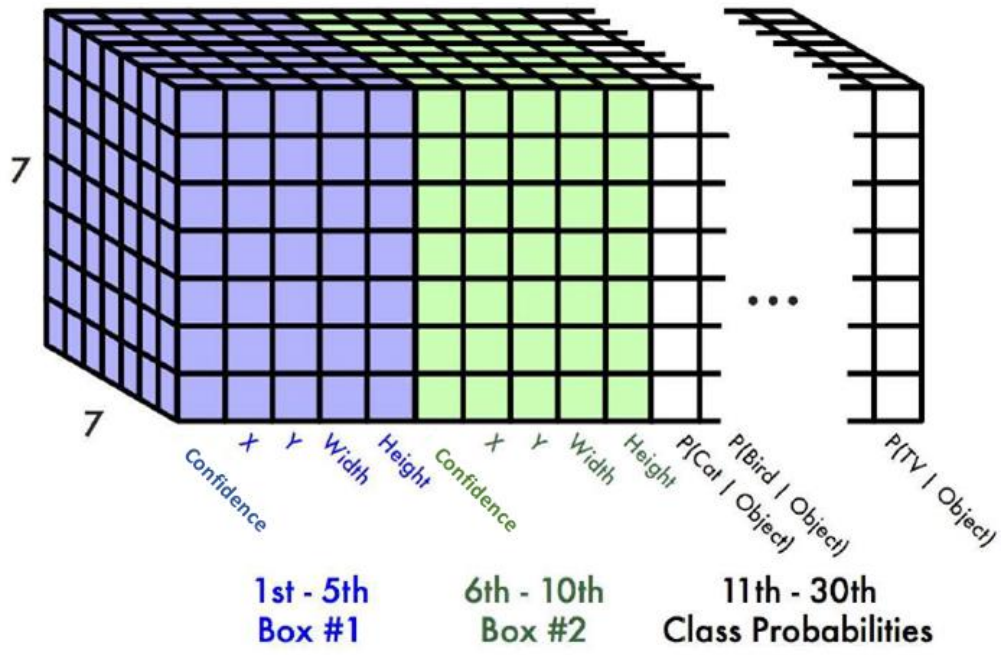
The input image is divided into the S×S grid (S=7). If the center of the object falls into the grid cell, that grid cell is responsible for detecting that object.

Each grid cell predict B bounding boxes (B=2) and confidence scores for those boxes. These confidence score reflect how confident model is that the box contains an object, i.e., any objects in the box, P(Objects).

Each bounding box consists of five predictions: x, y, w, h, and confidence.

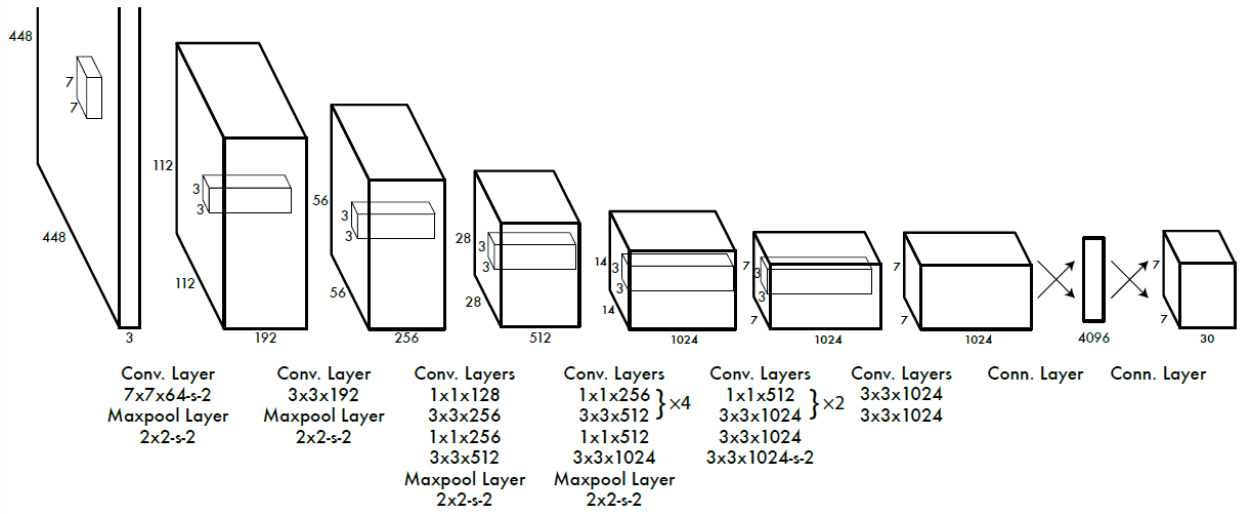
- The (x, y) coordinates represent center of the box relative to the bound of the grid cell.
- The height h and width w are predicted relative to whole image.
- The confidence represents the IOU (Intersection Over Union) between the predicted box and any ground truth box.

Each grid cell also predicts conditional class probabilities, P(Class|Object). (Total number of classes=20)



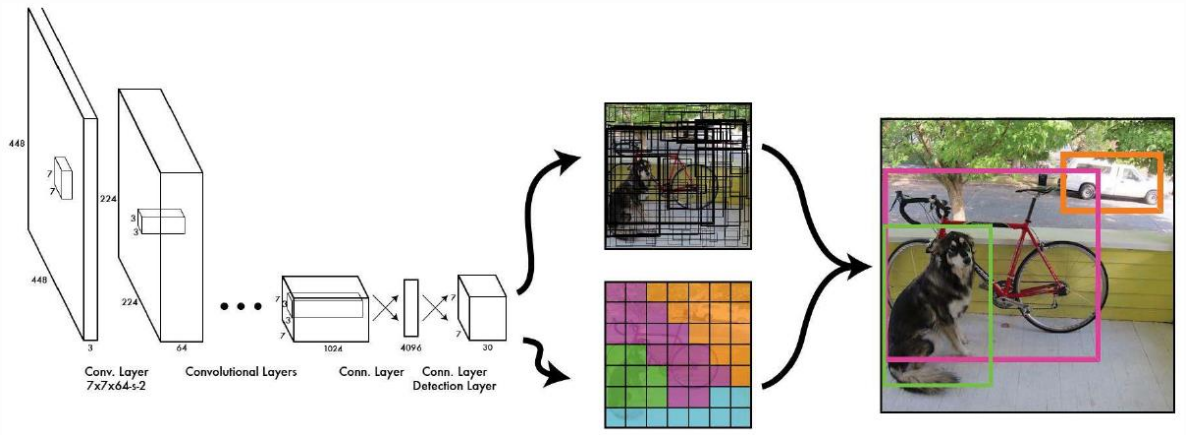
The output size becomes: 7×7×(2×5+20)=1470

Network Architecture of YOLOv1



The model consists of 24 convolutional layers, followed by two fully connected layers. Alternating 1×1 convolutional layers reduce features space from preceding layers. (1×1)Conv has been used in GoogLeNet for reducing the number of parameters.)

Fast YOLO fewer convolutional layers (9 instead of 24) and fewer filters in those layers. The network pipeline is summarized like below:



Therefore, we can see that the input image goes through network once and then objects can be detected. And we can have end-to-end learning.

Q7. Adversarial Examples Improve Image Recognition

Answer:

Adversarial examples crafted by adding imperceptible perturbations to images can lead to (ConvNets)Convolutional Neural Networks to make wrong predictions. The existence of adversarial examples not only reveal limited generalization ability of ConvNets, but also poses security threats on the real-world deployment of these models. Since the first discovery of the vulnerability of ConvNets to adversarial attacks, many efforts have been made to improve network robustness.

Above Fig.: AdvProp improves image recognition. By training model on ImageNet, AdvProp helps EfficientNet-B7 to achieve 85.2% accuracy on ImageNet, 52.9% mCE (mean corruption error, lower is better) on ImageNet-C, 44.7% accuracy on ImageNet-A and 26.6% accuracy on Stylized-ImageNet, beating its vanilla counterpart by 0.7%, 6.5%, 7.0% and 4.8%, respectively. Theses sample images are randomly selected from category "goldfinch."

In this paper, rather than focusing on defending against adversarial examples, we shift our attention to leveraging adversarial examples to improve accuracy. Previous works show that training with adversarial examples can enhance model generalization but are restricted to certain situations—the improvement is only observed either on small datasets (*e.g.*, MNIST) in the fully-supervised setting [5], or on larger datasets but in the semi-supervised setting [21, 22]. Meanwhile, recent works [15, 13, 31] also suggest that training with adversarial examples on large datasets, *e.g.*, ImageNet [23], with supervised learning results in performance degradation on clean images. To summarize, it remains an open question of how adversarial examples can be used effectively to help vision models.

We observe all previous methods jointly train over clean images and adversarial examples without distinction, even though they should be drawn from different underlying distributions. We hypothesize this distribution mismatch between fresh examples and adversarial examples is a key factor that causes performance degradation in previous works.

Q8. Advancing NLP with Cognitive Language Processing Signals

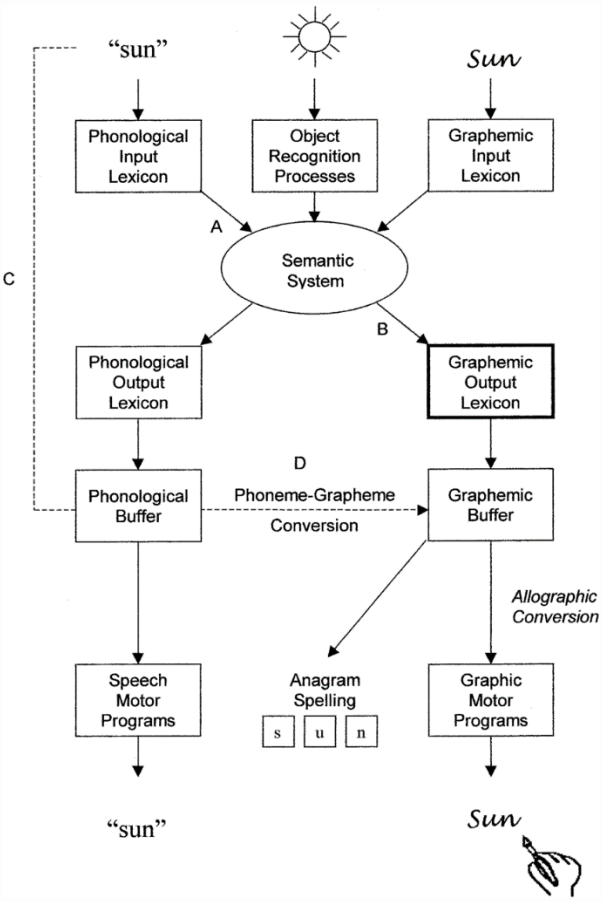
Answer:

When reading, humans process language "automatically" without reflecting on each step — Humans string words together into sentences, understand the meaning of spoken and written ideas, and process language without overthinking about how the underlying cognitive process happens. This process generates cognitive signals that could potentially facilitate natural language processing tasks.

In recent years, collecting these signals has become increasingly accessible and less expensive Papoutsaki et al. (2016); as a result, using cognitive features to improve NLP tasks has become more popular. For example, researchers have proposed a range of work that uses eye-tracking or gaze signals to improve part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2017), named entity recognition Hollenstein and Zhang (2019), among other tasks. Moreover, these signals have been used successfully to regularize attention in neural networks for NLP Barrett et al. (2018).

However, most previous work leverages only eye-tracking data, presumably because it is the most accessible form of cognitive language processing signal. Also, most state-of-the-artwork(SOTA) focused on improving a single task with a single type of cognitive signal. But can cognitive processing signals bring consistent improvements across modality (e.g., eye-tracking and EEG) and across various NLP

tasks? And if so, does the combination of different sources of cognitive signals bring incremental improvements?



Q8. Do you have any idea how can we use NLP on News headlines to predict index trends?

Answer:

Traders generally look up information about the company they are looking to buy shares into, for long and short trading. A frequent source of information is news media, which provides updates about the company's activities, such as expansion, better or worse revenues than expected, new products and much more. Depending on the news, trader can determine a bearish or bullish trend and decide to invests in it.

We may be able to correlate overall public sentiments towards as the company and its stock price: Apple is generally well-liked by the public, receives daily news coverage of its new product and financial stability, and its stock has been growing steadily. These facts may be correlated but first may not cause the second; we will analyze if news coverage can be used to predict the market trend. To do so, we will examine the top 25 news headlines of each open-market day from 2008 to late 2015 and try to predict

the end-of-day value of DJIA index for the same day. The theory behind predicting same day value is that traders will respond to news quickly and thus, the market will adjust within an hour of release. Therefore in the single business day, if the news is spread during business hours, its effect may be measured before closing bell of the market.

The motivation behind this analysis is that humans take decision using most of the available information. This usually takes several minutes to discover new information and take the decision. An algorithm is capable of processing gigabytes of texts from multi-source streams in second. We could potentially exploit this difference in order to create a trading strategy.

NLP (Natural Language Processing) techniques can be used to extract different information from headlines such as sentiment, subjectivity, context and named entities. We obtain an indicator vector using each of these techniques, which allows us to train different algorithms to predict a trend. To predict these values, we can use several methods that should be well suited for this type of information: Linear regression, Support Vector Machine(SVM), Long Short-Term Memory(LSTM) recurrent neural network, and a dense feed-forward (MLP) neural network. We included techniques used by Bollen et al (2010), which resulted in state-of-the-art(SOTA) results. We will also analyze the method used in other studies with a similar context

Information in headlines

Latent Sentiment Analysis is done by building up a corpus of labeled words which usually connote a degree of +ve or -ve sentiment. We can extend the corpus to include emoticons (i.e. ":-)") and expression, which often correlates to strong emotions. Naive sentiment analysis consists of a lookup of each word in sentence to be analyzed and evaluation of a score for sentence overall. This approach is limited by its known vocabulary, which can be mitigated by context analysis and introduction of synonyms.Second limitation is sarcasm, which is prevalent in twitter feed analysis. The sentiment inferred by words is opposed to the sentiment assumed by the user. This is mitigated by technique detecting sarcasm which

lead to a polarity flip of such tweets.

Sentiment analysis gives insight on how favorable the media is and maybe the bias traders may have towards buying or selling.

Another NLP technique which gave promising results was context analysis. This is a recent deep learning approach where you rely on a large corpus of text in order to learn and predict the words around a target. You can then deduce in what context it usually appears. The result is a vector representing each word. Other vectors with little distance are usually synonyms. The representation also allows us to do algebra, such as the famous "king - man + woman = queen"

Learning this representation offers the possibility of associating a specific context with a bullish or bearish market.

iNeuron

DATA SCIENCE INTERVIEW PREPARATION

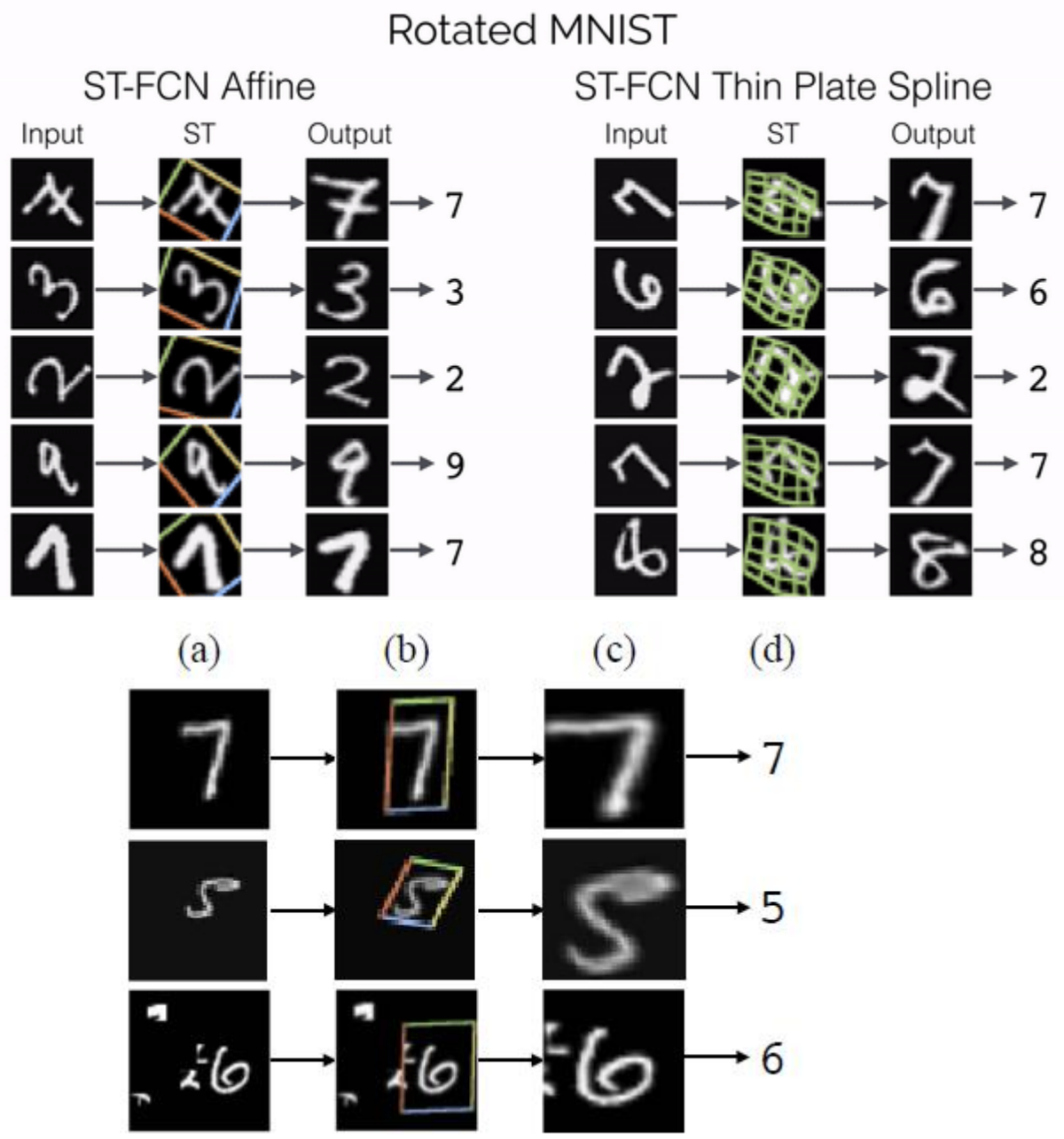
(30 Days of Interview Preparation)

Day24

Q1.What is STN?

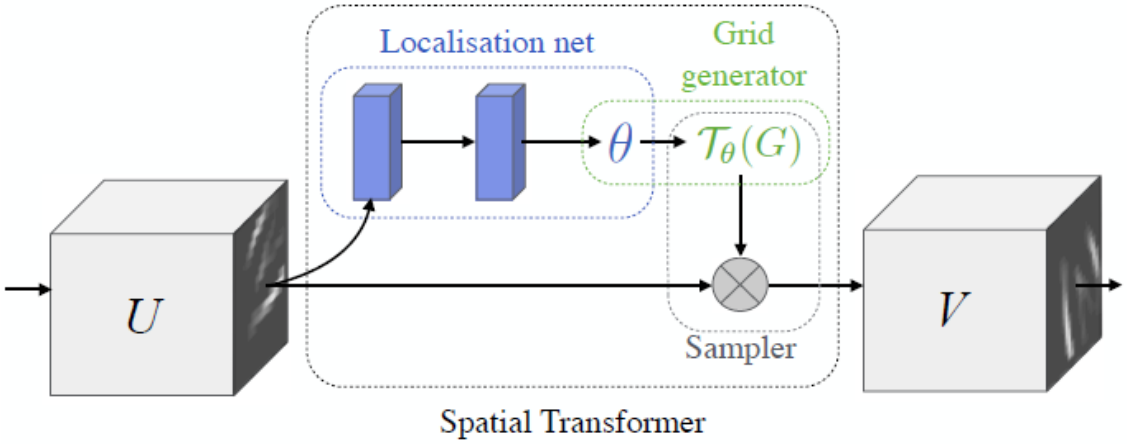
Answer:

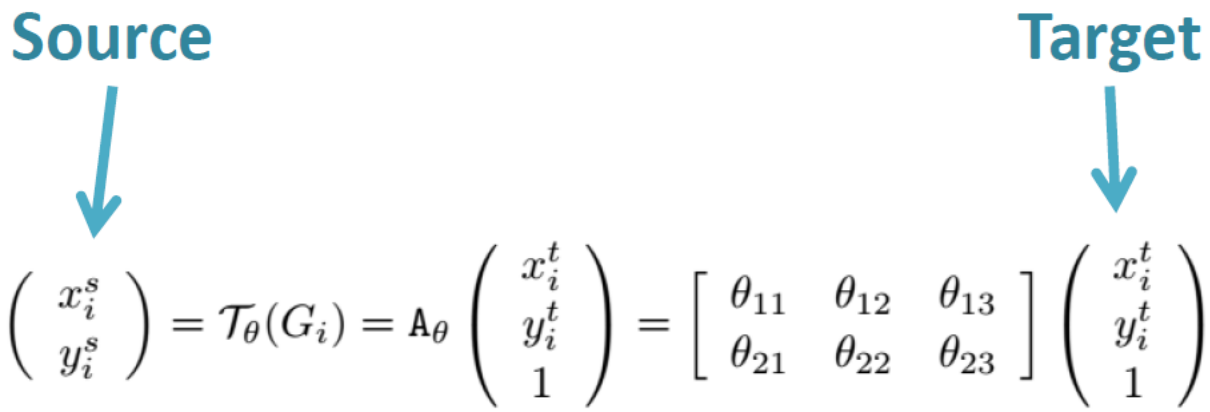
STN stands for Spatial Transformer Network for image classification. Google Deepmind briefly reviews it. STN helps to crop out and scale-normalizes appropriate region, which can simplify the subsequent classification task and lead to better classification performance as below:



(a) Input Image with Random Translation, Scale, Rotation, and Clutter, (b) STN Applied to Input Image, (c) Output of STN, (d) Classification Prediction

Spatial Transformer Network (STN)





- STN is composed of **Localisation Net**, **Grid Generator,** and **Sampler**.

Localization Net

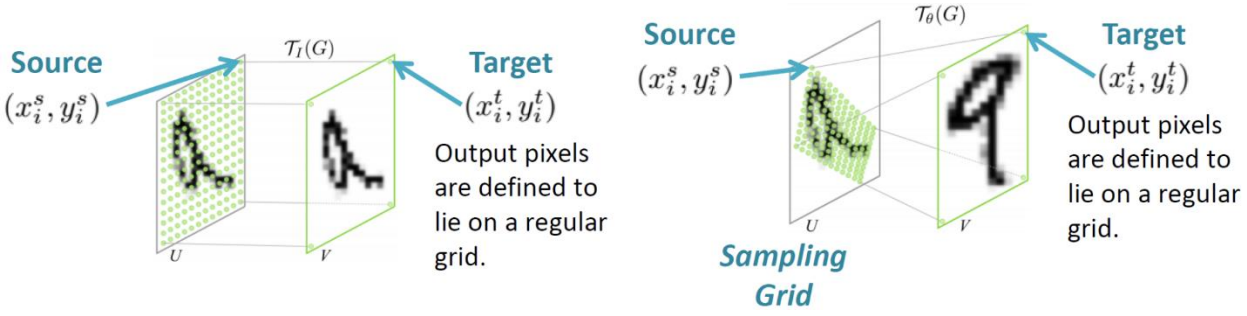
- With **input feature map** U, with (width) W, (height) H, and C channels, **outputs are** ϑ, parameters of transformation $T\vartheta$. It can be learned as affine transform as above. Or to be more constrained, such as the used for attention which only contains scaling and translation as below:

$$\mathbf{A}_\theta = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix}$$

Grid Generator

- Suppose we have a regular grid G, this G is a set of points with **source coordinates (xs_i, ys_i)**, which act as **input**.
- Then we **apply transformation T_ϑ on G**, i.e., $T_\vartheta(G)$.
- After $T_\vartheta(G)$, a set of points with **destination coordinates (xt_i, yt_i) is outputted**. These points have been altered based on the transformation parameters. It can be Translation, Scale, Rotation or More Generic Warping depending on how we set ϑ as mentioned above.

Sampler



- **Based on the new set of coordinates (xt_i, yt_i), we generate a transformed output feature map *V***. This *V* is translated, scaled, rotated, warped, projective transformed or affined, whatever.
- It is noted that STN can be applied to not only input image but also intermediate feature maps.

Q2.What is decaNLP?

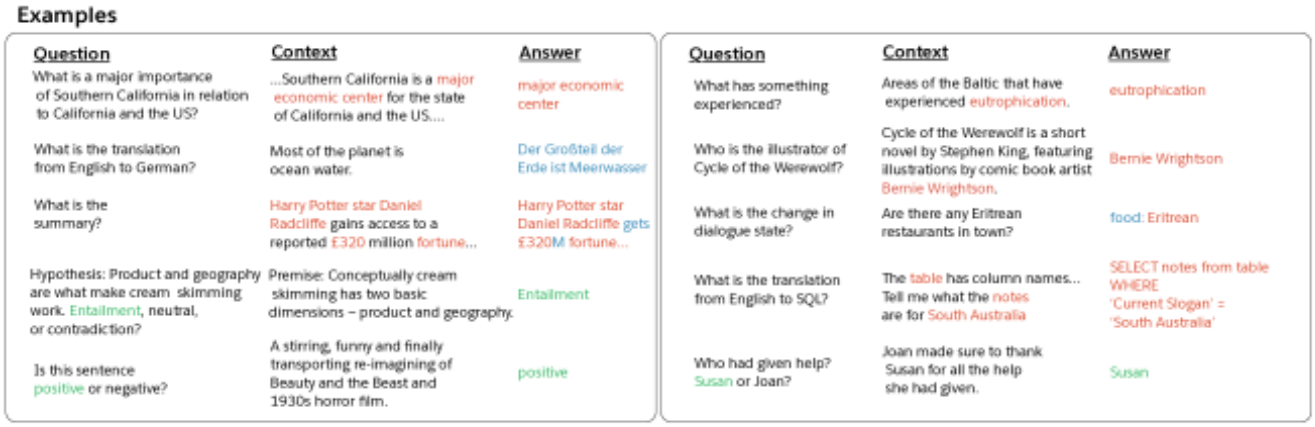
Answer:

We introduced the Natural Language Decathlon (decaNLP) to explore models that generalize to many different kinds of Natural Language Processing(NLP) tasks. decaNLP encourages single model to

simultaneously optimize for 10 tasks: question answering, machine translation, document summarization, semantic parsing, sentiment analysis, natural language inference(NLI), semantic role labeling, relation extraction, goal-oriented dialogue, and pronoun resolution.

We frame all the tasks as question answering [Kumar et al., 2016] by allowing task specification to take the form of a natural language question q: all inputs have a context, question, and answer (Fig. 1). Traditionally, NLP examples have inputs x and output y, and the underlying task t is provided through explicit modeling constraints. Meta-learning approaches include t as additional input. Our approach does not use the single representation for any t but instead uses natural language questions that describe underlying tasks. This allows single models to multitask effectively and makes them more suitable as pre-trained models for transfer learning and meta-learning: natural language questions allow a model to generalize to entirely new tasks through different but related task descriptions.

The MQAN (multitask question answering network) is designed for decaNLP and makes use of a novel dual attention and multi-pointer-generator decoder to multitask across all tasks in decaNLP. Our results represent that training the MQAN jointly on all tasks with the right anti-curriculum strategy can achieve performance comparable to that of ten separate MQANs, each trained separately. An MQAN pretrained on decaNLP shows improvements in transfer learning for machine translation and named entity recognition(NER), domain adaptation for sentiment analysis and natural language inference(NLI), and zero-shot capabilities for text classification. Though not explicitly designed for any one job, MQAN proves to be a robust model in a single-task setting as well, achieving state-of-the-art results on the semantic parsing component of decaNLP.



In the above figure: Overview of the decaNLP dataset with one example from each decaNLP task in the order presented in Section 2. They show how the datasets were pre-processed to become question

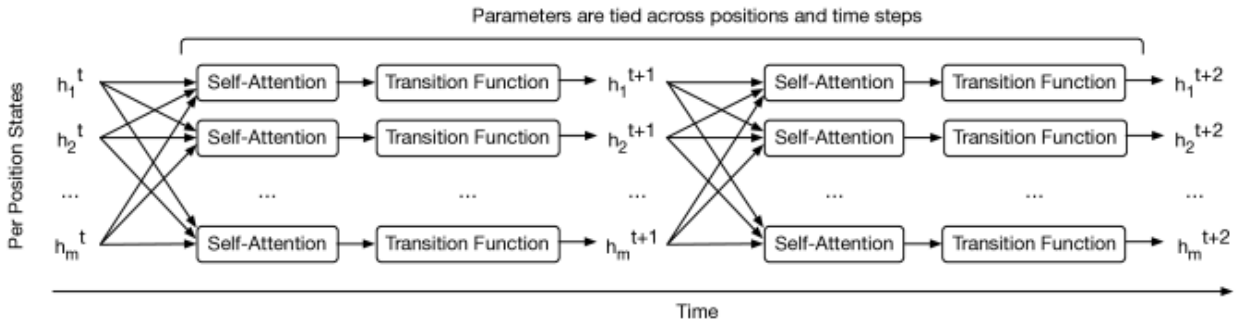
answering problems. Answer words in red are generated by pointing to the context, in green from the issue, and in blue if they are made from a classifier over the output vocabulary.

Q3.Universal Transformers

Answer:

Convolutional and fully-attentional feed-forward architectures such as the Transformer model have recently emerged as viable alternatives to RNNs(Recurrent neural networks) (for the range of sequence modeling tasks, notably machine translation (JonasFaceNet2017; transformer,). These architectures address the significant shortcoming of RNNs, namely their inherently sequential computation, which prevents parallelization across elements of input sequence while still addressing vanishing gradients problem (vanishing-exploding-gradient). Transformer model, in particular, achieves this by relying entirely on the self-attention mechanism (decomposableAttnModel; lin2017structured) to compute series of context-informed vector-space representations of symbols in its input and output, which are then used to predict distributions over subsequent symbols as the model predicts output sequence symbol-by-symbol. Not only in this mechanism straightforward to parallelize, but as each symbol's representation is also directly informed by all other symbols representations, this results in an active global receptive field. This stands, in contrast, to, e.g., convolutional architecture, which typically has limited receptive field.

Notably, however, Transformer foregoes the (Recurrent Neural Network)RNN's inductive bias towards learning recursive or iterative transformations. Our experiments indicate that this inductive bias may be important for several algorithmic and language understanding tasks of varying complexity: in contrast to models such as the Neural Turing Machine, the Neural GPU, or Stack RNNs, the Transformer does not generalize well to input lengths not encountered during training.



In this paper, we propose a *Universal Transformer*. It combines the parallelizability and global receptive field of a Transformer model with the recurrent inductive bias of RNNs, which seems to be better suited to range of algorithmic and natural language understanding(NLU) sequence-to-sequence problems. As the name implies, in contrast to standard Transformer, under certain assumptions, a Universal Transformer can be shown to be computationally universal.

In each step, the Universal Transformer iteratively refine its representations for all positions in sequence in parallel with self-attention mechanism decomposableAttnModel (); lin2017structured (), followed by the recurrent transformation consisting of a depth-wise separable convolution (xception2016) or a position-wise fully-connected layer (see above Fig). We also extended the Universal Transformer by employing an adaptive computation time mechanism at each position in sequence (graves2016adaptive), allowing model to choose the required number of refinement steps for each symbol dynamically.

When running for fixed number of steps, the Universal Transformer is equivalent to a multi-layer Transformer with a tied parameter across its layers. However, another, and possibly more informative, way of characterizing Universal Transformer is as recurrent function evolving per-symbol hidden states in parallel, based at each step on a sequence of the previous unknown state. In this way, it is similar to architectures such as Neural GPU and the Neural Turing Machine. The Universal Transformer thereby retains the attractive computational efficiency of original feed-forward Transformer model, but with an added recurrent inductive bias of RNNs. In its adaptive form, we show that the Universal Transformer can effectively interpolate between the feed-forward, fixed-depth Transformer, and a gated, recurrent architecture running for several steps depending on the input data.

Our experimental results show that its recurrence improve results in machine translation, where Universal Transformer outperforms the standard Transformer with a same no.of parameters. In experiments on several algorithmic tasks, Universal Transformer consistently improves significantly over LSTM(Long Short Term Memory) RNNs and the standard Transformer. Furthermore, on bAbI and LAMBADA text understanding data sets, the Universal Transformer achieves a new state of the art.

Q4. What is StarSpace in NLP?

Answer:

We introduce StarSpace, the neural embedding model that is general enough to solve a wide variety of problems:

- Other labeling tasks, or Text classification, e.g., sentiment classification.

- Ranking of the set of entities, e.g., a classification of web documents given a query.
- Collaborative filtering-based recommendation, e.g., recommending documents, videos or music.
- Content-based recommendation where content is defined with discrete features, e.g., words of documents.
- Embedding graphs, e.g., multi-relational graphs such as Freebase.
- Learning word, sentence, or document embeddings.

It can be viewed as a straight-forward and efficient strong baseline for any of these tasks. In experiment, it is shown to be on par with or outperforming several competing methods while being generally applicable to cases where many of that method are not.

The method works by learning entity embeddings with discrete feature representation from relations among collections of those entities directly for the task of ranking or classification of interest. In the general case, StarSpace embeds objects of different types into a vectorial embedding space; hence, the "star" ("*," meaning all types) and "space" in a name and in that familiar space compares them against each other. It learns to rank the set of entities, documents, or objects given a query entity, document, or object, where the query is not necessarily of the same type as the items in the set.

Q5. TransferTransfo in NLP

Answer:

Non-goal-oriented dialogue systems (chatbots) are interesting test-bed for interactive Natural Language Processing (NLP) systems and are also directly useful in wide range of applications ranging from technical support services to entertainment. However, building intelligent conversational agent remains an unsolved problem in artificial intelligence(AI) research. Recently, recurrent neural network(RNN) based models with sufficient capacity and access to large datasets attracted large interest when first attempted. It showed that they were capable of generating meaningful responses in some chit-chat settings. Still, further inquiries in the capabilities of these neural network

architectures and developments indicated that they were limited which made communicating with them a rather unsatisfying experience for human beings.

The main issues with these architectures can be summarized as:

- (i) the wildly inconsistent outputs and the lack of a consistent personality (Li and Jurafsky, 2016),
- (ii) the absence of long-term memory as these models have difficulties in taking into account more than the last dialogue utterance; and
- (iii) a tendency to produce consensual and generic responses that are vague and not engaging for humans (Li, Monroe, and Jurafsky, 2016).

In this work, we take a step toward more consistent and relevant data-driven conversational agents by proposing a model architecture, associated training and generation algorithms which are able to significantly improve over the traditional seq-2-seq and information-retrieval baselines in terms of (i) relevance of the answer (ii) coherence with a predefined personality and dialog history, and (iii) grammaticality and fluency as evaluated by auto.

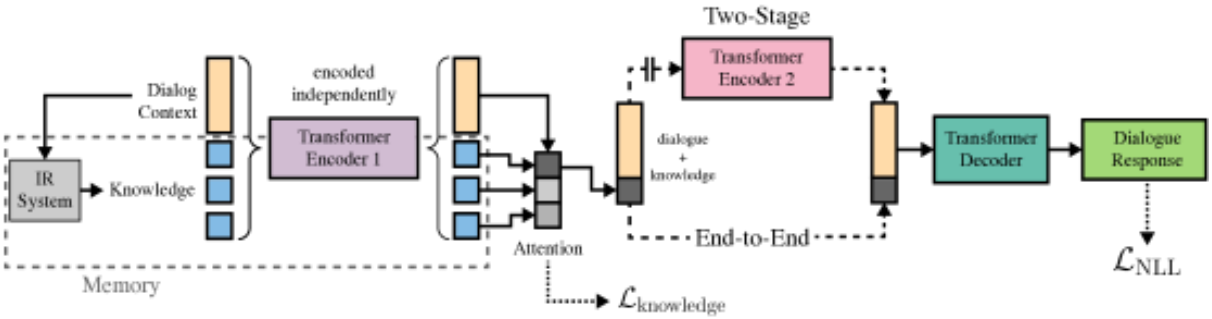
Q6. Wizard of Wikipedia: Knowledge-Powered Conversational Agents

Answer:

Arguably, one of the critical goals of AI and the ultimate goal of natural language research is for the human to be able to talk to the machine. In order to get close to this goal, machines must master the no. of skills: to be able to comprehend language, employ memory to retain and recall knowledge, to reason about these concept together, and finally output a response that both fulfills functional goals in conversation while simultaneously being captivating to their human speaking partner. The current state-of-the-art(SOTA) approaches, sequence to sequence(seq2seq) models of various kinds (Sutskever et al., 2014; Vinyals & Le, 2015; Serban et al., 2016; Vaswani et al., 2017) attempt to address some of these skills, but generally suffer from inability to bring memory and knowledge to bear; as indicated by their name, they involve encoding input sequence, providing limited reasoning by transforming their hidden state given input, and then decoding to the output. To converse intelligently on the given topic, the speaker needs knowledge of that subject, and it is our contention here that more direct knowledge memory mechanisms need to be employed. In this work, we consider setups where this can be naturally measured and built.

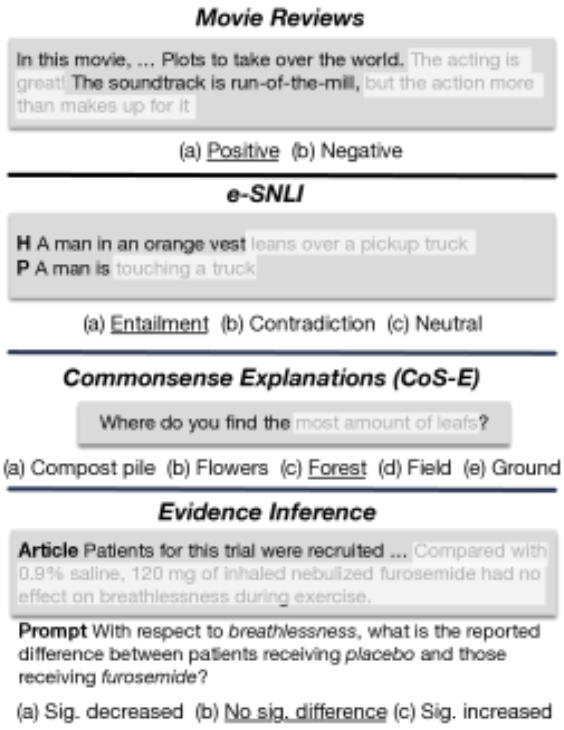
We consider the task of open-domain dialogue, where two speakers conduct open-ended chit-chat given an initial starting topic, and during the conversation, the topic can broaden or focus on related themes. During such conversations, an interlocutor can glean new information and personal points of view from

their speaking partner, while providing themselves similarly. This is a challenging task as it requires several components not found in many standard models. We design a set of architectures specifically for this goal that combine elements of Memory Network architectures (Sukhbaatar et al., 2015) to retrieve knowledge and read and condition on it, and Transformer architectures (Vaswani et al., 2017) to provide state-of-the-art text representations and sequence models for generating outputs, which we term Transformer Memory Networks.



Q7. ERASER: A Benchmark to Evaluate Rationalized NLP Models

Answer:



Interest has recently grown in interpretable (Natural Language Processing) NLP systems that can reveal **how** and **why** model make their predictions. But work in this direction has been conducted on the

different dataset with correspondingly different metrics, and inherent subjectivity in defining what constitute 'interpretability' has translated into researcher using different metrics to quantify performance. We aimed to facilitate measurable progress on designing interpretable NLP(Natural Language Processing) models by releasing the standardized benchmark of datasets — augmented and repurposed from pre-existing corpora, and spanning the range of NLP tasks — and associated metrics for measuring the quality of rationales. We refer to this as ERASER(Evaluating Rationales And Simple English Reasoning) benchmark.

In curating and releasing ERASER we take inspiration from stickiness of GLUE (Wang et al., 2019b) and SuperGLUE Wang et al. (2019a) benchmarks for evaluating progress in natural language understanding(NLU) tasks. These have enabled rapid growth in models for inclusive language representation learning. We believe still somewhat nascent subfield of interpretable NLP(Natural Language Processing) stands to similarly benefit from the analogous collection of standardized datasets or tasks and metric.

'Interpretability' is the broad topic with many possible realizations Doshi-Velez and Kim (2017); Lipton (2016). In ERASER, we focuses specifically on *rationales*, i.e., snippets of text from the source document that support a specific categorization. All datasets contained in ERASER include such rational, explicitly marked by annotators as supporting specific classifications. By definition, rationales should be *sufficient* to categorize document, but they may not be comprehensive. Therefore, for some dataset, we have collected *complete* rationales, i.e., in which *all* evidence supporting the classification has been marked.

How one measures 'quality' of extracted rationales will invariably depend on their intended use. With this in mind, we propose the suite of metrics to evaluate rationales that might be appropriate for different scenarios. Widely, this includes measures of agreement with human-provided rationales and assessment of *faithfulness*. The latter aim to capture extent to which rationales provided by the model, in fact, informed its prediction.

While we propose metrics that we think are reasonable, we view a problem of designing metrics for evaluating rationales-especially for capturing faithfulness — as a topic for further research that we hope that ERASER will help facilitate. We plan to revisit metrics proposed here in future iterations of benchmark, ideally with input from community. Notably, while we provide a 'leaderboard,' this is perhaps better viewed as the 'results board'; we do not privilege any particular metric. Instead, we hope that ERASER permits comparison between models that provide rationales wrt different criteria of interest.

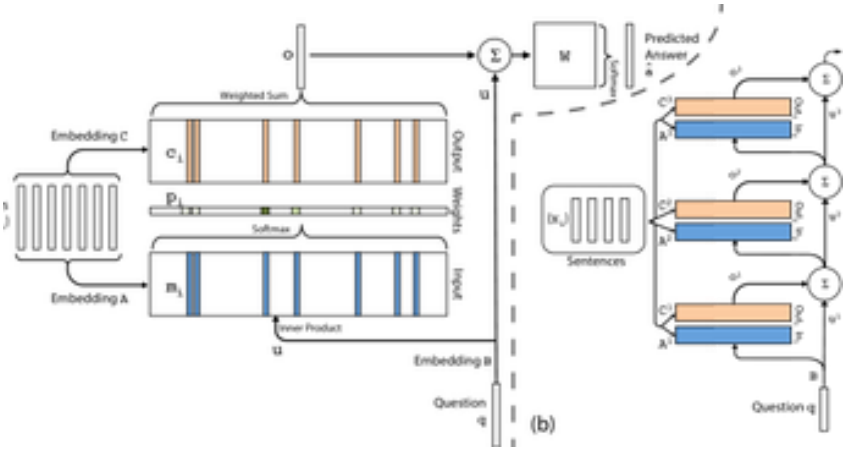
Q8. End to End memory networks

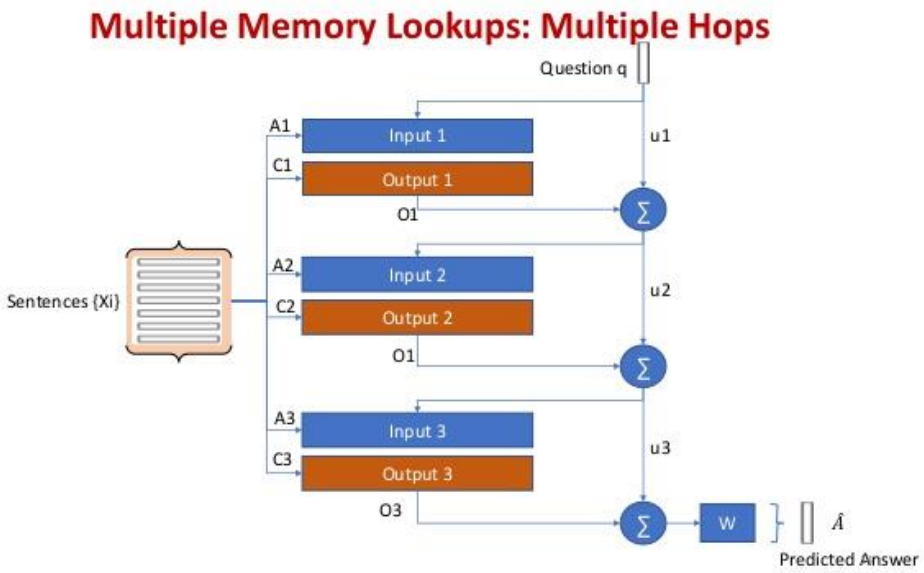
Answer:

Two grand challenges in artificial intelligence(AI) research have been to build a model that can make multiple computational step in the service of answering the question or completing the task, and models that can describe long term dependencies in sequential data.

Recently there has been the resurgence in models of computation using explicit storage and a notion of attention; manipulating such storage offers an approach to both of these challenges. In, the storage is endowed with continuous representation; reads from and writes to storage, as well as other processing steps, are modeled by actions of neural networks.

In this work, we present the new recurrent neural network (RNN) architecture where recurrence reads from possibly large external memory multiple times before outputting symbol. Our model can be considered the continuous form of the Memory Network implemented in. The model in that work was not easy to train via back-propagation and required supervision at each layer of a network. The continuity of model we present here means that it can be trained end-to-end from input-output pairs, and so applies to more tasks, i.e., tasks where such supervision is not available, like in language modeling or realistically supervised question answering tasks. Our model can also be seen as version of RNNsearch with multiple computational steps per output symbol. We will show experimentally that various hops over the long-term memory are crucial to excellent performance of our model on these tasks, and that training the memory representation can be integrated in a scalable manner into our end-to-end neural network model.





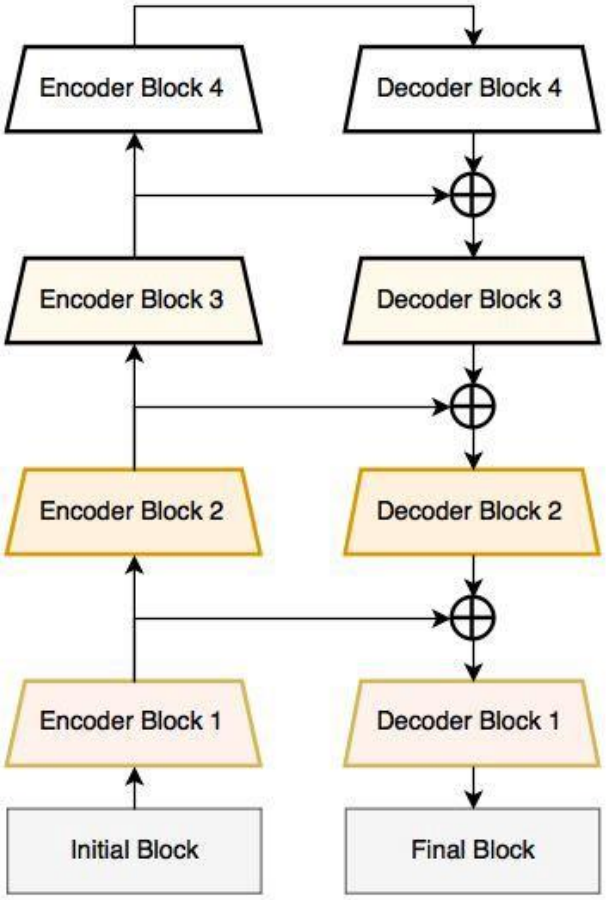
Q9. What is LinkNet?

Answer:

From my experience, LinkNet is lightning fast, which is one of the main improvements the authors site in their summary. LinkNet is a relatively lightweight network with around 11.5 million parameters; networks like VGG have more than 10x that amount.

The structure of LinkNet is to use a series of encoder and decoder blocks to break down the image and build it back up before passing it through a few final convolutional layers. The structure of the network was designed to minimize the number of parameters so that segmentation could be done in real-time.

I performed some tests of the LinkNet architecture but did not spend too much time iterating to improving the models.


Encoder Block 4
Decoder Block 4
Encoder Block 3
Decoder Block 3
Encoder Block 2
Decoder Block 2
Encoder Block 1
Decoder Block 1
Initial Block
Final Block

DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

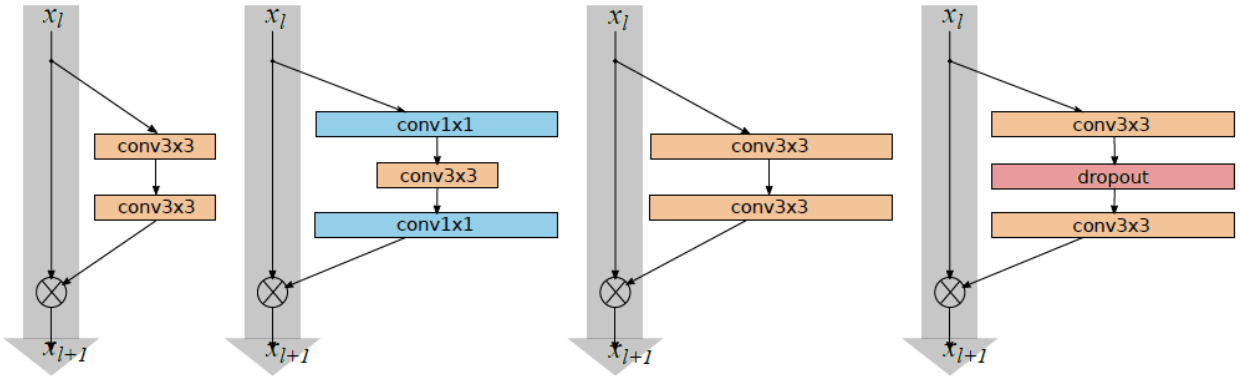
Day-25

Q1. What is WRN?

Answer:

WRN: It stands for Wide Residual Networks is presented. By widening Residual Network (ResNet), the network can be more superficial or shallow with same accuracy or improved accuracy. More external network means:

- the number of layers can be reduced.
- Training time can be shorter, as well.



(a) basic (b) bottleneck (c) basic-wide (d) wide-dropout

Problems on Residual Network (ResNet)

Circuit Complexity Theory

The authors of residual networks(ResNet) tried to make them as thin as possible in favor of increasing their depth and having less parameters and even introduced a «bottleneck» block, which makes ResNet blocks even thinner.

Diminishing Feature Reuse

However, As gradient flows through network, there is nothing to force it to go through residual block weights, and it can avoid learning anything during training, so there may be either only few blocks that learn useful representations or many blocks share very little information with a small contribution to the final goal. This problem was formulated as a diminishing feature reuse.

WRNs (Wide Residual Networks)

In WRNs, plenty of parameters are tested like the design of ResNet block, how deep (deepening factor l), and how extensive (widening factor k) within the ResNet block.

When *k*=1, it has the same width as the *ResNet*. While *k*>1, it is *k* time wider than *ResNet*.

WRN-*d*-*k*: means the WRN has a depth of *d* and with widening factor *k*.

- *Pre-Activation ResNet* is used in CIFAR-10, CIFAR-100, and SVHN datasets. Original *ResNet* is used in the ImageNet dataset.
- The significant difference is that *Pre-Activation ResNet* has the structure of performing batch norm and ReLU before convolution (i.e., BN-ReLU-Conv) while original *ResNet* has the structure of Conv-BN-ReLU. And *Pre-Activation ResNet* is generally better than the original one, but it has no visible improvement in ImageNet when layers are only around 100.

The design of the ResNet block

block type	depth	# params	time,s	CIFAR-10
$B(1,3,1)$	40	1.4M	85.8	6.06
$B(3,1)$	40	1.2M	67.5	5.78
$B(1,3)$	40	1.3M	72.2	6.42
$B(3,1,1)$	40	1.3M	82.2	5.86
$B(3,3)$	28	1.5M	67.5	5.73
$B(3,1,3)$	22	1.1M	59.9	5.78

-

- **B(3;3)**: Original «basic» block, in the above figure a.
- **B(3;1;3)**: With one extra (1×1) layer in between the two 3×3 layers
- **B(1;3;1)**: With the same dimensionality of all convolutions, bottleneck
- **B(1;3)**: The network has the alternating (1×1, 3×3) convolutions.
- **B(3;1)**: The network has the alternating(3×3, 1×1) convolutions.
- **B(3;1;1)**: This is Network in Network style block.

B(3;3) has the smallest error rate (5.73%).

Note: The Number of depths (layers) is different is to keep the number of parameters close to each other.

Q2.What is SIMCO: SIMilarity-based object Counting?

Answer:

Most approaches for counting similar objects in images assume a single object class; when is not, ad-hoc learning is necessary. None of them are genuinely agnostic and multi-class, i.e., able to capture repeated patterns of different types without any tuning. Counting approaches are based on density or regression estimation; here, we focus on counting by detection, so the counted objects are individually detected first.

Research on agnostic counting is vital in many fields. It serves for obvious question answering, where counting questions could be made on too-specific entities outside the semantic span of the available classes (e.g., "What is the most occurrent thing?" in below Fig.). In representation learning, unsupervised counting of visual primitives (i.e., visible "things") is crucial to obtain a rich image representation. Counting is a hot topic in cognitive robotics, where autonomous agents learn by separating sensory input into the finite number of classes (without a precise semantics), building the classification system that

counts on each of them.

Application-wise, agnostic counting may help the manual tagging of training images, providing a starting guess for the annotator on single- or multi-spectral images. Inpainting filters may benefit from a magic wand capturing repeated instances to remove.

Figure: SIMCO on obvious question answering: the most occurrent object? SIMCO finds 47 LEGO heads.

In this paper, we present the SIMCO (SIMilarity-based object COunting) approach, which is entirely agnostic, i.e., with no need for any *ad-hoc* class-specific fine-tuning, and multi-class, i.e., finding different types of repeated patterns. Two main ideas characterize SIMCO.

First, every object to be counted is considered as a specialization of a basic 2D shape: this is particularly true with many and small objects (see in above Fig: LEGO heads can be approximated as circles). SIMCO incorporates this idea building upon the novel Mask-RCNN-based classifier, fine-tuned just once on a novel synthetic shape dataset, *InShape*.

The second idea is that leveraging on the 2D shape approximation of objects; one can naturally perform unsupervised grouping of the detected objects (grouping circles with circles, etc.), discovering *different* types of repeated entities (without resorting to a particular set of classes). SIMCO realizes this with a head branch in the network architecture implementing triplet losses, which provides a 64-dim embedding that maps objects close if they share the same shape class plus some appearance attributes. Affinity propagation clustering finds groups over this embedding.

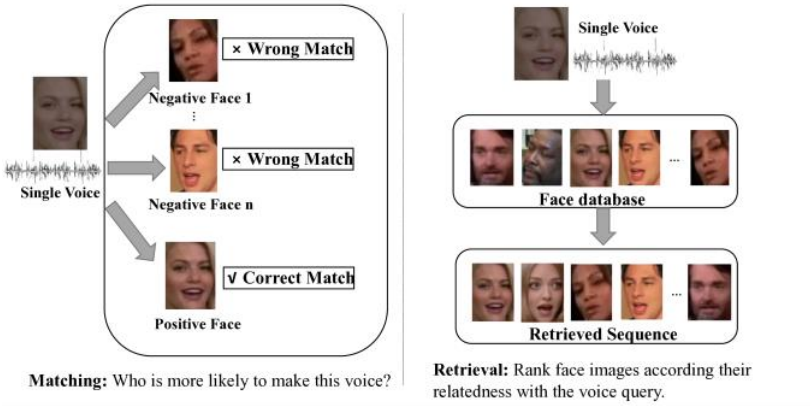
Q3. What is Voice-Face Cross-modal Matching and Retrieval?

Answer:

Studies in biology and neuroscience have shown that human's appearances are associated with their voices. Both the facial features and voice-controlling organs of individuals are affected by hormones and genetic information. Human beings can recognize this association. For example: when hearing from the phone call, we can guess the gender, the approximate age of the person on the other end of the line. When watching an unvoiced TV show. We can imagine an approximate voice by observing the face movement of the protagonist. With the recent advances of deep learning, face recognition models, and speaker recognition models have achieved exceptionally high precision. Can the associations between voices and faces be discovered algorithmically by machines? The research on this problem can benefit a lot of applications such as synchronizing video faces and talking sound, generating faces according to voices.

In recent years, much research attention has been paid on the voice-face cross-modal learning tasks, which have shown the feasibility of recognizing voice-face associations. This problem is generally formulated as a voice-face matching task and the voice-face retrieval task, as shown in Figure 1. Given a set of voice audios and faces, voice-face matching is to tell which look makes the voice when machine hearing voice audio. Voice-face retrieval is to present a sorted sequence of faces in the order of the

estimated match from a query of voice recording.



SVHF is the prior of voice-face cross-modal learning, which studies the performance of CNN-based deep network on this problem. The human's baseline for the voice-face matching task is also proposed in the paper. Both the "voice to face" and the "face to voice" matching tasks are studied in the Pins and Horiguchi's work, which exhibits similar performance on these two tasks. The curriculum learning schedule is introduced in Pins for hard negative mining. Various visualizations of the embedding vectors are presented to show the learned audio-visual associations in Kim's work. DIMNet learns the common representations for faces and voices by leveraging their relationship with some covariates such as gender and nationality. DIMNet obtains an accuracy of 84.12% on the 1:2 matching, which exceeds the human level.

Research on this problem is still in the early stage. Datasets used by previous research are always tiny, which can't evaluate the generalization ability of models sufficiently. Traditional test schemes based on random tuple mining tend to have low confidence. The benchmark for this problem needs to be established. This paper presents the voice-face cross-modal matching and retrieval framework, a dataset from Chinese speakers and a data collection tool. In the frame, cross-modal embeddings are learned with CNN-based networks, and triplet loss in a voice anchored metric space with L2-Norm constraint. An identity-based example sampling method is adopted to improve the model efficiency. The proposed framework achieves state-of-the-art performance on multiple tasks. For example, the result of 1:2 matching tested on 10 million triplets (thousands of people) reached 84.48%, which is also higher than DIMNet tested on 189 people. We have evaluated the various modules of the CNN-based framework and provided our recommendations. Even matching and retrieval based on the average of multiple voices and multiple faces are also attempted, which can further improve the performance. This task is the simplest way of analyzing video data. Large-scale datasets are used in this problem to ensure the generalization ability required in a real application. The cross-language transfer capability of the model is studied on the voice-face dataset of Chinese speakers we constructed. The series of performance

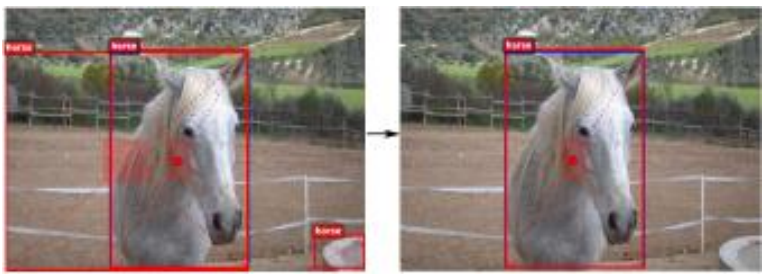
metrics are presented on the tasks by extensive experiments. The source code of the paper and the dataset collection tool will be published along with the article.

Q4. What is CenterNet: Object Detection with Keypoint Triplets?

Answer:

Object detection has been significantly improved and advanced with the help of deep learning, especially convolutional neural networks (CNNs). In the current era, one of the most popular flowcharts is anchor-based, which placed the set of rectangles with pre-defined sizes, and regressed them to the desired place with the help of the ground-truth objects. These approaches often need a large number of anchors to ensure the sufficiently high IoU (intersection over union) rate with the ground-truth objects, and the size and aspect ratio of each anchor box needs to be manually designed. Also, anchors are usually not aligned with the ground-truth boxes, which is not conducive to bounding box classification tasks.

To overcome the drawbacks of anchor based approaches, a keypoint-based object detection pipeline named CornerNet was proposed. It represented each object by a pair of corner key points, which bypassed the need for anchor boxes and achieved the state-of-the-art one-stage object detection accuracy. Nevertheless, the performance of CornerNet is still restricted by its relatively weak ability to refer to the global information of an object. That is to say since a pair of corners construct each object, the algorithm is sensitive to detect the boundary of objects, meanwhile not being aware of which pairs of critical points should be grouped into the objects. Consequently, as shown in Figure a, it often generates some incorrect bounding boxes, most of which could be easily filtered out with complementary information, *e.g.*, the aspect ratio.



To address this issue, we equip CornerNet with an ability to perceive the visual patterns within each proposed region, so that it can identify the correctness of each bounding box by itself. In this paper, we present the low-cost yet effective solution named **CenterNet**, which explores the central part of the proposal, *i.e.*, the region that is close to the geometric center, with one extra keypoint. Our intuition is that, if the predicted bounding box has a high IoU with the ground-truth box, then the probability that the center key point in its central region is predicted as the same class is high, and vice versa. Thus, during inference, after the proposal is generated as a pair of corner keypoints, we determine if the plan is indeed an object by checking if there is a crucial central point of the same class falling within its central

region. The idea, as shown in Figure a, is to use a triplet instead of a pair of key points to represent each object.

Accordingly, for better detecting the center keypoints and corners, we propose two strategies to enrich center and corner information, respectively. The first strategy is named as **center pooling**, which is used in the branch for predicting the center keypoints. Center pooling helps the center keypoints obtain more recognizable visual patterns within objects, which makes the central part of the proposal be better perceived. We achieve this by getting out the max summed response in both horizontal and vertical directions of the center key point on a feature map for predicting center keypoints. The second strategy is named **cascade corner pooling**, which equips the original corner pooling module with the ability to perceive internal information. We achieve this by getting out the max summed response in both boundary and inner directions of objects on a feature map for predicting corners. Empirically, we verify that such the two-directional pooling method is more stable, *i.e.,* being more robust to the feature-level noises, which contributes to the improvement of both precision and recall.

We evaluate the proposed CenterNet on the MS-COCO dataset, one of the most popular benchmarks for large scale object detection. CenterNet, with both center pooling and the cascade corner pooling incorporated, reports an AP of 47.0% on the test-dev set, which outperforms all existing one-stage detectors by the extensive margin. With an average inference time of 270ms using a 52-layer hourglass backbone and 340ms using a 104-layer hourglass backbone per image, CenterNet is quite efficient yet closely matches the state-of-the-art performance of the other two-stage detectors.

Q5. What is Task2Vec: Task Embedding for Meta-Learning?

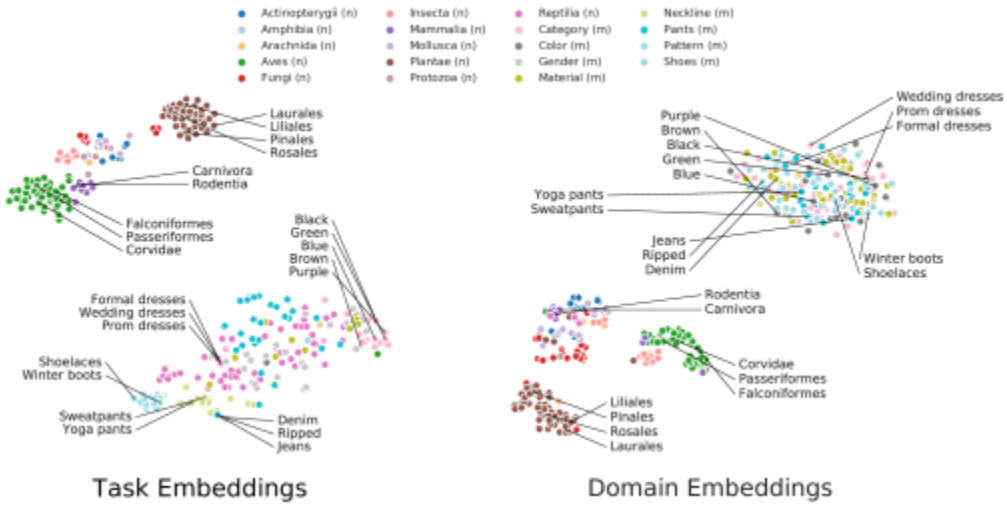
Answer:

The success of Deep Learning hinges in part on the fact that models learned for one task can be used on the other related tasks. Yet, no general framework exists to describe and learn relations between tasks. We introduce task2vec embedding, the technique to represent tasks as elements of the vector space is based on the Fisher Information Matrix. The norms of the embedding correlates with the complexity of the task, while the distance between embeddings captures the semantic similarities between tasks (Fig. 1). When other natural distances are available, such as taxonomical distance in the biological classification, we find that the embedding distance correlates positively with it (Fig. 2). Moreover, we introduce an asymmetric distance on tasks that correlates with the transferability between tasks.

Computation of the embedding leverages the duality between network parameters (weights) and outputs (activations) in a deep neural network (DNN): Just as the activations of a DNN trained on the complex visual recognition task are the rich representation of the input images, we show that the gradients of the weights relative to a task-specific loss are the rich representation of the task itself. Specifically, given a task defined by the dataset D={(xi,yi)}Ni=1 of labeled samples, we feed the data through a pre-trained

reference convolutional neural network which we call "*probe network*", and compute the diagonal Fisher Information Matrix (FIM) of the network filter parameters, to capture the structure of task. Since the architecture and weights of the probe network are fixed, the FIM provides the fixed-dimensional representation of the tasks. We show this embedding encodes the "difficulty" of the tasks, characteristics of the input domain, and features of the probe network are useful to solve it.

Our task embedding can be used to reason about the space of the tasks and solve meta-tasks. As a motivating example, we study the problem of selecting the best pre-trained feature extractor to solve a new task. This can be particularly valuable when there is insufficient data to train or fine-tune a generic model, and the transfer of knowledge is essential. task2vec depends solely on the task and ignores interactions with the model, which may, however, play an essential role. To address this, we learn about the joint task and model embedding, called model2vec, in such a way that models whose embeddings are close to a task exhibit excellent performance on the task. We use this to select an expert from the given collection, improving performance relative to fine-tuning a generic model trained on ImageNet and obtaining close to the ground-truth optimal selection.



Q6. What is GLMNet: Graph Learning-Matching Networks for Feature Matching?

Answer:

Many problems of interest in computer vision and pattern recognition area can be formulated as a problem of finding consistent correspondences between two sets of features, which are known as feature matching problem. Feature set that incorporates the pairwise constraint can be represented via an attribute

graph whose nodes represent the unary descriptors of feature points, and edges encode the pairwise relationships among different feature points. Based on this graph representation, feature matching can then be reformulated as a graph node matching problem.

Graph matching generally first operates with both node and edge affinities that encode similarities between the node and edge descriptors in two graphs. Then, it can be formulated mathematically as an Integral Quadratic Programming (IQP) problem with permutation constraint on related solutions to encode the one-to-one matching constraints. It is known to be NP-hard. Thus, many methods usually solve it approximately by relaxing the discrete permutation constraint and finding locally optimal solutions. Also, to obtain better node/edge affinities, learning methods have been investigated to determine the more optimal parameters in node/edge affinity computation. Recently, deep learning methods have also been developed for matching problems. The main benefit of deep learning matching methods is that they can conduct visual feature representation, node/edge affinity learning, and matching optimization together in an end-to-end manner. Zanfir et al. propose an end-to-end graph matching model, which makes it possible to learn all the parameters of the graph matching process. Wang et al. recently aim to explore graph convolutional networks (GCNs) for graph matching which conducts graph node embedding and matching simultaneously in a unified system.

Inspired by recent deep graph matching methods, in this paper, we propose a novel Graph Learning-Matching Network (GLMNet) for graph matching problems. Overall, the main contributions of this paper are three aspects.

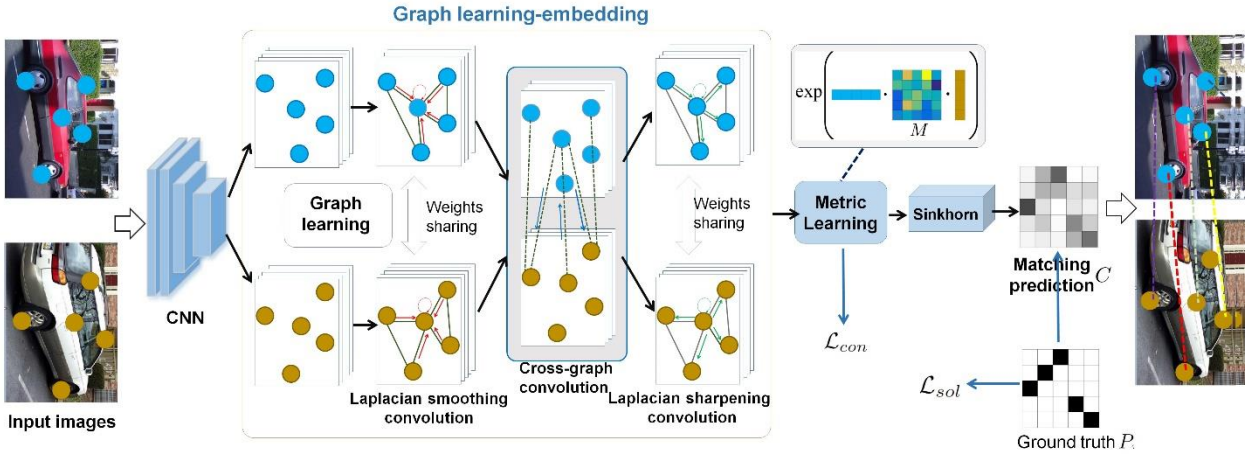
First, a critical aspect of (feature) graph matching is the construction of two matching graphs. Existing deep graph matching models generally use fixed structure graphs, such as k-NN, Delaunay graph, etc., which thus are not guaranteed to serve the parallel task best. To address this issue, we propose to adaptively learn a pair of optimal graphs for the matching task and integrate *graph learning* and *graph matching* simultaneously in a unified end-to-end network architecture.

Second, the existing GCN based graph matching model adopts the general smoothing based graph convolution operation for graph node embedding, which may encourage the feature embedding of each node becoming more similar to those of its neighboring nodes. This is desirable for graph node labeling or classification tasks, *but* undesirable for the matching task because extensive smoothing convolution may dilute the discriminatory information. To alleviate this effect, we propose to incorporate a Laplacian sharpening based graph convolution operation for graph node embedding and matching tasks. Laplacian sharpening process can be regarded as the counterpart of Laplacian smoothing which encourages the embedding of each node farther away from its neighbors.

Third, existing deep graph matching methods generally utilize a doubly stochastic normalization for the final matching prediction. This usually ignores the discrete one-to-one matching constraints in matching

optimization/prediction. To overcome this issue, we develop a novel constraint regularized loss to further incorporate the one-to-one matching constraints in matching prediction.

Experimental results, including ablation studies, demonstrate the effectiveness of our GLMNet and advantages of devised components, including graph learning-matching architecture, Laplacian sharpening convolution for discriminative embedding, and constraint regularized loss to encode one-to-one matching constraints.

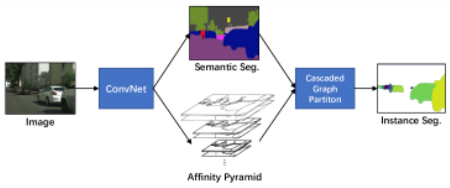


Q7. What is SSAP: Single-Shot Instance Segmentation With Affinity Pyramid?

Answer:

The rapid development of Convolutional networks has revolutionized various vision tasks, enabling us to move towards a more fine-grained understanding of images. Instead of classic bounding-box level object detection or class-level semantic segmentation, instance segmentation provides in-depth knowledge by segmenting all the objects and distinguish different object instances. Researchers are showing increasing interests in instance segmentation recently.

Current state-of-the-art solutions to this challenging problem can be classified into the *proposal-based* and *proposal-free* approaches. The proposal-based methods regard it as an extension to the classic object detection task. After localizing each object with a bounding box, the foreground mask is predicted within each bounding box proposal. However, the performances of the scheme based methods are highly limited by the quality of the bounding box predictions, and the two-stage pipeline also limits the speed of systems. By contrast, the proposal-free approach has the advantage of its efficient and straightforward design. This work also focuses on the proposal-free paradigm.



The proposal-free methods mostly start by producing instance-agnostic pixel-level semantic class labels, followed by clustering them into the different object instances with particularly designed instance-aware features. However, previous methods mainly treat the two sub-processes as the two separate stages and employ multiple modules, which is suboptimal. The mutual benefits between the two sub-tasks can be exploited, which will further improve the performance of the instance, segmentation. Moreover, employing multiple modules may result in additional computational costs for real-world applications.

To cope with the above issues, this work proposes a single-shot proposal-free instance segmentation method, which jointly learns the pixel-level semantic class segmentation and object instance differentiating in a unified model with a single backbone network, as shown in Fig. 1. Specifically, for distinguishing different object instances, an affinity pyramid is proposed, which can be jointly learned with the labeling of semantic classes. The pixel-pair affinity computes the probability that two pixels belong to the same instance. In this work, the short-range relationships for pixels close to each other are derived with dense small learning windows. Simultaneously, the long-range connections for pixels distant from each other are also required to group objects with large scales or nonadjacent parts. Instead of enlarging the windows, the multi-range relationships are decoupled, and long-range connections are sparsely derived from the instance maps with lower resolutions. After that, we propose learning the affinity pyramid at multiple scales along the hierarchy of a U-shape network, where the short-range and long-range affinities are effectively learned from the feature levels with the higher and lower resolutions respectively. Experiments in Table 3 show that the pixel-level semantic segmentation and the pixel-pair affinity pyramid based grouping are indeed mutually benefited from the proposed joint learning scheme. The overall instance of segmentation is thus further improved.

Then, to utilize the cues about global context reasoning, this work employs a graph partition method to derive instances from the learned affinities. Unlike previous time-consuming methods, the cascaded graph partition module is presented to incorporate the graph partition process with the hierarchical manner of the affinity pyramid and finally provides both acceleration and performance improvements. Concretely, with the learned pixel-pair affinity pyramid, the graph is constructed by regarding each pixel

as the node and transforming affinities into the edge scores. Graph partition is then employed from higher-level lower-resolution layers to the lower-level higher-resolution layers progressively. Instance segmentation predictions from the lower resolutions produce confident proposals, which significantly reduce node numbers at higher resolutions. Thus the whole process is accelerated.

Q8. What is TENER: Adapting Transformer Encoder for Name Entity Recognition?

Answer:

The named entity recognition (NER) is the task of finding the start and end of an entity in the sentence and assigning a class for this entity. NER has been widely studied in the field of natural language processing (NLP) because of its potential assistance in question generation Zhou et al. (2017), relation extraction Miwa and Bansal (2016), and coreference resolution Fragkou (2017). Since Collobert et al. (2011), various neural models have been introduced to avoid the hand-crafted features Huang et al. (2015); Ma and Hovy (2016); Lample et al.

NER is usually viewed as a sequence labeling task, the neural models typically contain three components: word embedding layer, context encoder layer, and decoder layer Huang et al. (2015); Ma and Hovy (2016); Lample et al. (2016); Chiu and Nichols (2016); Chen et al. Zhang et al. (2018); Gui et al. (2019b). The difference between various NER models mainly lies in the variance in these components.

Recurrent Neural Networks (RNNs) are widely employed in NLP tasks due to its sequential characteristic, which is aligned well with the language. Specifically, bidirectional extended short-term memory networks (BiLSTM) Hochreiter and Schmidhuber (1997) is one of the most widely used RNN structures. (Huang et al., 2015) was the first one to apply the BiLSTM and the Conditional Random Fields (CRF) Lafferty et al. (2001) to sequence the labeling tasks. Owing to BiLSTM's high power to learn the contextual representation of words, it has been adopted by the majority of the NER models as the encoder Ma and Hovy (2016); Lample et al. (2016); Zhang et al. (2018); Gui et al.

Recently, Transformer Vaswani et al. (2017) began to prevail in the various NLP tasks, like machine translation Vaswani et al. (2017), language modeling Radford et al. (2018), and pretraining models Devlin et al. (2018). The Transformer encoder adopts the fully-connected self-attention structure to model the long-range context, which is the weakness of RNNs. Moreover, the Transformer has better parallelism ability than RNNs. However, in the NER task, Transformer encoder has been reported to perform poorly Guo et al. (2019), our experiments also confirm this result. Therefore, it is intriguing to explore the reason why the Transformer does not work well in the NER task.

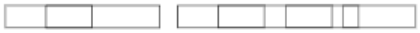


Figure 1: An example for NER. The relative direction is important in the NER task, because words before "Inc." are mostly to be an organization, words after "in" are more likely to be time or location. Besides, the relative distance between words is also important, since only continuous words can form an entity, the former "Louis Vuitton" can not form an entity with the "Inc.".

The first is that the sinusoidal position embedding used in the vanilla Transformer is relative distance sensitive and direction-agnostic, but this property will lose when used in the vanilla Transformer. However, both the direction and relative distance information are essential in the NER task. For example, words after "in" are more likely to be a location or time than words before it, and words before "Inc." is most likely to be of the entity type "ORG." Besides, an entity is a continuous span of words. Therefore, the awareness of relative distance might help the word better recognizes its neighbor. To endow the Transformer with the ability of directionality and relative distance awareness, we adopt direction-aware attention with the relative positional encoding Shaw et al. (2018); Huang et al. (2019); Dai et al. (2019). We propose a revised relative positional encoding that uses fewer parameters and performs better.

The second is an empirical finding. The attention distribution of the vanilla Transformer is scaled and smooth. But for NER, sparse attention is suitable since not all words are necessary to be attended. Given the current word, a few contextual words are enough to judge its label. The smooth attention could include some noisy information. Therefore, we abandon the scale factor of dot-production consideration and the use of un-scaled and sharp attention.

With the above improvements, we can significantly boost the performance of the Transformer encoder for NER.

Q9. What is Subword ELMo?

Answer:

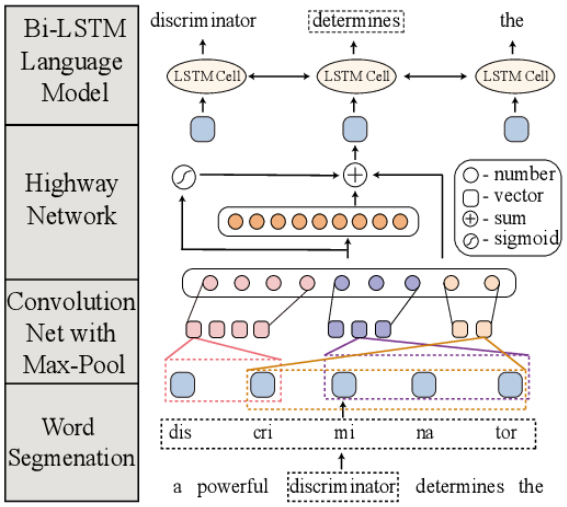
Recently, pre-trained language representation has shown to be useful for improving many NLP tasks. Embeddings from Language Models is one of the most outstanding works, which uses the character-aware language model to augment word representation.

An essential challenge in training word-based language models is how to control the vocabulary size for better rare word representation. No matter how large the vocabulary is, unique words are always insufficiently trained. Besides, an extensive vocabulary takes too much time and computational resources for the model to converge. Whereas, if the dictionary is too small, the out-of-vocabulary (OOV) issue will harm the model performance slowly. To obtain effective word representation, Jozefowicz et al. (2016) introduce character-driven word embedding using the convolutional neural network (CNN).

However, potential insufficiency when modeling word from characters which hold little linguistic sense, especially, the morphological source. Only 86 roles are adopted in English writing, making the input too coarse for embedding learning. As we argue that for the better representation from a refined granularity, the word is too large, and character is too small, it is natural for us to consider subword units between character and the word levels.

Splitting the word into subwords and using them to augment the word representation may recover the latent syntactic or semantic information. For example, *uselessness* could be divided into the following subwords: Previous work usually considers linguistic knowledge-based methods to tokenize each word into the subwords (namely, morphemes). However, such treatment may encounter the three main inconveniences. First, the subwords from linguistic knowledge, typically including the morphological suffix, prefix, and stem, may not be suitable for the targeted NLP task Banerjee and Bhattacharyya or mislead the representation of some words, like the meaning of *understanding* cannot be formed by *under* and *stand*. Second, linguistic knowledge, including related annotated lexicons or corpora, may not even be available for the specific low-resource language. Due to these limitations, we focus on the computationally motivated subword tokenization approaches in this work.

In this paper, we propose Embedding from Subword-aware Language Models (ESuLMo), which takes subword as input to augment word representation and release a sizeable pre-trained language model research communities. Evaluations show that the pre-trained language models of ESuLMo outperform all RNN-based language models, including ELMo, in terms of PPL and ESuLMo beats state-of-the-art results in three of four downstream NLP tasks.



iNeuron

DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day26

Q1.What is DCGANs (Deep Convolutional Generative Adversarial Networks)?

Answer:

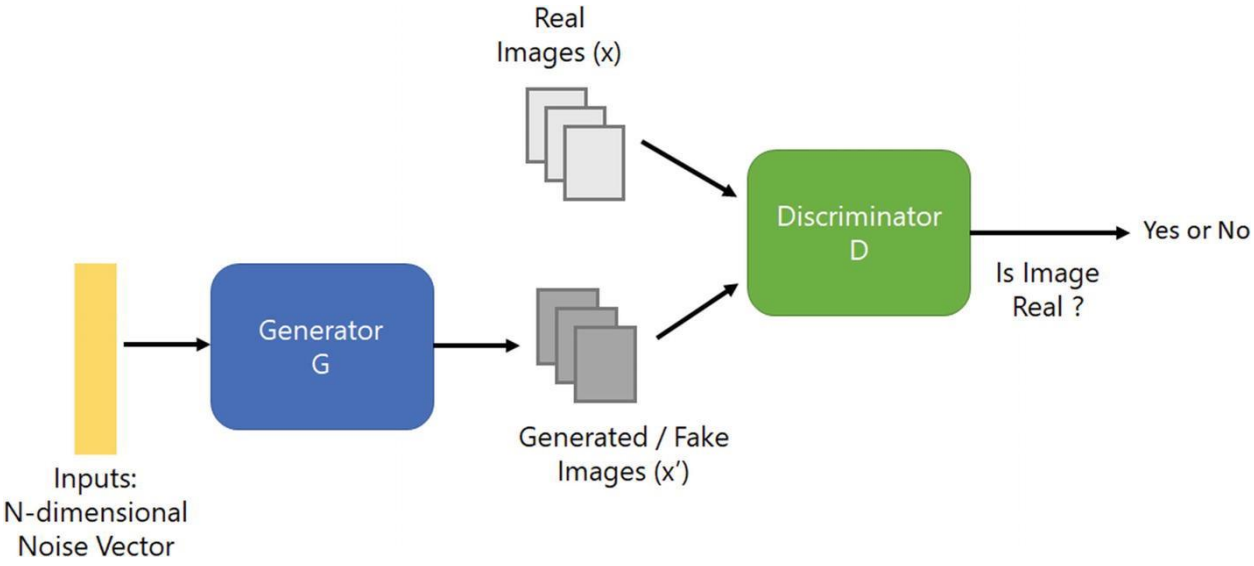
GANs stands for Generative adversarial networks, which is introduced by Ian Goodfellow in 2014. GANs is entirely new way of teaching computers how they do complex tasks through a generative process.

GANs have two components.

- A **Generator (**An artist**)** neural network.
- **A Discriminator (**An art critic**)** neural network.

Generator (An artist**)** generates an image. **Generator** does not know anything about the real images and learns by interacting with the **Discriminator**. The **Discriminator (**An art critic**)** determines whether an object is *"real"* and *"fake"* (usually represented by a value close to 1 or 0).

High-Level DCGAN Architecture Diagram



Original DCGAN architecture (Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*)* have four convolutional layers for **Discriminator** and "four fractionally-strided convolutions" layers for **Generator.**

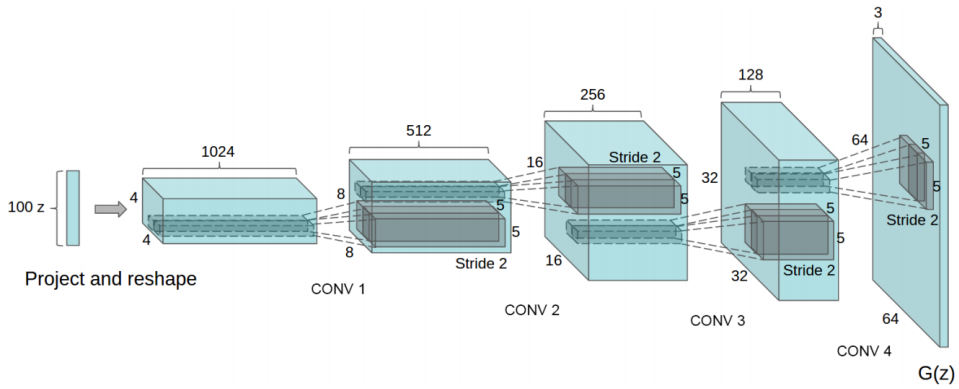


Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution Z is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a 64×64 pixel image. Notably, no fully connected or pooling layers are used.

The Discriminator Network

The **Discriminator** is a *"art critic"* who tries to distinguish between "real" and "fake" images. This is a convolutional neural network(CNN) for image classification.

The **Discriminator** is 4 layers strided convolutions with batch normalization (except its input layer) and leaky ReLU activations. **Leaky ReLU** helps the gradients flow easier through the architecture.

The Generator Network

The **Generator** is "An artist" who tries to create an image that looks as "*real*" as possible, to fool **Discriminator**.

The **Generator** is four layers fractional-strided convolutions with batch normalization (except its input layer) and use **Hyperbolic Tangent (*tanh*)**activation in the final output layer and **Leaky ReLU** in rest of the layers.

Training of DCGANs

The following steps are repeated in training

- First **Generator** creates some new examples.
- And the **Discriminator** is trained using real data and generated data.

- After **Discriminator** has been trained, models are trained together.
- The **Discriminator**'s weights are frozen, but its gradients are used in **Generator** model so that **Generator** can update it's weights.

Q2. Explain EnAET.

Answer:

EnAET: Self-Trained Ensemble AutoEncoding Transformations for Semi-Supervised Learning

Deep neural network has shown its sweeping successes in learning from large-scale labeled datasets like ImageNet. However, such successes hinge on the availability of large amount of labeled examples that are expensive to collect. Moreover, deep neural networks usually have large number of parameters that are prone to over-fitting. Thus, we hope that semi-supervised learning can not only deal with limited labels but also alleviate the over-fitting problem by exploring unlabeled data. In this paper, we successfully prove that both goals can be achieved by training the semi-supervised model built upon self-supervised representations.

Semi-Supervised Learning (SSL) has been extensively studied due to its great potential for addressing the challenge with limited labels. Most state-of-the-art approaches can be divided into two categories. One is confident predictions, which improves a model's confidence by encouraging low entropy prediction on unlabeled data. The other category imposes consistency regularization by minimizing discrepancy among the predictions by different models. The two approaches employ reasonable objectives since good models should make confident predictions that are consistent with each other.

On the other hand, a good model should also recognize the object even if it is transformed in different ways. With deep networks, this is usually achieved by training a model with augmented labeled data. However, unsupervised data augmentation is preferred to explore effect of various transformations on unlabeled data. For this reason, we will show that self-supervised representations learned from auto-ending the ensemble of spatial and non-spatial transformations can play a key role in significantly enhancing semi-supervised models. To this end, we will present an Ensemble of Auto-Encoding Transformations (AETs) to self-train semi-supervised classifiers with various transformations by combining the advantages of both existing semi-supervised approaches and the newly developed self-supervised representations.
Our contributions are summarized as follows:

- We propose first method that employs ensemble of both spatial and non-spatial transformations from both labeled and unlabeled data in the self-supervised fashion to train a semi-supervised network.
- We apply an ensemble of AETs to learn robust features under various transformations, and improve the consistency of the predictions on transformed images by minimizing their KL divergence.
- We demonstrate EnAET outperforms the state-of-the-art models on all benchmark datasets in both semi-supervised and fully-supervised tasks.
- We show in the ablation study that exploring an ensemble of transformations plays a key role in achieving new record performances rather than simply applying AET as a regularize.

Q3. What is Data Embedding Learning?

Answer:

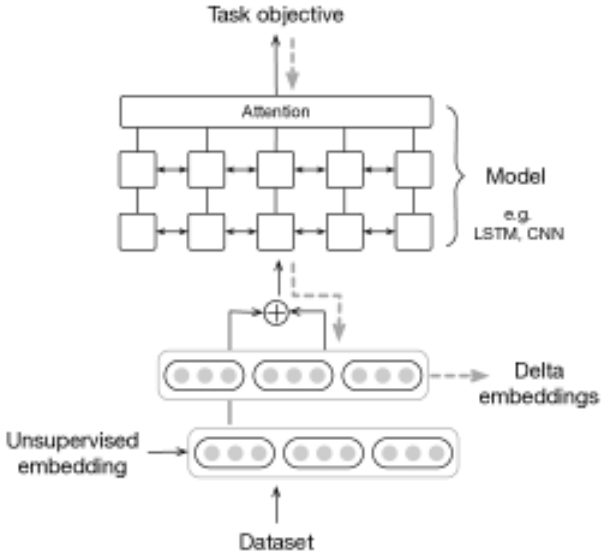
Unsupervised word embeddings have become basis for word representation in NLP tasks. Models such as skip-gram and Glove capture statistics of large corpus and have good properties that corresponds to semantics of word. However there are certain problems with unsupervised word embeddings, such as difficulty in modeling some fine-grained word semantics. For e.g., words in the same category but with different polarities are often confused because those words share common statistics in the corpus.

In supervised NLP(Natural Language Processing) tasks, these unsupervised word embeddings are often used in one of 2 ways: keeping fixed or using as initialization (fine-tuning). The decision is made based on amount of available training data in order to avoid overfitting. Nonetheless, underfitting with keeping fixed and certain degree of overfitting with fine-tuning is inevitable. Because this all are none optimization of word embeddings lacks control over the learning process, embeddings are not trained to an optimal point, which can result in suboptimal task performance.

In this paper, we propose delta embedding learning, the novel method that aims to address a above problems together: using regularization to find optimal fine-tuning of word embeddings. Better task performance can be reached with properly optimized embeddings. At the same time, regularized fine-tuning effectively combines semantics from supervised learning and unsupervised learning, which addresses some limitations in unsupervised embeddings and improves quality of embeddings.

Unlike retrofitting, which learns directly from lexical resources, our method provides the way to learn word semantics from supervised NLP(Natural Language Processing) tasks. Embeddings usually become task-specific and lose its generality when trained along with the model to maximize a task objective. Some approach tried to learn reusable embeddings from NLP(Natural Language Processing) tasks include multi-task learning, where one predict context words and external labels at the same time, and

specially designed gradient descent algorithms for fine-tuning .Our method learns reusable supervised embeddings by fine-tuning unsupervised embeddings on supervised task with a simple modification. The method also makes it possible to examine and interpret the learned semantics.



The aim of a method is to combine the benefits of unsupervised learning and supervised learning to learn better word embeddings. Unsupervised word embeddings like skip-gram, trained on large corpus (like Wikipedia), gives good-quality word representations. We use such embedding Wunsup as the starting point and learn a delta embedding wΔ on top of it:

$$\mathbf{w} = \mathbf{w}_{unsup} + \mathbf{w}_{\Delta}. \quad (1)$$

The unsupervised embedding Wunsup is fixed to preserve good properties of the embedding space and word semantics learned from large corpus. Delta embedding wΔ is used to capture discriminative word semantics from supervised NLP(Natural Language Processing) tasks and is trained together with model for the supervised task. In order to learn useful word semantics rather than task-specific peculiarities that results from fitting (or overfitting) the specific task, we impose L21 loss, one kind of structured regularization on wΔ:

$$loss = loss_{task} + c \sum_{i=1}^{n} (\sum_{j=1}^{d} w_{\Delta ij}^2)^{\frac{1}{2}} \quad (2)$$

The regularization loss is added as extra term to loss of supervised task.

The effect of L21 loss on wΔ has straightforward interpretation: to minimize total moving distance of word vectors in embedding space while reaching optimal task performance. The L2part of a regularization keeps change of word vectors small, so that it does not lose its original semantics. The L1 part of regularization induces sparsity on delta embeddings, that only small number of words get non-zero delta embeddings, while majority of words are kept intact. The combined effect is selective fine-tuning with moderation: delta embedding capture only significant word semantics that is contained in the training data of a task while absent in the unsupervised embedding.

Q4. Do you have any idea about Rookie?

Answer:

Rookie: A unique approach for exploring news archives

News archives offer the rich historical record. But if the reader or the journalist wants to learn about new topic with a traditional search engine, they must enter query and begin reading or skimming old articles one-by-one, slowly piecing together intricate web of people, organizations, events, places, topics, concepts and social forces that make up "the news."

We propose Rookie, which began as attempt to build a useful tool for journalists. With Rookie, a user's query generates an interactive timeline, a list of important related subjects, and summary of matching articles—all displayed together as a collection of interactive linked views. Users click and drag along the timeline to select certain date ranges, automatically regenerating the summary and subject list at interactive speed. The cumulative effect: users can fluidly investigate complex news stories as they evolve across time. Quantitative user testing shows how this system helps users better understand complex topics from documents and finish a historical sensemaking task 37% faster than with a traditional interface. Qualitative studies with student journalists also validate the approach.

We built the final version of Rookie following eighteen months of iterative design and development in consultation with reporters and editors. Because the system aimed to help real-world journalists, the software which emerged from the design process is dramatically different from similar academic efforts. Specifically, Rookie was forced to cope with limitations in the speed, accuracy and interpretability of current natural language processing techniques. We think that understanding and designing around such limitations is vital to successfully using NLP in journalism applications; a topic which, to our knowledge, has not been explored in prior work at the intersection of two fields.

How it works?

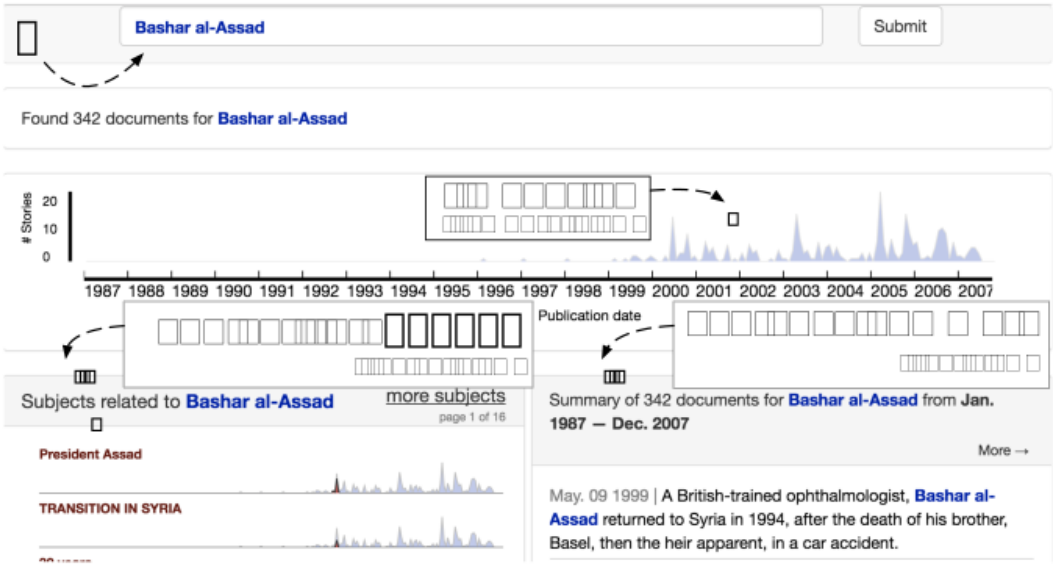
At any given time, Rookie's state is defined with the **user selection state**, the triple (**Q**,**F**,**T**) where:

- **Q** is the free text query string (e.g. “Bashar al-Assad”)
- **F** is related subject string (e.g. “30 years”) or is null
- **T** is time-span (e.g. Mar. 2000–Sep. 2000); by default, this is set to span of publication dates in the corpus.

Users first interact with Rookie by entering a query, **Q** into a search query bar using a web browser. For example, in below figure, a user seeking to understand the roots of the Syrian civil war has entered **Q** = “Bashar al-Assad”. In response, Rookie renders interactive time series visualization showing the frequency of matching documents from the corpus, a list of subjects in the matching documents, called subjects-sum and a textual summary of those documents, called sentence-sum.1In this example, the corpus is the collection of *New York Times* world news articles from 1987 to 2007 that contain the string “Syria”. All of the country-specific examples in this study are subsets of the same *New York Times* LDC corpus.

After entering **Q**, user might notice that “Bashar al-Assad” is mainly mentioned from 1999 onwards. To investigate, they might adjust time series slider to a spike in early mentions of Bashar al-Assad, **T** =Mar. 2000–Sep. 2000.

When user adjusts **T** to Mar. 2000–Sep. 2000, sentence-sum and subjects-sum change to reflect the new timespan below figure(c). subjects-sum now shows subjects like “TRANSITION IN SYRIA”,2Formatting from NYT section header style. “President Assad”, “eldest son” and “30 years” which are important to **Q** during **T**. (Bashar al-Assad’s father ruled for 30 years).



a) A user enters Q =”Bashar al-Assad” in order to learn more about the Syrian civil war.

At this point, user might explore further by investigating related subject, **F** ="President Assad"—clicking to select. sentence-sum now attempts to summarize the relationship between **Q**="Bashar al-Assad" and **F** ="President Assad" during **T** =Mar. 2000–Sep. 2000 figure(d). For instance, sentence-sum now shows the sentence: "Bashar al-Assad was born on Sept. 11, 1965, in Damascus, the third of President Assad's five children." If a user wants to understand this sentence in context, they can click sentence—which opens underlying document in the modal dialog.

F and **Q** are assigned red and blue colors throughout interface, allowing user to quickly scan for information. Bolding **Q** and **F** give additional clarity, and helps ensure that Rookie still work for colorblind users.

This example demonstrates how Rookie's visualization and summarization techniques work together to offer linked views of the underlying corpus. Linked views (a.k.a. multiple coordinated views) interfaces are common tools for structured information: each view displays the same selected data in a different dimension (e.g. a geographic map of a city which also shows a histogram of housing costs when a user selects a neighborhood). In Rookie's case, linked views display different levels of resolution. The time series visualization offers a **temporal view** of query-responsive documents, subjects-sum displays a medium-level **lexical view** of important subjects within the documents, and sentence-sum displays a low-level **text view** of parts of the underlying documents. The documents themselves, available by clicking extracted sentences, offer the most detailed level of zoom. Thus Rookie supports the commonly advised visualization pathway: "overview first, zoom and filter, and details on demand" (Shneiderman1996).

Note that we use term **summarization** to mean selecting short text, or sequence of short texts, to represent a body of text. By this definition, both subjects-sum and sentence-sum are a form of summarization, as each offers a textual representation of the corpus—albeit at two different levels of resolution, phrases and sentences.

Q5.SECRET: Semantically Enhanced Classification of Real-world Tasks

Answer:

Significant progress has been made in NLP(natural language processing) and supervised machine learning (ML) algorithms over the past 2 decades. NLP successes include machine translation, speech or emotion or sentiment recognition, machine reading, and social media mining. Hence, NLP(Natural Language Processing) is beginning to become widely used in real-world applications that include either text or speech. Supervised Machine Learning(ML) algorithms excel at modeling the data-label relationship while maximizing performance and minimizing energy consumption and latency.

Supervised ML algorithms train on feature-label pairs to model the application of interest and predict labels. The label involves semantic information. use this information through vector representations of words to find novel class within the dataset. Karpathy and Fei-Fei generate figure captions based on the collective use of image datasets and word embeddings. Such studies indicate that data feature and semantic relationship correlate well. However, current supervised (Machine Learning)ML algorithms do not utilize such correlations in the decision-making (or prediction) process. Their decisions are based on the feature-label relationship, while neglecting significant information hidden in labels, i.e., meaning-based (semantic) relationships among label. Thus, they are not able to exploit synergies between feature and semantic space.

In this article, we show above synergies can be exploited to improve prediction performance of Machine Learning(ML) algorithms. Our method, called SECRET, combines vector representations of label in semantic space with available data in feature space within various operations (e.g., ML hyperparameter optimization and confidence score computation) to make final decisions (assign labels to the datapoints). Since SECRET does not target any particular Machine Learning(ML) algorithm or data structure, it is widely applicable.

The main contributions of this article are as follows:

- We introduce the dual-space Machine Learning(ML) decision process called SECRET. It combines new dimension (semantic space) with traditional (single-space) classifiers that operate in feature space. Thus, SECRET not only utilizes available data-label pairs, but also take advantage of meaning-based (semantic) relationships among labels to perform classification for the given real-world task.
- We demonstrate the general applicability of SECRET on various supervised Machine Learning(ML) algorithms and wide range of datasets for various real-world tasks.
- We demonstrate advantages of SECRET's new dimension (semantic space) through detailed comparisons with traditional Machine Learning(ML) approaches that have same processing and information (except semantic) resources.
- We compare the semantic space Machine Learning(ML) model with traditional approaches. We shed light on how SECRET builds semantic space component and its impact on overall classification performance.

Q6. Semantic bottleneck for computer vision tasks

Answer:

Image-to-text tasks have made tremendous progress since the advent of deep learning(DL) approaches. The work presented in this paper builds on these new types of image-to-text functions to evaluate capacity of textual representations to semantically and fully encode visual content of images for

demanding applications, in order to allow prediction function to host semantic bottleneck somewhere in its processing pipeline. The main objective of semantic bottleneck is to play role of *explanation* of the prediction process since it offers opportunity to examine meaningfully on what ground will further predictions be made, and potentially decide to reject them either using human common-sense knowledge and experience, or automatically through dedicated algorithms. Such the explainable semantic bottleneck instantiates good tradeoff between prediction accuracy and interpretability.

Reliably evaluating the quality of explanation is not straightforward. In this work, we propose to evaluate the explainability power of the semantic bottleneck by measuring its capacity to detect the failure of the prediction function, either through an automated detector as, or through human judgment. Our proposal to generate surrogate semantic representation is to associate the global and generic textual image description (caption) and visual quiz in the form of small list of questions and answers that are expected to refine contextually the generic caption. The production of this representation is adapted to vision task and learned from the annotated data.

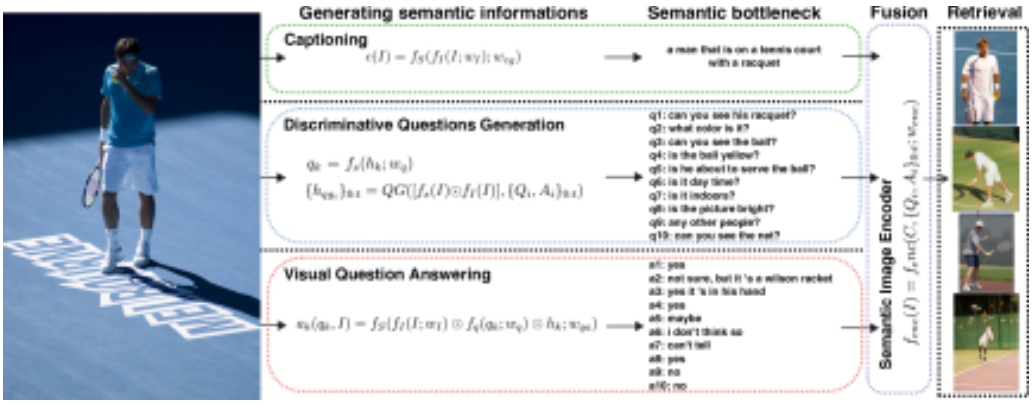


Figure : Semantic bottleneck approach: images are replaced by purely but rich textual representations, for tasks such as multi-label classification or image retrieval.

Q7. Gender Bias in Neural Natural Language Processing

Answer:

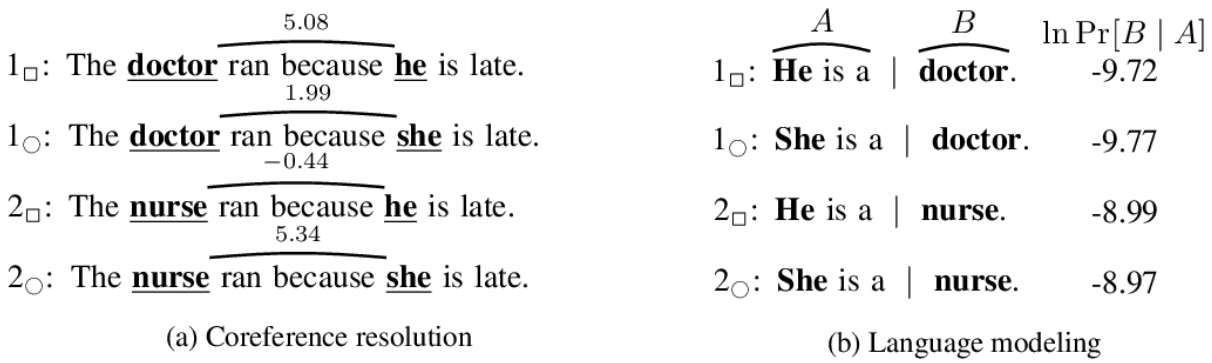
Natural language processing (NLP) with neural networks has grown in importance over last few years. They provide state-of-the-art(SOTA) models for tasks like coreference resolution, language modeling, and machine translation. However, since these models are trained on human language texts, natural question is whether they exhibit bias based on gender or other characteristics, and, if so, how should this bias be mitigated. This is a question that we address in this paper.

Prior work provides evidence of bias in autocomplete suggestions and differences in accuracy of speech recognition based on gender and dialect on popular online platforms. Word embeddings, initial pre-processors in many (Natural Language Processing)NLP tasks, embed words of a natural language into a vector space of limited dimension to use as their semantic representation. Observed that popular word embeddings including *word* exhibit gender bias mirroring stereotypical gender associations such as the eponymous "Man is to computer programmer as a Woman is to homemaker".

Yet the question of how to measure bias in general way for neural (Natural Language Processing)NLP tasks has not been studied. Our first contribution is general benchmark to quantify gender bias in variety of neural (Natural Language Processing)NLP tasks. Our definition of bias loosely follows idea of causal testing: matched pairs of individuals that differ in only the targeted concept (like gender) are evaluated by the model and the difference in outcomes (or scores) is interpreted as causal influence of the concept in scrutinized model. The definition is parametric in scoring function and the target concept. Natural scoring functions exist for number of neural natural language processing(NLP) tasks.

We instantiate definition for two important tasks—coreference resolution and language modeling. Coreference resolution is a task of finding words and expressions referring to the same entity in the natural language text. The goal of language modeling is to model distribution of word sequences. For neural coreference resolution models, we measure gender coreference score disparity between gender-neutral words and gendered words like the disparity between "doctor" and "he" relative to "doctor" and "she" pictured as edge weights in below Fig.. For language models, we measure disparities of emission log-likelihood of gender-neutral words conditioned on gendered sentence prefixes as is shown in below Fig. Our empirical evaluation with state-of-the-art(SOTA) neural coreference resolution and textbook RNN-based language models trained on benchmark datasets finds gender bias in these models.Note that these results have practical significance. Both coreference resolution and language modeling are core natural language processing(NLP) tasks in that they form the basis of many practical systems for information extraction, text generation, speech recognition and machine translation.

Next we turn our attention to mitigating the bias. Bolukbasi et al. (2016) introduced the technique for *debiasing* word embeddings which has been shown to mitigate unwanted associations in analogy tasks while preserving embedding's semantic properties. Given their spread use, a natural question is whether this technique is sufficient to eliminate bias from downstream tasks like coreference resolution and language modeling. As our 2nd contribution, we explore this question empirically. We find that while technique does reduce bias, residual bias is considerable. We further discover that debiasing models that make use of embeddings that are co-trained with their other parameters exhibit a significant drop in accuracy.



(a) Coreference resolution (b) Language modeling

Figure 1: Examples of gender bias in coreference resolution and language modeling as measured

Q8. DSReg: Using Distant Supervision as a Regularizer

Answer:

Consider the following sentences in a text classification task, in which we want to identify sentences containing revenue values:

- *S1: The revenue of education sector is 1 million. (positive)*
- *S2: The revenue of education sector increased a lot. (hard-negative)*
- *S3: Education is a fundamental driver of global development. (easy-negative)*

S1 is a positive example since it contains precise value for the revenue, while both S2 and S3 are negative because they do not have the concrete information of revenue value. However, since S2 is highly similar to S1, it is hard for a binary classifier to make a correct prediction on S2. As another example, in reading comprehension tasks like NarrativeQA (Kočiskỳ et al., 2018) or MS-MARCO (Nguyen et al., 2016), truth answers are human-generated ones and might not have exact matches in the original corpus. A commonly adopted strategy is to first locate similar sequences from the original corpus using a ROUGE-L threshold and then treat these sequences as a positive training examples. Sequences that are semantically similar but right below this threshold will be treated as negative examples and thus inevitably introduce massive noise in training.

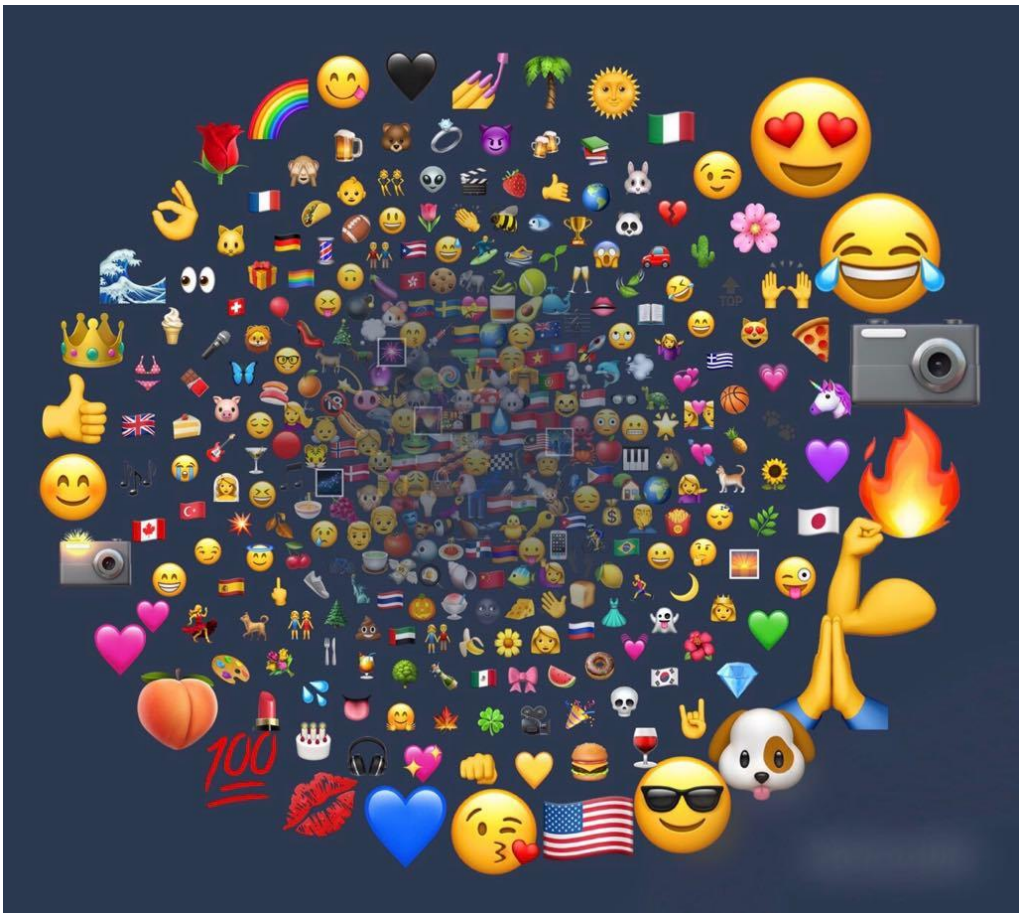
This problem is ubiquitous in a wide range of NLP tasks, i.e., when some of the negative examples are highly similar to the positive examples. We refer to these negative examples as *hard-negative examples* for the rest of this paper. Also, we refer to those negative examples that are not similar to the positive examples as *easy-negative examples*. If hard-negative examples significantly outnumber positive ones, features that they share in common will contribute significantly to the negative example category.

To tackle this issue, we propose using the idea of distant supervision to regulate the training. We first harvest hard-negative examples using distant supervision. This process can be done by the method as simple as using word overlapping metrics (e.g., ROUGE, BLEU or whether a sentence contains a certain keyword). With the harvested hard-negative examples, we transform the original binary classification task to a multi-task learning task, in which we jointly optimize the original target objective of distinguishing positive examples from negative examples along with an auxiliary objective of distinguishing hard-negative examples plus positive examples from easy-negative examples. For a neural network model, this goal can be easily achieved by using different softmax functions to readout the final-layer representations. In this way, both the difference and the similarity between positive examples and hard-negative examples can be captured by the model. It is worth noting that there are several key differences between this work and the mainstream works in distant supervision for relation extraction, at both the setup level and the model level. In traditional work on distant supervision for relation extraction, there is no training data initially and the distant supervision is used to get positive training data. In our case, we do have a labeled dataset, from which we retrieve hard-negative examples using the distant supervision.

Q9. What is Multimodal Emotion Classification?

Answer:

Emotion is any experience characterized by intense mental activity and certain degree of pleasure or displeasure. It primarily reflects all aspects of our daily lives, playing the vital role in our decision-making and relationships. In recent years, there have been growing interest in a development of technologies to recognize emotional states of individuals. Due to escalating use of social media, emotion-rich content is being generated at increasing rate, encouraging research on automatic emoji classification techniques. Social media posts are mainly composed of images and captions. Each of modalities has very distinct statistical properties and fusing these modalities helps us learn useful representations of data. Emotion recognition is the process that uses low-level signal cues to predict high-level emotion labels. With rapid increase in usage of emojis, researchers started using them as labels to train classification models.A survey conducted by secondary school students suggested that use of emoticons can help reinforce the meaning of the message. Researchers found that emoticons when used in conjunction with written message, can help to increase the “intensity” of its intended meaning.

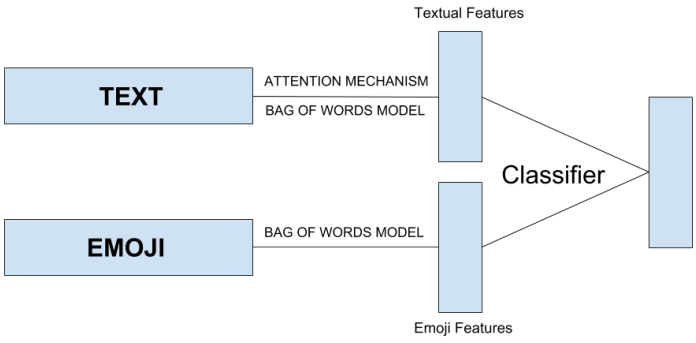


Emojis are being used for visual depictions of human emotions. Emotions help us to determine interactions among human beings. The context of emotions specifically brings out complex and bizarre social communication. These social communications are identified as judgment of other person's mood based on his emoji usage (Rajhi, 2017). According to the study made by Rajhi et al. (Rajhi, 2017), the real-time use of emojis can detect the human emotions in different scenes, lighting conditions as well as angles in real-time. Studies have shown that emojis when embedded with text to express emotion make tone and tenor of message clearer. This further helps in reducing or eliminating the chances of misunderstanding, often associated with plain text messages. A recent study proved that co-occurrence allows users to express their sentiment more effectively.

Psychological studies conducted in the early '80s provide us strong evidence that human emotion is closely related to the visual content. Images can both express and affect people's emotions. Hence it is intriguing and important to understand how emotions are conveyed and how they are implied by visual content of images. With this as the reference, many computer scientists have been working to relate and learn different visual features from images to classify emotional intent. Convolutional Neural Networks (CNNs) have served as baselines for major Image processing tasks. These deep Convolutional Neural

Networks(CNNs) combine the high and low-level features and classify images in an end-to-end multi layer fashion.

Earlier most researchers working in field of social Natural Language Processing(NLP) have used either textual features or visual features, but there are hardly any instances where researchers have combined both these features. Recent works by Barbieri et al.'s, Illendula et al.'s, on multimodal emoji prediction and Apostolova et al. work on information extraction fusing visual and textual features have shown that combining both modalities helps in improving the accuracies. While a high percentage of social media posts are composed of both images and caption, researchers have not looked at multimodal aspect for emotion classification. Consider post in above Firgure where a user is sad and posts the image when a person close to him leaves him. The image represents the disturbed heart and has the textual description "sometimes tough if your love leaves you #sad #hurting" conveys a sad emotion. Similarly emoji used conveys emotion of being depressed. We hypothesize that all the modalities from the social media post including visual, textual, and emoji features, contribute to predicting emotion of the user. Consequently, we seek to learn importance of different modalities towards emotion prediction task.



DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)

Day27

Q1. Learning to Caption Images with Two-Stream Attention and Sentence Auto-Encoder

Answer:

Understanding the world around us via visual representations, and communicating this extracted visual information via language is one of the fundamental skills of human intelligence. The goal of recreating a similar level of intellectual ability in artificial intelligence(AI) has motivated researchers from computer vision and natural language communities to introduce the problem of automatic image captioning. Image captioning, which is to describe the content of an image in the natural language, has been an active area of research and widely applied to video and image understanding in multiple domains. The ideal model for this challenging task must have two characteristics: understanding of an image content well and generating descriptive sentences which is coherent with the image content. Many image captioning methods propose various encoder-decoder models to satisfy these needs where encoder extracts the embedding from an image, and decoder generates the text based on the embedding. These two parts are typically built with a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN), respectively.

Fig.: This Image captioning decoder with two-stream attention and the Auxillary decoder “finds” and “localizes” relevant words better than general caption-attention baselines

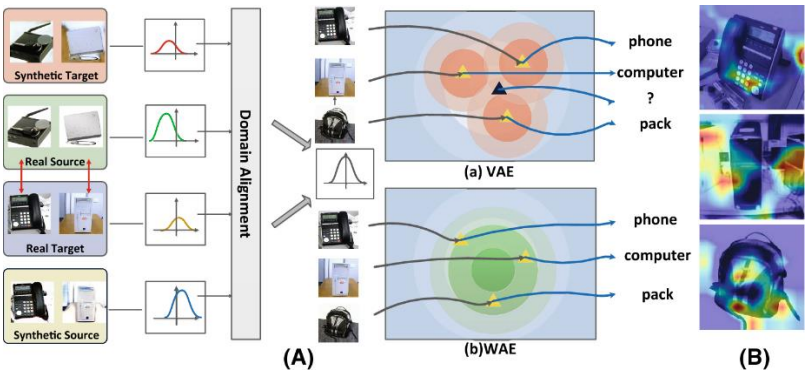
One of the challenging question in encoder-decoder architectures is how to design interface that controls the information flow between a CNN and RNN. While early work employs static representation for interface such that the CNN compresses an entire image into a fixed vector, and an RNN decodes representation into natural language sentences, this strategy is shown to perform poorly when target sentence is prolonged, and the image is reasonably cluttered. Inspired from, Xu *et al*. propose the powerful dynamic interface, namely attention mechanism, that identifies relevant parts of a image embedding to estimate the next word. RNN model then predicts the word based on the context vector associated with the related image regions and the previously generated words. The attentional interface is shown to obtain significant performance improvements over static one, and

since then, it has become the key component in all state-of-the-art(SOTA) image captioning models. Although this interface is substantially effective and flexible, it comes with critical shortcoming.

Nevertheless, visual representations that are learned by Convolutional Neural Network(CNNs) have been rapidly improving the state-of-the-art(SOTA) recognition performance in various image recognition tasks in past few years. They can still be inaccurate when applied to noisy images and perform poorly to describe their visual contents. Such noisy representations can lead to incorrect association between words and images regions and potentially drive the language model to poor textual descriptions. To address these shortcomings, we propose two improvements that can be used in standard encoder-decoder based image captioning framework.

First, we propose the novel and powerful attention mechanism that can more accurately attend to relevant image regions and better cope with ambiguities between words and image regions. It automatically identifies *latent categories* that capture high-level semantic concepts based on visual and textual cues, as illustrated in the second fig. The two-stream attention is modeled as a neural network where each stream specializes in orthogonal tasks: the first one soft-labels each image region with the latent categories, and the second one finds the most relevant area for each group. Then their predictions are combined to obtain a context vector that is passed to a decoder.

Second, inspired by sequence-to-sequence (seq2seq) machine translation methods, we introduce a new regularization technique that forces the image encoder coupled with the attention block to generate a more robust context vector for the following RNN model. In particular, we design and train an additional seq2seq sentence auto-encoder model (“SAE”) that first reads in a whole sentence as input, generates the fixed dimensional vector, then the vector is further used to reconstruct input sentence. SAE is trained to learn structure of the input (sentence) space in an offline manner, Once it is trained, we freeze its parameters and incorporate *only* its decoder part (SAE-Dec) to our captioning model (“IC”) as the auxiliary decoder branch. SAE-Dec is employed along with the original image captioning decoder (“IC-Dec”) to output target sentences during training and removed in test time. We show that the proposed SAE-Dec regularizer improves the captioning performance for IC-Dec and does not bring any additional computation load in test time.



Q2.Explain PQ-NET.

Answer:

PQ-NET: A Generative Part Seq2Seq Network for 3D Shapes.Learning generative models of 3D shapes is a crucial problem in both computer vision and computer graphics. While graphics are mainly concerned with 3D shape modeling, in inverse graphics, a significant line of work in computer vision, one aims to infer, often from a single image, a disentangled representation wrt 3D shape and scene structures. Lately, there has been a steady stream of works on developing deep neural networks for 3D shape generation using different shape representations, e.g., voxel grids, point clouds, meshes, and, most recently, implicit functions. However, most of these works produce *unstructured* 3D shapes, even though object perception is generally believed to be a process of a *structural understanding*, i.e., to infer shape parts, their compositions, and inter-part relations.

In this paper, we introduce a deep neural network that represents and generates 3D shapes via *sequential part assembly*, as shown in both Fig. In a way, we regard assembly sequence as a "sentence," which organizes and describes the parts constituting the 3D shape. Our approach is inspired, in part, by the resemblance between speech and shape perception, as suggested by the seminal work of Biederman on recognition-by-components (RBC). Another related observation is that the phase structure rules for language parsing, first introduced by Noam Chomsky, take on the view that sentence is both a linear string of words and a hierarchical structure with phrases nested in phrases. In the context of shape structure presentations, our network adheres to linear part orders, while other works have opted for *hierarchical* part organizations.

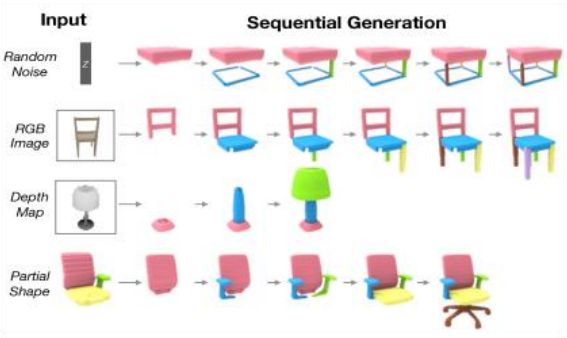


Fig 1: Our network, PQ-NET, learns 3D shape representation as a *sequential part assembly*. It can be adapted to generative tasks such as random 3D shape generation, single-view 3D reconstruction (from RGB or depth images), and shape completion.

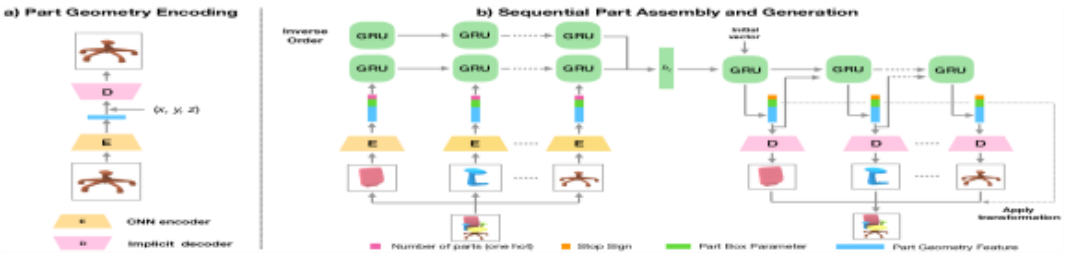


Fig2: The architecture of PQ-NET: our part Seq2Seq generative network for 3D shapes.

The input to our network is a 3D shape segmented into parts, where each part is first encoded into a feature representation using a part autoencoder; see Fig2(a). The core component of our network is a *Seq2Seq* autoencoder, which encodes a sequence of part features into the latent vector of fixed size, and the decoder reconstructs the 3D shape, one part at the time, resulting in sequential assembly; see Fig 2(b). With its part-wise Seq2Seq architecture, our network is coined *PQ-NET*. The latent space formed by Seq2Seq encoder enables us to adapt the decoder to perform several generative tasks,

including shape autoencoding, interpolation, new shape generation, and single-view 3D reconstruction, where all generated shapes are composed of meaningful parts.

As training data, we take the segmented 3D shapes from PartNet, which was built on ShapeNet. It is important to note that we do not enforce any particular part order or consistency across input shapes. The shape parts are always specified in the file following some linear order in the dataset; our network takes whatever part order that is in a shapefile. We train the part and Seq2Seq autoencoders of PQ-NET separately, either per shape category or across all shape categories, of PartNet.

Our part autoencoder adapts IM-NET to encode shape parts, rather than whole shapes, with the decoder producing an implicit field. The part Seq2Seq autoencoder follows a similar architecture as the original Seq2Seq network developed for machine translation. Specifically, the encoder is a bidirectional stacked recurrent neural network (RNN) that inputs two sequences of part features, in opposite orders, and outputs a latent vector. The decoder is also a stacked RNN, which decodes the latent vector representing the whole shape into a sequential part assembly.

PQ-NET is the first *fully generative* network that learns a 3D shape representation in the form of sequential part assembly. The only prior part sequence model was 3D-PRNN, which generates part boxes, not their geometry — our network jointly encodes and decodes part structure and geometry. PQ-NET can be easily adapted to various generative tasks, including shape autoencoding, novel shape generation, structured single-view 3D reconstruction from both RGB and depth images, and shape completion. Through extensive experiments, we demonstrate that performance and output quality of our network is comparable or superior to state-of-the-art generative models, including 3D-PRNN, IM-NET, and StructureNet.

Q3. What is EDIT?

Answer:

EDIT: Exemplar-Domain Aware Image-to-Image Translation

A scene can be expressed in various manners using sketches, semantic maps, photographs, and painting,artworks, to name just a few. The way that one portrays the scene and expresses his/her vision is the so-called style, which can reflect the characteristic of either a class/domain or a specific case. Image-to-image translation (I2IT) refers to the process of converting an image I of a particular style to another of the target style St with the content preserved. Formally, seeking the desired translator T can be written in the following form:

$$\min \quad \mathcal{C}(I_t, I) + \mathcal{S}(I_t, S_t) \quad \text{with} \quad I_t := \mathcal{T}(I, S_t), \tag{1}$$

where $\mathcal{C}(I_t, I)$ is to measure the content difference between the translated I_t and the original I, while $\mathcal{S}(I_t, S_t)$ is to enforce the style of I_t following that indicated by S_t.

Figure 1: Several results by the proposed EDIT. Our EDIT can take arbitrary exemplars as reference for translating images across multiple domains, including photo-painting, shoe-edge, and semantic map-facade in *one* model.

With the emergence of deep techniques, a variety of I2IT strategies have been proposed with excellent progress made over the last decade. In what follows, we briefly review contemporary works along two main technical lines, *i.e.*, one-to-one translation and many-to-many translation.

One-to-one Translation. Methods in this category aim at mapping images from a source domain to a target domain. Benefiting from the generative adversarial networks (GANs), the style of translated results satisfies the distribution of the target domain Y, achieved by S(It,St):=D(It,Y), where D(It, Y) represents a discriminator to distinguish if It is real with respect to Y. An early attempt by Isola *et al.* uses conditional GANs to learn mappings between two domains. The paired data supervise the content preservation, *i.e.*, C(It, I):=C(It, Igtt) with Igtt, the ground-truth target. However, in real-world situations, acquiring such paired datasets, if not impossible, is impractical. To alleviate the pressure from data, inspired by the concept of cycle consistency, cycleGAN in an unsupervised fashion was proposed, which adopts C(It,I):=C(FY⟶X(FX⟶Y(I)),I) with FX⟶Y the generator from domain X to Y and FY⟶X the reverse one. Afterward, StarGAN further extends the translation between two domains that cross multiple areas in a single model. Though the effectiveness of the mentioned methods has been witnessed by a broad spectrum of specific applications such as photo-

caricature, making up-makeup removal, and face manipulation, their main drawback comes from the nature of deterministic (uncontrollable) one-to-one mapping.

Many-to-many Translation. The goal of approaches in this class is to transfer the style controlled by an exemplar image to a source image with content maintained. Arguably, the most representative work goes to, which uses the pre-trained VGG16 network to extract the content and style features, then transfer style information by minimizing the distance between Gram matrices constructed from the generated image and the exemplar E, say S(It,St):=S(Gram(It),Gram(E)). Since then, numerous applications on the 3D scene, face swap, portrait stylization and font design have been done. Furthermore, several schemes have also been developed towards relieving limitations in terms of speed and flexibility. For example, to tackle the requirement of training for every new exemplar (style), Shen *et al.* built a meta-network, which takes in the style image and produces a corresponding image transformation network directly. Risser *et al.* proposed the histogram loss to improve the training instability. Huang and Belongie designed a more suitable normalization manner, *i.e.*, AdaIN, for style transfer. Li *et al.* replaced the Gram matrices with an alternative distribution alignment manner from the perspective of domain adaption. Johnson *et al.* trained the network with a specific style image and multiple content images while keeping the parameters at the inference stage. Chen *et al.* introduced a style-bank layer containing several filter-banks, each of which represents a specific style. Gu *et al.* proposed a style loss to make parameterized, and non-parameterized methods complement each other. Huang *et al.*designed a new temporal loss to ensure the style consistency between frames of a video. Also, to mitigate the deterministic nature of one-to-one translation, several works, for instance, advocated to separately take care of content c(I) and style s(I) subject to $I \simeq c(I) \circ s(I)$ with $\circ$ the combined operation. They manage to control the translated results by combining the content of an image with the style of the target, *i.e.*, $c(I) \circ s(E)$. Besides, one domain pair requires one independent model, their performance, as observed from comparisons, is inferior to our method in visual quality, diversity, and style preservation. Please see the above Fig. , For example produced by our approach that handles multiple domains and arbitrary exemplars in a unified model.

Q4. What is Doctor2Vec?

Answer:

Doctor2Vec: Dynamic Doctor Representation Learning for
Clinical Trial Recruitment

The rapid growth of electronic health record (EHR) data and other health data enables the training of complex deep learning models to learn patient representations for disease diagnosis, risk prediction,

patient subtyping, and medication recommendation. However, almost all current works focus on modeling patients. Deep neural networks for doctor representation learning are lacking.

Doctors play pivotal roles in connecting patients and treatments, including recruiting patients into clinical trials for drug development and treating and caring for their patients. Thus an effective doctor representation will better support a broader range of health analytic tasks. For example, identifying the right doctors to conduct the trial *site selection* to improve the chance of completion of the trials [hurtado2017improving] and doctor recommendation for patients.

In this work, we focus on studying the *clinical trial recruitment* problem using doctor representation learning. Current standard practice calculates the median enrollment rate. Enrollment rate of a doctor is the number of patients enrolled by a doctor to the trial. For the therapeutic area as the predicted enrollment success rate for whole participating doctors, which is often incorrect. Also, some develop a multi-step manual matching process for site selection, which is labor-intensive. Recently, deep neural networks were applied on site selection tasks via static medical concept embedding using only frequent medical codes and simple term matching to trials. Despite the success, two challenges remain open.

1. Existing works do not capture the time-evolving patterns of doctors experience and expertise encoded in EHR data of patients that the doctor have seen;
2. Current jobs learn a static doctor representation. However, in practice, given a trial for a particular disease, the doctor's experience of relevant diseases are more important. Hence the doctor representation should change based on the corresponding trial representation.

To fill the gap, we propose Doctor2Vec, which simultaneously learns i) doctor representations from longitudinal patient EHR data and ii) trial embedding from the multimodal trial description. In particular, Doctor2Vec leverages a dynamic memory network where the observations of patients seen by the doctor are stored as memory while trial embedding serves as queries for retrieving from the mind. Doctor2Vec has the following contributions.

1. **Patient embedding as a memory for dynamic doctors representation learning**. We represent doctors' evolving experience based on the representations from the doctors' patients. The patient representations are stored as a memory for dynamic doctor representation extraction.
2. **Trial embedding as a query for improved doctors selection**. We learn hierarchical clinical trial embedding where the unstructured trial descriptions were embedded using BERT [devlin2018bert]. The trial embedding serves as queries of the memory network and will attend over patient representation and dynamically assign weights based on the relevance of doctor experience and trial representation to obtain the final context vector for an optimized doctor representation for a specific test.

We evaluated Doctor2Vec using large scale real-world EHR and trial data for predicting trial enrollment rates of doctors. Doctor2Vec demonstrated improved performance in the site selection task over the best baselines by up to 8.7% in PR-AUC. We also showed that the Doctor2Vec embedding could be transferred to benefit data insufficiency settings, including trial recruitment in less populated/newly explored countries or for rare diseases. Experimental results show for the country transfer, Doctor2Vec achieved 13.7% relative improvement in PR-AUC over the best baseline. While for embedding transfer to unique disease trials, Doctor2Vec made 8.1%relative improvements in PR-AUC over the best benchmark.

Q5. Explain PAG-Net.

Answer:

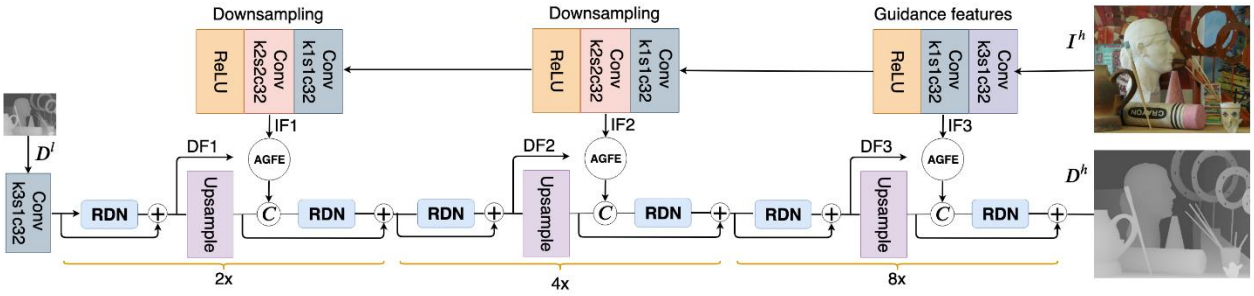
PAG: Progressive Attention Guided Depth Super-resolution Network

A geometric description of a scene, the high-quality depth map is useful in many computer vision applications, such as 3D reconstruction, virtual reality, scene understanding, intelligent driving, and robot navigation. Literature mainly contains two classes of techniques for depth information acquisition, which are passive methods and active sensors. Firstly, passive methods infer depth maps from the most widely used dense stereo matching algorithms, but they are time-consuming. Despite the advances in technology, the depth information from passive methods is still inaccurate in occluded and low-texture regions. The acquisition of high-quality depth maps is more challenging to obtain than RGB images.

Depth acquisition from active sensors has become increasingly popular in our daily life and ubiquitous to many consumer applications, due to their simplicity, portability, and inexpensive. Unlike passive methods, the depth of a scene can be acquired in real-time, and they are more robust in low-textured regions by low-cost sensors such as Time-of-Flight camera and Microsoft Kinect. Current sensing techniques measure depth information of a scene by using echoed light rays from the stage. Time-of-Flight sensor (ToF) is one of the mainstream types which computes depth at each pixel between camera and subject, by measuring the round trip time. Although depth-sensing technology has attracted much attention, it still suffers from several quality degradations.

Depth information captured by ToF sensors suffers from low-spatial resolutions (e.g., 176×144, 200×200 or 512×424) and noise when compared with the corresponding color images. Due to the offset between projector and sensor, depth maps captured by Microsoft Kinect sensors contain structural missing along discontinuities and random missing at homogeneous regions. These issues restrict the use of depth maps in the development of depth-dependent applications. High-quality depth is significant in many computer vision applications. Therefore, there is a need for restoration of

depth maps before using in applications. In this work, we consider the problem of depth map super-resolution from a given low-resolution depth map and its corresponding high-resolution color image.



Existing depth super-resolution (DSR) methods can be roughly categorized into three groups: filter design-based, optimization-based, and learning-based algorithms. Many of the existing techniques assumed that a corresponding high-resolution color image helps to improve the quality of depth maps and used aligned RGB image as guidance for depth SR. However, significant artifacts including texture copying and edge blurring, may occur when the assumption violated. The color texture will be transferred to the super-resolved depth maps if the smooth surface contains rich textures in the corresponding color image. Secondly, depth and color edges might not align in all the cases. Subsequently, it leads to ambiguity. Hence, there is a need for optimal guidance for the high-resolution depth map.

Although there have been many algorithms proposed in the literature for the depth super-resolution (DSR), most of them still suffer from edge-blurring and texture copying artifacts. In this paper, we offer a novel method for attention guided depth map super-resolution. It is based on dense residual networks and involves a unique attention mechanism. The attention used here to suppress the texture copying problem arises due to improper guidance by RGB images and transfer only the salient features from the guidance stream. The attention module mainly involves providing spatial attention to the guidance image based on the depth features. The entire architecture for the example of super-resolution by the factor of 8 is shown in Above Fig.

Q6. An End-to-End Audio Classification System based on Raw Waveforms and Mix-Training Strategy

Answer:

Sound is the indispensable medium for information transmission of surrounding environment. When some sounds happen, such as baby crying, glass breaking, and so on, we usually expect that we can "hear" sounds immediately, even if we are not around. In this case, an audio classification that aims to

predict whether an acoustic event appears has gained significant attention in recent years. It has many practical applications in remote surveillance, home automation, and public security.

In real life, an audio clip usually contains multiple overlapping sounds, and types of sounds are various, ranging from natural soundscapes to human activities. It is challenging to predict a presence or absence of audio events in an audio clip. Audio Set is the common large-scale dataset in this task, which contains about two million multi-label audio clips covering 527 classes. Recently, some methods have been proposed to learn audio tags on this dataset. Among them, a multi-level attention model achieved state-of-the-art(SOTA) performance, which outperforms Google's baseline. However, the shortcoming of these models is that the input signal is the published bottleneck feature, which causes information loss. Considering that the actual length of sound events is different and the handcrafted features may throw away relevant information at a short time scale, raw waveforms containing more valuable information is a better choice for multi-label classification. In the audio tagging task of DCASE 2017, 2018 challenge, some works [5, 6] combined handcrafted features with raw waveforms as input signal on a small dataset consisting of 17 or 41 classes. To our knowledge, none of the works proposes an end-to-end network taking raw waveforms as input in the Audio Set classification task.

In this paper, we propose a classification system based on two variants of ResNet, which directly extracts features from raw waveforms. Firstly, we use one-dimension (1D) ResNet for feature extraction. Then, two-dimension (2D) ResNet with multi-level prediction and attention structure is used for classification. For obtaining better classification performance further, a mix-training strategy is implemented in our system. In this training process, the network is trained with mixed data, which extends training distribution and then transferred to the target domain using raw data.

In this work, the main contributions are as follows:

1. The novel end-to-end Audio Set classification system is proposed. To best of our knowledge, it is first time to take raw waveforms as input on Audio Set and combine 1D ResNet with 2D ResNet for feature extraction at different time scales.
2. A mix-training strategy is introduced to improve the utilization of limited training data effectively. Experiments show that it is robust in multi-label audio classification compared to the existing data augmentation methods.

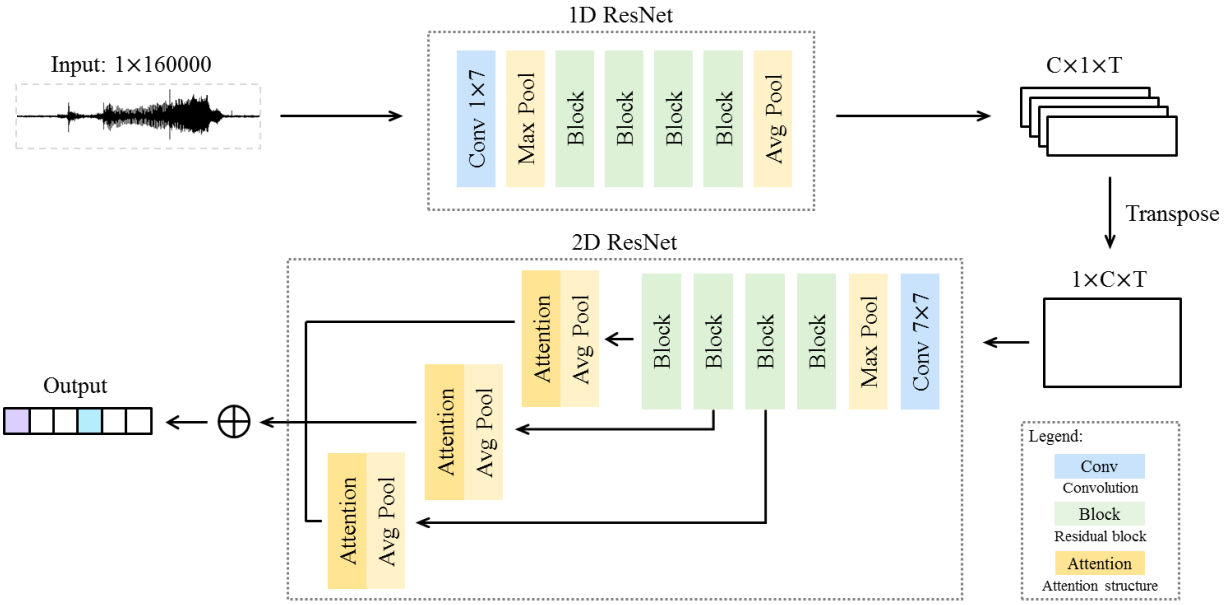


Figure 1: Architecture of the end-to-end audio classification network. The raw waveform (1D vector) is the input signal. First, the 1D ResNet is applied to extract audio features. Then, the elements are transposed from C×1×Tto 1×C×T. Finally, a 2D ResNet with a multi-level prediction structure performs audio classification. The output of network has multiple labels and is the mean of multi-level prediction results. The Block is composed of n bottleneck blocks, where n is related to a number of layers in ResNet.

Q7. What is Cnak? : Cluster Number Assisted *K-means*

Answer:

Cnak stands for Cluster Number Assisted K-means

In cluster analysis, it is required to group the set of data points in a multi-dimensional space so that data points in same group are more similar to each other than to those in other groups. These groups are called clusters. Various distance functions may be used to compute degree of dissimilarity or similarity among these data points. Typically Euclidean distance function is widely used in clustering. This unsupervised technique aims to increase homogeneity in the group and heterogeneity between groups. Several clustering methods with different characteristics have been proposed for different purposes. Some well-known methods include partition-based clustering, hierarchical clustering [Hierarchy1963], spectral clustering [onspectral2001], density-based clustering [DBSCAN]. However, they require the knowledge of cluster number for a given dataset a
priori [Lloyd57; onspectral2001; DBSCAN; DBCLASD; DENCLUE].

Nevertheless, estimation of the number of clusters is difficult problem as the underlying data distribution is unknown. Readers can find several existing techniques for determining cluster number in [survey_cluster_number2017; R3_Chiang2010]. We have followed the terminology used in R3_Chiang2010 for categorizing different methods for the prediction of cluster numbers. In this work, we choose to focus only on three approaches: 1) variance-based approach, 2) Structural approach, and 3) the Resampling approach. Variance-based plans are based on measuring compactness within a cluster. Structural approaches include between-cluster separation as well as within-cluster variance. We have chosen these approaches as they are either more suitable for handling big data, or appear in a comparative study by several researchers. Some well-known approaches are Calinski-Harabaz [CH], Silhouette Coefficient [sil], Davies-Bouldin [DB], Jump [jump], Gap statistic [gap], etc. These approaches are not appropriate for handling big data, as they are computationally intensive and require ample storage space. It requires a scalable solution [kluster2018; ISI_LL_LML2018] for identifying the number of clusters. Resampling-based approaches can be considered in such scenario. Recently, the concept of stability in clustering has become popular. A few methods [instability2012; CV_A] utilize the idea of clustering robustness against the randomness in the choice of sampled datasets to explore clustering stability.

Q8. What is D3S?

Answer:

D3S – A Discriminative Single Shot Segmentation Tracker.Visual object tracking is one of the core computer vision problems. The most common formulation considers the task of reporting the target location in each frame of the video given a single training image. Currently, the dominant tracking paradigm, performing best in evaluations [kristan_vot2017, kristan_vot2018], is correlation bounding box tracking where the target represented by a multi-channel rectangular template is localized by cross-correlation between the template and a search region.

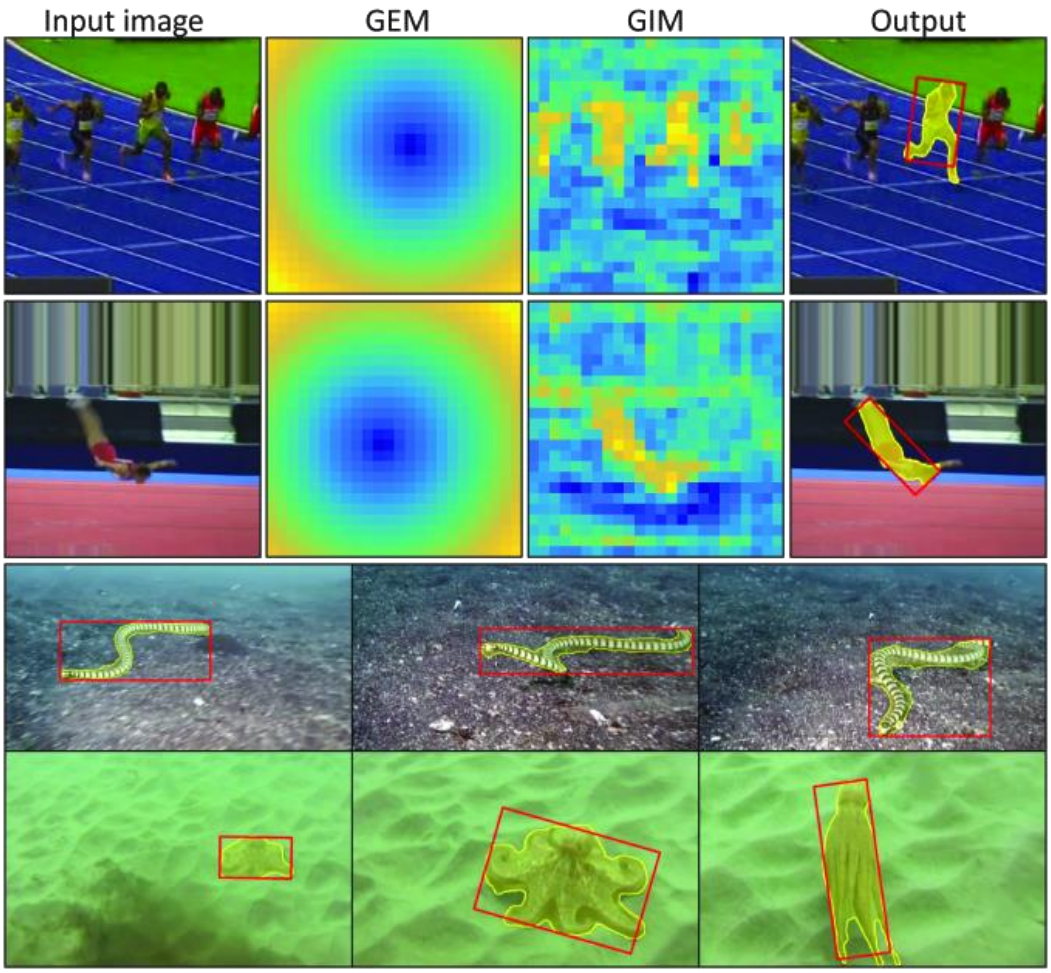


Figure 1: The D3S tracker represents the target by two models with complementary geometric properties, one invariant to a wide range of transformations, including non-rigid deformations (GIM - geometrically invariant model), the other assuming a rigid object with motion well approximated by a euclidean change (GEM - geometrically constrained Euclidean model). The D3S, exploiting the complementary strengths of GIM and GEM, provides both state-of-the-art localization and accurate segmentation, even in the presence of substantial deformation.

State-of-the-art template-based trackers apply an efficient brute-force search for target localization. Such a strategy is appropriate for low-dimensional transformations like translation and scale change but becomes inefficient for more general situations, e.g. such that induce an aspect ratio change and rotation. As a compromise, modern trackers combine approximate exhaustive search with sampling and bounding box refinement/regression networks for aspect ratio estimation. However, these approaches are restricted to axis-aligned rectangles.

Estimation of high-dimensional template-based transformation is unreliable when a bounding box is a sparse approximation of the target. This is common – consider, e.g. elongated, rotating, deformable

objects, or a person with spread out hands. In these cases, the most accurate and well-defined target location model is a binary per-pixel segmentation mask. If such output is required, tracking becomes the video object segmentation task recently popularized by DAVIS and YoutubeVOS challenges.

Unlike in tracking, video object segmentation challenges typically consider large target observed for less than 100 frames with low background distractor presence. Top video object segmentation approaches thus fare poorly in short-term tracking scenarios where the target covers a fraction of the image, substantially changes its appearance over a more extended period, and moves over a cluttered background. Best trackers apply visual model adaptation, but in the case of segmentation errors, it leads to irrecoverable tracking failure. Because of this, in the past, segmentation has played only an auxiliary role in template-based trackers, constrained DCF learning and tracking by 3D model

construction.

Recently, the SiamRPN tracker has been extended to produce high-quality segmentation masks in two stages – SiamRPN branches first localize the target bounding box, and then segmentation mask is computed only within this region by another branch. The two-stage processing misses the opportunity to treat localization and segmentation jointly to increase robustness. Another drawback is that a fixed template is used that cannot be discriminatively adapted to the changing scene.

We propose a new single-shot discriminative segmentation tracker, D3S, that addresses the limitations as mentioned above. Two discriminative visual models encode the target – one is adaptive and highly discriminative but geometrically constrained to a euclidean motion (GEM), while the other is invariant to a broad range of transformation (GIM, geometrically invariant model), see above Fig.

GIM sacrifices spatial relations to allow target localization under significant deformation. On the other hand, GEM predicts the only position but discriminatively adapts to the target and acts as a selector between possibly multiple target segmentations inferred by GIM. In contrast to related trackers [siammask_cvpr19, siamrpn_cvpr2019, atom_cvpr19], the primary output of D3S is the segmentation map computed in a single pass through the network, which is trained end-to-end for segmentation only.

Some applications and most tracking benchmarks require reporting the target location as a bounding box. As a secondary contribution, we propose an effective method for interpreting the segmentation mask as a rotated rectangle. This avoids an error-prone greedy search and naturally addresses changes in location, scale, aspect ratio, and rotation.

D3S outperforms all state-of-the-art trackers on most of the significant tracking benchmarks [kristan_vot2016, kristan_vot2018, got10k, muller_trackingnet] despite not being trained for bounding box tracking. In video object segmentation benchmarks [davis16, davis17], D3S

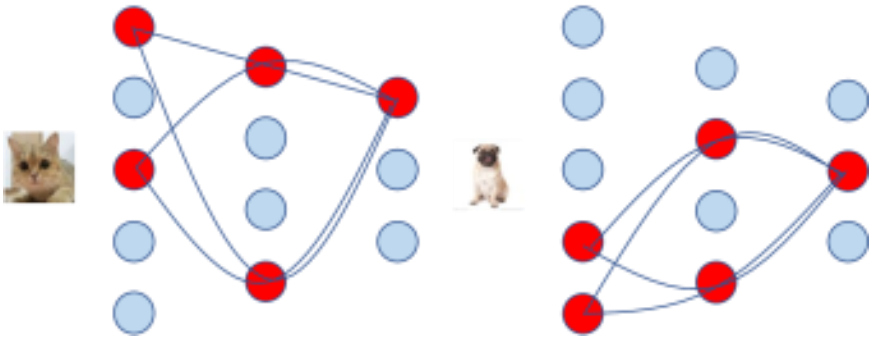
outperforms the leading segmentation tracker [siammask_cvpr19] and performs on par with top video object segmentation algorithms (often tuned to a specific domain), yet running orders of magnitude faster. Note that the D3S is not re-trained for different benchmarks – a single pre-trained version shows remarkable generalization ability, and versatility.PyTorch implementation will be made available.

Q9. What is DRNet?

Answer:

DRNet stands for Dissect and Reconstruct the Convolutional Neural Network via Interpretable Manners Convolutional neural networks (CNNs) have been broadly applied on various visual tasks due to its superior performance ([vgg], [resnet], [densenet]). But the huge computation burden prevents convolutional neural networks from running on mobile devices. Some works had been done to prune neural networks into smaller ones ([slimming], [pruning1], [pruning2]). Also, there are too many lightweight network structures ([mobilenet], [mobilenetv2], [shufflenet]) were proposed to adapt convolutional neural networks to computational limited mobile devices. However, these methods usually require running a whole pre-trained network, whatever the task is. i.e., the first task requires the discrimination power of cats and dogs, and the second task requires the discrimination power of apples and watermelons. If one has a CNN which was pre-trained on ImageNet, he must run the whole CNN on each task, which is usually time-consuming and computation wasted.

Our work focuses on an underlying problem, i.e., can we run only parts of a CNN? To achieve this goal, we need to find a method to dissect the whole network into pieces and reconstruct some of these pieces according to specific tasks. The reconstructed CNN should have a smaller computation cost and better performance. Meanwhile, the process of generating this substructure should be quick and easy. Therefore this technology can be applied on mobile devices and small robots such as cell-phones and uncrewed aerial vehicles. Using these technologies, these devices only need to store one complete CNN and some information about the substructure generating program. When specific tasks come, these devices can create a smaller substructure in an instant and run on it, rather than run the whole original CNN.



In this paper, we proposed a novel and interpretable algorithm to generate these smaller substructures. An interpretable way of CNN inspires our method. As shown in Figure 1: the original CNN has many channels, but not every channel is useful for the discrimination of every class. What we need to do is to find the channels relevant to every type and combine them for the specific task. This method looks similar to the previous work: structured network pruning ([slimming], [pruning3], [pruning4]). However, all of these pruning methods need fine-tuning, which is time-consuming and not allowed on mobile devices. And these pruning methods are usually lack of interpretability which is much needed by human-beings when using CNNs. Therefore, we do not mean to propose a pruning method and make CNN smaller, but to find the best channels for each class, and combine them for specific tasks. Our approach not only can be used on VGG and ResNet but also on some light structures such as MobileNetV2. Also, we make this process quick and interpretable.

DATA SCIENCE INTERVIEW PREPARATION

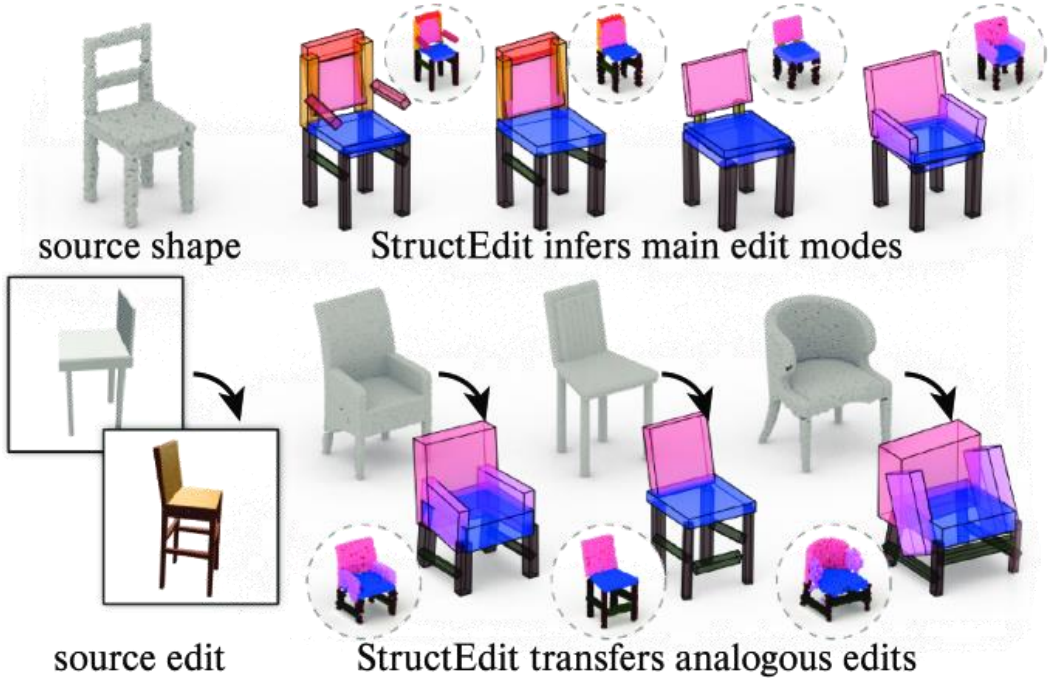
(30 Days of Interview Preparation)

Day28

Q1. Explain StructEdit(Learning Structural Shape Variations).

Answer:

The shapes of the 3D objects exhibit remarkable diversity, both in their compositional structure in terms of parts, as well as in geometry of the elements themselves. Yet human are remarkably skilled at imagining meaningful shape variation even from the isolated object instances. For example, having seen a new chair, we can easily imagine its *natural* changes with the different height back, a wider seat, with or without armrests, or with a diverse base. In this article, we investigate how to learn such shape variations directly from the 3D data. Specifically, given the shape collection, we are interested in two sub-problems: first, for any given shape, we want to discover main modes of edits, which can be inferred directly from shape collection; and second, given an example edit on one shape, we want to transfer edit to another shape in the group, as a form of analogy-based edit transfer. This ability is useful in several settings, including the design of individual 3D models, the consistent modification of the 3D model families, and the fitting of CAD models to noisy and incomplete 3D scans.



Above Fig: **Edit generation and transfer with StructEdit.**

We present the StructEdit, a method that learns the distribution of *shape differences* between structured objects that can be used to generate an ample variety of edits (in a first row); and accurately transfer edits between different purposes and across different modalities (on the second row). Edits can be both geometric and topological.

There are many challenges in capturing space of shape variations. First, individual shape can have different representations as image, surface meshes, or point clouds; second, one needs the unified setting for representing both continuous deformations as well as structural changes; third, shape edits are not directly expressed but are only implicitly contained in shape collections; and finally, learning the space of structural variations that is applicable to more than the single shape amounts to learning mappings between different shape edit distributions, since different shapes have various types and numbers of parts (like tables with or without leg bars).

In much of the existing literature on 3D machine learning(ML), 3D shapes are mapped to points in the representation space whose coordinates encode latent features of each shape. In such representation, shape edits are encoded as vectors in that same space – in other words, as differences between points representing shapes. Equivalently, we can think of forms as "anchored" vectors rooted at origin, while shape differences are "floating" vectors that can be transported around in shape space. This type of vector space arithmetic is commonly used [wu2016learning, achlioptas2017learning, wang2018global, gao2018automatic, xia2015realtime, Villegas_2018_CVPR], for example, in performing analogies, where the vector that is the difference of possible point A from point B is added to point C to produce an analogous point D. The challenge with this view in our setting is that while Euclidean spaces are perfectly homogeneous and vectors can be comfortably transported and added to points anywhere, shape spaces are far or less so. While for continuous variations, a vector space model has some plausibility, this is not so for structural variations: the "add arms" vector does not make sense for the point representing a chair that already has arms. We take the different approach. We consider embedding shapes differences or deltas *directly in their own latent space*, separate from general shape embedding space. Encoding and decoding such shape differences is always done through a VAE(variational autoencoder), in the context of the given source shape, itself encoded through the part hierarchy. This has the number of key advantages: (i) allows compact encodings of shape deltas, since in general, we aim to describe local variation; (ii) encourages network to abstract commonalities in shape variations across shape space; and (iii) adapts the edit to the provided source shape, suppressing the mode that are semantically implausible.

We have extensively evaluated the *StructEdit* on publicly available shape data sets. We introduce the new synthetic dataset with ground truth shape edits to quantitatively evaluate our method and compare it against baseline alternative. We then provide evaluation results on PartNet dataset [mo2019partnet] and provide ablation studies. Finally, we demonstrates that extension of our method allows the handling of both images and point cloud as shape sources, can predict plausible edit modes from single shape examples, and can also transfer example shape edit on one shape to other shapes in the collection.

Q2. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation

Answer:

As a vital part of human intelligence, emotional perceptivity is playing elemental role in various social communication scenarios, such as., education and healthcare systems. Recently, sensitive conversation generation has received an increasing amount of attention to address emotion factors in an end-to-end framework. However, as li2018syntactically revealed, that conventional emotional conversation system aims to produce more emotion-rich responses according to the specific user-input emotion, which inevitably leads to the psychological inconsistency problem.

Studies on social psychology suggest that empathy is the crucial step towards a more humanized human-machine conversation, which improves emotional perceptivity in emotion-bonding social activities. To design the intelligent automatic dialogue system, it is essential to make a chatbot empathetic within dialogues. Therefore, in this paper, we focus on a task of *empathetic dialogue generation*, which automatically tracks and understands the user's emotion at each turn in multi-turn dialogue scenarios.

Despite the achieved successes, obstacles to establishing the empathetic conversational system are still far beyond current signs of progress:

- Merely considering the sentence-level emotion while neglecting more precise token-level feelings may lead to insufficient emotion perceptivity. It is challenging to capture nuances of human emotion accurately without modeling multi-granularity emotion factors in the dialogue generation.
- Merely relying on the dialogue history but overlooking the potential of user feedback for the generated responses further aggravates the deficiencies above, which causes undesirable reactions.

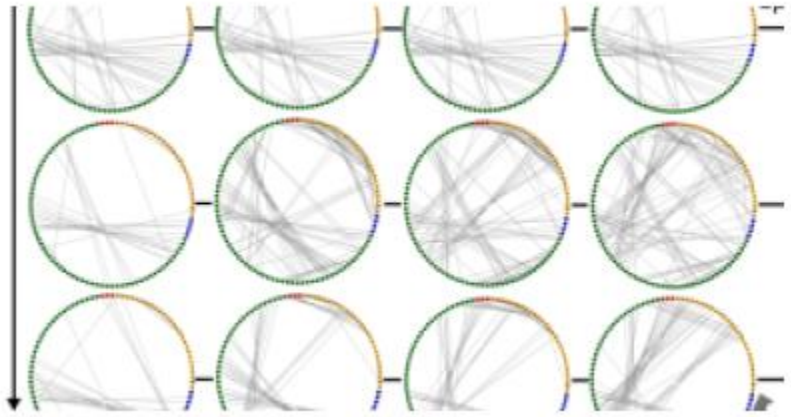
In this paper, we propose the multi-resolution adversarial empathetic dialogue generation model, named EmpGAN, to address the above challenges through generating more empathetic and appropriate responses. To capture nuances of user feelings sufficiently, EmpGAN make responses by taking both coarse-grained sentence-level and fine-grained token-level emotions into account. The response generator in EmpGAN dynamically understands sentiments along with a conversation to perceive a user's emotion states in multi-turn conversations. Furthermore, an interactive adversarial learning framework is augmented to take user feedback into account thoughtfully, where two interactive discriminators identify whether the generated responses evoke emotion perceptivity regarding both the dialogue history and user emotions.

In particular, the EmpGAN contains the empathetic generator and two interactive inverse discriminators. The empathetic generator is composed of three components: (i) A semantic understanding module based on Seq2Seq(sequence to sequence) neural networks that maintain the multi-turn semantic context. (ii) A multi-resolution emotion perception model captures the fine and coarse-grained emotion factors of each dialogue turn to build the emotional framework. (iii) An empathetic response decoder combines semantic and emotional context to produce appropriate responses in terms of both meaning and emotion. The two interactive inverse discriminator additionally incorporate the user feedback and corresponding emotional feedback as inverse supervised signal to induce the generator to produce a more empathetic response.

Q3. G-TAD: Sub-Graph Localization for Temporal Action Detection

Answer:

Video understanding has gained much attention from both academia and industry over recent years, given the rapid growth of videos published in online platforms. Temporal action detection is one of exciting but challenging tasks in this area. It involves detecting start and the end frames of action instances, as well as predicting their class label. This is onerous, especially in long untrimmed videos.



Video context is an important cue to detect actions effectively. Here, we refer to mean as frames that are outside the target action but carry valuable indicative information of it. Using video context to infer potential actions is natural for human beings. Empirical evidence shows that human can reliably predict or guess the occurrence of the specific type of work by only looking at short video snippets where the action does not happen. Therefore, incorporating context into temporal action detection has become important strategy to boost detection accuracy in the recent literature. Researchers have proposed various ways to take advantage of the video context, such as extending temporal action boundaries by the pre-

defined ratio, using dilated convolution to encode meaning into features, and aggregating definition feature implicitly by way of the Gaussian curve. All these methods only utilize temporal context, which follows or precedes an action instance in its immediate secular neighborhood. However, real-world videos vary dramatically in temporal extent, action content, and even editing preferences. The use of such temporal contexts does not fully exploit precious merits of the video context, and it may also impair detection accuracy if not adequately designed for underlying videos.

So, what properties characterize the desirable video context for accurate action detection? First, setting should be semantically or grammatically correlated to the target action other than merely temporally located in its vicinity. Imagine a case where we manually stitch an action clip into some irrelevant frames; the abrupt scene change surrounding the action would not benefit the action detection. On the other hand, snippets located at a distance from an operation but containing similar semantic content might provide indicative hints for detecting the action. Second, context should be content-adaptive rather than manually pre-defined. Considering the vast variation of videos, a framework that helps to identify different action instances could be changed in lengths and locations based on the video content. Third, context should be based on multiple semantic levels, since using only one form/level of meaning is unlikely to generalize well.

In this paper, we endow video context with all the above properties by casting action detection as a sub-graph localization problem based on a graph convolutional network (GCN). We represent each video sequence as the graph, each snippet as a node, each snippet-snippet correlation as an edge, and target actions associated with context as sub-graphs, as shown in Fig. 1. The meaning of a snippet is considered to be all snippets connected to it by an edge in a video graph. We define two types of edges — temporal corners and semantic edges, each corresponding to temporal context and grammatical context, respectively. Temporal edges exist between each pair of neighboring snippets, whereas semantic edges are dynamically learned from the video features at each GCN layer. Hence, the multi-level context of each snippet is gradually aggregated into the features of the snippet throughout the entire GCN. ResNeXt inspires the structure of each GCN block, so we name this GCN-based feature extractor GCNeXt.

The pipeline of our proposed Graph-Temporal Action Detection method, dubbed G-TAD, is analogous to faster R-CNN in object detection. There are two critical designs in G-TAD. First, GCNeXt, which generates context-enriched features, corresponds to the backbone network, analogous to a series of Convolutional Neural Network (CNN) layers in faster R-CNN. Second, to mimic RoI(region of interest) alignment in faster R-CNN, we design the sub-graph alignment (SGAlign) layer to generate a fixed-size representation for each sub-graph and embed all sub-graphs into same Euclidean space. Finally, we apply a classifier on the features of each sub-graph to obtain detection results. We summarize our contributions as follows.

(1) We present a novel GCN-based video model to exploit video context for effective temporal action detection fully. Using this video GCN representation, we can adaptively incorporate multi-level semantic meaning into the features of each snippet.

(2) We propose G-TAD, a new sub-graph detection framework, to localize actions in video graphs. G-TAD includes two main modules: GCNeXt and SGAlign. GCNeXt performs graph convolutions on video graphs, leveraging both temporal and semantic context. SGAlign re-arranges sub-graph features in the embedded space suitable for detection.

(3) G-TAD achieves state-of-the-art(SOTA) performance on two popular action detection benchmarks. On ActityNet-1.3, it achieves an average mAP of 34.09%. On THUMOS-14, it reaches 40.16%mAP@0.5, beating all contemporary one-stage methods.

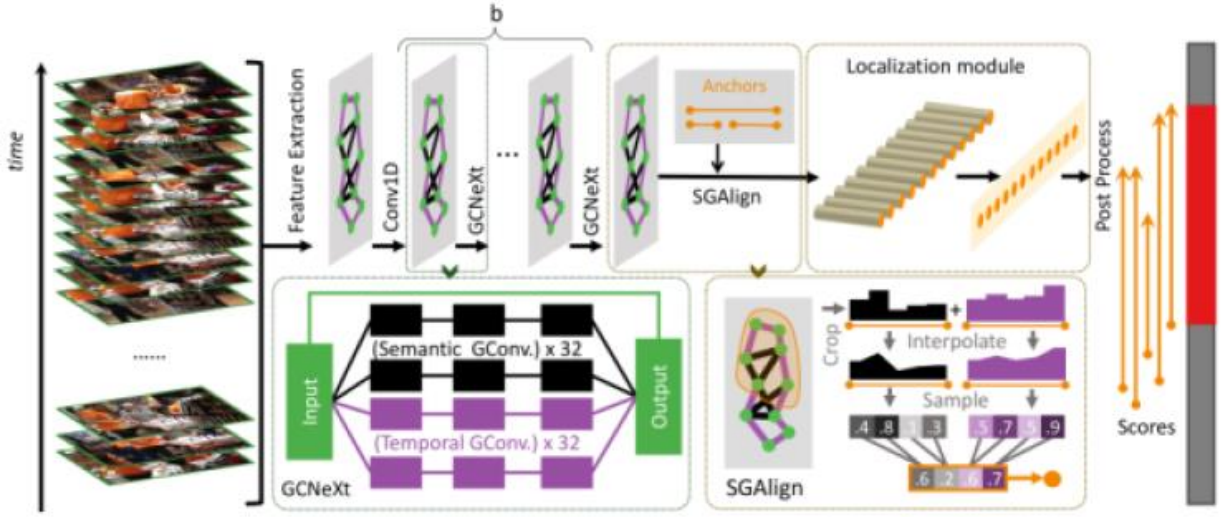
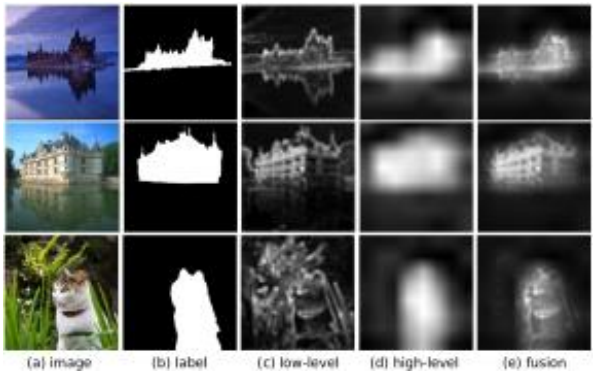


Fig: **Overview of G-TAD architecture.** The input of G-TAD is the sequence of snippet features. We first extract features using b=3 GCNeXt blocks, which gradually aggregate both temporal and multi-level semantic context. Semantic context, encoded in semantic edges, is dynamically learned from elements at each GCNeXt layer. Then we feed extracted features into the SGAlign layer, where sub-graphs defined by the set of anchors are transformed to a fixed-size representation in the Euclidean space. Finally, the localization module scores and ranks the sub-graphs for detection.

Q4. What is F3Net?

Answer:



F3Net is a combination of Fusion, Feedback, and Focus for Salient object detection (SOD) aims to estimate the significant visual regions of images or videos and often serves as the pre-processing step for many downstream vision tasks. Earlier SOD algorithms mainly rely on heuristic priors (*e.g.,* color, texture and contrast) to generate saliency maps. However, these hand-craft features can hardly capture high-level semantic relations and context information. Thus they are not robust enough to complex scenarios. Recently, convolutional neural networks (CNNs) have demonstrated its powerful feature extraction capability in visual feature representation. Many CNNs-based models have achieved remarkable progress and pushed the performance of SOD to a new level. These models adopt the encoder-decoder architecture, which is simple in structure and computationally efficient. The encoder usually is made up of a pre-trained classification model (e.g., ResNet and VGG), which can extract multiple features of different semantic levels and resolutions. In the decoder, extracted features are combined to generate saliency maps.

However, there remain two significant challenges in accurate SOD. First, features of different levels have different distribution characteristics. High-level features have rich semantics but lack precise location information. Low-level features have rich details but full of background noises. To generate better saliency maps, multi-level features are combined. However, without delicate control of the information flow in the model, some redundant features, including noises from low-level layers and coarse boundaries from high-level layers, will pass in and possibly result in performance degradation. Second, most of the existing models use binary cross-entropy that treats all pixels equally. Intuitively, different pixels deserve different weights, *e.g.*, pixels at the boundary are more discriminative and should be attached with more importance. Various boundary losses have been proposed to enhance the boundary detection accuracy, but considering only the boundary pixels is not comprehensive enough since there are lots of pixels near the boundaries prone to wrong predictions. These pixels are also essential and should be assigned with

larger weights. In consequence, it is essential to design a mechanism to reduce the impact of inconsistency between features of different levels and assign larger weights to those significant pixels.

To address the above challenges, we proposed a novel SOD framework, named F3Net, which achieves remarkable performance in producing high-quality saliency maps. First, to mitigate the discrepancy between features, we design a cross-feature module (CFM), which fuses elements of different levels by element-wise multiplication. Different from addition and concatenation, CFM takes a selective fusion strategy, where redundant information will be suppressed to avoid the contamination between features, and important features will complement each other. Compared with traditional fusion methods, CFM can remove background noises and sharpen boundaries, as shown in Fig. 1. Second, due to downsampling, high-level features may suffer from information loss and distortion, which can not be solved by CFM. Therefore, we develop the cascaded feedback decoder (CFD) to refine these features iteratively. CFD contains multiple sub-decoders, each of which includes both bottom-up and top-down processes. For the bottom-up method, multi-level features are aggregated by CFM gradually. For the top-down process, aggregated features are feedback into previous features to refine them. Third, we propose the pixel position-aware loss (PPA) to improve the commonly used binary cross-entropy loss, which treats all pixels equally. Pixels located at boundaries or elongated areas are more complicated and discriminating. Paying more attention to these hard pixels can further enhance model generalization. PPA loss assigns different weights to different pixels, which extends binary cross-entropy. The weight of each pixel is determined by its surrounding pixels. Hard pixels will get larger weights, and easy pixels will get smaller ones.

To demonstrate the performance of F3Net, we report experimental results on five popular SOD datasets and visualize some saliency maps. We conduct a series of ablation studies to evaluate the effect of each module. Quantitative indicators and visual results show that F3Net can obtain significantly better local details and improved saliency maps. Codes have been released. In short, our main contributions can be summarized as follows:

- We introduce the cross feature module to fuse features of different levels, which can extract the shared parts between features and suppress each other's background noises and complement each other's missing parts.
- We propose the cascaded feedback decoder for SOD, which can feedback features of both high resolutions and high semantics to previous ones to correct and refine them for better saliency maps generation.
- We design pixel position-aware loss to assign different weights to different positions. It can better mine the structure information contained in the features and help the network focus more on detail regions.

- Experimental results demonstrate that the proposed model F3Net achieves the state-of-the-art performance on five datasets in terms of six metrics, which proves the effectiveness and superiority of the proposed method.

Q5.Natural Language Generation using Reinforcement Learning with External Rewards

Answer:

We aim to develop models that are capable of generating language across several genres of text, conversational texts, and restaurant reviews. After all, humans are adept at both. Extant NLG(natural language generation) models work on either conversational text (like movie dialogues) or longer text (e.g., stories, reviews) but not both. Also, while the state-of-the-art(SOTA) in this field has advanced quite rapidly, current model is prone to generate language that is short, dull, off-context. More importantly, a generated language may not adequately reflect affective content of the input. Indeed, humans are already adept at this task, as well. To address these research challenges, we propose the RNN-LSTM architecture that uses an encoder-decoder network. We also use reinforcement learning(RL) that incorporates internal and external rewards. Specifically, we use emotional appropriateness as an internal reward for the NLG(Natural Language Generation) system – so that the emotional tone of generated language is consistent with the emotional tone of prior context fed as input to model. We also effectively incorporate usefulness scores as external rewards in our model. Our main contribution is the use of distantly labeled data in architecture that generates coherent, affective content and we test the architecture across two different genres of text.

What are the problem statement and their intuition?

Our aim is to take advantage of reinforcement learning(RL) and external rewards during the process of language generation. Complementary to this goal, we also aim to generate language that has same emotional tone as the other input. Emotions are recognized as functional in decision-making by influencing motivation and action selection. However, the external feedback and rewards are hard to come by for language generation; these would need to be provided through crowdsourcing judgment on generated responses *during* generation process, which makes process time-consuming and impractical. To overcome this problem, we look for distance labeling and use labels provided in training set as a proxy for human judgment on generated responses. Particularly, we incorporate usefulness scores in a restaurant review corpus as the proxy for external feedback.

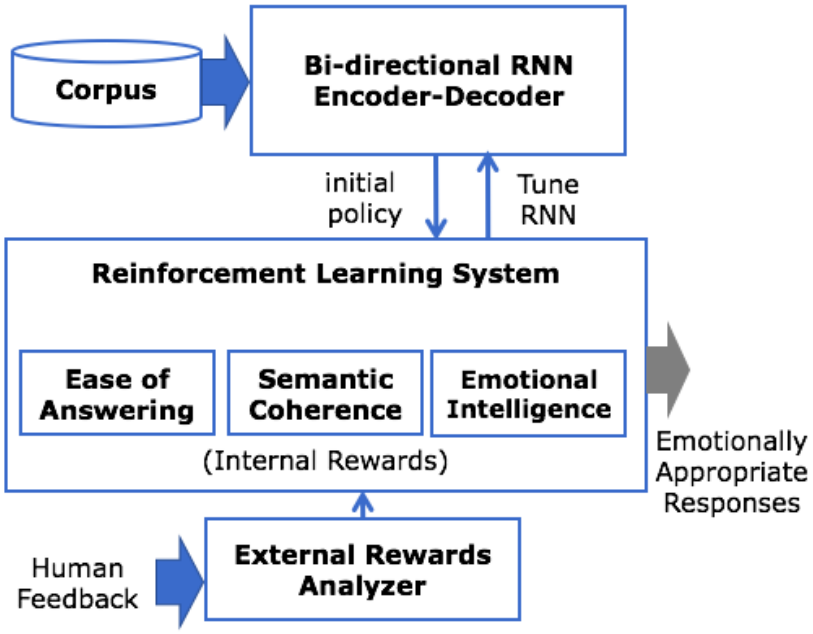


Fig. 1: Overall Architecture of the system showing internal and external rewards using reinforcement learning

Q6. LaFIn: Generative Landmark Guided Face Inpainting

Answer:

Image inpainting (*a.k.a.* image completion) refers to the process of reconstructing lost or deteriorated regions of images, which can be applied to, as a fundamental component, various tasks such as image restoration and editing. Undoubtedly, one expects the completed result to be realistic so that the reconstructed regions can be hardly perceived. Compared with natural scenes like oceans and lawns, manipulating faces, the focus of this work, is more challenging. Because the faces have much stronger topological structure and attribute consistency to preserve. Figure 1 shows three such examples. Very often, given the observed clues, human beings can easily infer what the lost parts possibly, although inexactly, look like. As a consequence, a slight violation of the topological structure and the attribute consistency in the reconstructed face highly likely leads to a significant perceptual flaw. The following defines the problem:

Definition:

Face Inpainting. Given a face image, I *with corrupted regions masked by* M. *Let* $\overline{\mathbf{M}}$ *designate the complement of* M *and* $\circ$ *the Hadamard product. The goal is to fill the target part with semantically meaningful and visually continuous information to the observed part. In other words, the completed*

result ^I:=M∘^I+¯¯¯M∘I *should preserve the topological structure among face components such as eyes, nose, and mouth, and the attribute consistency on like pose gender, ethnicity, and expression.*

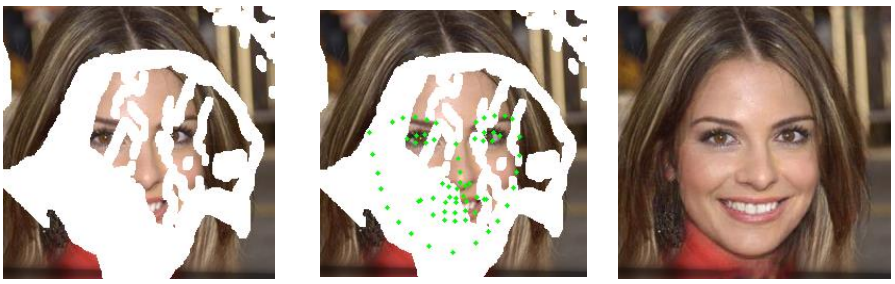


Figure 1: Three face completion results by our method. From left to right: corrupted inputs, plus landmarks predicted from the inputs, and our final results, respectively.

Q7. Image2StyleGAN++: How to Edit the Embedded Images?

Answer:

(i) (ii) (iii) (iv)

From above fig: (i) and (ii): input images; (iii): the "two-face" generated by naively copying the left half from (i) and the right half from (ii); (iv): the "two-face" created by our Image2StyleGAN++ framework.

Recent GANs demonstrated that synthetic images could be generated with very high quality. This motivates research into embedding algorithms that embed a given photograph into a GAN latent space. Such embedding algorithms can be used to analyze the limitations of GANs, do image inpainting, local image editing, global image transformations such as image morphing and expression transfer, and few-shot video generation.

In this paper, we propose to extend a very recent embedding algorithm, Image2StyleGAN. In particular, we would like to improve this previous algorithm in three aspects. First, we noticed that the embedding quality could be further improved by including Noise space optimization into embedding framework. The key insight here is that stable Noise space optimization can only be conducted if optimization is done sequentially with W+ space and not jointly. Second, we would like to improve capabilities of the embedding algorithm to increase the local control over the embedding. One way to improve local authority is to include mask in embedding algorithm with undefined content. The goal of the embedding algorithm should be to find a plausible embedding for everything outside the mask, while filling in reasonable semantic content in the masked pixels.

Similarly, we would like to provide the option of approximate embeddings, where the specified pixel colors are only a guide for the embedding. In this way, we aim to achieve high-quality embeddings that can be controlled by user scribbles. In the third technical part of the paper, we investigate the combination of embedding algorithm and direct manipulations of the activation maps (called activation tensors in our article).

Q8. oops! Predicting Unintentional Action in Video

Answer:

From just a glance at the video, we can often tell whether a person's action is intentional or not. For example, the Below figure shows a person attempting to jump off a raft, but unintentionally tripping into the sea. In a classic series of papers, developmental psychologist Amanda Woodward demonstrated that children learn this ability to recognize the intentionality of action during their first year. However, predicting the intention behind action has remained elusive for machine vision. Recent advances in action recognition have primarily focused on predicting the physical motions and atomic operations in the video, which captures the means of action but not the intent of action.

We believe a key limitation for perceiving visual intentionality has been the lack of realistic data with natural variation of intention. Although there are now extensive video datasets for action recognition, people are usually competent, which causes datasets to be biased towards successful outcomes. However, this bias for success makes discriminating and localizing visual intentionality challenging for both learning and quantitative evaluation.

Fig: **The oops! Dataset: Each pair of frames shows an example of intentional and unintentional action in our dataset. By crawling publicly available ‘‘fail’’ videos from the web, we can create a diverse and in-the-wild dataset of accidental action. For example, at the bottom-left corner shows a man failing to see gate arm, and at the top-right shows two children playing competitive games where it is inevitable; one person will fail to accomplish their goal.**

We introduce a new annotated video dataset that is abundant with unintentional action, which we have collected by crawling publicly available ‘‘fail’’ videos from the web. From the above figure shows some examples, which cover in-the-wild situations for both intentional and unintentional action. Our video dataset, which we will publicly release, is both large (over 50 hours of video) and diverse (covering hundreds of scenes and activities). We annotated videos with the temporal location at which the video transitions from intentional to unintentional action. We define three tasks on this dataset: classifying the intentionality of action, localizing the change from intentional to unintentional, and forecasting onset of unintentional action shortly into the future.

To tackle these problems, we investigate several visual clues for learning with minimal labels to recognize intentionality. First, we propose a novel self-supervised task to learn to predict the speed of the video, which is incidental supervision available in all unlabeled videos for learning the action representation. Second, we explore the predictability of temporal context as a clue to learn features, as unintentional action often deviates from expectation. Third, we study an order of events as a clue to recognize intentionality, since intentional action usually precedes unintentional action.

Experiments and visualizations suggest that unlabeled video has intrinsic perceptual clues to recognize intentionality. Our results show that, while each self-supervised task is useful, and learning to predict the speed of video helps the most. By ablating model and design choices, our analysis also suggests that models do not rely solely on low-level motion clues to solve unintentional action prediction. Moreover, although human's consistency in our dataset is high, there is still a large gap in performance between our models and human agreement, underscoring that analyzing human goals from videos remains the

fundamental challenge in computer vision(OpenCV). We hope this dataset of unintentional and unconstrained action can provide the pragmatic benchmark of progress.

Q9. FairyTED: A Fair Rating Predictor for TED Talk Data

Answer:

In recent times, artificial intelligence is being used for inconsequential decision making. Governments make use of it in the criminal justice system to predict recidivism [brennan2009evaluating, tollenaar2013method], which affects the decision about bail, sentencing, and parole. Various firms are also using machine learning algorithms to examine and filter resumes of job applicants [nguyen2016hirability, chen2017automated, naim2016automated], which is crucial for the growth of a company. Machine learning algorithms are also being used to evaluate human's social skills, such as presentation performance [Chen2017a, Tanveer2015], essay grading.

To solve such decision-making problems, machine learning algorithms are trained on massive datasets that are usually collected in the wild. Due to difficulties in the manual curation or adjustment over large datasets, the data likely capture unwanted bias towards the underrepresented group based on race, gender, or ethnicity. Such bias results in unfair decision-making systems, leading to unwanted and often catastrophic consequences to human life and society. For example, the recognition rates of pedestrians in autonomous vehicles are reported to be not equally accurate for all groups of people [wilson2019predictive]. Matthew et al. [kay2015unequal]showed that societal bias gets reflected in the machine learning algorithms through a biased dataset and causes representational harm for occupations. Face recognition is not as useful for people with different skin tones. Dark-skinned females have 43 times higher detection error than light-skinned males.

In this work, we propose a predictive framework that tackles the issue of designing a fair prediction system from biased data. As an application scenario, we choose the problem of fair rating prediction in the TED talks. TED talks cover a wide variety of topics and influence the audience by educating and inspiring them. Also, it consists of speakers from a diverse community with imbalances in age, gender, and ethnic attributes. The ratings are provided by spontaneous visitors to the TED talk website. A machine learning algorithm trained solely from the audience ratings will have a possibility of the predicted score being biased by sensitive attributes of the speakers.

It is a challenging problem because numerous factors drive human behavior and hence have huge variability. It is challenging to know the way these factors interact with each other. Also, uncovering the true interaction model may not be feasible and often expensive. Even though the sharing platforms such as YouTube, Massive Open Online Courses (MOOC), or ted.com make it possible to collect a large amount of observational data, these platforms do not correct for bias and unfair ratings.

In this work, we utilize *causal models* [pearl2009causal] to define possible dependencies between attributes of the data. We then address the problem of not knowing true interaction model by averaging outputs of predictors across several possible causes. Further, using these causal models, we generate *counterfactual samples* of sensitive attributes. These counterfactual samples are the key components in our fair prediction framework (adapted from kusner2017counterfactual russell2017worlds) and help reducing bias in ratings wrt sensitive attributes. Finally, we introduce the novel metric to quantify degree of fairness employed by our FairyTED pipeline. To best of our knowledge, FairyTED is first fair prediction pipeline for public speaking datasets and can be applied to any dataset of similar grounds. Apart from theoretical contribution, our work also has practical implications in helping both the viewers and organizers make informed and unbiased choices for the selection of talks and speakers.

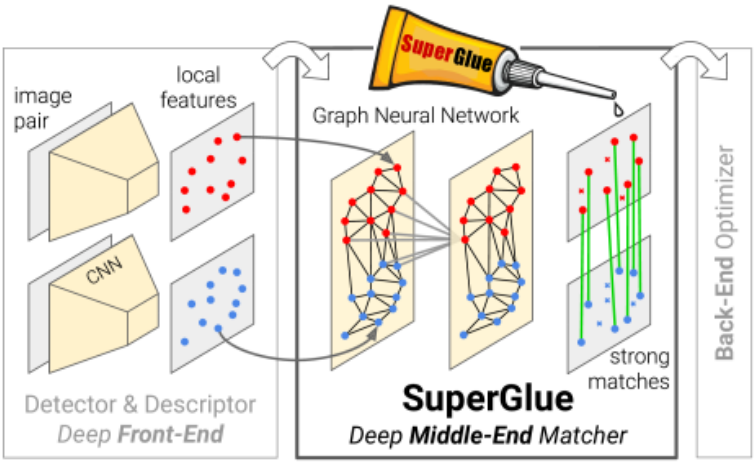
DATA SCIENCE INTERVIEW PREPARATION

(30 Days of Interview Preparation)

Day29

Q1. What is SuperGlue?

Answer:



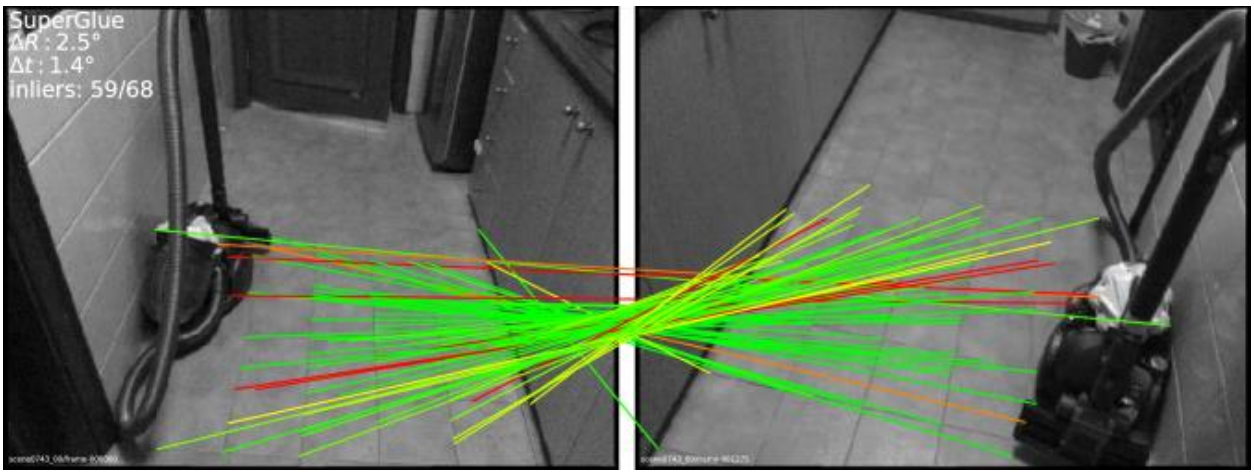
SuperGlue is a Learning Feature Matching with Graph Neural Networks. Correspondences between points in images are essential for estimating 3D structure and camera poses in geometric computer vision(OpenCV) tasks such as SLAM(Simultaneous Localization and Mapping) and SfM(Structure-from-Motion). Such correspondences are generally estimated by matching local features, the process called as data association. Broad viewpoint and lighting changes, occlusion, blur, and lack of texture are factors that make 2D-to-2D data association particularly challenging.

In this paper, we present new way of thinking about feature matching problem. Instead of learning better task-agnostic local features followed by simple matching heuristics and tricks, we propose to determine the matching process from pre-existing local features using a novel neural architecture called SuperGlue. In the context of SLAM, which typically decomposes the problem into the visual feature extraction *front-end* and the bundle adjustment or poses estimation *back-end*, our network lies directly in middle – SuperGlue is a learnable *middle-end* (see in above Figure).

In this work, *learning feature matching* is viewed as finding partial assignment between two sets of local feature. We revisit classical graph-based strategy of matching by solving the linear assignment problem, which, when relaxed to the optimal transport problem, can be solved differentiably. The cost function of this optimization is predicted by a GNN(Graph Neural Network). Inspired by success of the Transformer, it uses self- (intra-image) and cross- (inter-image) attention to leveraging both spatial relationships of

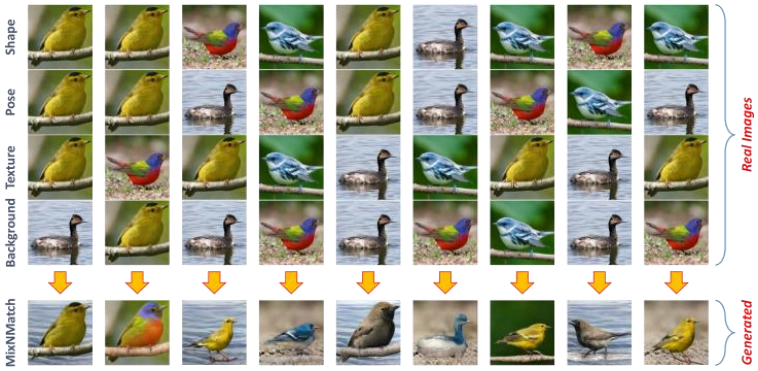
keypoints and their visual appearance. This formulation enforces assignment structure of the prediction while enabling cost to learn complex priors, handling occlusion, and non-repeatable keypoints. Our method is trained end-to-end from images pair – we learn priors for pose estimation from large annotated dataset, enabling SuperGlue to reason about 3D scene and assignment. Our work can be applied to a variety of multiple-view geometry problems that require high-quality features correspondences (see in below Figure).

We show superiority of SuperGlue compared to both handcrafted matches and learned inlier classifiers. When combined with SuperPoint, a deep front-end, SuperGlue advances the state-of-the-art on the tasks of indoor and outdoor pose estimation and paves the way towards end-to-end deep SLAM.



Q2. What is MixNMatch?

Answer:



It is a Multifactor Disentanglement and Encoding for Conditional Image Generation. Consider the real image of the yellow bird in the above Fig, First column. What would a bird look like in a different background, say that of a duck? How about in the different texture, perhaps that of the rainbow textured bird in the second column? What if we wanted to keep its texture but changes its shape to that of rainbow bird and background and pose to that of duck, as in the 3rd column? How about sampling shape, pose, texture, and experience from 4 different reference images and combining them to create entirely new image (last column)

Problem.

While research in conditional image generation has made tremendous progress, no actual work can simultaneously disentangle *background*, *object pose*, *shape*, and *texture* with minimal supervision, so that these factors can be combined from *multiple real images* for fine-grained controllable image generations. Learning disentangled representations with minimal supervision is the extremely challenging problem since the underlying factors that give rise to the data are often highly correlated and intertwined. Work that disentangles *two* such factors, by taking as input 2 reference images, e.g., one for appearance and another for pose, do exist [huang-eccv2018, joo-cvpr18, lee-eccv18, lorenz-cvpr2019, xiao-iccv2019], but they cannot disentangle other factor such as pose vs. shape or foreground vs. background appearance. Since only two factors can be controlled, these approaches cannot arbitrarily change,e.g., the object's background, shape, and texture, while keeping its pose the same. Others require intense supervision in the form of keypoint or pose or mask annotations [peng-iccv2017, Balakrishnan-cvpr2018, ma-cvpr2018, esser-cvpr2018], which limit their scalability and still fall short of disentangling all of four factors outlined above.

Our proposed conditional generative model, *MixNMatch*, aim to fill this void. MixNMatch learns to disentangle and encode background, object pose, shape, and texture latent factors from the real images, and importantly, does so with minimal human supervision. This allows, e.g., each factor to be extracted from a different actual image, and then combined for mix-and-match image generation; see in above fig. During training, MixNMatch only requires a loose bounding box around the object to the model background but requires no other supervision for modeling the object's pose, shape, and texture.

Q3. FAN: Feature Adaptation Network

Answer:

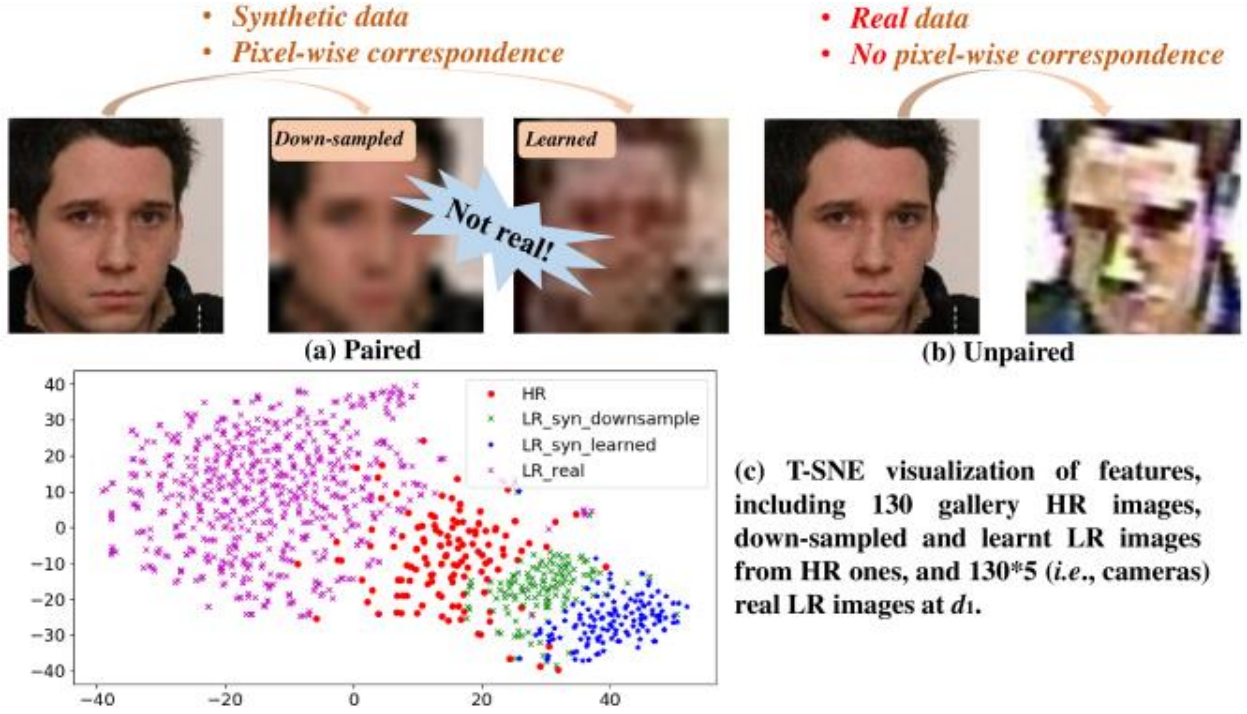
Figure: Visual results on four datasets. Vertically we show input in row first and our results in row third. For LFW and SCface datasets, we show the ground truth and gallery images in second row, respectively. For WIDER FACE and QMUL-SurFace datasets which do not have ground truth high-resolution images, we compare with two state-of-the-art(SOTA) methods: Bulat et al. [bulatyang2018learn] and FSRGAN [CT-FSRNet-2018] in row 2, respectively.

It is used for Surveillance Face Recognition and Normalization. Surveillance Face Recognition (FR) is a challenge and a significant problem yet less studied. The performance on conventional benchmarks such as LFW [LFWTech] and IJB-A have been greatly improved by state-of-the-art (SOTA) (Face Recognition(FR) methods [wang2018cosface, wen2016discriminative, deng2019arcface], which still suffer when applied to surveillance Face Recognition(FR). One intuitive approach is to perform Face Super-Resolution (FSR) on surveillance face to enhance facial details. However, existing Face Super-Resolution(FSR) methods are problematic to handle surveillance faces, because they usually ignore the *identity* information and require to *paired* training data. Preserving identity information is more crucial for surveillance of all face than recovering other information, e.g., background, Pose, Illumination, Expression (PIE).

In this work, we study surveillance face recognition(FR) and normalization. Specifically, given the surveillance face image, we aim to learn robust identity features for Face recognition(FR). Meanwhile, the feature are used to generate a normalized face with enhanced facial details and neutral PIE. Our normalization is performed mainly on the aspect of the resolution. While sharing same goal as traditional SR, it differs in removing the pixel-to-pixel correspondence between original and super-resolved images, as required by conventional SR. Therefore, we term it as face normalization. For same reason, we

compare ours to FSR instead of prior normalization methods operating on pose or expression. To the best of our knowledge, this is a *first* work to study surveillance face normalization.

We propose the novel Feature Adaptation Network(FAN) to jointly perform face recognition and normalization, which has 3 advantages over conventional FSR. i) Our joint learning scheme can benefit each other, while most FSR methods do not consider a recognition task. ii) Our framework enables training with both paired and unpaired data while conventional SR methods only support paired training. iii) Our approach simultaneously improves resolution and alleviates the background and PIE from real surveillance faces while traditional methods only act on recommendation. Examples in below Fig. One demonstrates the superiority of FAN over SOTA SR methods.



(c) T-SNE visualization of features, including 130 gallery HR images, down-sampled and learnt LR images from HR ones, and 130*5 (*i.e.*, cameras) real LR images at d_1.

Our Feature Adaptation Network (FAN) consists of 2 stages. In first stage, we adopt disentangled features learning to learns both identity and non-identity characteristics mainly from high-resolution(HR) images, which are combined as input to the decoder for pixel-wise face recovering. In second stage, we propose feature adaptation to facilitate the feature further learning from the low-resolution (LR) images by approximating feature distribution between the low-resolution and high-resolution identity encoders. There are two advantages to use Feature Adaption Network(FAN) for surveillance facial-recognition(FR) and normalization. First, Feature Adaption Network (FAN) focuses on learning disentangled identity features from Low-resolution(LR) images, which is better for facial recognition (FR) than extracting features from super-resolved faces [tran2017disentangled, zhang2018facesr, wu2016j]. 2nd, our adaptation is performed in disentangled identity feature space, which enables training with unpaired data without pixel-to-pixel

correspondences. As shown in the last fig., the synthetic paired data used in prior works [CBN_ECCV16, CT-FSRNet-
2018, bulatyang2018learn, wu2016j, zhang2018facesr, DRRN, MemNet_ICCV17, rad2019srobb]can not accurately reflect difference between real low-resolution(LR) and high-resolution(HR in-the-wild faces, which is also observed in [cai2019toward].

Furthermore, to better handle surveillance faces with the unknown and diverse resolution, we propose the Random Scale Augmentation (RSA) method that enables the network to learn all kinds of scales during training. Prior FSR [CT-FSRNet-2018, CBN_ECCV16, URDGN_ECCV16] methods either *artificially* generate the LR images from the HR ones by simple *down-sampling*, or *learn* the degradation mapping via a Convolutional Neural Network (CNN). However, their common drawback is to learn reconstruction under *fixed* scales, which may greatly limit their applications to surveillance faces. In contrast, our RSA efficiently alleviates the constraint on scale variation.

Q5. WSOD with PSNet and Box Regression

Answer:

The object detection task is to find objects belonging to specified classes and their locations in images. Benefiting from the rapid development of deep learning(DL) in recent years, the fully supervised object detection task has made significant progress. However, fully supervised task requires instance-level annotation for training, which costs lot of time and resources. Unlabeled or images labeled datasets cannot be effectively used by fully supervised method. On another hand, image-level annotated datasets are easy to generate and can even be automatically generated by web search engines. To effectively utilize these readily available datasets, we focus on weakly-supervised object detection(WSOD) tasks. The WSOD task only takes image-level annotations to train instance-level object detection network, which is different from the fully supervised object detection task.

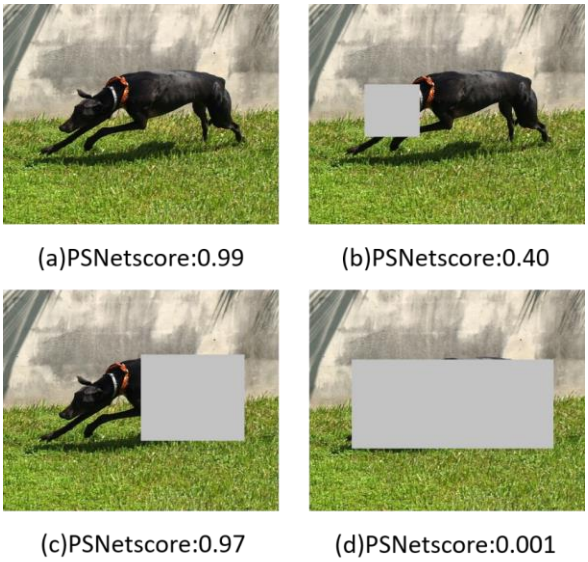
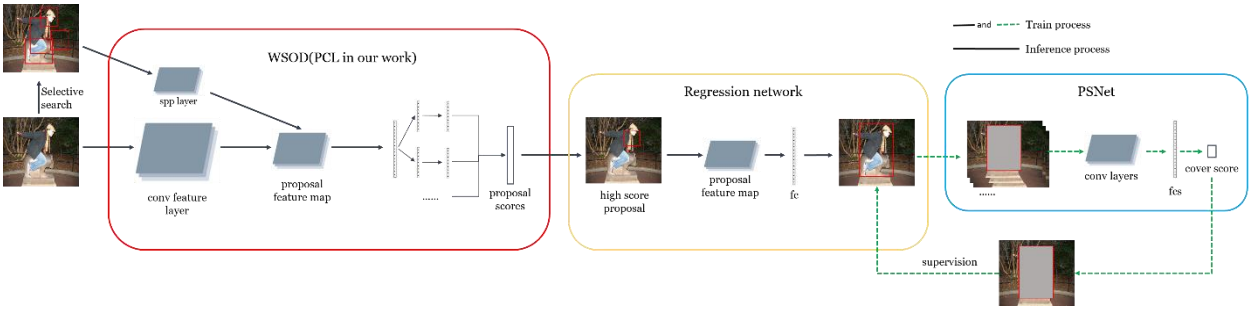


Fig.: Examples of PSNet outputs: (i) a dog without proposal occlusion, (ii) a dog whose head is occluded by the proposal box, (iii) a dog that proposal covers part of the body, and (iv) proposal completely cover the entire dog. If proposal does not completely include the whole dog, PSNet gives a high score. If proposal ultimately consists of the whole dog, PSNet gives a low score.

There are 3 main methods for weakly supervised object detection: The first is to update detector and pseudo labels from inaccurate pseudo labels iteratively; The second is to construct an end-to-end network that can take image-level annotation as supervision to train this object detection network. The third two-stage method is that taking an algorithm to optimize pseudo labels from other WSOD networks and training a fully supervised object detection network. In addition, according to different modes of proposing proposals, each of above methods can be divided into 2 classes: one is to propose proposals based on feature map that predicts probability of each pixel belonging to each class, and then get the possible instances and their locations in image; The second is detector-based method that uses the trained detector to identify multiple proposals and determines whether each proposal belongs to a specific object class or not. Comparing the effects of these methods, the end-to-end detector-based approach performs well, and our work follows this series of process.

The earliest end-to-end detector-based WSOD network is WSDDN Bilen and Vedaldi (2016), which trains a two-streams network to predict the classification accuracy of each proposal and its contributions to each class. The results of the two streams are combined to get the image classification score so that the WSDDN can take advantage of image-label annotations for training. Subsequent other work aims to improve performance of this network, like adding more classification streams, using the clustering method, adding a fully supervised module, and so on. The end-to-end detector-based approach has 2 drawbacks: one is that context information cannot be fully used to classify proposal; The second is that the most discriminative parts of the object may be detected instead of the entire object.



To make full use of the context information of the proposal and avoid finding only the most discriminative part, we design a new network structure that adds a box regression branch to the traditional WSOD network. In the previous WSOD network, there is usually no box regression part, while this branch plays an essential role in fully supervised object detection networks. The box regression network can adjust position and scale of proposal, make it closer to the ground truth. In the fully supervised object detection task, we can use the instance-level label as supervision to train box regression network; but in WSOD task, network cannot obtain the instance-level annotation and thus cannot train this branch. To obtain reliable instance annotation to train the regression network, we designed the proposal scoring network named PSNet that can detect whether proposal completely covers the object. The PSNet is specially trained multi-label classification network. Even if the object in the image is occluded or incomplete, the PSNet can detect the presence of the object. The PSNet can be used to evaluate images without proposal area. If the proposal completely covers whole object, rest of the image will not contain information about it. We use PSNet to evaluate the output of the WSOD network, and then select appropriate proposals as pseudo labels to train box regression network. Examples of the output of PSNet are shown in the above Figure.

Q6. Autonomous Driving Assistance Systems (ADAS) and Vehicle Automation.

Answer:

Vehicles are being equipped with increasingly complex autonomous driving assistance systems (ADAS) that take over parts of driving tasks previously performed by the human driver. There are several different ADAS technologies in vehicles, starting from basics that have been in vehicles for several years, such as automatic windscreen wipers and anti-lock braking systems. More advanced techniques are already on

the road today, where both the longitudinal (braking/accelerating, e.g., adaptive cruise control) and lateral (steering, e.g., assisted lane-keeping) control of the vehicle is shifting to ADAS. Further enhanced levels of automated driving functionality include autopilot (Tesla), intellisafe (Volvo), and Distronic plus steering assist (Mercedes). Overall this fast pace of market penetration of ADAS in vehicles has not allowed drivers to develop understanding of new systems over an extended period.

The most common taxonomy to capture the development of ADAS technology in cars are SAE's levels of automation sae. This approach is based on six levels of automation, ranging from no automation (level 0) to full automation (level 5). In particular, in levels 2/3, the automated system can take partial control of vehicle, where level 2 expectations of the human driver are to monitor the system and intervene appropriately, while the level 3 expectation of the human driver is to intervene appropriately upon a request from the system. Today most ADAS technology equipped cars are at level 1, in which progression to partial/semi-automation (level 2/3) with in-built ADAS technology in even lower-priced car models is becoming more common. Also, level 2/3 automation will likely be reality for some time to come, given that fuller automation (4/5) is emerging slowly without clear market deployment roadmap.

One of main challenges that arise in level 2/3 automation is transition of control from the ADAS to the human driver, often referred to as the "handover problem." This transition is, according to social factors and safety research, a phase where human attention and reliability is critical, but where humans tend to underperform in those respects son2017situation. E.g., research has indicated that automatic cruise control technology leads to a reduction in mental workload and, thus, to problems with regaining control of the vehicle in failure scenarios stanton1998vehicle. Additionally, a common misconception concerning ADAS technology is that when more automation is introduced, human error will disappear atlantic2015save, which may give rise to the problematic idea that driver training is not necessarily needed. However, social factors research advises against not training for the use of new sophisticated automation technology lee2006human; salas2006design; saetren2015effects, as humans in the technology loop will still be needed for use, maintenance or design of the technology. It may even be that increased automation increases the level of competence required for the driver, as the driver must know both how to handle system manually, for instance, if the sensors in a car stop working due to bad weather, in addition to knowing how to control and supervise the advanced automation technology.

In our previous work rismani2018qualitative, we performed a qualitative survey and found that the handover problem is challenging, and it is unclear to drivers how this could best be handled securely. Furthermore, drivers were worried about the implications of vehicle automation due to lack of knowledge and experience of level 2/3 systems and seemed concerned about the kind of training and licensing that accompanies these developments in vehicle automation. The lack of certainty around training and licensing concerning emerging ADAS technologies is a relevant ethical concern, as it exposes a gap in regulation and industry best practices that have not been the focus of much research to date.

This lack of certainty around driver training and licensing wrt level 2/3 automation systems underscores the need to understand better the following research questions: (i) What are drivers' awareness of ADAS in their vehicles, (ii) How knowledgeable are drivers about ADAS in their vehicles, and (iii) How willing are drivers to engage or use ADAS in their vehicles? Overall we expect to see people's engagement or use pattern of ADAS technologies in their vehicle correlate to their awareness and knowledge of those techniques.

Previous work has looked at driver perception of ADAS and vehicle automation, including understanding learner drivers' perspective of Blind Spot Detection(BSD) and Adaptive Cruise Control(ACC) systems. That work found that driver's awareness, use, and perceived safety of Blind Spot Detection(BSD) was higher than that of ACC tsapi_introducing_2015, and contributed to a greater understanding of driver preparation and acceptance of ADAS crump2016differing, and how drivers learn and prefer to learn about ADAS, and what their expectations are regarding ADAS and vehicle automation hoyos2018consumer.

To answer our research questions, we performed a quantitative public survey of issues specific to the public's awareness, knowledge, and use of ADAS technologies in level 2/3 automation. Also, based on previous work tsapi_introducing_2015; crump2016differing; hoyos2018consumer, we analyzed gender and age relationships as well as income and type of training with regards to our research questions above.

Q7. Robot Learning and Execution of Collaborative Manipulation Plans from YouTube Videos.

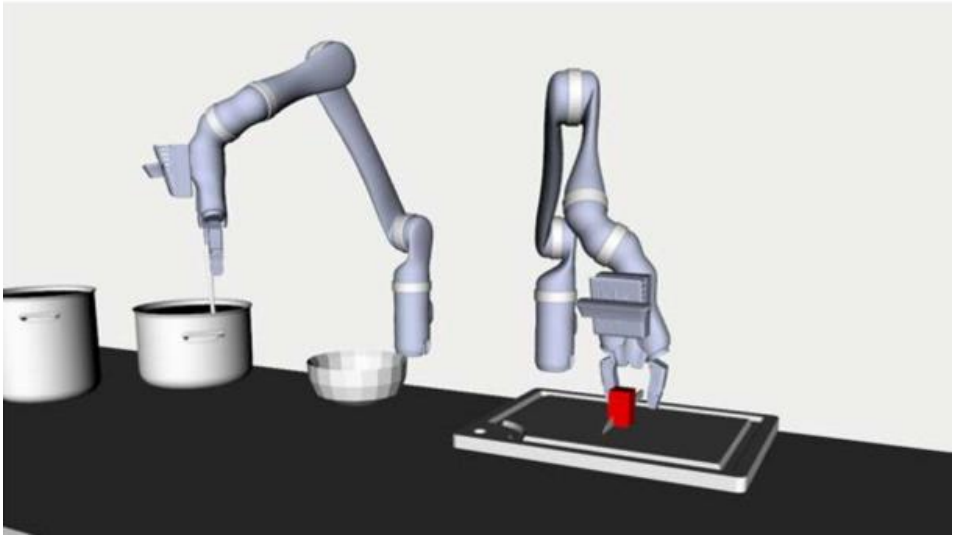
Answer:

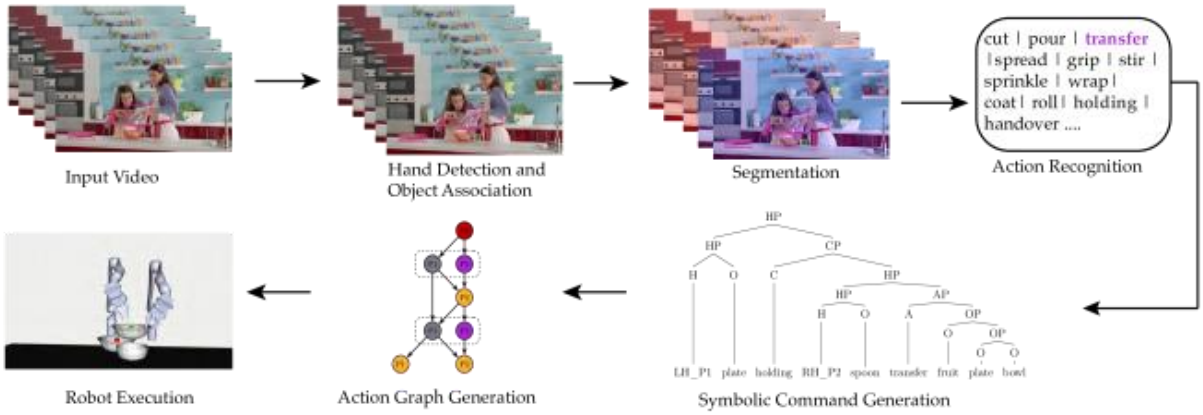
We focus on problem of learning collaborative action plans for robot. Our goal is to have robot "watch" unconstrained videos on web, extract the action sequences shown in the videos and convert them to an executable plan that it can perform either independently or as part of a human-robot or robot-robot team.

Learning from online videos is hard, particularly in collaborative settings: it requires recognizing the actions executed, together with manipulated tools and objects. In many collaborative tasks, these actions include handing objects over or holding object for the other person to manipulate. There is a very large variation in how the actions are performed and collaborative actions may overlap spatially and temporally.

In our previous work [hejia_isrr19], we proposed a system for learning activities performed by two humans collaborating at the cooking task. The system implements a collaborative action grammar built upon action grammar initially proposed by Yang et al. [yang2015robot]. Qualitative analysis in 12 clips showed that parsing these clips with grammar results in human-interpretable tree structures representing

a variety of single and collaborative actions. The clips were manually segmented and were approximate 100 frames each.





In this paper, we generalize this work with a framework for *generating single and collaborative action trees from full-length YouTube videos lasting several minutes* and *concatenating the trees in an action graph that is executable by one or more robotic arms*.

The framework takes as input YouTube video showing collaborative tasks from start to end. We assumes that objects in video are annotated with label and bounding boxes, e.g., by running the YOLOv3 algorithm. We also think a skill library that associates a detected action with skill-specific motion primitives. We focus on cooking tasks because of the variety of manipulation actions and their importance in-home service robotics.

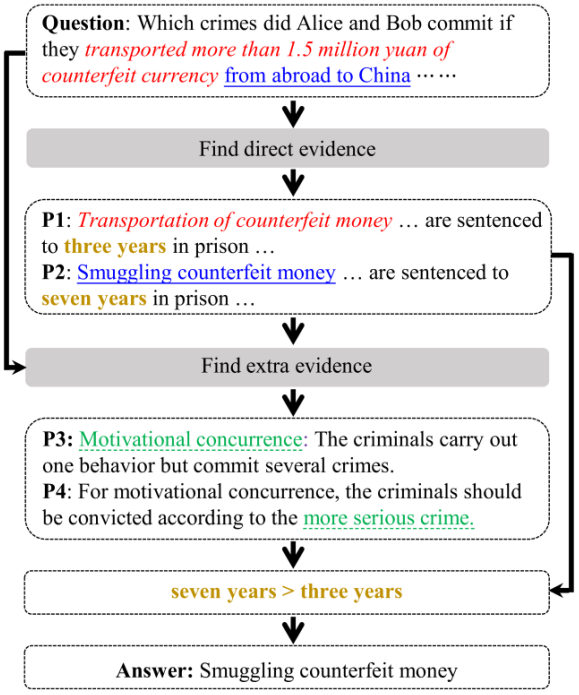
In second fig. shows the components of proposed framework. We rely on insight that hands are main driving force of manipulation actions. We detect the human hands in the video and use the hand trajectories to split the video into clips. We then associate objects and hands spatially and temporally to recognize the actions and generate human-interpretable robot commands. Finally, we propose an open-sourced platform for creating and executing an action graph. We provide a quantitative analysis of performance in two YouTube videos of 13401 frames in total and a demonstration in the simulation of robots learning and performing the actions of the third video of 2421 frames correctly.

While the extracted action sequences are executed in the open-loop manner and thus do not withstand real-world failures or disturbances, we find that this work brings us the step closer to having robots generate and execute variety of semantically meaningful plans from watching videos online.

Q8. JEC-QA: A Legal-Domain Question Answering Dataset

Legal **Q**uestion **A**nswering (**LQA**) aims to provide explanations, advice, or solutions for legal issues. A qualified LQA system can not only demonstrates a professional consulting service for unskilled humans but also help professionals to improve work efficiency and analyze real cases more accurately, which makes LQA an important NLP application in the legal domain. Recently, many researchers attempt to

build LQA systems with machine learning techniques and neural networks. Despite these efforts in employing advanced NLP models, LQA is still confronted with the following two significant challenges. The first is that there is less qualified LQA dataset, which limits the research. The second is that the cases and questions in the legal domain are very complex and rigorous. As shown in Table 1, most problems in LQA can be divided into two typical types: the knowledge-driven questions (KD-questions) and case-analysis questions (CA-questions). KD-questions focus on the understanding of specific legal concepts, while CA-questions concentrate more on the analysis of real cases. Both types of questions require sophisticated reasoning ability and text comprehension ability, which makes LQA a hard task in NLP.



To get a better understanding of these reasoning abilities, we show a question of JEC-QA in Fig. 1describing a criminal behavior that results in two crimes. The models must understand "Motivational Concurrence" to reason out further evidence rather than lexical-level semantic matching. Moreover, the models must have the ability of multi-paragraph reading and multi-hop reasoning to combine the direct evidence and the additional evidence to answer the question, while numerical analysis is also necessary for comparing which crime is more dangerous. We can see that answering one question will need multiple reasoning abilities in both retrieving and answering, makes JEC-QA a challenging task.

To investigate the challenges and characteristics of LQA, we design a unified OpenQA framework and implement seven representative neural methods of reading comprehension. By evaluating the

performance of these methods on JEC-QA, we show that even the best approach can only achieve about 25% and 29% on KD-questions and CA-questions, respectively, while skilled humans and unskilled humans can reach 81% and 64% accuracies on JEC-QA. The experimental results show that existing OpenQA methods suffer from the inability of complex reasoning on JEC-QA as they cannot well understand legal concepts and handle multi-hop logic.

Q9. SpoC: Spoofing Camera Fingerprints

Answer:

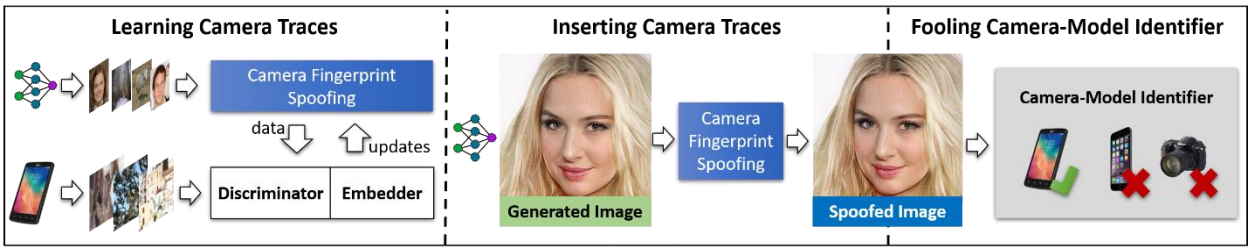


Figure 1: *SpoC* learns to spoof camera fingerprints. It can be used to insert camera traces to a generated image. Experiments show that we can fool state-of-the-art camera-model identifiers that were not seen during training.

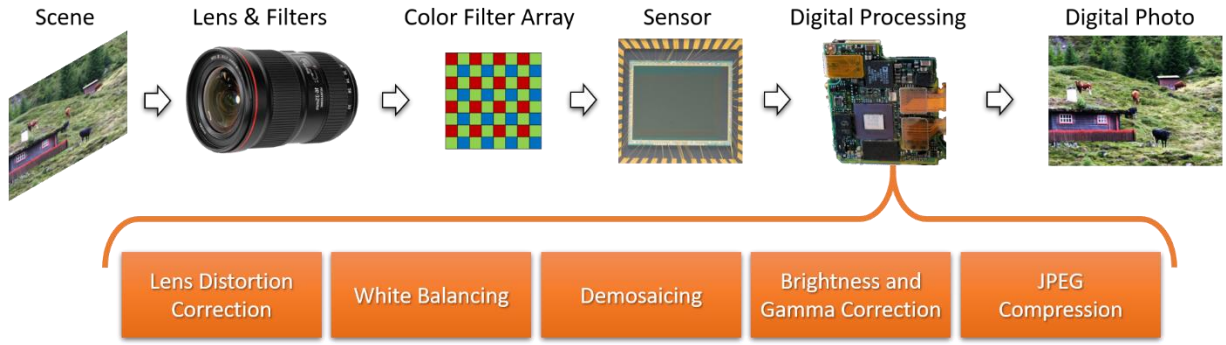


Figure 2: A digital image of a scene contains camera-related traces of the image formation process that could act as a fingerprint of a camera model. The used lenses and filters, the sensor, and the manufacturer-specific digital processing pipelines result in unique patterns. These patterns can be used to identify camera models.

There have been astonishing advances in synthetic media generation in the last few years, thanks to deep learning, and in particular to Generative Adversarial Networks (GANs). This technology-enabled a

significant improvement in the level of realism of generated data, increasing both resolution and quality. Nowadays, powerful methods exist for creating an image from scratch, and for changing its style or only some specific attributes. These methods are beneficial, especially on faces, and allow one to change the expression of a person easily or to modify its identity through face-swapping. This manipulated visual content can be used to build more effective fake news. It has been estimated that the average number of reposts for a report containing an image is 11 times larger than for those without images. This raises serious concerns about the trustworthiness of digital content, as testified by the growing attention to the profound fake phenomenon.

The research community has responded to this threat by developing several forensic detectors. Some of them exploit high-level artifacts, like asymmetries in the color of the eyes, or anomalies arising from an imprecise estimation of the underlying geometry. However, technology improves so fast that these visual artifacts will soon disappear. Other approaches rely on the fact that any acquisition device leaves distinctive traces on each captured image, because of its hardware, or its signal processing suite. They allow associating a media with its acquisition device at various levels, from the type of source (camera, scanner, etc.), to its brand/model (e.g., iPhone6 vs. iPhone7), to the individual device. A primary impulse to this field has been given by the seminal work of Lukàs et al., where it has been shown that reliable device identification is possible based on the camera photo-response non-uniformity (PRNU) pattern. This pattern is due to tiny imperfections in the silicon wafer used to manufacture the imaging sensor and can be considered as a type of device fingerprint.

Beyond extracting fingerprints that contain device-related traces, it is also possible to recover camera model fingerprints. These are related to the internal digital acquisition pipeline, including operations like demosaicing, color balancing, and compression, whose details differ according to the brand and specific model of the camera (See Fig.2). Such differences help attribute images to their source camera, but can also be used to highlight better anomalies caused by image manipulations. The absence of such traces, or their modification, is a strong clue that the image is synthetic or has been manipulated in some way. Detection algorithms, however, must confront with the capacity of an adversary to fool them. This applies to any classifier and is also very well known in forensics, where many counter-forensics methods have been proposed in the literature. Indeed, forensics and counter-forensics go hand in hand, a competition that contributes to improving the level of digital integrity over time.

In this work, we propose a method to synthesize traces of cameras using a generative approach that is agnostic to the detector (i.e., not just targeted adversarial noise). We achieve this by training a conditional generator to jointly fool an adversarial discriminator network as well as a camera embedding network. To this end, the proposed method injects the distinctive traces of a target camera model in synthetic images, while reducing the first generation traces themselves, leading all tested classifiers to attribute such images to the target camera (’targeted attack’).

DATA SCIENCE INTERVIEW PREPARATION

(30 Days of Interview Preparation)

#FinaleDay30

Most important questions Related to Project

Disclaimer: The answers given here are not generic ones. These answers are given based on the attendance system that we have developed to do face detection. The answers will vary based on the projects done, methodologies used and based on the person being interviewed.

Face Recognition and Identification system Project

Q1. Tell me about your current project.

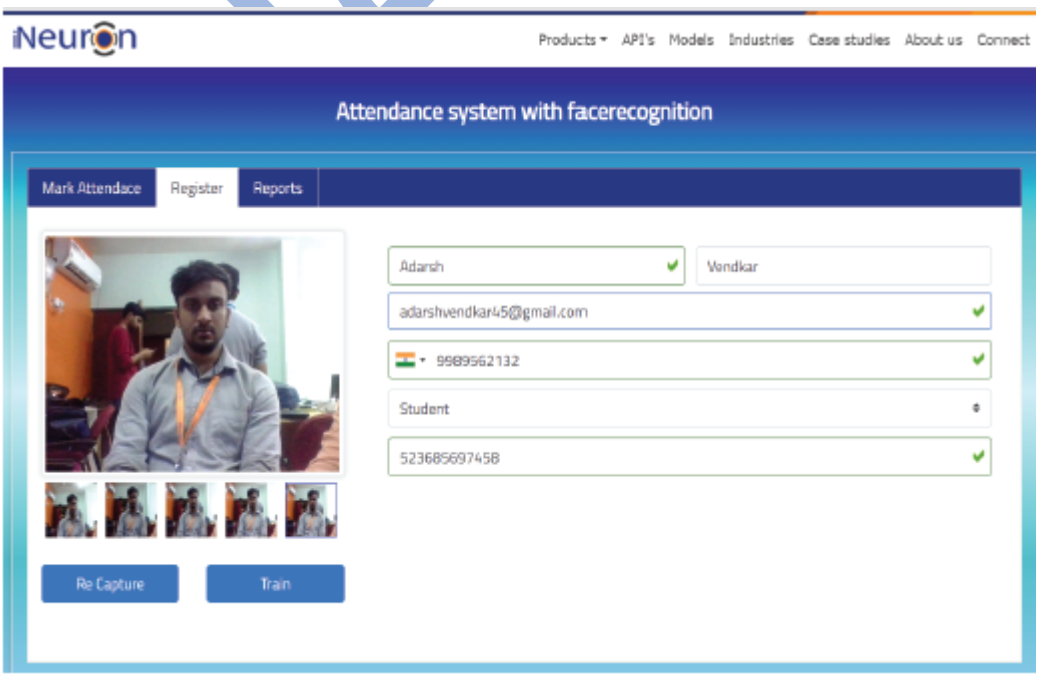
Answer:

The project is called Attendance System using facial recognition.

The goal of the project is to identify the person and mark their attendance. First, the user has to register himself/herself in the application providing the required details. The application takes multiple snaps of the user and then stores it into the database. Once the same user comes before the camera again, the application captures the image, references it against the already stored images in the database, and then marks the attendance, if the user is present in the database. Reports can be generated for a particular duration based on the user requirement.

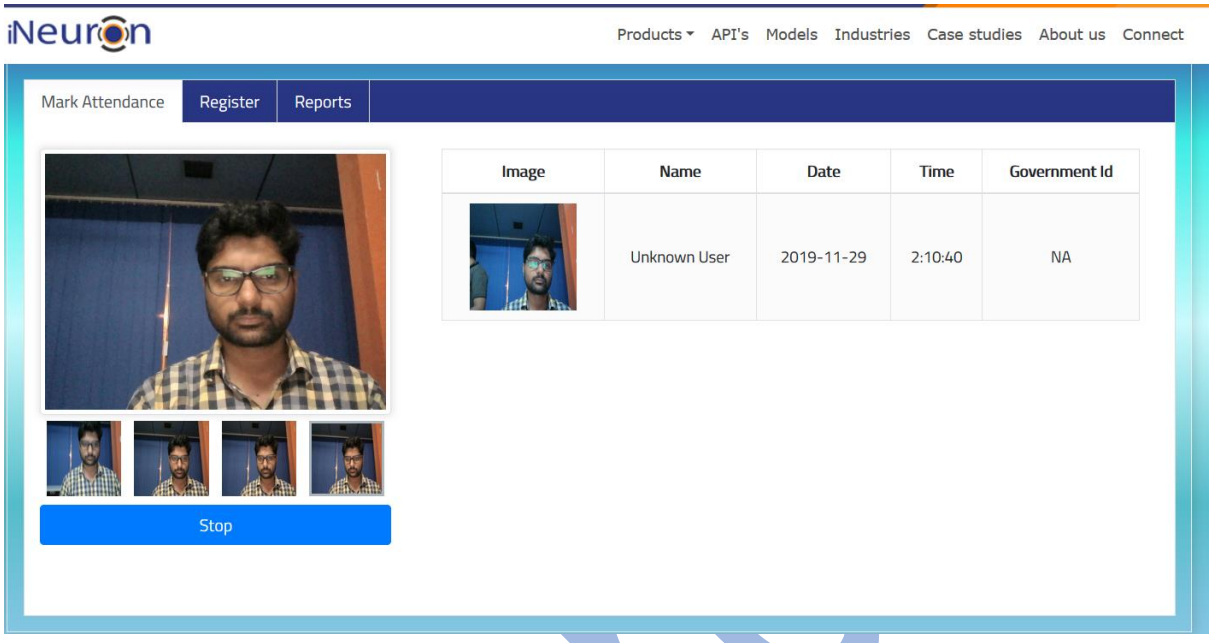
Some snaps from the project are as follows:

1st-time registration:

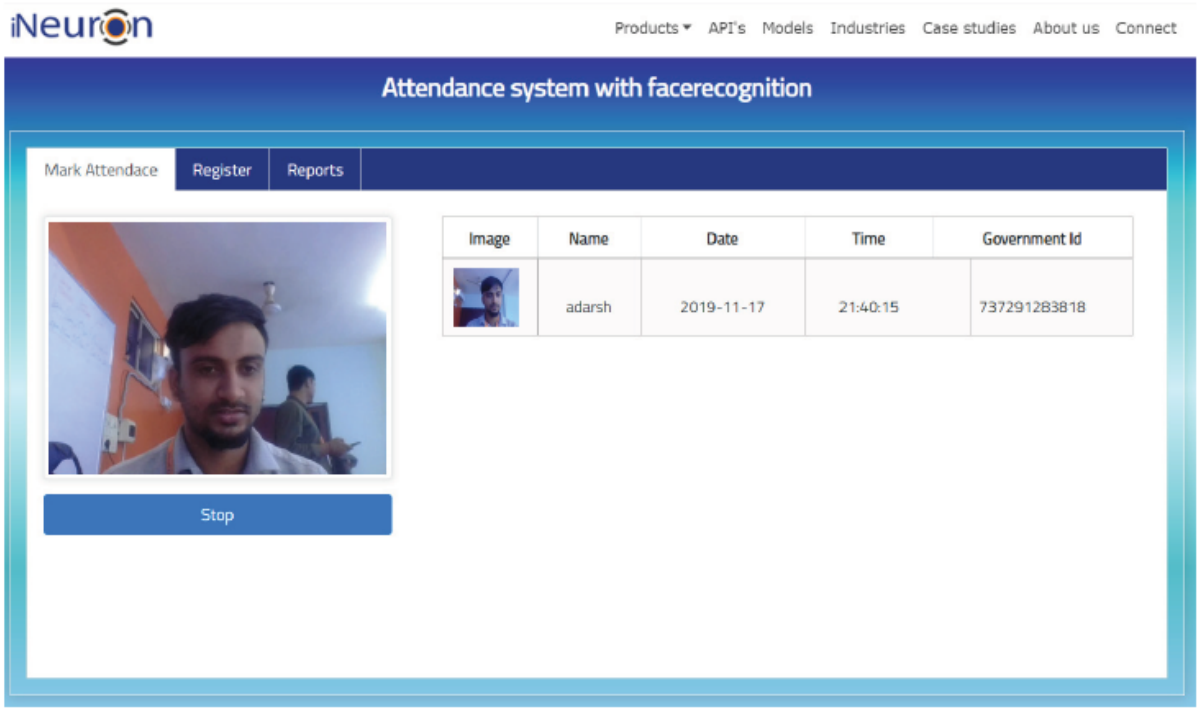


Marking the Attendance:

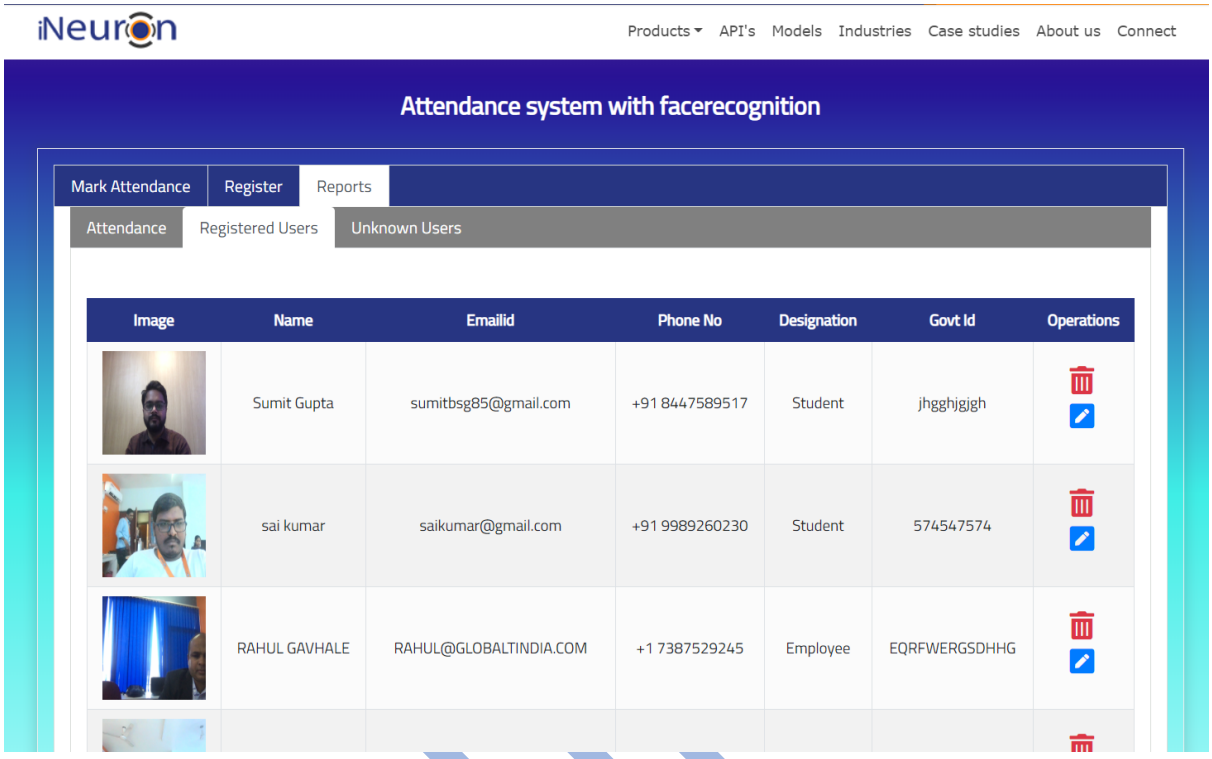
With un-registered user:



With a registered user:



Seeing the reports:



Features:

- Works with generic IP cameras with good quality.
- Works even with PC, you don't need high-end systems.
- Works in both indoor as well as outdoor environments.
- Works with limited pose changes.
- Works with spectacles.
- Works for people of different ethnicity.
- Works for tens of thousands of registered faces.
- Works with limited lighting conditions.
- Works with partial facial landmarks.
- Non-recognition of static input images when provided by the user.

Functionalities in the Attendance System

- Registration of users in the system.
- Capturing the user details during registration using Passport, Adhar Card, and Pan Card.
- All details will be extracted using the in-house OCR technique.
- Tracking of the login and logout timings of the users from the system.
- Generation of user logs on a temporal basis.
- Generation of timely reports.

Deployment/Installation

- The application can be easily installed as a web-based API on any cloud platform. This installation is similar to a plug and play scenario.
- The application can also be installed in an edge device (like the Google Coral). This installation provides realtime streaming capabilities to the application.

Q2. What was the size of the data?

Answer:

The number of images used for training was 12,313.

Q3. What was the data type?

Answer:

The data used for training this model consisted of thousands of images; the images then are converted to tensor objects, which have a float 32 representation.

Q4. What was the team size and distribution?

Answer:

The team consisted of:

- 1 Product Manager,
- 1 Solution Architect,
- 1 Lead,
- 2 Dev-Ops engineers,
- 2 QA engineers,
- 2 UI developers, and
- 3 Data Scientists.

Q5.What Hadoop distribution were you using?

Answer:

The Hadoop distribution from Cloudera was used as it provides many of the much-needed capabilities out of the box like multi-function analytics, shared data experience with optimum security and governance, hybrid capabilities for support to clouds, on-premise servers as well as multi-clouds.

Q6.What is the version of distribution?

Answer:

CDH – 5.8.0

Q7.What was the size of the cluster?

Answer:

The cluster(production setup) consisted of 15 servers with

- Intel i7 processors
- 56 GB of RAM
- 500 GB of Secondary storage each
- Mounted NAS locations

Q8. How many nodes were there in all the Dev, UAT, and Prod environments?

Answer:

The necessary coding was done on one development server. But as a standalone machine won't give enough speed to train the model in a short time, once we saw that the model's loss is decreasing for a few numbers of epochs in the standalone machine, the same code was deployed to a cloud-based GPU machine for training. Once the model was trained there, we used the saved model file for prediction/classification. The same model file was deployed to the cloud UAT and Production environments.

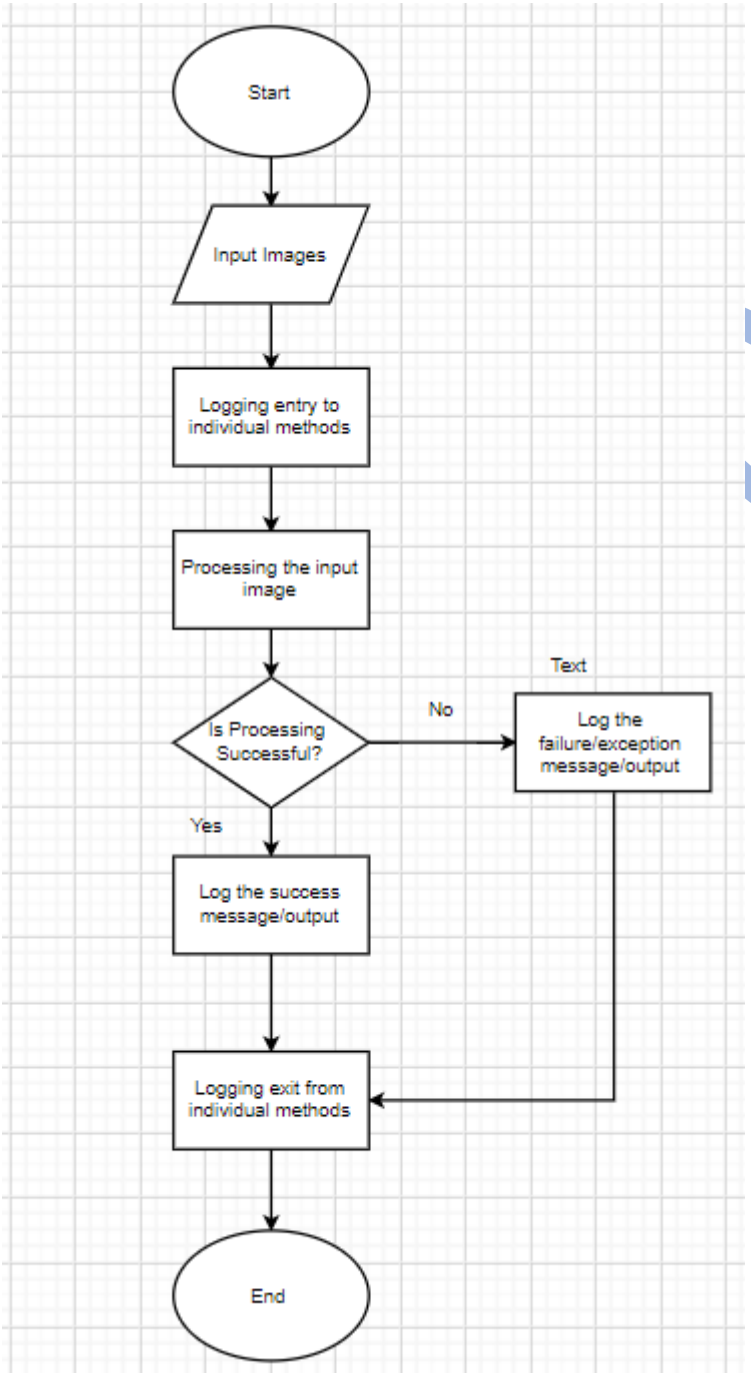
In total, we had:

- 5 nodes in the dev environment,
- 5 nodes in UAT, and
- 15 nodes in production.

Q9.How were you creating and maintaining the logs?

Answer:

The logs are maintained using MongoDB. The logging starts with the start of the application. The start time of the application gets logged. After that, there are loggings for entry and exits to the individual methods. There are loggings for the error scenarios and exception block as well.



Q10.What techniques were you using for data pre-processing for various data science use cases and visualization?

Answer:

There are multiple steps that we do for data preprocessing, like data cleaning, data integration, data scaling, etc. Some of them are listed as follows:

➔ **For Machine Learning**:

- While preparing data for a model, data should be verified using multiple tables or files to ensure data integrity.
- Identifying and removing unnecessary attributes.

  For example,

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1
0	1000001	P00069042	F	0-17	10	A	2	0	3
1	1000001	P00248942	F	0-17	10	A	2	0	1
2	1000001	P00087842	F	0-17	10	A	2	0	12
3	1000001	P00085442	F	0-17	10	A	2	0	12
4	1000002	P00285442	M	55+	16	C	4+	0	8

  Here, the user_ID column does not contribute to the customer behavior for purchasing the products. So, it can be dropped from the dataset.

- Identifying, filling or droping the rows/columns containing missing values based on the requirement.

Checking for columnwise null values

```
: df.isnull().sum()
```

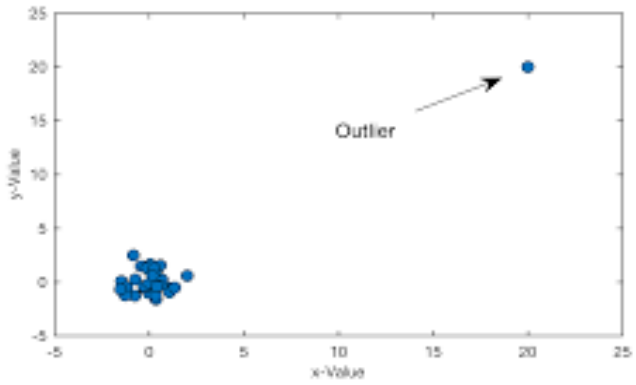
```
: Product_ID                        0
  Gender                            0
  Age                               0
  Occupation                        0
  City_Category                     0
  Stay_In_Current_City_Years        0
  Marital_Status                    0
  Product_Category_1                0
  Product_Category_2           245982
  Product_Category_3           545809
  Purchase                     233599
  B                                 0
  C                                 0
  dtype: int64
```

Here, the Product_Category_3 has about 5.5 lac missing values. It can be dropped using the command→ df.drop('Product_Category_3',axis=1, inplace=True)

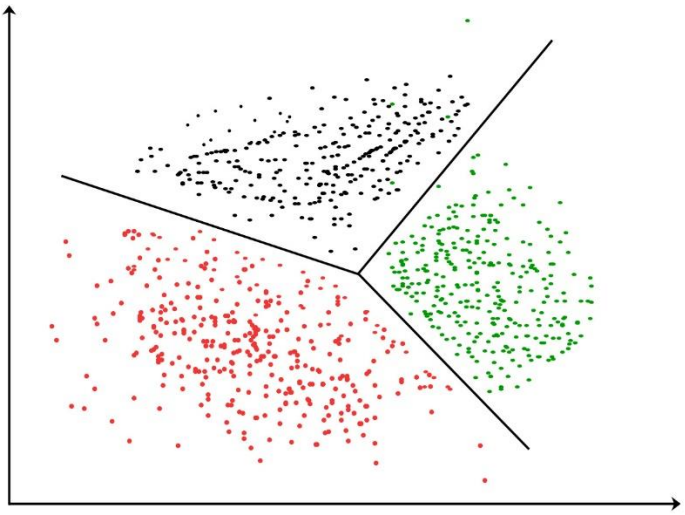
Or, if the count of null values have been lower, they could have been imputed using→

df['Purchase'] = df['Purchase'].fillna(df['Purchase'].mean())

- Identifying and removing outliers



- In the image above, one point lies very far from the other data points, i.e., it's an outlier that is not following the general trend of the data. So, that point can be dropped.
- Based on the requirement, form clusters of data to avoid an overfitted model.



Contrary to the example in the previous point, there can be several points that do not follow a particular pattern or which have a pattern of their own. If those points are too many, they can't be considered as outliers. Then we need to consider those points separately. In that kind of scenario, we create the clusters of similar points, and then we try and train our model on those clusters.

- Scaling the data so that the difference between the magnitudes of the data points in different columns are not very big.

	Gender	Age	Occupation	Stay_In_Current_City_Years	Marital_Status	cat1	cat2	cat3	Purchase	B	C
0	0	1	10	2	0	3	8.0	16.0	8370.0	0	0
1	0	1	10	2	0	1	6.0	14.0	15200.0	0	0
2	0	1	10	2	0	12	8.0	16.0	1422.0	0	0
3	0	1	10	2	0	12	14.0	16.0	1057.0	0	0
4	1	7	16	4	0	8	8.0	16.0	7969.0	0	1
5	1	3	15	3	0	1	2.0	16.0	15227.0	0	0
6	1	5	7	2	1	1	8.0	17.0	19215.0	1	0
7	1	5	7	2	1	1	15.0	16.0	15854.0	1	0
8	1	5	7	2	1	1	16.0	16.0	15686.0	1	0
9	1	3	20	1	1	8	8.0	16.0	7871.0	0	0

In the diagram above, the magnitude of the values in the 'Purchase' column is way larger than the other columns. This kind of data makes our model sensitive. To rectify this, we can do→

```
# Feature Scaling So that data in all the columns are to the same scale
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
```

After scaling the data looks like:

	Gender	Age	Occupation	Stay_In_Current_City_Years	Marital_Status	cat1	cat2	cat3	Purchase	B	C
0	0.572754	-0.367452	0.600884	-0.666350	1.199047	-0.094207	-0.293215	0.369371	0.000026	1.173655	-0.672287
1	0.572754	-0.367452	-1.239139	1.660866	1.199047	-1.125331	-1.688120	0.369371	0.540162	-0.852039	-0.672287
2	0.572754	1.109957	-0.165793	-1.442089	-0.833995	-0.094207	-0.293215	0.369371	0.000026	1.173655	-0.672287
3	0.572754	2.587366	-1.085804	0.885128	1.199047	0.679136	-0.293215	0.369371	-0.815744	-0.852039	-0.672287
4	0.572754	1.848662	0.754219	-0.666350	1.199047	-0.351988	-0.990668	1.098466	-1.520074	1.173655	-0.672287
5	0.572754	-1.106157	-0.932469	0.885128	-0.833995	-0.867550	-1.223152	0.369371	-1.359912	1.173655	-0.672287
6	0.572754	-0.367452	-0.932469	-1.442089	1.199047	-1.125331	-0.293215	0.369371	0.000026	-0.852039	-0.672287
7	0.572754	-0.367452	0.600884	0.885128	-0.833995	-0.094207	-0.758184	-0.724271	0.000026	-0.852039	1.487460
8	0.572754	-0.367452	-0.165793	1.660866	-0.833995	0.679136	0.869206	0.369371	0.175403	1.173655	-0.672287
9	0.572754	-1.106157	1.827566	1.660866	1.199047	-0.094207	-0.293215	0.369371	-0.501124	-0.852039	-0.672287

- Converting the categorical data into numerical data.

  For example, gender data (Male or Female) is a categorical one. It can be converted to numeric values, as shown below:

  ```
  df['Gender']=df['Gender'].map({'F':0, 'M':1})
  ```

- Replacing or combining two or more attributes to generate a new attribute which serves the same purpose.

  For example, if we use one-hot encoding in the example above, it will generate two separate columns for males and females. But if we observe, a person who is not a male is automatically a female(if we consider only two genders). So, the two columns essentially convey the same information in that case. This is called the *dummy variable trap*. So, one column can be conveniently dropped.

- Trying out dimensionality reduction techniques like PCA(Principal Component Analysis), which tries to represent the same information but in a space with reduced dimensions.

➔ **For Deep Learning:**

- Data augmentation strategies followed by image annotation. Data augmentation consists of image rotation, contrast, and color adjustments, lighting variations, random erasing, etc.
- Then all the images are made of identical size.
- Then image annotation is done.

Q11. How were you maintaining the failure cases?

Answer:

Let's say that our model was not able to make a correct prediction for an image. In that case, that image gets stored in the database. There will be a report triggered to the support team at the end of the day with all the failed scenarios where they can inspect the cause of failure. Once we have a sufficient number of cases, we can label and include those images while retraining the model for better model performance.

Q12.What kind of automation have you done for data processing?

Answer:

We had a full-fledged ETL pipeline in place for data extraction. Employers already have images of their employees. That data can be easily used after doing pre-processing for training the image identification model.

Q13.Have you used any scheduler?

Answer:

Yes, a scheduler was used for retraining the model after a fixed time(20 days).

Q14.How are you monitoring your job?

Answer:

There are logging set-ups done. We regularly monitor the logs to see for any error scenarios. For fatal errors, we had email notifications in place. Whenever a specific error code, which has been classified as a fatal error occurs, email gets triggered to the concerned parties.

Q15. What were your roles and responsibilities in the project?

Answer:

My responsibilities consisted of gathering the dataset, labeling the images for the model training, training the model on the prepared dataset, deploying the trained model to the cloud, monitoring the deployed model for any issues, providing QA support before deployment and then providing the warranty support post-deployment.

Q16.What was your day to day task?

Answer:

My day to day tasks involved completing the JIRA tasks assigned to me, attending the scrum meetings, participating in design discussions and requirement gathering, doing the requirement analysis, data validation, image labeling, Unit test for the models, providing UAT support, etc.

Q17.In which area you have contributed the most?

Answer:

I contributed the most to image labeling and model training areas. Also, we did a lot of brainstorming for finding and selecting the best algorithms for our use cases. After that, we identified and finalized the best practices for implementation, scalable deployment of the model, and best practices for seamless deployments as well.

Q18.In which technology you are most comfortable?

Answer:

I have worked in almost all the fields viz. Machine Learning, Deep Learning, and Natural Language Processing, and I have nearly equivalent knowledge in these fields. But if you talk about personal preference, I have loved working in Deep Learning and NLP the most.

Q19.How you rate yourself in big data technology?

Answer:

I gave worked often in the big data computing technology with ample knowledge in distributed and cluster-based computing. But my focus and extensive contribution have been as a data scientist.

Q20. In how many projects you have already worked?

Answer:

It's difficult to give a number. But I have worked in various small and large scale projects, e.g., object detection, object classification, object identification, NLP projects, chatbot building, machine learning regression, and classification problems.

Q21. How were you doing deployment?

Answer:

The mechanism of deployment depends on the client's requirement. For example, some clients want their models to be deployed in the cloud, and the real-time calls they take place from one cloud application to another. On the other hand, some clients want an on-premise deployment, and then they do API calls to the model. Generally, we prepare a model file first and then try to expose it through an API for predictions/classifications. The mechanism in which he API gets called depends on the client requirement.

Q22.What kind of challenges have you faced during the project?

Answer:

The biggest challenge that we face is in terms of obtaining a good dataset, cleaning it to be fit for feeding it to a model, and then labeling the prepared datasets. Labeling is a rigorous task and it burns a lot of hours. Then comes the task of finding the correct algorithm to be used for that business case. Then that model is optimized. If we are exposing the model as an API, then we need to work on the SLA for the API as well, so that it responds in optimum time.

Q23.What will be your expectations?

Answer:

It’s said that the best learning is what we learn on the job with experience. I expect to work on new projects which require a broad set of skills so that I can hone my existing skills and learn new things simultaneously.

Q24. What is your future objective?

Answer:

The field of data science is continuously changing. Almost daily, there is a research paper that changes the way we approach an AI problem. So, it really makes it exciting to work on things that are new to the entire world. My objective is to learn new things as fast as possible and try and implement that knowledge to the work that we do for better code, robust application and in turn, a better user/customer experience.

Q25. Why are you leaving your current organization?

Answer:

I was working on similar kinds of projects for some time now. But the market is rapidly changing, and the skill set required to be relevant in the market is changing as well. The reason for searching a new job is to work on several kinds of projects and improve my skill set. ***<Mention about the company profile and if you have the project name that you are being interviewed for as new learning opportunities for you>.***

Q26. How did you do Data validation?

Answer:

Data validation is done by looking at the images gathered. There should be ample images for the varied number of cases like change in the lighting conditions, distance from the camera, movement of the user, the angle at which camera is installed, the position at which the camera is installed, the angle at which the snap of the user has been taken, the alignment of the image, the ratio of the face and the other areas in the image etc.

Q27. How did you do Data enrichment?

Answer:

Data enrichment in vision problems mostly consists of image augmentation. Apart from image augmentation, we tried to train the model with images with different lighting conditions, with b/w and colored images, images from different angles, etc.

Q28. How would you rate yourself in machine learning?

Answer:

Well, honestly, my 10 and your 10 will be a lot different as we have different kinds of experiences. On my scale of 1 to 10, I'll rate myself as an 8.2.

Q29. How would you rate your self in distributed computation?

Answer:

I'd rate myself a 7.7 out of 10.

Q30. What are the areas of machine learning algorithms that you already have explored?

Answer:

I have explored various machine learning algorithms like Linear Regression, Logistic Regression, L1 and L2 Regression, Polynomial Regression, Multi Linear Regression,Decision Trees, Random Forests, Extra Trees Classifier, PCA, TSnE, UMAP, XG Boost,CAT Boost, ADA Boost, Gradient Boosting, Light Boost, K-Means,K-Means ++,LDA, QDA, KNN, SVM, SVR,Naïve Bayes, Agglomerative clustering, DBScan, Hierarchical clustering, TFIDF, Word to Vec, Bag of words, Doc to Vec, Kernel Density Estimation are some of them.

Q31. In which part of machine learning have you already worked on?

Answer:

I have worked on both supervised and unsupervised machine learning approaches and building different models using the as per the user requirement.

Q32.How did you optimize your solution?

Answer:

Well, model optimization depends on a lot of factors.

- Train with better data(increase the quality), or do data pre-processing steps more efficiently.
- Keep the resolution of the images identical.
- Increase the quantity of data used for training.
- Increase the number of epochs for which the model was trained
- Tweak the batch input size, the number of hidden layers, the learning rate, rate of decay, etc. to produce the best results.
- If you are not using transfer learning, then you can alter the number of hidden layers, activation function.
- Change the function used in the output layer based on the requirement. The sigmoid functions work well with binary classification problems, whereas for multi-class problems, we use a sigmoid model.
- Try and use multithreaded approaches, if possible.
- Reduce Learning Rate in plateau reasons optimizes the model even further.

Q33.How much time did your model take to get trained?

Answer:

With a batch size of 128 and the number of epochs 100000 with 7000 images, it took around 110 hours to train the model using Nvidia Pascal Titan GPU.

Q34. At what frequency are you retraining and updating your model?

Answer:

The model gets retrained every 20 days.

Q35. In which mode have you deployed your model?

Answer:

I have deployed the model both in cloud environments as well in the on-premise ones based on the client and project requirements.

Q36. What is your area of specialization in machine learning?

Answer:

I have worked on various algorithms. So, It's difficult to point out one strong area. Let's have a discussion on any specific requirement that you have, and then we can take it further from there.