

# Performance Evaluation Metrics

- General
  - ▶ True positives, false positives, true negatives, false negatives
  - ▶ Precision/recall
  - ▶ F-measure, F1
  - ▶ ROC, AUC
  - ▶ Retrieval: Rank@K, Prec@K, mAP
- NLP: BLEU, ROUGE, METEOR, CIDEr
- Speech: Word Error Rate (WER)
- Signal quality: MSE, SNR, PSNR
- Image quality: Inception score
- Segmentation: Dice coefficient, Jaccard index

## Performance metrics

- **Classification**

- Results of a prediction model can be summarized using a 2x2 confusion matrix depicting four possible scenarios

		<i>gold standard labels</i>		
		gold positive	gold negative	
<i>system output labels</i>	system positive	<b>true positive</b>	<b>false positive</b>	$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$
	system negative	<b>false negative</b>	<b>true negative</b>	
		$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$		$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}}$

## Performance metrics

- Classification of data samples in two categories e.g. spam detection, medical diagnosis, fire detector etc.
- We define the positive (fire) and negative (no fire) classes.
- Confusion matrix.** Results of fire-prediction model summarized in a 2x2 confusion matrix:

<b>True Positive</b> Truth: a flame rose up The alarm rang Result: people have been saved	<b>False Positive</b> Truth: no fire The alarm rang Result: people made complaints
<b>False Negative</b> Truth: a flame rose up The alarm did not ring Result: people died in the fire	<b>True Negative</b> Truth: no fire. The alarm did not ring. Result: everyone is safe

- True positive/negative.** The number of positive/negative samples which the model correctly predicts.
- False positive/negative.** The number of positive/negative samples which the model incorrectly predicts.

## Performance metrics

- **Precision.** it attempts to answer the following question: "what fraction of positive identification was correct?"

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

- Precision for fire-prediction model:

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 3} = 0.25 \quad (2)$$

<b>True Positive=1</b> Truth: a flame rose up The alarm rang Result: people have been saved	<b>False Positive=4</b> Truth: no fire The alarm rang Result: people made complaints
<b>False Negative=5</b> Truth: a flame rose up The alarm did not ring Result: people died in the fire	<b>True Negative=50</b> Truth: no fire. The alarm did not ring. Result: everyone is safe

## Performance metrics

- **Recall.** It attempts to answer the following question: "what proportion of actual positives was identified correctly?"

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- Recall for fire-prediction model:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1}{1 + 5} = 0.16 \quad (4)$$

<b>True Positive=1</b> Truth: a flame rose up The alarm rang Result: people have been saved	<b>False Positive=4</b> Truth: no fire The alarm rang Result: people made complaints
<b>False Negative=5</b> Truth: a flame rose up The alarm did not ring Result: people died in the fire	<b>True Negative=50</b> Truth: no fire. The alarm did not ring. Result: everyone is safe

# Performance metrics

- **F-measure, F1.** It is the harmonic mean between precision and recall.

► especially useful in case classes are imbalanced as it provides a weighted average of precision and recall.

$$F_1 = \left( \frac{Recall^{-1} + Precision^{-1}}{2} \right)^{-1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

- $F_1$  for fire-prediction model:

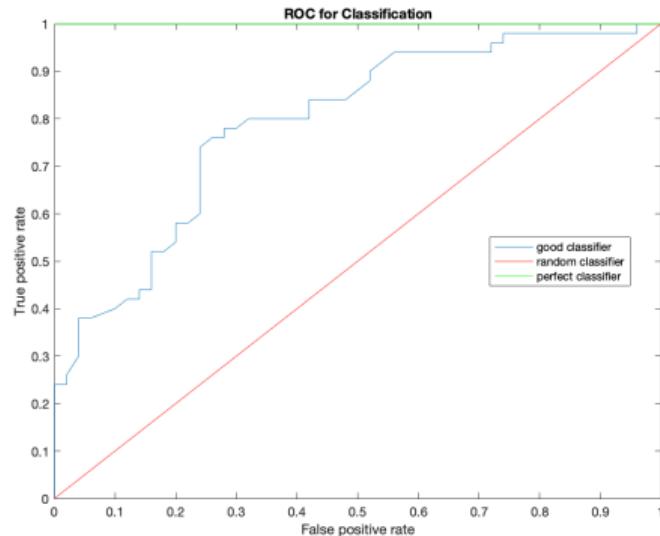
$$F_1 = \frac{Precision \cdot Recall}{Precision + Recall} = \frac{0.25 \cdot 0.16}{0.25 + 0.16} = 0.0976 \quad (6)$$

<b>True Positive=1</b> Truth: a flame rose up The alarm rang Result: people have been saved	<b>False Positive=4</b> Truth: no fire The alarm rang Result: people made complaints
<b>False Negative=5</b> Truth: a flame rose up The alarm did not ring Result: people died in the fire	<b>True Negative=50</b> Truth: no fire. The alarm did not ring. Result: everyone is safe

## Performance metrics

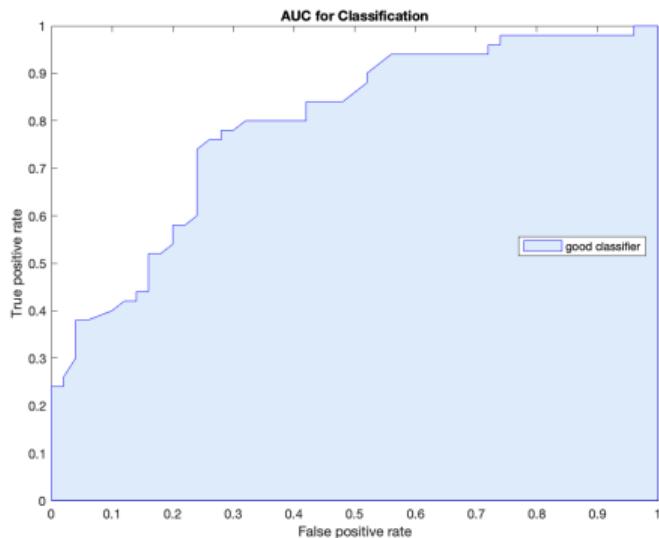
- **Receiver Operating Characteristic (ROC).** A curve plotting the performance of a classification model over different classification threshold in terms of:

- ▶ True Positive Rate  $TPR = \frac{TP}{TP+FN}$
- ▶ False Positive Rate  $FPR = \frac{FP}{FP+TN}$



## Performance metrics

- **Area Under the ROC Curve: (AUC).** Given a classifier  $C$ , AUC value can be interpreted as the probability a random positive sample  $x^+$  is ranked by  $C$  higher than a random negative sample  $x^-$ , i.e.  $\text{auc}(C) = P[C(x^+) > C(x^-)]$ 
  - $AUC \in [0, 1]$
  - $AUC = 1$  it is the perfect classifier where all positives come after all negatives
  - $AUC = 0.5$  is a random classifier



# Performance metrics

- Recall and precision in retrieval task.

- ▶ Suppose you run a query in a retrieval system, e.g. arXiv.org, to search a document you need for your research. The output results can be summarized in the confusion matrix below.

<b>True Positive</b> Truth: the document is relevant The document is retrieved Result: you got the required document	<b>False Positive</b> Truth: the document is not relevant The document is retrieved Result: you got a useful document
<b>False Negative</b> Truth: the document is relevant The document is rejected. Result: you can't get the required document	<b>True Negative</b> Truth: the document is not relevant. The document is rejected. Result: you avoided useless documents

- Recall. The fraction of relevant documents that were retrieved.  $Recall = \frac{RelevantRetrieved}{RelevantRetrieved + RelevantRejected}$

- Precision. The fraction of retrieved documents that is relevant.  $Precision = \frac{RelevantRetrieved}{RelevantRetrieved + NotRelevantRetrieved}$

# Performance metrics

## Ranking in retrieval system.

- Suppose you are looking for "local descriptors in arXiv.org". Once the system processes your query, it will output an ordered list of results and you will start to examine it until you are satisfied or give up.

The screenshot shows the arXiv search interface with the query 'local descriptors'. The results page displays 50 items per page, sorted by relevance. Each result includes a title, authors, abstract, and a link to the full paper. The results are as follows:

- arXiv:1901.00027 [pdf, other]** [\[tex, CV\]](#)  
Discriminative and Contrast Invertible Descriptors  
Authors: Zhenwei Mao, Kim-Hui Yap, Xudong Jiang, Subhhrama Sridha, Zhenhua Wang  
Abstract: Local feature extraction and matching is the basis for solving many tasks in the area of computer vision, such as 3D registration, modeling, recognition and retrieval. However, this process commonly draws into false correspondences, due to missing features, occlusion, incomplete surface and etc. In order to estimate accurate transformation based on... [More](#)  
Submitted 31 December, 2018; originally announced January 2019.
- arXiv:1901.00788 [pdf, other]** [\[cond-mat.mtrl-sci\]](#)  
Classification of Local Chemical Environments from X-ray Absorption Spectra using Supervised Machine Learning  
Authors: Matthew R. Carbone, Shriya Yos, Mehmet Topakal, Deyu Lu  
Abstract: - is a premier, element-specific technique for materials characterization. Specifically, the x-ray absorption near edge structure (XANES) encodes important information about the local chemical environment of an absorbing atom, including coordination number, symmetry and oxidation state. Interpreting XANES spectra is a key step towards understanding the structure... [More](#)  
Submitted 3 January, 2019; originally announced January 2019.  
Comments: 11 numbered pages and 6 figures
- arXiv:1901.00937 [pdf, other]** [\[cs.CV\]](#)  
Local Area Transform for Cross-Modality Correspondence Matching and Deep Scene Recognition  
Authors: Seungjib Hyo  
Abstract: - a non-linearly deformed image pair induced by different modality conditions is a challenging problem. This paper describes a efficient but powerful image transform called local area transform (LAT) for modality-obtaining correspondence estimation. Specifically, LAT transforms an image from the intensity domain to the... [More](#)  
Submitted 3 January, 2019; originally announced January 2019.  
Comments: Doctoral Dissertation
- arXiv:1901.05104 [pdf, other]** [\[cs.CV\]](#)  
A Comprehensive Performance Evaluation for 3D Transformation Estimation Techniques  
Authors: Bao Zhao, Xiaobo Chen, Xinyi Li, Junting Xi  
Abstract: 3D local feature extraction and matching is the basis for solving many tasks in the area of computer vision, such as 3D registration, modeling, recognition and retrieval. However, this process commonly draws into false correspondences, due to missing features, occlusion, incomplete surface and etc. In order to estimate accurate transformation based on... [More](#)  
Submitted 15 January, 2019; originally announced January 2019.
- arXiv:1901.05104 [pdf, other]** [\[cs.CV\]](#)  
A Comprehensive Performance Evaluation 3D Transformation Estimation Techniques  
Authors: Bao Zhao, Xiaobo Chen, Xinyi Li, Junting Xi  
Abstract: 3D local feature extraction and matching is the basis for solving many tasks in the area of computer vision, such as 3D registration, modeling, recognition and retrieval. However, this process commonly draws into false correspondences, due to missing features, occlusion, incomplete surface and etc. In order to estimate accurate transformation based on... [More](#)  
Submitted 15 January, 2019; originally announced January 2019.
- arXiv:1901.06915 [pdf, ps, other]** [\[cond-mat.soft\]](#) [\[physics.comp-ph\]](#) [\[q-bio.BM\]](#)  
A deep learning approach to the structural analysis of proteins  
Authors: Marco Giulini, Raffaele Potestio  
Abstract: - to express the information contained in the molecule's atomic positions and distances in a set of input quantities that the network can process. Many of the molecular descriptors derived insofar are effective and manageable for relatively small structures, but become complex and cumbersome for larger ones. Furthermore, most of them are defined... [More](#)  
Submitted 1 January, 2019; originally announced January 2019.
- arXiv:1812.03780 [pdf, other]** [\[cs.CV\]](#)  
Efficient Condition-based Representations for Long-Term Visual Localization  
Authors: Hugo German, Guillaume Bourmaud, Vincent Lepetit  
Abstract: We propose an approach to localization from images that is designed to explicitly handle the strong variations in appearance happening when capturing conditions change throughout the day or across seasons. As revealed by recent long-term... [More](#)  
Submitted 10 December, 2018; originally announced December 2018.
- arXiv:1812.03590 [pdf, other]** [\[cs.CV\]](#)  
From Coarse to Fine: Robust Hierarchical Localization at Large Scale  
Authors: Paul-Eduard Sarlin, Cesar Cadena, Roland Siegwart, Martin Dymczyk  
Abstract: Robust and accurate visual localization is a fundamental capability for numerous applications, such as autonomous driving, mobile robotics, or augmented reality. It remains, however, a challenging task, particularly for large-scale environments and in presence of significant appearance changes. State-of-the-art methods not only struggle with such scenarios... [More](#)  
Submitted 9 December, 2018; originally announced December 2018.  
Comments: v1
- arXiv:1812.07050 [pdf, other]** [\[cs.CV\]](#)  
3D Point Cloud Learning for Large-scale Environment Analysis and Place Recognition  
Authors: Zhe Liu, Shurbo Zhou, Chuanhe Sun, Yingtan Liu, Hesheng Wang, Yun-Hai Liu  
Abstract: In this paper, we develop a new deep neural network which can extract discriminative and generalizable global descriptors from the raw 3D point cloud. Specifically, two novel modules, Adaptive... [More](#)  
Submitted 10 December, 2018; originally announced December 2018.
- arXiv:1812.04174 [pdf, other]** [\[eess.IV\]](#) [\[dsc\]](#) [\[10.1109/CASPP.2017.7952389\]](#)  
10

- P@K. Proportion of retrieved top-k papers which are relevant.  $P@K = \frac{\text{RelevantRetrievedTop}K}{\text{RelevantRetrievedTop}K + \text{NotRelevantRetrievedTop}K}$
- R@K. Proportion of relevant papers which are retrieved in top-k.  $R@K = \frac{\text{RelevantRetrievedTop}K}{\text{RelevantRetrievedTop}K + \text{RelevantRejected}}$

# Performance metrics

- **Average precision.**

- ▶ Given a query, analyze the ranking list at each relevant entry at a time, and measure the precision for the relevant entry  $P@K = \frac{\text{RelevantRetrievedTopK}}{\text{RelevantRetrievedTopK} + \text{NotRelevantRetrievedTopK}}$ .
- ▶ Average precision is the average of all  $P@K$  values.

Title	Relevant	R@k	P@k
Paper 1		0.10	1.00
Paper 2		0.10	0.50
Paper 3		0.20	0.67
Paper 4		0.30	0.76
Paper 5		0.40	0.80
Paper 6		0.50	0.83
Paper 7		0.60	0.86
Paper 8		0.60	0.75
Paper 9		0.70	0.78
Paper 10		0.70	0.70
Paper 11		0.80	0.73
Paper 12		0.80	0.67
Paper 13		0.80	0.62
Paper 14		0.90	0.64
Paper 15		0.90	0.60
Paper 16		0.90	0.56
Paper 17		0.90	0.53
Paper 18		0.90	0.50
Paper 19		0.90	0.47
Paper 20		1.00	0.50
		average-precision	0.76

● P@K: Proportion of retrieved top k papers which are relevant.  $P@K =$

$\frac{\text{RelevantRetrievedTopK}}{\text{RelevantRetrievedTopK} + \text{NotRelevantRetrievedTopK}}$

# Performance metrics

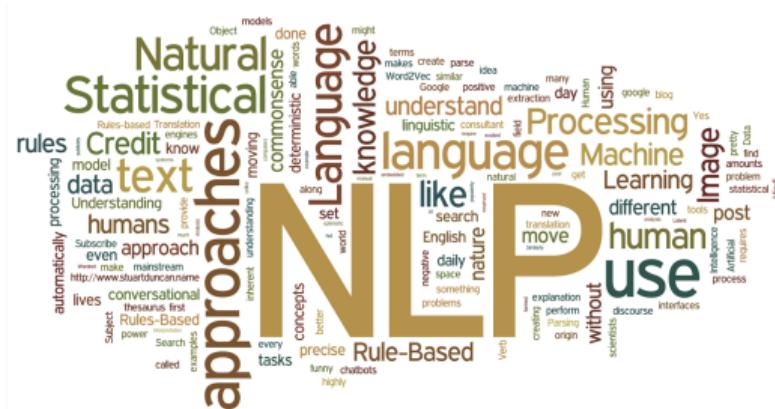
- Mean average precision.

- Given a set of queries, mean average precision is the average of all average precision values.

Title	Relevant	R@k	P@k	Title	Relevant	R@k	P@k	Title	Relevant	R@k	P@k
Paper 1		0.10	1.00	Paper 1		0.10	1.00	Paper 1		0.00	0.00
Paper 2		0.10	0.50	Paper 2		0.20	1.00	Paper 2		0.00	0.00
Paper 3		0.20	0.67	Paper 3		0.30	1.00	Paper 3		0.00	0.00
Paper 4		0.30	0.76	Paper 4		0.40	1.00	Paper 4		0.00	0.00
Paper 5		0.40	0.80	Paper 5		0.50	1.00	Paper 5		0.00	0.00
Paper 6		0.50	0.83	Paper 6		0.60	1.00	Paper 6		0.00	0.00
Paper 7		0.60	0.86	Paper 7		0.70	1.00	Paper 7		0.00	0.00
Paper 8		0.60	0.75	Paper 8		0.80	1.00	Paper 8		0.00	0.00
Paper 9		0.70	0.78	Paper 9		0.90	1.00	Paper 9		0.00	0.00
Paper 10		0.70	0.70	Paper 10		1.00	1.00	Paper 10		0.00	0.00
Paper 11		0.80	0.73	Paper 11		1.00	0.91	Paper 11		0.10	0.09
Paper 12		0.80	0.67	Paper 12		1.00	0.83	Paper 12		0.20	0.17
Paper 13		0.80	0.62	Paper 13		1.00	0.77	Paper 13		0.30	0.23
Paper 14		0.90	0.64	Paper 14		1.00	0.71	Paper 14		0.40	0.29
Paper 15		0.90	0.60	Paper 15		1.00	0.67	Paper 15		0.50	0.33
Paper 16		0.90	0.56	Paper 16		1.00	0.63	Paper 16		0.60	0.38
Paper 17		0.90	0.53	Paper 17		1.00	0.59	Paper 17		0.70	0.41
Paper 18		0.90	0.50	Paper 18		1.00	0.56	Paper 18		0.80	0.44
Paper 19		0.90	0.47	Paper 19		1.00	0.53	Paper 19		0.90	0.47
Paper 20		1.00	0.50	Paper 20		1.00	0.50	Paper 20		1.00	0.50
	average-precision	0.76			average-precision	1.00			average-precision	0.33	
				mean-average-precision	0.69						

# Performance metrics in Natural Language Processing

- **Evaluation metrics for text generation**, e.g, translation from one language to another language, speech synthesis:
    - ▶ BLEU (Bilingual Evaluation Understudy Score)
    - ▶ METEOR (Metric for Evaluation of Translation with Explicit Ordering)
    - ▶ WER (Word Error Rate)
  - **Evaluation metrics for text summarizing**, e.g, automatic paper summary:
    - ▶ ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - **Evaluation metrics for image description**, e.g, describing one image with a sentence:
    - ▶ CIDEr (Consensus-based Image Description Evaluation)



# Performance metrics in NLP

- **BLEU (Bilingual Evaluation Understudy Score)** the main idea is "the closer a machine translation is to a professional human translation, the better it is" \*1
  - ▶ It requires two inputs: 1) a machine translation; 2) a set of human translations.
    - ★ Candidate: the the the the the the.
    - ★ Reference 1. The cat is on the mat.
    - ★ Reference 2. There is a cat on the mat.
  - ▶ **n-gram precision**  $P = m/w_t$ , where  $m$  is number of words (n-gram) in candidate translation which occur in any reference translations and  $w_t$  is the total number of words (n-gram) in the candidate translation. Thus the precision of the candidate text is  $P = 7/7 = 1$
  - ▶ **modified n-gram precision**  $P = m_{max}/w_t$ , where  $m_{max}$  is maximum total count of unigrams in any reference translation  $P = 2/7$
  - ▶ **modified n-gram precision** on full blocks of text.  $p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{Count(n-gram')}$

\*1. Papineni, K., et al. BLEU: a method for automatic evaluation of machine translation (PDF) in ACL 2002:

## Performance metrics

- METEOR (Metric for Evaluation of Translation with Explicit Ordering)<sup>\*2</sup>

- ▶ The metric is an extension of BLEU metric providing stemming and synonymy matching. It's based on the harmonic mean of n-gram precision and recall with recall weighted higher than precision.
- ▶ **n-gram recall**  $R = m/w_r$ , where  $m$  is number of words (n-gram) in candidate translation which occur in any reference translations and  $w_r$  is the number of words (n-gram) in the reference translation.
- ▶ **harmonic mean**  $F_{mean} = \frac{10PR}{R+9P}$ , precision and recall are combined using a type of harmonic mean
- ▶ **chunk** is a set of unigrams that are adjacent in the hypothesis and in the reference. The longer the adjacent mappings between the candidate and the reference, the fewer chunks there are.
- ▶ **penalty**  $0.5 \cdot \left(\frac{c}{u_m}\right)^3$ , where  $c$  is the number of chunks, and  $u_m$  is the number of mapped n-grams.
- ▶ **meteor**  $M = F_{mean}(1 - P)$ , the final meteor metric where the penalty has the effect of reducing the  $F_{mean}$  by up to 50%

\*<sup>2</sup>. Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments in ACL 2005.

## Performance metrics

- **Evaluation metrics for text summarizing,**
- **Rouge (Recall-Oriented Understudy for Gisting Evaluation)<sup>\*3</sup>**
  - ▶ The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.
- **Evaluation metrics for image description,**
- **CIDEr (Consensus-based Image Description Evaluation)<sup>\*4</sup>**
  - ▶ The metrics evaluate for image how well a candidate caption matches the consensus of a set of image descriptions.

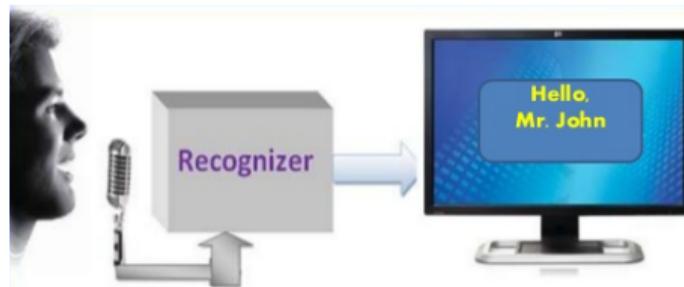
<sup>\*3</sup>. Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In WAS 2004.

<sup>\*4</sup>. Vedantam et al. CIDEr: Consensus-based image description evaluation. In: CVPR. (2015)

## Performance metrics

- Evaluation measure for Speech recognition system. WER (Word Error Rate)<sup>\*6</sup>

- ▶  $WER = (S + D + I)/N$ , where **S** is number of substitutions, **D** number of deletions and **I** insertions, **N** number of words in the reference.
- ▶ **Deletion** input: I read interesting papers → result: I read papers
- ▶ **Insertion** input: I read papers → result: I read interesting papers
- ▶ **Substitution** input: I read interesting papers → result: I read relevant papers

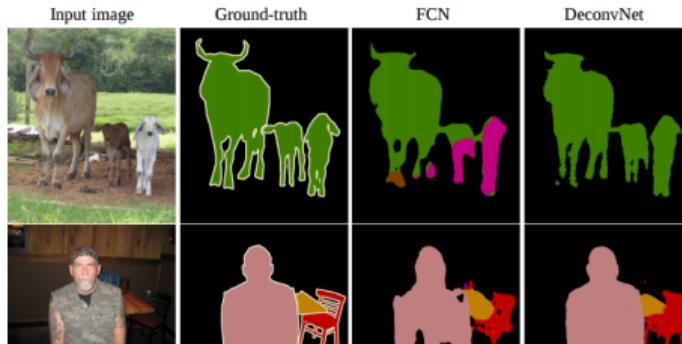


<sup>\*6</sup>. Klakow Dietrich and Jochen Peters. Testing the correlation of word error rate and perplexity. In Speech Communication 2013.

## Performance metrics

- **Evaluation measure for image segmentation.** Given  $A$  the automatic segmented image and  $G$  the Grand Truth.
- **Dice coefficient.** It measures of the extent of spatial overlap between two binary images. Typically used for training.
  - ▶ 
$$D = \frac{2(A \cap G)}{(A \cap G + A \cup G)}$$
- **Jaccard index.** Also referred to as the intersection over union (IoU) metric. It measures the number of pixels common between the grand truth and prediction masks divided by the total number of pixels present across both masks.

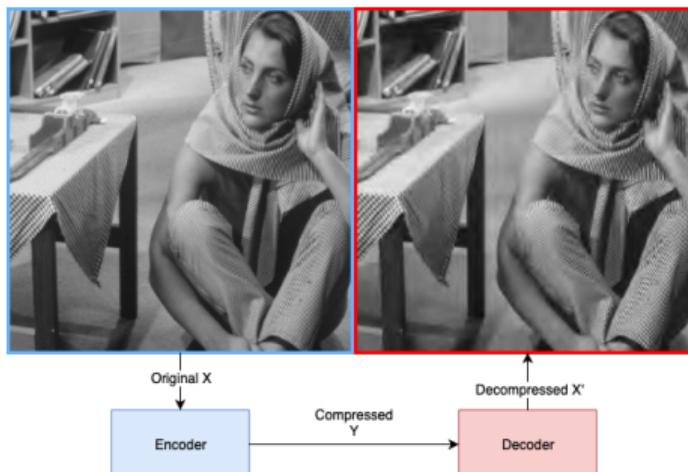
- ▶ 
$$IoU = \frac{G \cap A}{G \cup A}$$



# Performance metrics

- **Evaluation measure for signal quality.** Suppose you have an original signal  $X$  and reconstructed one  $X'$  with size e.g. for image  $M \times N$

- $MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} ||X(i,j) - X'(i,j)||^2$
- $PNR = 20 * \log_{10} \frac{\text{MAX}(X)}{\sqrt{MSE}}$
- $SNR = 20 * \log_{10} \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} X(i,j)^2}{\sqrt{MSE}}$



## Performance metrics

- Evaluation measure for quality of images.
- Inception score.\*<sup>5</sup>. it encodes two desirable properties:
  - ▶ the images generated should contain clear objects
  - ▶ it should output a high diversity of images from all the different classes in ImageNet.
- $ISG(G) = \exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|k)||p(y)))$
- $x \sim p_g$  indicates that  $x$  is an image sampled from  $p_g$ , which is the distribution encoded by the generative model  $G$ .
- $D_{KL}$  is the KL-divergence between the distributions  $p$  and  $q$ , where  $q$  is the distribution of real images.
- $p(y|x)$  is the conditional class distribution indicating the probability that the Inception v3 Network assigns the class  $y$  to the sample  $x$ .
- $p(y) = \int_x p(y|x) * p_g(x)$  is the marginal class distribution.



\*<sup>5</sup>. Salimans et al. Improved techniques for training GANs. In NIPS 2016.

# Performance Metrics

- General
  - True positives, false positives, true negatives, false negatives
  - Precision/recall
  - F-measure, F1
  - ROC, AUC
  - Retrieval: Rank@K, Prec@K, mAP
- NLP: BLEU, ROUGE, METEOR, CIDEr
- Speech: Word Error Rate (WER)
- Signal quality: MSE, SNR, PSNR
- Image quality: Inception score
- Segmentation: Dice coefficient, Jaccard index

