Day 8

Speaker : Prof Vijay Namboodiri

Title : Multimodal Learning

Learning with Multiple Modalities

Vinay P. Namboodiri
vpn22@bath.ac.uk

UNIVERSITY OF BATH

UNIVERSITY OF BATH

Motivation

- Consider that we are interested in an individual (for e.g., your Grandmother or a favorite actor) how would our brains represent information about that particular individual?
- As per research our brains respond to all aspects of that individual, for instance the picture of that individual, or the name or voice. Thus multiple modalities of information are used by our brains to represent that individual.

* Our brains inherently capture multimodal scenes

https://www.newscientist.com/article/dn7567-why-your-brain-has-a-jennifer-aniston-cell/
https://en.wikipedia.org/wiki/Grandmother_cell

McGurk Effect.

UNIVERSITY OF BATH

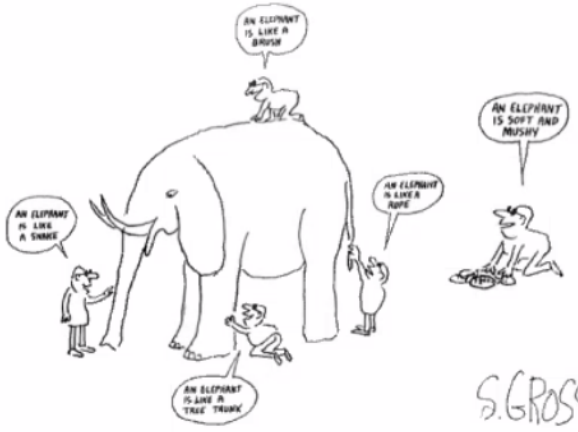
McGurk Effect

McGurk, H & MacDonald, J (1976); "Hearing lips and seeing voices," Nature, Vol 264(5588), pp. 746–748.
https://www.youtube.com/watch?v=PWGeUztTkRA

* Same sound ⟶ affected by lip reading what we perceive

UNIVERSITY OF BATH

Learning with multiple modalities



Each community has its own interpretation and is a very active area of research

All communities like CV, NLP, RL ⟶ coming closer

Solving complex tasks

- Imagine that you would like to make an application that could help a visually impaired person using deep learning
- This would naturally involve complex tasks like vision based question answering. Could include vision, language and audio modalities
- Much progress for each individual modality
  - object detection, machine translation and speech recognition,
- We naturally aim to move towards multimodal tasks that are practically useful

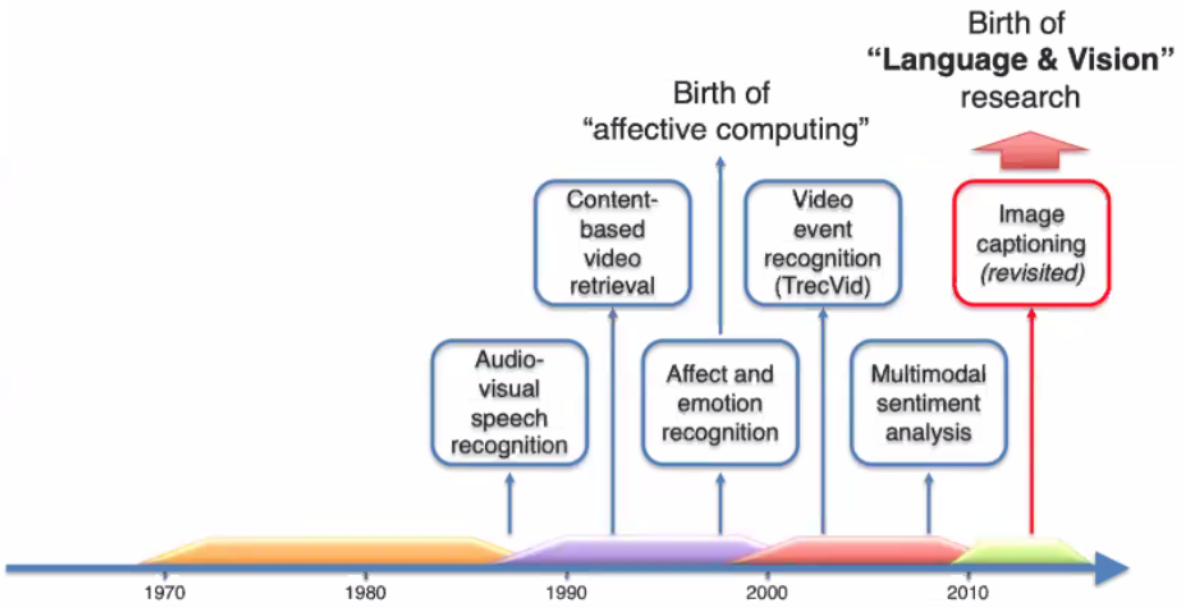


Fig credit: Louis Philippe Morency: Multimodal Machine Learning Course
https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

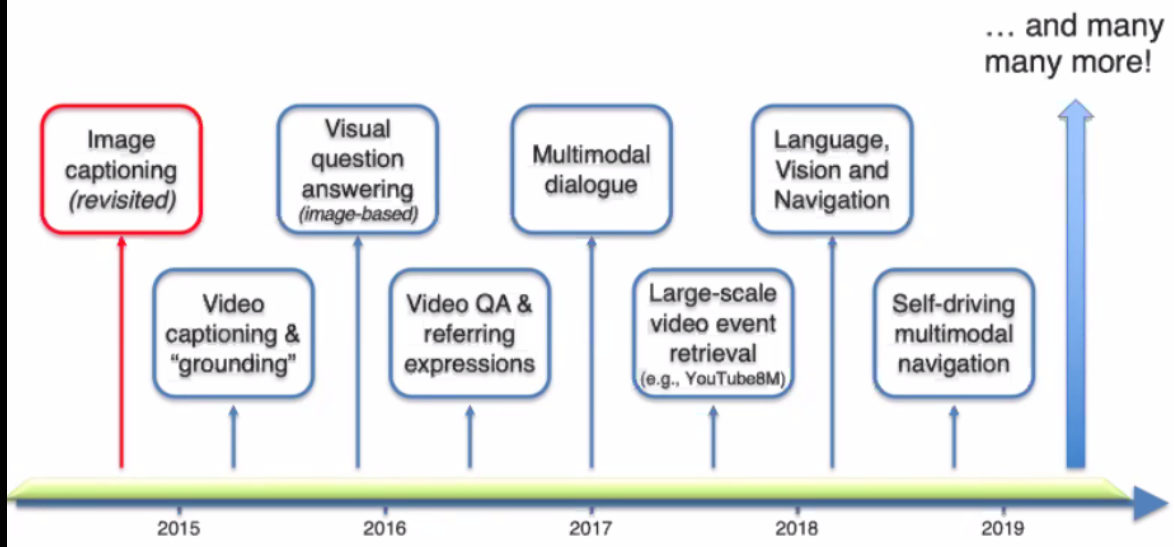
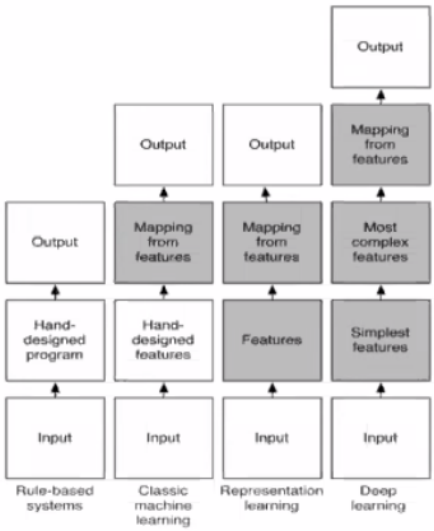


Fig credit: Louis Philippe Morency: Multimodal Machine Learning Course
https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

* Multimodalities are universal

UNIVERSITY OF BATH

Deep representation learning



Y. Bengio et al, ``Deep Learning'', MIT Press, 2015

Deep representation learning

- A number of techniques that take in input data and learn through a hierarchical series of transformations to map to an output data that is target
- The representation and the series of transformations at each layer are determined by the algorithm
- Can be thought of as learning a series of invariances at each level - (cf: Work by Stephen Mallat)

Layer 1 → Eg. phone at loc x = phone at loc y {translational invariance}

Layer 2 → Rotational Invariance

How about the different Modalities?

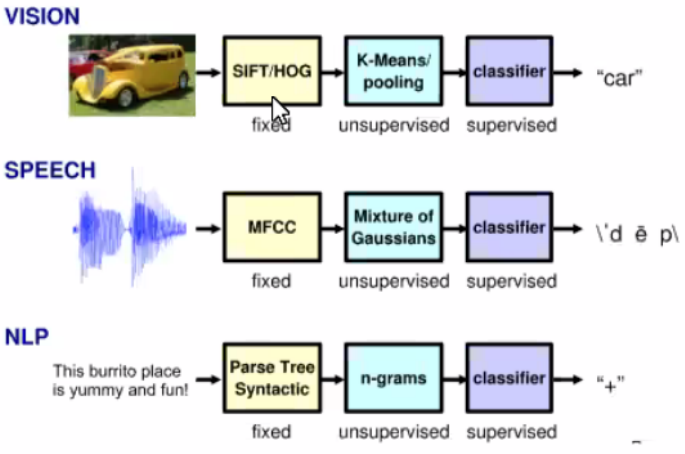
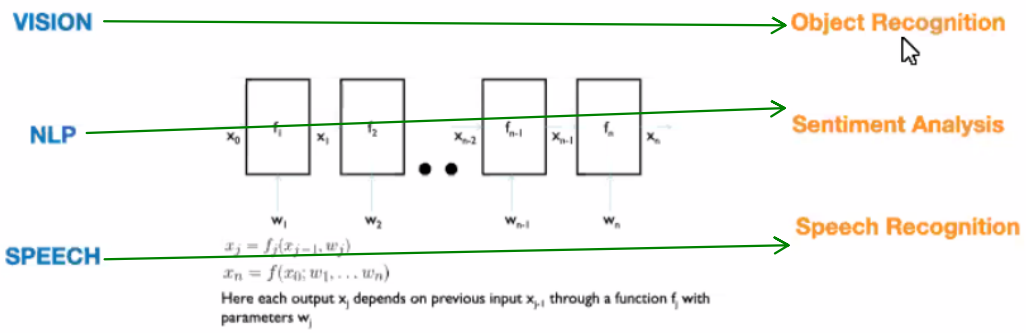
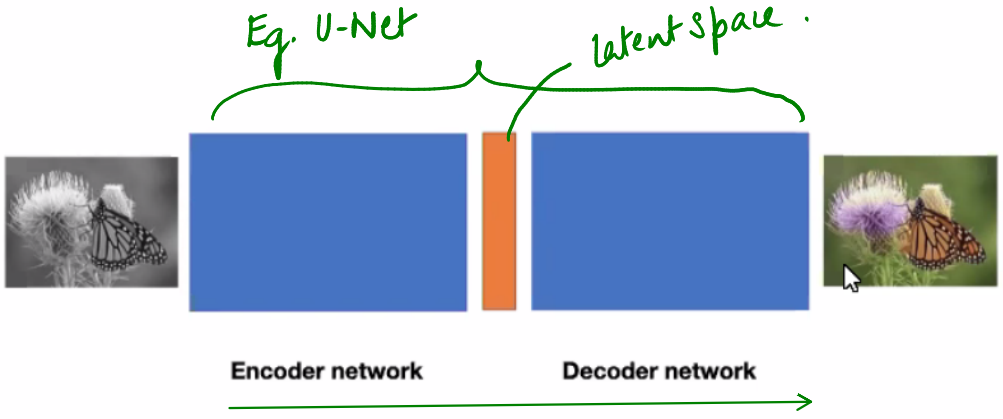


Fig credit: Marc Aurelio Ranzato, Tutorial, CVPR 2014

A consistent set of transformations for all modalities

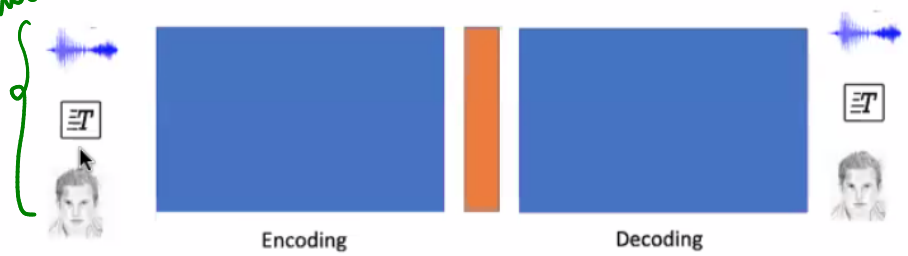


Encoder-Decoder formulation



Encoder-decoder architectures

Multimodal Task



Inherently multimodal and cross-modal

* Encoder-Decoder n/w's are inherently multimodal

Visual Question Answering

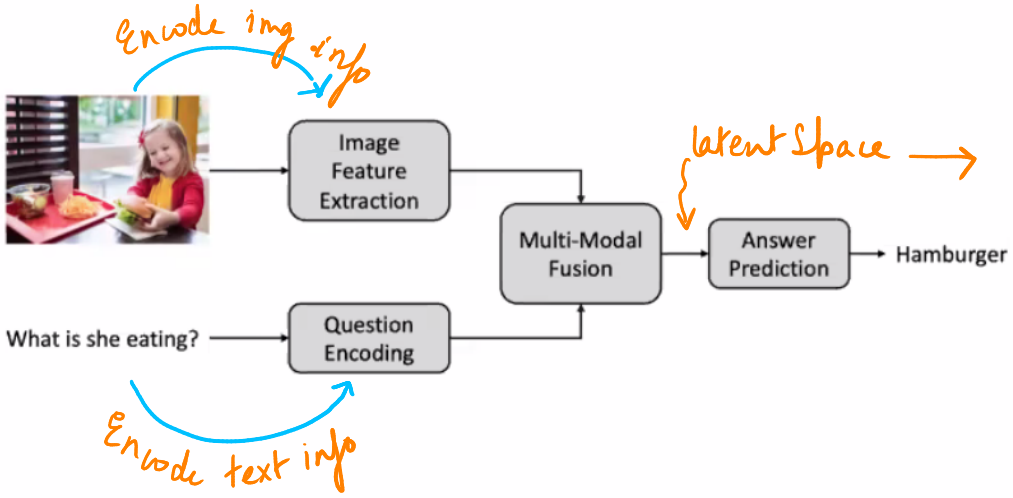
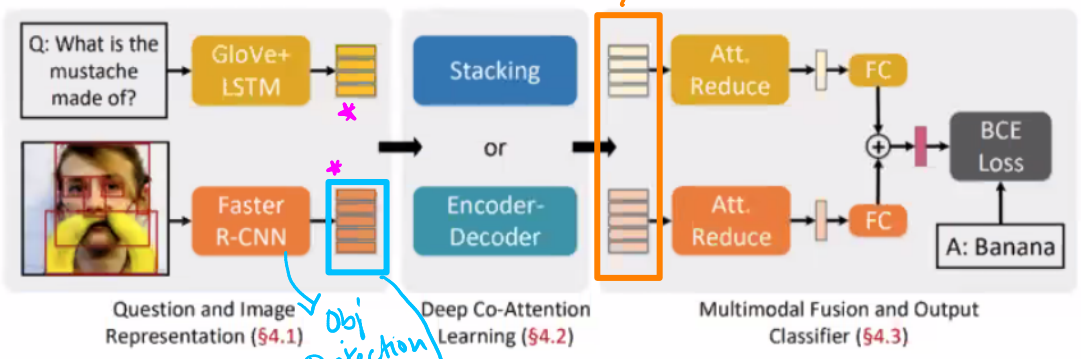


Fig credit: Zhe Gan, Tutorial: From VQA to VLN: Recent Advances in Vision-and-Language Research, CVPR 2021

UNIVERSITY OF BATH

MCAN: Deep Modular Co-Attention Network



[1] Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019
[2] Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering, CVPR 2019

attended features

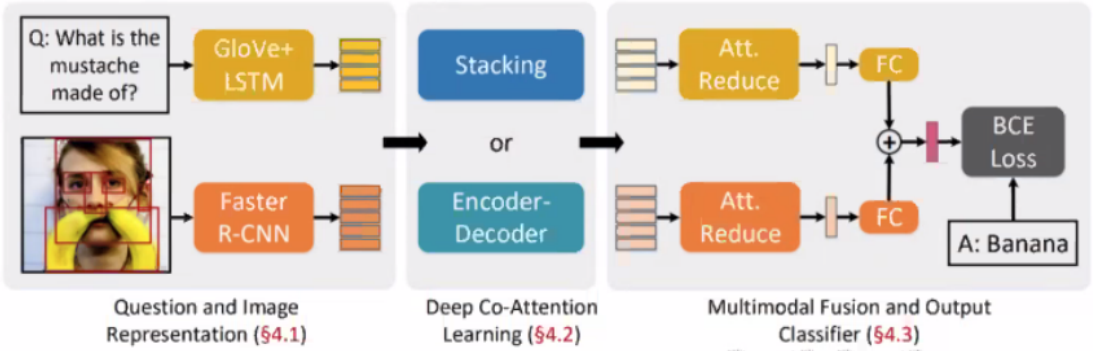
To read.

Obj Detection

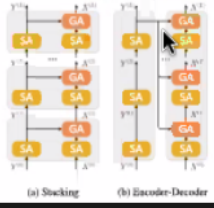
feature representation

UNIVERSITY OF BATH

MCAN: Deep Modular Co-Attention Network



[1] Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019
[2] Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering, CVPR 2019



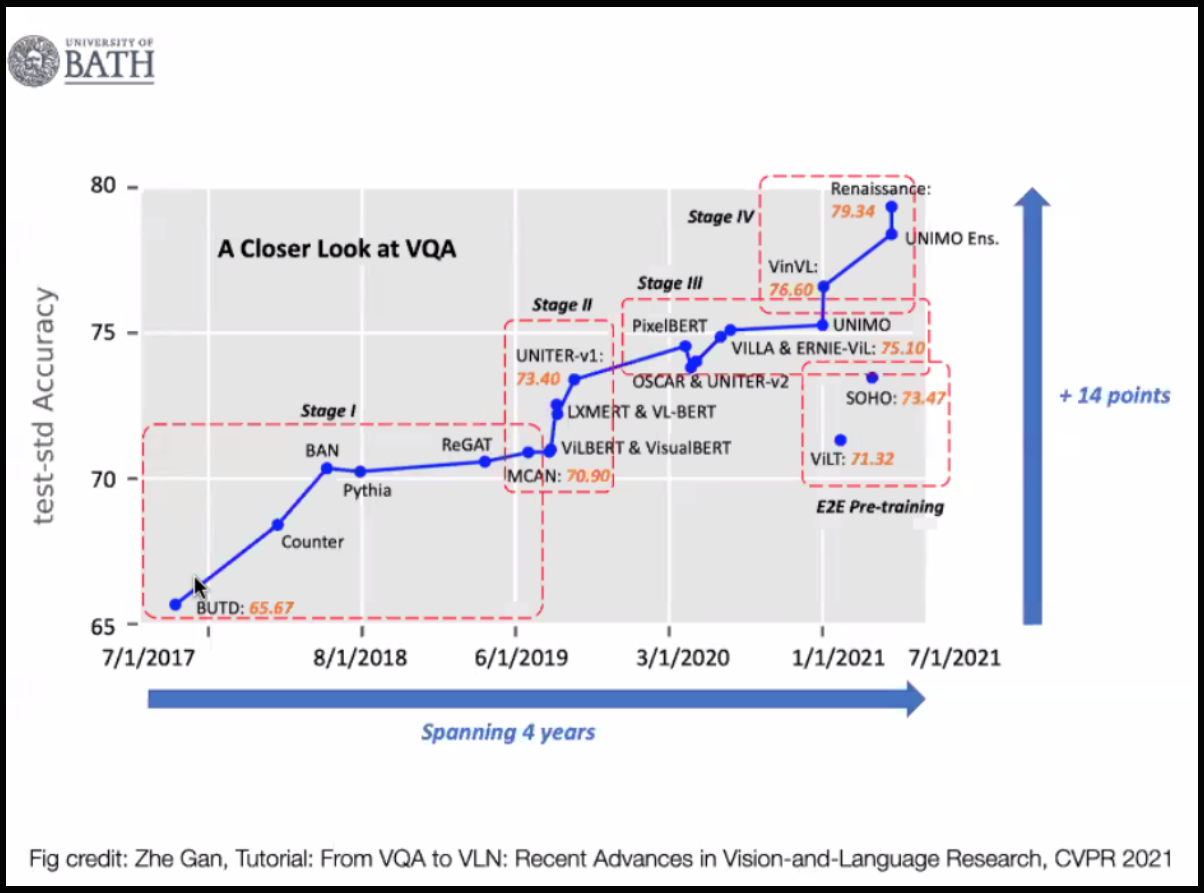


Fig credit: Zhe Gan, Tutorial: From VQA to VLN: Recent Advances in Vision-and-Language Research, CVPR 2021

UNIVERSITY OF BATH

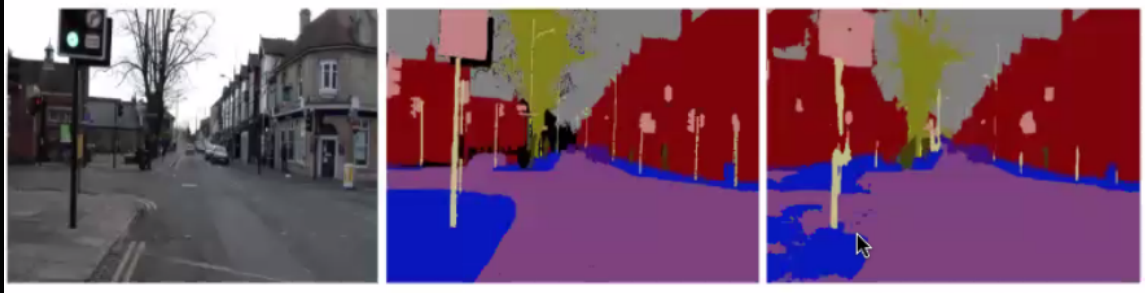
How to ensure we are not doing a 'Clever Hans'

* But actually horse was observing his trainer to utter the answer.

Motivation for Probabilistic Deep Learning

- A way to tell what our model knows and what it does not know (based on probability estimates)
- Suitability for conditioning data and reducing uncertainty estimates
- Principled framework with ability to incorporate semi-supervision based on need
- Ability to generalise to multiple modalities or multiple cues

Motivation Example



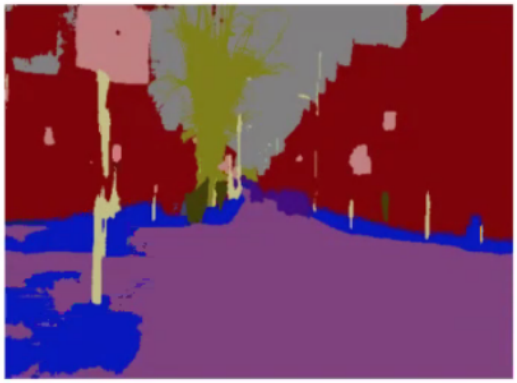
(a) Input Image (b) Ground Truth (c) Semantic Segmentation

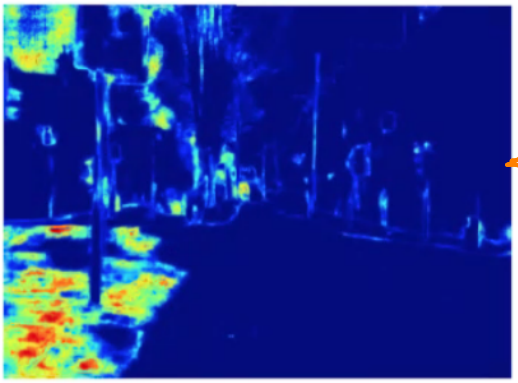
Alex Kendall and Yarin Gal.
What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?
Advances in Neural Information Processing Systems (NIPS), 2017

How do we train the model & segmentation ?

UNIVERSITY OF BATH

Motivation Example





knowing the uncertainties

Alex Kendall and Yarin Gal.
What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?
Advances in Neural Information Processing Systems (NIPS), 2017

UNIVERSITY OF BATH

Deep Learning

Conceptually simple models

Data: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$
Model: given matrices $\mathbf{W}$ and non-linear func. $\sigma(\cdot)$, define "network"

$$\tilde{\mathbf{y}}_i(\mathbf{x}_i) = \mathbf{W}_2 \cdot \sigma(\mathbf{W}_1 \mathbf{x}_i)$$

Objective: find $\mathbf{W}$ for which $\tilde{\mathbf{y}}_i(\mathbf{x}_i)$ is close to $\mathbf{y}_i$ for all $i \leq N$.

X → Y

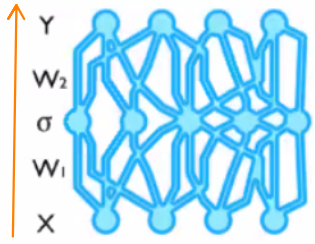


Fig credit: Yarin Gal:
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
Yarin Gal, Zoubin Ghahramani
Proceedings of The 33rd International Conference on Machine Learning, PMLR 48:1050-1059, 2016.

UNIVERSITY OF BATH

Deep Learning

Conceptually simple models

Data: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$
Model: given matrices $\mathbf{W}$ and non-linear func. $\sigma(\cdot)$, define "network"

$$\tilde{\mathbf{y}}_i(\mathbf{x}_i) = \mathbf{W}_2 \cdot \sigma(\mathbf{W}_1 \mathbf{x}_i)$$

Objective: find $\mathbf{W}$ for which $\tilde{\mathbf{y}}_i(\mathbf{x}_i)$ is close to $\mathbf{y}_i$ for all $i \leq N$.

Deep learning is awesome ✓

- Simple and modular
- Huge attention from practitioners and engineers
- Great software tools
- Scales with data and compute
- Real-world impact

... but has many issues ✗

- What does a model not know?
- Uninterpretable black-boxes
- Easily fooled (AI safety)
- Lacks solid mathematical foundations (mostly ad hoc)
- Crucially relies on big data

Fig credit: Yarin Gal:
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
Yarin Gal, Zoubin Ghahramani

UNIVERSITY OF BATH

Need for Uncertainty

- We need a way to tell **what our model knows** and what not.
  - We train a model to recognise dog breeds

Fig credit: Yarin Gal:
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
Yarin Gal, Zoubin Ghahramani

UNIVERSITY OF BATH

Need for Uncertainty

- We need a way to tell **what our model knows** and what not.
  - We train a model to recognise dog breeds
  - And are given a cat to classify
  - What would you want your model to do?

feed forward n/w → o/p — ??

Fig credit: Yarin Gal:
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
Yarin Gal, Zoubin Ghahramani

UNIVERSITY OF BATH

Need for Uncertainty

- We need a way to tell **what our model knows** and what not.
- Uncertainty gives insights into the black-box when it fails —where am I not certain?

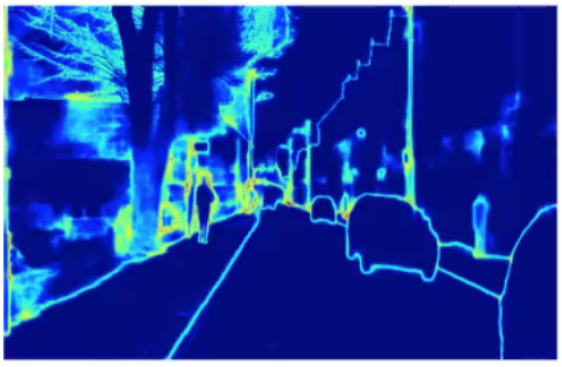


Fig credit: Yarin Gal:

Main Idea

- Can be shown that dropout training is identical to *approximate inference in Bayesian modelling* [Gal, 2016],
- Connecting **Deep Learning to Bayesian probability theory**.
- The **mathematically grounded** connection gives a treasure trove of new research opportunities:
  - **uncertainty** in deep learning, e.g. interpretability and AI safety
  - **principled extensions** to deep learning
  - enable deep learning in **small data** domains

Fig credit: Yarin Gal:

From Bayesian neural networks to Dropout

- Place **prior** $p(\mathbf{W})$ dist. on weights, making these r.v.s
- Given dataset $\mathbf{X}, \mathbf{Y}$, the r.v. $\mathbf{W}$ has a **posterior**: $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$
- Which is difficult to evaluate—many great researchers tried
- Can define **simple distribution** $q_{\mathbf{M}}(\cdot)$ and approximate

$$q_{\mathbf{M}}(\mathbf{W}) \approx p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$$

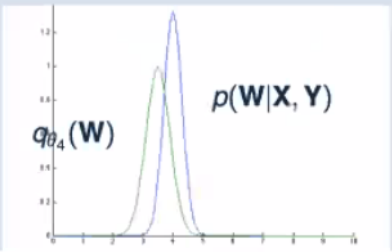


Fig credit: Yarin Gal:

* Variational Approach ↙ ↘ Monte Carlo Sampling

UNIVERSITY OF BATH

Main Idea

We would like to estimate the following

$$p(y^*|x^*, X, Y) = \int p(y^*|x^*, w)p(w|X, Y)dw$$

This is however, expensive. This can be approximated by the following

$$\begin{aligned} p(y^* = c|x^*, X, Y) &= \int p(y^* = c|x^*, w)p(w|X, Y)dw \\ &\approx \int p(y^* = c|x^*, w)q_\theta(w)dw \\ &\approx \frac{1}{M}\sum_{m=1}^{M} p(y^* = c|x^*, \hat{w}_m) \end{aligned}$$

with $\hat{w}_m \sim q_\theta(w)$, where $q_\theta(w)$ is called the dropout distribution.

→ learned wt. vector

{Learning with uncertainty}

UNIVERSITY OF BATH

Theorem (Dropout as approximate variational inference)

Define

$$q_{\mathbf{M}}(\mathbf{W}) := \mathbf{M} \cdot diag(Bernoulli)$$

with variational parameter $\mathbf{M}$.
The optimisation objective of (stochastic) variational inference with $q_{\mathbf{M}}(\mathbf{W})$ *is identical to the objective of a dropout neural network.*

Proof.

See Gal [2016]. □

Implementing **inference** with $q_{\mathbf{M}}(\mathbf{W})$

=

Implementing **dropout training**.

Line to line.

Fig credit: Yarin Gal:

Practical Implementation

In practical terms[1], given point x:

- drop units **at test time**
- **repeat 10 times**
- and look at **mean and sample variance**.
- Or in Python:

```
y = []
for _ in xrange(10):
    y.append(model.output(x, dropout=True))
y_mean = numpy.mean(y)
y_var = numpy.var(y)
```

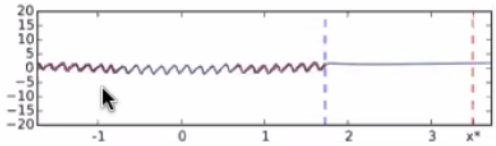
[1] Friendly introduction given in `yarin.co/blog`

Fig credit: Yarin Gal:

Example

What would be the CO_2 concentration level in Mauna Loa, Hawaii, in 20 years' time?

- Normal dropout:



- Same network, Bayesian perspective:

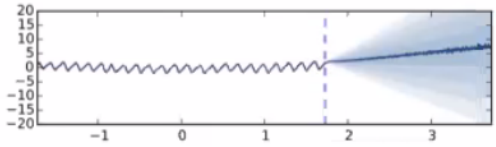
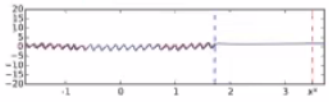


Fig credit: Yarin Gal:

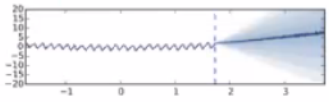
Example

What would be the CO_2 concentration level in Mauna Loa, Hawaii, in 20 years' time?

Normal dropout:



Bayesian perspective:



What can we do with this?

- Interpretability & AI safety
- Principled deep learning extensions
- Deep learning in small data domains

Fig credit: Yarin Gal:

Types of Uncertainty

- Epistemic Uncertainty (Model Uncertainty)
- Aleatoric Uncertainty (Observation Uncertainty)

UNIVERSITY OF BATH

Epistemic Uncertainty

Epistemic uncertainty captures our ignorance about which model generated our collected data. This uncertainty can be explained away given enough data, and is often referred to as *model uncertainty*. Epistemic uncertainty is really important to model for:

- Safety-critical applications, because epistemic uncertainty is required to understand examples which are different from training data,
- Small datasets where the training data is sparse.

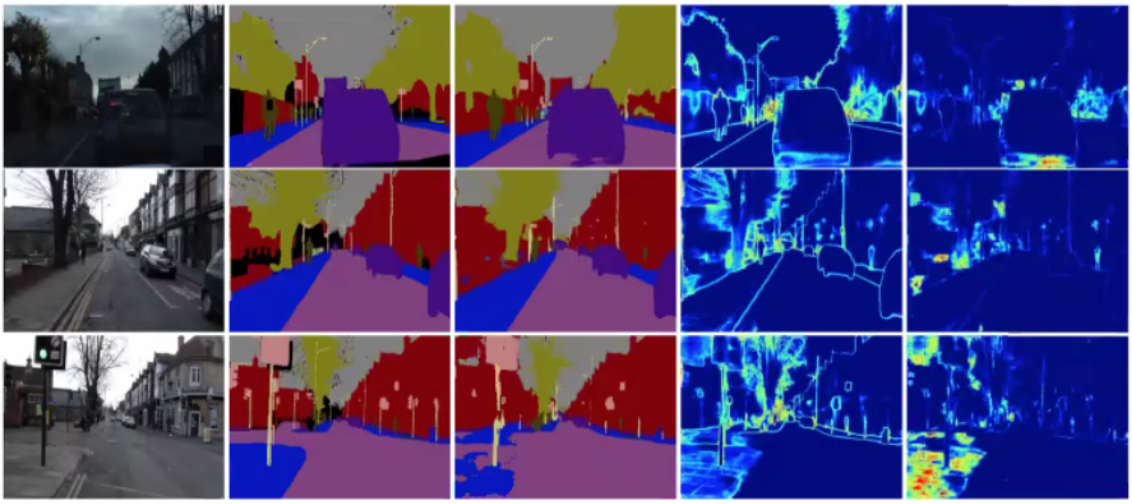
UNIVERSITY OF BATH

Aleatoric Uncertainty

Aleatoric uncertainty captures our uncertainty with respect to information which our data cannot explain. For example, aleatoric uncertainty in images can be attributed to occlusions (because cameras can't see through objects) or lack of visual features or over-exposed regions of an image, etc. It can be explained away with the ability to observe all explanatory variables with increasing precision. Aleatoric uncertainty is very important to model for:

- Large data situations, where epistemic uncertainty is mostly explained away,
- Real-time applications, because we can form aleatoric models as a deterministic function of the input data, without expensive Monte Carlo sampling.

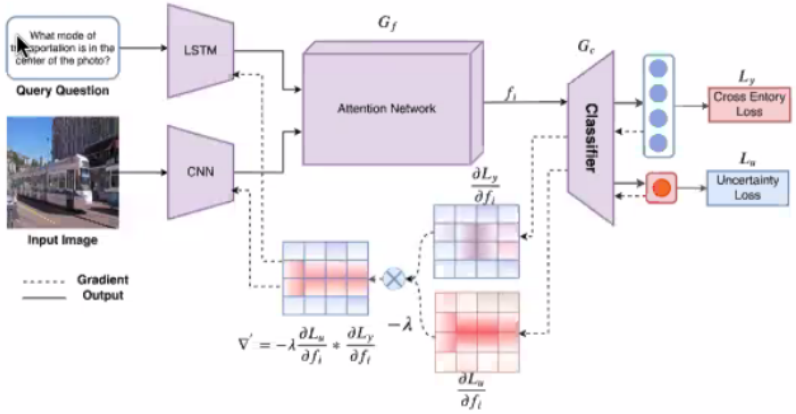
Types of Uncertainty



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

UNIVERSITY OF BATH

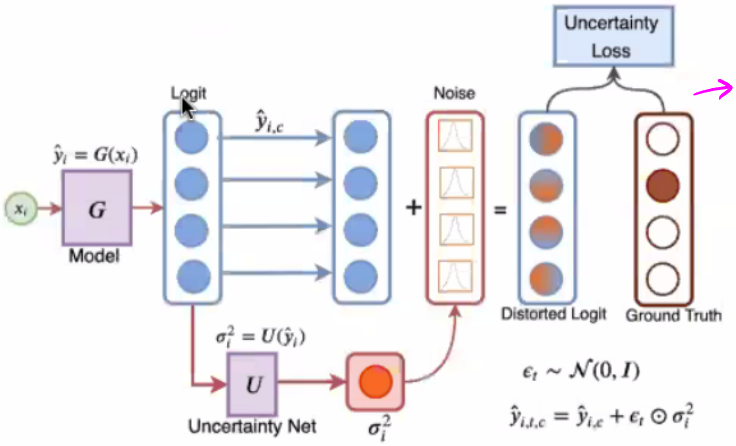
Introduction



Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

UNIVERSITY OF BATH

U-CAM: Visual Explanation using Uncertainty based Class Activation Maps

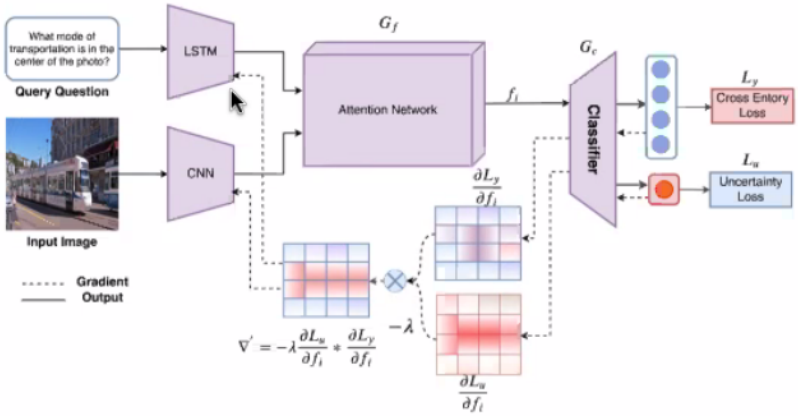


→ Captures Variance

Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

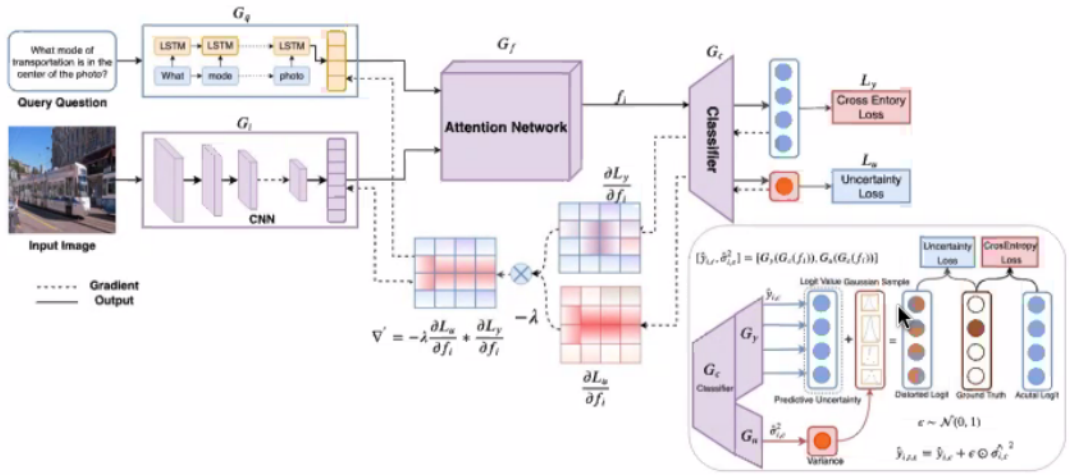
UNIVERSITY OF BATH

Introduction



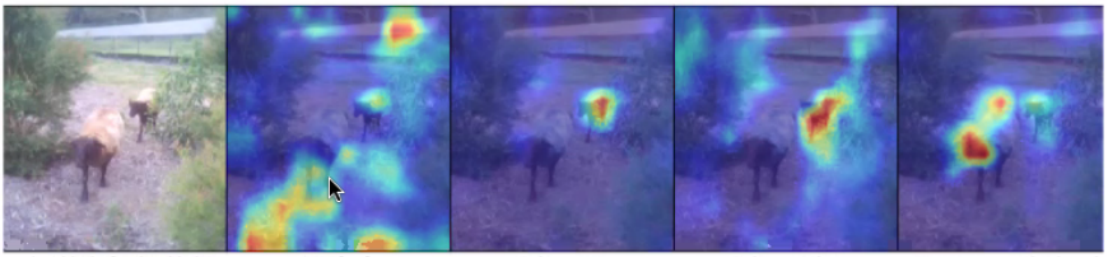
Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

Model



Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

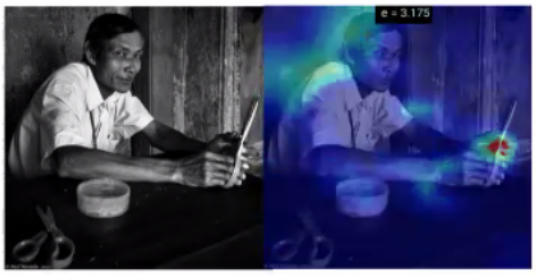
Observation



Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

Visualisation

What is the girl eating?



SOTA

Models	All	Y/N	Num	Oth
DPPnet [36]	57.2	80.7	37.2	41.7
SMem[[50]]	58.0	80.9	37.3	43.1
SAN [53]	58.7	79.3	36.6	46.1
DMN[49]	60.3	80.5	36.8	48.3
QRU(2)[32]	60.7	82.3	37.0	47.7
HieCoAtt [33]	61.8	79.7	38.9	51.7
MCB [14]	64.2	82.2	37.7	54.8
MLB [27]	65.0	84.0	37.9	54.7
DVQA[37]	65.4	83.8	38.1	55.2
P-GCA + SAN (ours)	60.4	80.7	36.6	47.9
A-GCA + MCB (ours)	66.3	84.2	38.0	55.5
P-GCA + MCB (ours)	**66.5**	**84.6**	**38.4**	**55.9**

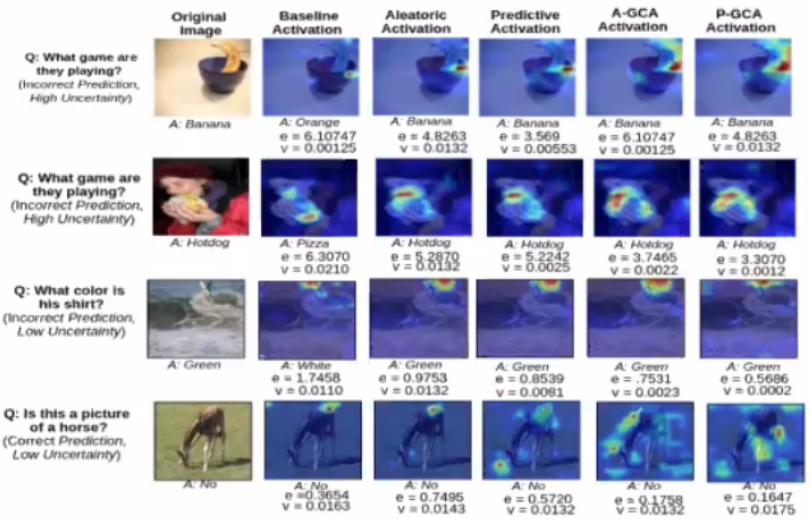
Table 4. SOTA: Open-Ended VQA1.0 accuracy on test-dev

Models	All	Y/N	Num	Oth
SAN-2[53]	56.9	74.1	35.5	44.5
MCB [14]	64.0	78.8	38.3	53.3
Bottom[[1]]	65.3	81.8	44.2	56.0
DVQA[37]	65.9	82.4	43.2	56.8
MLB [27]	66.3	83.6	44.9	56.3
DA-NTN [4]	67.5	84.3	47.1	57.9
Counter[54]	68.0	83.1	51.6	58.9
BAN[26]	**69.5**	85.3	**50.9**	**60.2**
P-GCA + SAN (ours)	59.2	75.7	36.6	46.8
P-GCA + MCB (ours)	65.7	79.6	40.1	54.7
P-GCA + Counter (ours)	69.2	**85.4**	50.1	59.4

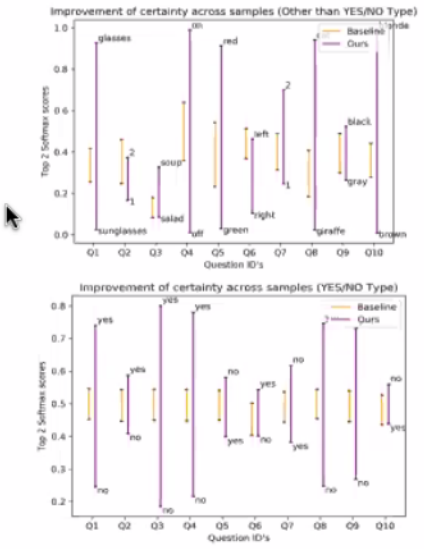
Table 5. SOTA: Open-Ended VQA2.0 accuracy on test-dev

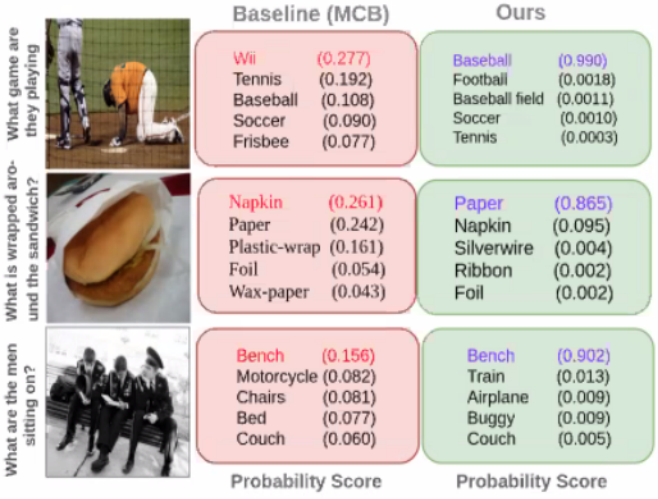
Badri N Patro, Mayank Lunayach, Vinay P Namboodiri (2020).Uncertainty Class Activation Map (U-CAM) using Gradient Certainty method'', IEEE Transactions on Image Processing,2020

Qualitative Result



Qualitative Results



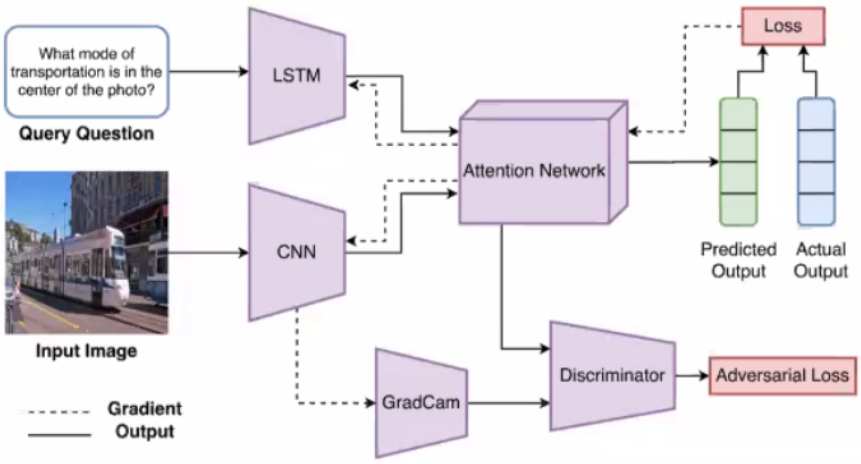


How do we ensure that the ‘right regions are attended to’

UNIVERSITY OF BATH

UNIVERSITY OF BATH

Introduction



Patro, B., Anupriy & Namboodiri, V. (2020, April). Explanation vs attention: A two-player game to obtain attention for vqa. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 11848-11855)

Does more information help?

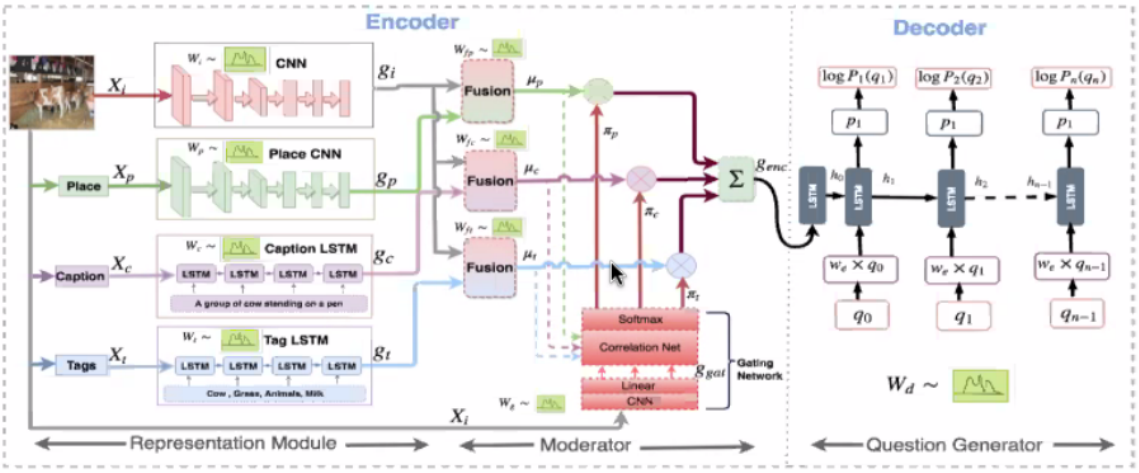
UNIVERSITY OF BATH

UNIVERSITY OF BATH

Incorporating multiple sources of information

Patro, B., Kurmi, V., Kumar, S., & Namboodiri, V. (2020). Deep Bayesian Network for Visual Question Generation. In The IEEE Winter Conference on Applications of Computer Vision (pp. 1566-1576).

Model Diagram



Patro, B., Kurmi, V., Kumar, S., & Namboodiri, V. (2020). Deep Bayesian Network for Visual Question Generation. In The IEEE Winter Conference on Applications of Computer Vision (pp. 1566-1576).

Evidence that certainty improves with more cues

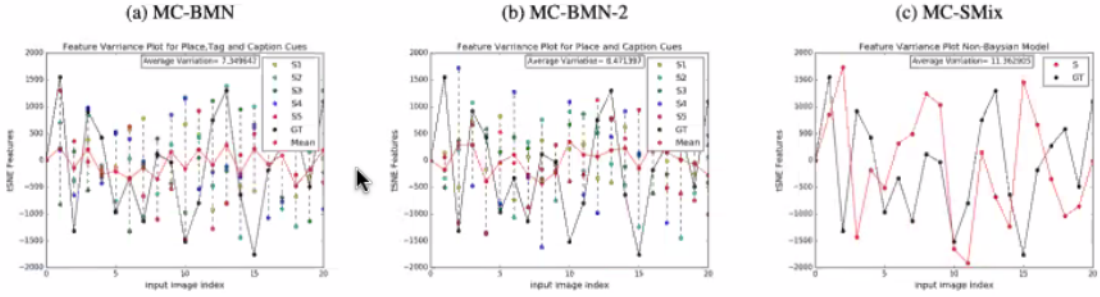


Figure 4. Variance plots for Bayesian and Non-Bayesian networks for a toy example of 20 images. We have drawn 5 samples of each image using Monte-Carlo sampling from a distribution (this is predictive posterior distribution for the Bayesian case) and then plot the mean features of these 5 samples along with the ground truth features. MC-BMN (3 cues) reduces normalized variance (difference in mean feature value & ground truth feature value) as compared to two cues(MC-BMN-2). Whereas for MC-SMix(Non-Bayesian network), the variance is too high as compared to MC-BMN.

Patro, B., Kurmi, V., Kumar, S., & Namboodiri, V. (2020). Deep Bayesian Network for Visual Question Generation. In The IEEE Winter Conference on Applications of Computer Vision (pp. 1566-1576).

UNIVERSITY OF BATH

Results

Methods	BLEU1		METEOR	
	Max	**Avg**	**Max**	**Avg**
Natural [38]	19.2	-	19.7	-
Creative [23]	35.6	-	19.9	-
MDN [43]	36.0	-	**23.4**	-
Img Only (Bernoulli Dropout (BD))	21.8	19.57 ±2.5	13.8	13.45±1.52
Place Only(BD)	26.5	25.36±1.14	14.5	13.60±0.40
Cap Only (BD)	27.8	26.40 ±1.52	18.4	17.60±0.65
Tag Only (BD)	20.3	18.13 ±2.09	12.1	12.10±0.61
Img+Place (BD)	27.7	26.96 ±0.65	16.5	16.00±0.41
Img+Cap (BD)	26.5	24.43 ±1.14	15.0	14.56 ±0.31
Img+Tag (BD)	31.4	29.96 ±1.47	20.1	18.96±1.08
Img+Place+Cap (BD)	28.7	27.86 ±0.74	18.1	15.56±1.77
Img+Place+Tag (BD)	30.6	28.46 ±1.58	18.5	17.60 ±0.73
Img+Cap+Tag (BD)	37.3	36.43 ±1.15	21.7	20.70±0.49
MC-SMN(Img+Place+Cap+Tag(w/o Dropout))	33.3	33.33 ±0.00	21.1	21.10 ±0.00
MC-BMN (Img+Place+Cap+Tag (Gaussian Dropout))	38.6	35.63 ±2.73	22.9	21.53 ±1.06
MC-BMN(Img+Place+Cap+Tag(BD)) (**Ours**)	**40.7**	**38.73** ±1.67	22.6	22.03 ±0.80
Humans[38]	86.0	–	60.8	–

Table 2. Comparison with **state-of-the-art** and different combination of **Cues**. The first block consists of the SOTA methods, second block depicts the models which uses only a single type of information such as Image or Place, third block has models which take one cue along with the Image information, fourth block takes two cues along with the Image information. The second last block consists of variations of our method. First is MC-SMN (Simple Moderator Network) in which there is no dropout (w/o Dropout) at inference time as explained in section 4.3 and the second one uses Gaussian dropout instead of the Bernoulli dropout (BD) which we have used across all the models.

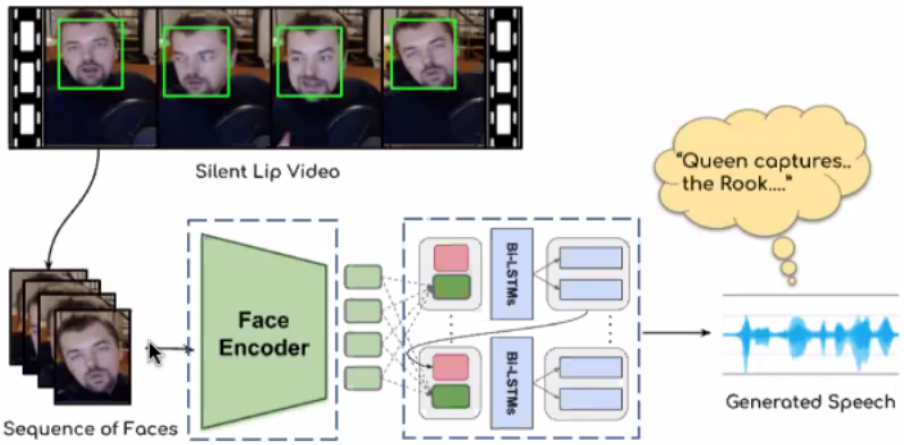
Audio-Visual Language Models

UNIVERSITY OF BATH

Encoder-decoder architectures

Inherently multimodal and cross-modal

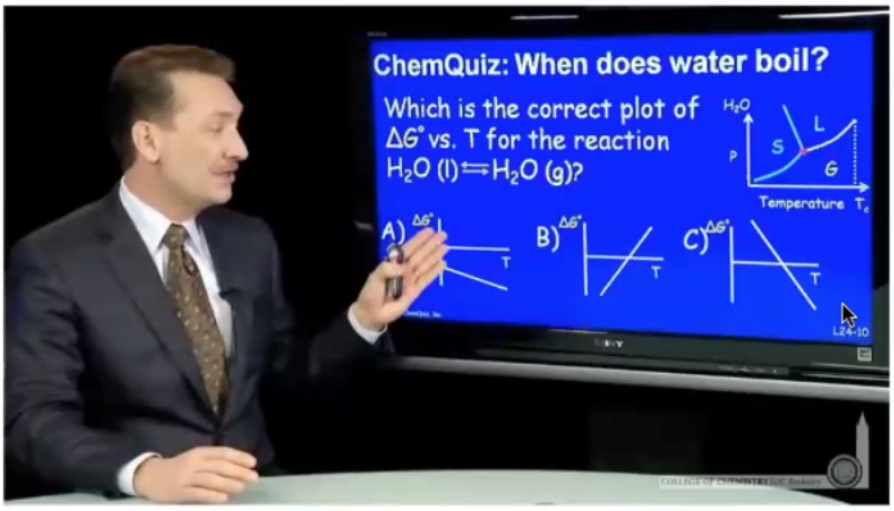
Sequence-to-Sequence Architecture for Lip to Speech



"Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis",Prajwal K.R., Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C.V. Jawahar, CVPR 2020

UNIVERSITY OF BATH

The speech you are hearing is fully generated from the lip movements using our model

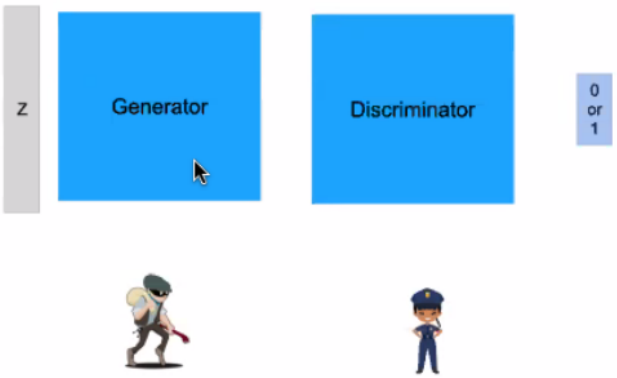


00:12 -00:00

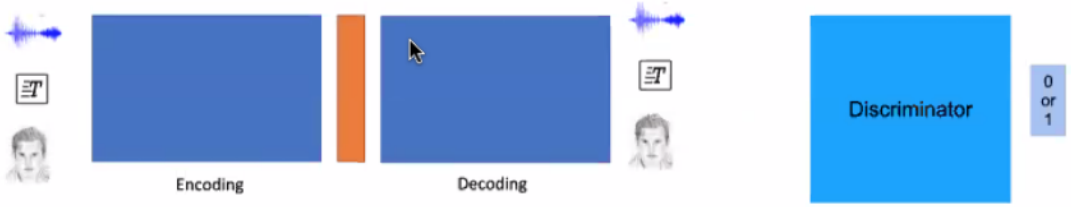
Can we go the other way?

UNIVERSITY OF BATH

Generative Adversarial Networks

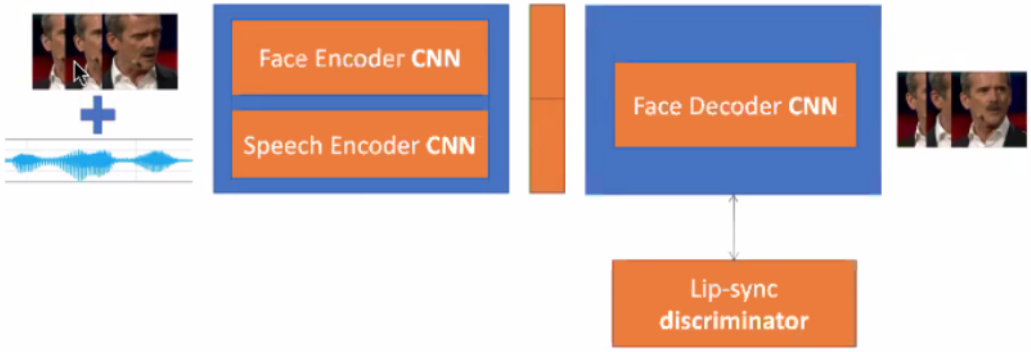


Use of Discriminators with Cross-modal generation



UNIVERSITY OF BATH

LipGAN for speech to lip generation



UNIVERSITY OF BATH

Correcting Lip Synchronization in Dubbed Movies

Background faded out for illustration purposes

Lip Sync corrected

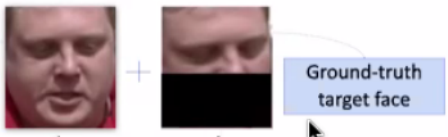
Unsynced

Movie: Harry Potter and chamber of secrets

Providing an additional Visual Prior

"Ask the model to in-paint just the lower half of the target face"

Train time



Reference face

Ground-truth target face

LipGAN

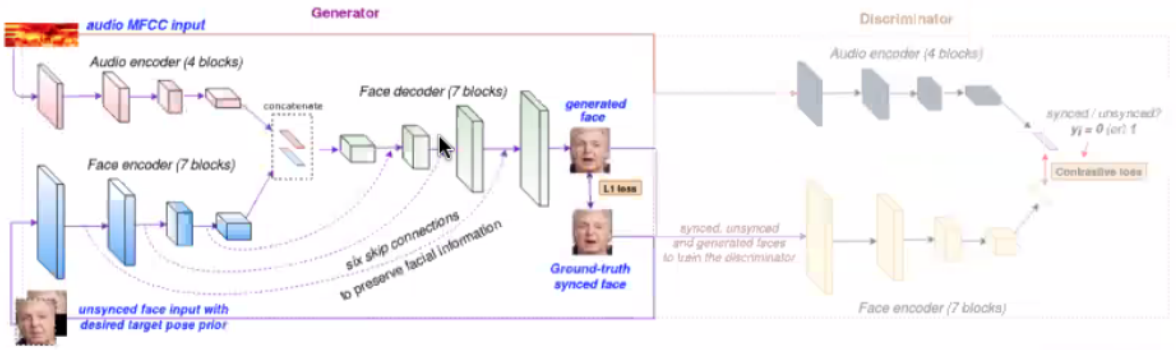
During inference on a real-world video

LipGAN

Same reference face as a prior! The network needs to change ONLY the lip-shape during generation

Significantly reduces artifacts!

LipGAN: The Generator



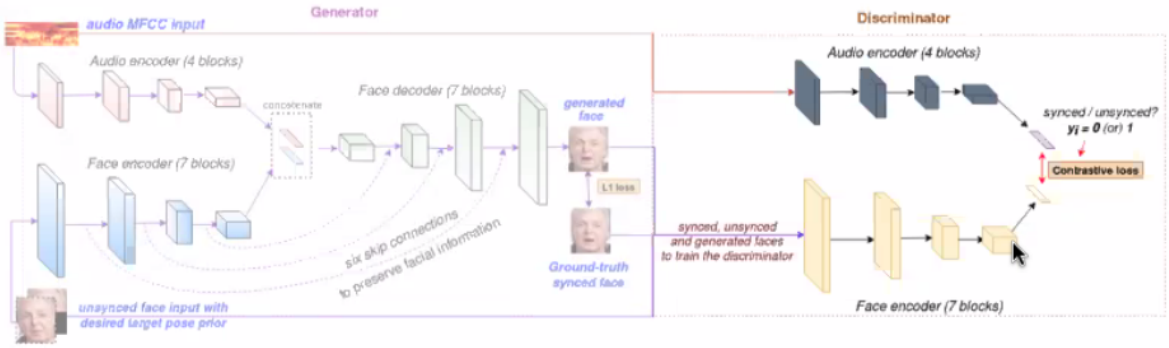
The generator is similar to the initial training framework discussed before, but with key design differences as mentioned previously.

Extensive skip connections across the encoder-decoder are used to preserve rich visual information.

A pose prior is concatenated channel-wise along with a reference frame, guiding the generator to generate a face in a desired pose.

The model generates lip-synced target faces in a desired pose, allowing for direct pasting into the video frame.

A discriminator for penalizing inaccurate lip-shapes



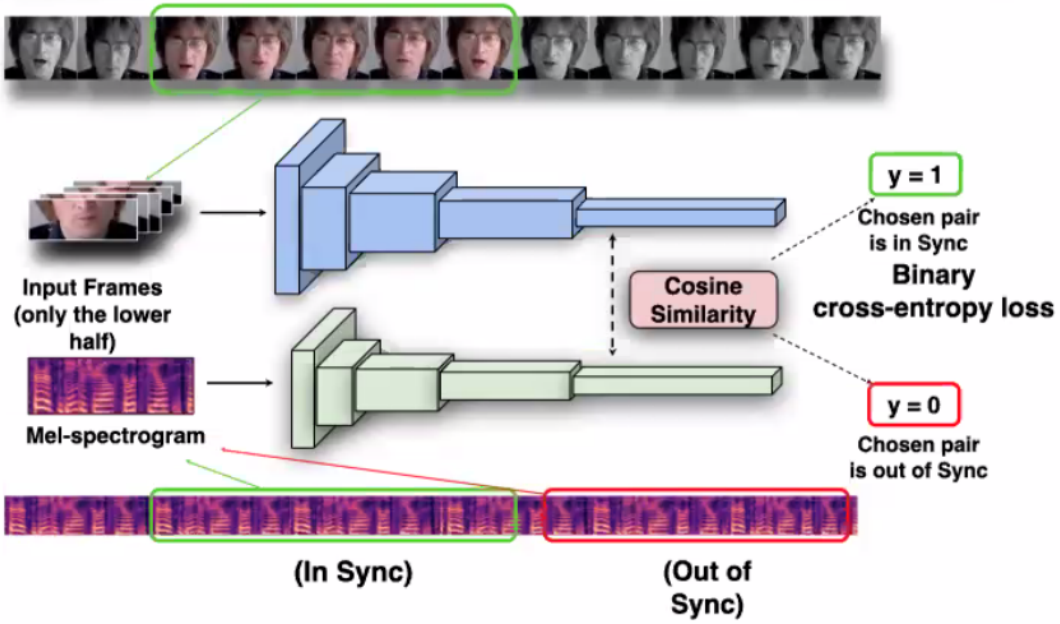
Lip region constitutes just 4% of the face, and is also influenced by large inter-person variations.

A large portion of the plain reconstruction loss is "used" for preserving the miscellaneous visual information. A discriminator is needed for a fine-grained attribute such as lip-sync.

We employ a discriminator that specifically checks if the generated faces are in-sync or out-of-sync with the audio.

The discriminator is trained in conjunction with the generator in a GAN setup. Details in the next slide.

Training a better expert to improve further - Lip-sync Expert



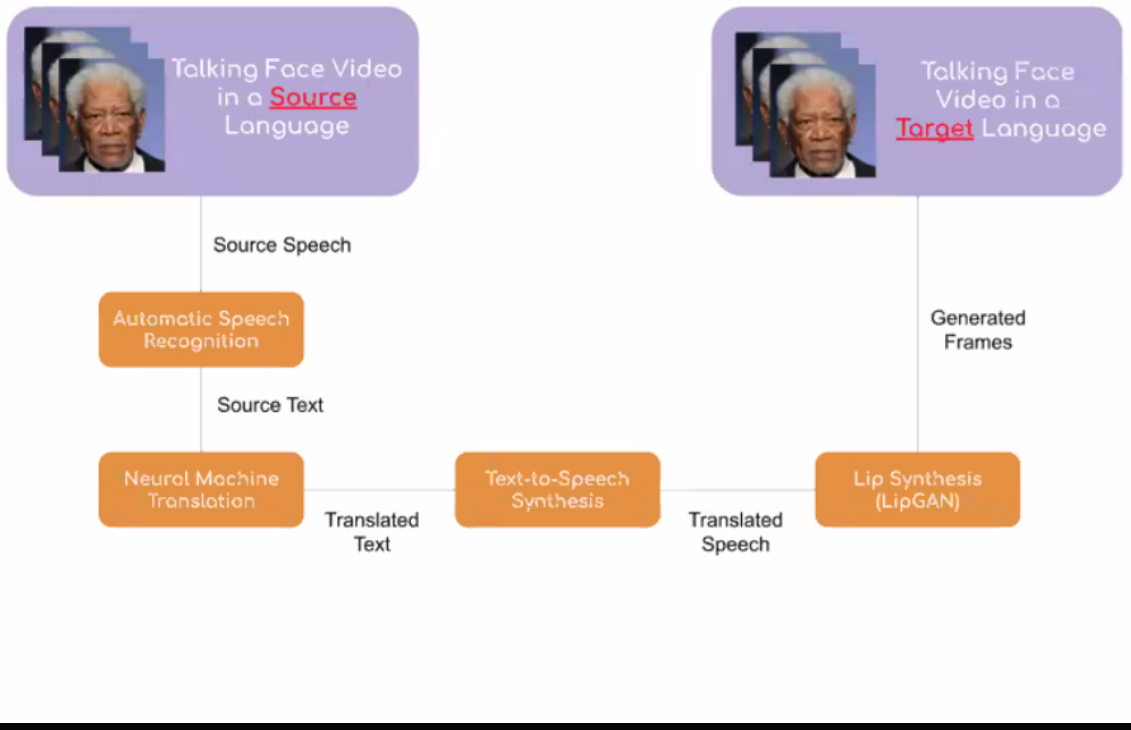
K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20)

UNIVERSITY OF BATH

Application #4: Animating CGI characters to speak with synchronized lip movements

Our model is able to lip-sync animated characters though we only train on human faces!

UNIVERSITY OF BATH



UNIVERSITY OF BATH

Potentially this can make these lectures available to thousands of non English viewers

Application #3

Original lecture is in English.

Our model produces accurate lip sync to the German speech

Note: Our model works for a TTS generated voice as well.

The original video is at: https://www.youtube.com/watch?v=ArPaAX_PhIs&list=PLkDaE6sCZn6GI29AoE31iwdVwSG-KnDzF

UNIVERSITY OF BATH

Main takeaways

- Integrating multiple modalities can enable numerous real-world applications
- In general, the distribution alignment task is an interesting challenge that can have many practical implications
- The probabilistic approach could help explainability
- Multiple components are involved and each of them needs to be optimized to perform well