Rise of Causality in Computer Vision

The repository contains lists of papers on causality and how relevant techniques are being used to further enhance deep learning era computer vision solutions.

The repository is organized by Maheep Chaudhary and Haohan Wang as an effort to collect and read relevant papers and to hopefully serve the public as a collection of relevant resources.

Causality

- The Seven Tools of Causal Inference with Reflections on Machine Learning
  - ▶ Maheep's notes
- On Pearl's Hierarchy and the Foundations of Causal Inference
  - ▶ Maheep's notes
- Unit selection based on counterfactual logic
  - ▶ Maheep's notes

    `argmax c βP (complier|c) + γP (always-taker|c) + θP (never-taker|c) + δP (defier|c)`
    where benefit of selecting a complier is β, the benefit of selecting an always-taker is γ, the benefit of selecting a never-taker is θ, and the benefit of selecting a defier is δ. Our objective, then, should be to find c.
    Theorem 4 says that if Y is monotonic and satisfies gain equallity then the benefit function may be defined as: -
    `(β − θ)P (y,x |z) + (γ − β)P (y,x' |z) + θ`
    "Third, the proposed approach could be used to evaluate machine learning models as well as to generate labels for machine learning models. The accuracy of such a machine learning model would be higher because it would consider the counterfactual scenarios."

    [Sundar et al., 1998; Blumenthal et al., 2001; Winer,

2001; Resnick et al., 2006; Lewis and Reiley, 2014]

- Unit Selection with Causal Diagram
  - ▶ Maheep's notes

    `W + σU ≤ f ≤ W + σL if σ < 0`, `W + σL ≤ f ≤ W + σU if σ > 0`, where "f" is the objective function. Previously in normal case the objective fucntion is bounded by the equation:
    `max{p 1 , p 2 , p 3 , p 4 } ≤ f ≤ min{p 5 , p 6 , p 7 , p 8 } if σ < 0,`
    `max{p 5 , p 6 , p 7 , p 8 } ≤ f ≤ min{p 1 , p 2 , p 3 , p 4 } if σ > 0`,
    In the extension of the same the author proposes the new situations which arise such as the when "z" is partially observable. and if "z" is a pure mediator.

The author then discusses about the availablity of the observational and experimantal data. If we only have experimantal data then we can simply remove the observationa terms in the theorem
$\max\{p_1, p_2\} \leq f \leq \min\{p_3, p_4\}$ if $\sigma < 0$,
$\max\{p_3, p_4\} \leq f \leq \min\{p_1, p_2\}$ if $\sigma > 0$,
but if we have only onservational data then we can take use of the observed back-door and front-door variables to generate the experimental data, but if we have partially observable back-dorr and front-door variables then we can use the equation:
$LB \leq P(y|do(x)) \leq UB$
The last topic which author discusses about is the reduciton of the dimensionality of the variable "z" which satisfies the back-door and front-door variable by substituting the causal graph by substiuting "z" by "W" and "U" which satisfies the condition that "no_of_states_of_W * no_of_states_of_U = no_of_states_of_z".

- The Causal-Neural Connection: Expressiveness, Learnability, and Inference
  - ▶ Maheep's notes

    The author proposes Neural Causal Models, that are a type of SCM but are capable of amending Gradient Descent. The author propses the network to solve two kinds of problems, i.e. "causal effect identification" and "estimation" simultaneously in a Neural Network as genrative model acting as a proxy for SCM.
    "causal estimation" is the process of identifying the effect of different variables
    "Identification" is obtained when we apply backdoor criterion or any other step to get a better insight. The power of identification has been seen by us as seen in the papers of Hanwang Zhang.
    Theorem 1: There exists a NCM that is in sync with the SCM on ladder 3

- The Causal Loss: Driving Correlation to Imply Causation(autonomus)
  - ▶ Maheep's notes

- Double Machine Learning Density Estimation for Local Treatment Effects with Instruments
  - ▶ Maheep's notes

    The author argues that by obtaining the PDF may give very valuable information as compared to only estimating the Cumlative Distribution Function.

Causality & Computer Vision

- Counterfactual Samples Synthesizing and Training for Robust Visual Question Answering
  - ▶ Maheep's notes

- How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness?
  - ▶ Maheep's notes

    $I(S; Y, T) = I(S; Y) + I(S; T|Y)$,
    The author proposes by this equation that the two models overlap, i.e. the objective model

and the pretrained model. S represents the features extracted the model by the objective model and T is the features extracted by the pretrained model.

- Counterfactual Zero-Shot and Open-Set Visual Recognition
  - ▶ Maheep's notes

    ZSL is usually provided with an auxiliary set of class attributes to describe each seen- and unseen-class whereas the OSR has open environment setting with no information on the unseen-classes [51, 52], and the goal is to build a classifier for seen-classes. Thae author describers previous works in which the generated samples from the class attribute of an unseen-class, do not lie in the sample domain between the ground truth seen and unseen, i.e., they resemble neither the seen nor the unseen. As a result, the seen/unseen boundary learned from the generated unseen and the true seen samples is imbalanced.

    The author proposes a technique using counterfactual, i.e. to generate samples using the class attributes, i.e. Y and sample attrubute Z by the counterfactual, i.e. X would be $\tilde{x}$, had Y been y, given the fact that Z = z(X = x) and the consistency rule defines that if the ground truth is Y then x_bar would be x. The proposed genrative causal model P(X|Z, Y) generate exapmles for ZSL and OSR.

- Counterfactual VQA: A Cause-Effect Look at Language Bias
  - ▶ Maheep's notes
- CONTERFACTUAL GENERATIVE ZERO-SHOT SEMANTIC SEGMENTATION
  - ▶ Maheep's Notes

    The model will contain a total of 4 variables R, W, F and L. The generator will to generate the fake features using the word embeddings and real features of the seen class and will generate fake images using word embeddings after learning. However, this traditional model cannot capture the pure effect of real features on the label because the real features R not only determine the label L by the link R !L but also indirectly influence the label by path R ! F ! L. This structure, a.k.a. confounder. Therefore they remove the R!F!L and let it be W!F!L, removing the confounding effect of F. Also they use GCN to generate the image or fake features from eord embeeddings using the GCN which alos provides to let the generator learn from similar classes.

- Adversarial Visual Robustness by Causal Intervention
  - ▶ Maheep's notes

    1. Augments the image with multiple retinoptic centres
    2. Encourage the model to learn causal features, rather than local confounding patterns.

    They propose the model to be such that max P (Y = $\hat{y}$|X = x + delta) - P(Y = $\hat{y}$|do(X = x + delta)), subject to P (Y = $\hat{y}$|do(X = x + delta)) = P (Y = $\hat{y}$|do(X = x)), in other words they focus on annhilating the confounders using the retinotopic centres as the instrumental variable.

- What If We Could Not See? Counterfactual Analysis for Egocentric Action Anticipation

- ▶ Maheep's notes

  They ask this question so as to only get the effect of semantic label. As the visual feature is the main feature the semantic label can act as a confouder due to some situations occuring frequently. Therfore the author proposes to get the logits "A" from the pipeline without making any changes to the model and then also getting the logits "B" when they provide a random value to visual feature denoting the question of counterfactual, i.e. “what action would be predicted if we had not observed any visual representation?" getting the unbiased logit by:

  Unbiased logit = A - B

- Transporting Causal Mechanisms for Unsupervised Domain Adaptation

  - ▶ Maheep's notes

- WHEN CAUSAL INTERVENTION MEETS ADVERSARIAL EXAMPLES AND IMAGE MASKING FOR DEEP NEURAL NETWORKS

  - ▶ Maheep's notes

  Effect(xi on xj , Z) = P (xj |do(xi_dash), Z_Xi) - P (xj |Z_Xi) .....(1) The excepted casual effect has been defined as: E_Xi[Effect(xi on xj , Z)] = (P(Xi = xi |Z)*(equation_1))

  The author proposes three losses to get the above equaitons, i.e. the effect of pixels. The losses are interpretability loss, shallow reconstruction loss, and deep reconstruction loss. Shallow reconstruction loss is simply the L 1 norm of the difference between the input and output of autoencoder to represent the activations of the network. For the second equation they applied the deep reconstruction loss in the form of the KL-divergence between the output probability distribution of original and autoencoder-inserted network.

  These losses are produced afer perturbtaing the images by maksing the images and inserting adverserial noise.

- Interventional Few-Shot Learning

  - ▶ Maheep's notes

  The author proposes the solution by proposing 4 variables, i.e. "D", "X", "C", "Y" where D is the pretrained model, X is the feature representaiton of the image, C is the low dimesion representation of X and Y are the logits. The author says the D affects both the X and C, also X affects C, X and C affects the logit Y. The autho removes the affect of D on X using backdoor.

- CLEVRER: COLLISION EVENTS FOR VIDEO REPRESENTATION AND REASONING

  - ▶ Maheep's notes

  The dataset is build on CLEVR dataset and has predicitive both predictive and counterfactual questions, i.e. done by, Predictive questions test a model’s capability of predicting possible occurrences of future events after the video ends. Counterfactual questions query the outcome of the video under certain hypothetical conditions (e.g. removing one of the objects). Models need to select the events that would or would not happen under the designated

condition. There are at most four options for each question. The numbers of correct and incorrect options are balanced. Both predictive and counterfactual questions require knowledge of object dynamics underlying the videos and the ability to imagine and reason about unobserved events.

The dataset is being prepared by using the pysics simulation engine.

- Towards Robust Classification Model by Counterfactual and Invariant Data Generation
  - ▶ Maheep's notes

    They augment using the augmentaions as: None, CF(Grey), CF(Random), CF(Shuffle), CF(Tile), CF(CAGAN) and the augmentaions which alter the invariant features using: F(Random) F(Shuffle) F(Mixed-Rand) F(FGSM)

- Unbiased Scene Graph Generation from Biased Training
  - ▶ Maheep's Notes
    1. To take remove the context bias, the author compares it with the counterfactual scene, where visual features are wiped out(containing no objects).

    The author argues that the true label is influenced by Image(whole content of the image) and context(individual objects, the model make a bias that the object is only to sit or stand for and make a bias for it) as confounders, whereas we only need the Content(object pairs) to make the true prediction. The author proposes the TDE = y_e - y_e(x_bar,z_e), the first term denote the logits of the image when there is no intervention, the latter term signifies the logit when content(object pairs) are removed from the image, therfore giving the total effect of content and removing other effect of confounders.

- Counterfactual Visual Explanations
  - ▶ Maheep's Notes
    1. He selects ‘distractor’ image I' that the system predicts as class c' and identify spatial regions in I and I' such that replacing the identified region in I with the identified region in I' would push the system towards classifying I as c'.

    The author proposes the implementation by the equation:
    `f(I*) = (1-a)*f(I) + a*P(f(I'))`
    where
    `I*` represents the image made using the `I` and `I' *` represents the Hamdard product.
    `f(.)` represents the spatial feature extractor
    `P(f(.))` represents a permutation matrix that rearranges the spatial cells of `f(I')` to align with spatial cells of `f(I)`

    The author implements it using the two greedy sequential relaxations – first, an exhaustive search approach keeping a and P binary and second, a continuous relaxation of a and P that replaces search with an optimization.

- Counterfactual Vision and Language Learning

- ▶ Maheep's Notes
  1. The author replaces the embedding of the question or image using another question or image so as to predict the correct answer and minimize counterfactual loss.

- Counterfactual Vision-and-Language Navigation via Adversarial Path Sampler

  - ▶ Maheep's Notes
    1. The author APS, i.e. adversarial path sampler which samples batch of paths P after augmenting them and reconstruct instructions I using Speaker. With the pairs of (P,I), so as to maximize the navigation loss L_NAV.
    2. The NAV, i.e. navigation model trains so as to minimize the L_Nav making the whole process more robust and increasing the performance.

    The APS samples the path based on the visual features v_t which are obtained using the attention on the feature space f_t and history h_t-1 and previous action taken a_t-1 to output the path using the predicted a_t and the features f_t.

- Beyond Trivial Counterfactual Explanations with Diverse Valuable Explanations

  - ▶ Maheep's Notes
    1. DiVE uses an encoder, a decoder, and a fixed-weight ML model.
    2. Encoder and Decoder are trained in an unsupervised manner to approximate the data distribution on which the ML model was trained.
    3. They optimize a set of vectors E_i to perturb the latent representation z generated by the trained encoder.

    The author proposes 3 main losses: `Counterfatual loss` : It identifies a change of latent attributes that will cause the ML model f to change it's prediction. `Proximity loss` : The goal of this loss function is to constrain the reconstruction produced by the decoder to be similar in appearance and attributes as the input, therfore making the model sparse. `Diversity loss` : This loss prevents the multiple explanations of the model from being identical and reconstructs different images modifing the different spurious correlations and explaing through them. The model uses the beta-TCVAE to obtain a disentangled latent representation which leads to more proximal and sparse explanations and also Fisher information matrix of its latent space to focus its search on the less influential factors of variation of the ML model as it defines the scores of the influential latent factors of Z. This mechanism enables the discovery of spurious correlations learned by the ML model.

- SCOUT: Self-aware Discriminant Counterfactual Explanations

  - ▶ Maheep's Notes

    The author implements using a network by giving a query image x of class y , a user-chosen counter class y' != y, a predictor h(x), and a confidence predictor s(x), x is then forwarded to get the F_h(x) and F_s(x). From F_h(x) we predict h_y(x) and h_y'(x) which are then combined with the original F_h(x) to produce the A(x, y) and A(x, y') to get the activation tensors and they are then combined with A(x, s(x)) to get the segmented region of the image which is discriminative of the counter class.

- Born Identity Network: Multi-way Counterfactual Map Generation to Explain a Classifier's Decision

- ▶ Maheep's Notes
    1. The author proposes Counterfactual Map Generator (CMG), which consists of an encoder E , a generator G , and a discriminator D . First, the network design of the encoder E and the generator G is a variation of U-Net with a tiled target label concatenated to the skip connections. This generator design enables the generation to synthesize target conditioned maps such that multi-way counterfactual reasoning is possible.
    2. The another main technique porposes is the Target Attribution Network(TAN) the objective of the TAN is to guide the generator to produce counterfactual maps that transform an input sample to be classified as a target class. It is a complementary to CMG.
       The author proposes 3 main losses:
       `Counterfatual Map loss` : The counterfactual map loss limits the values of the counterfactual map to grow as done by proximity loss in DiVE.
       `Adverserial loss` : It is an objective function reatained due to its stability during adversarial training.
       `Cycle Consistency loss` : The cycle consistency loss is used for producing better multi-way counterfactual maps. However, since the discriminator only classifies the real or fake samples, it does not have the ability to guide the generator to produce multi-way counterfactual maps.

- Introspective Distillation for Robust Question Answering

    - ▶ Maheep's Notes
        1. The author proposes to have a causal feature to teach the model both about the OOD and ID data points and take into account the `P_OOD` and `P_ID`, i.e. the predictions of ID and OOD.
        2. Based on the above predictions the it can be easily introspected that which one of the distributions is the model exploiting more and based on it they produce the second barnch of the model that scores for `S_ID` and `S_OOD` that are based on the equation `S_ID = 1/XE(P_GT, P_ID)`, where `XE` is the cross entropy loss. further these scores are used to compute weights `W_ID` and `W_OOD`, i.e. `W_OOD = S_OOD/(S_OOD + S_ID)` to train the model to blend the knowledge from both the OOD and ID data points.
        3. The model is then distilled using the knowledge distillation manner, i.e. `L = KL(P_T, P_S)`, where `P_T` is the prediction of the teacher model and the `P_S` is the prediction of the student model.

- Counterfactual Explanation and Causal Inference In Service of Robustness in Robot Control

    - ▶ Maheep's Notes

      The additonal component in the model is the predictor takes the modified image and produces real-world output. The implementation of it in mathematics looks like:
      `min d_g(x, x') + d_c(C(x'), t_c)`, where d_g is the distance b/w the modified and original image, d_c is the distance b/w the class space and C is the predictor that x' belongs to t_c class.
      The loss defines as: `total_loss = (1-alpha)*L_g(x, x') + (alpha)*L_c(x, t_c)`, where L_c is the loss `x` belongs to `t_c` class

- Counterfactual Explanation Based on Gradual Construction for Deep Networks
  - ▶ Maheep's Notes

    The proposed also focuses on 2 things, i.e. Explainability and Minimality. while implementing the techniue the authors observe the target class which were being generated were getting much perturbated so as to come under asverserial attack and therfore they propose the logit space of x' to belong to the space of training data as follows:
    `argmin(sigma(f_k'(x') - (1/N)*sigma(f_k'(X_i,c_t))) + lambda(X' - X))`
    where `f'` gives the logits for class k, `X_i,c_t` represents the i-th training data that is classified into c_k class and the N is the number of modifications.

- CoCoX: Generating Conceptual and Counterfactual Explanations via Fault-Lines
  - ▶ Maheep's Notes

    The proposed model is implemented by taking the CNN captured richer semantic aspect and construct xconcepts by making use of feature maps from the last convolution layer. Every feature map is treated as an instance of an xconcept and obtain its localization map using the Grad-CAM and are spatially pooled to get important weights, based on that top p pixels are selected and are clustered using K-means. The selection is done using the TCAV tecnique.

Algorithm 1: Generating Fault-Line Explanations

Input: input image I, classification model M, predicted class label c_{pred}, alternate class label c_{alt} and training dataset χ

1. Find semantically meaningful superpixels in χ,

$$\alpha_{m,L}^{c} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\partial y^{c}}{\partial A_{ij}^{m,L}}$$

2. Apply K-means clustering on superpixels and obtain xconcepts (Σ).
3. Identify class specific xconcepts (Σ_{pred} and Σ_{alt}) using TCAV,

$$S_{c,X} = \nabla g_c(f(I)) \cdot v_X$$

4. Solve Equation 4 to obtain fault-line Ψ,

$$\Psi \leftarrow \min_{\delta_{pred}, \delta_{alt}} \alpha D(\delta_{pred}, \delta_{alt}) + \beta \left\|\delta_{pred}\right\|_1 + \lambda \left\|\delta_{alt}\right\|_1$$

return Ψ.

- CX-ToM: Counterfactual Explanations with Theory-of-Mind for Enhancing Human Trust in Image Recognition Models
  - ▶ Maheep's Notes

Algorithm 1 Generating Fault-Line Explanations

input image I, classification model M, predicted class label c_{pred}, alternate class label c_{alt} and training dataset χ

1. Find semantically meaningful superpixels in χ,

$$\alpha_{m,L}^{c} = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\partial y^{c}}{\partial A_{ij}^{m,L}}$$

2. Apply K-means clustering on superpixels and obtain xconcepts (Σ).
3. Identify class specific xconcepts (Σ_{pred} and Σ_{alt}) using TCAV,

$$S_{c,X} = \nabla g_c(f(I)) \cdot v_X$$

4. Solve Equation 5 to obtain fault-line Ψ,

$$\Psi \leftarrow \min_{\delta_{pred}, \delta_{alt}} \alpha D(\delta_{pred}, \delta_{alt}) + \beta \left\|\delta_{pred}\right\|_1 + \lambda \left\|\delta_{alt}\right\|_1$$

return Ψ.

- DeDUCE: Generating Counterfactual Explanations At Scale
  - ▶ Maheep's Notes

Algorithm 1 DeDUCE

Inputs: original input $\mathbf{x} \in \mathcal{X}$, target class t, trained model f with feature extractor f_Z, training data $X_{tr} \in \mathcal{X}^N$, coefficient λ, step size δ, max iterations max_iter, max pixel changes p, number of pixels m, target confidence γ.
Output: counterfactual $\mathbf{x}' \in \mathcal{X}$

(before deployment) apply f to X_{tr} and fit the GMM $p(\mathbf{z}) = \frac{1}{|C|} \sum_{c \in C} p_c(\mathbf{z})$
$\mathbf{x}' \leftarrow \mathbf{x}$
$k \leftarrow 0$
$\mathbf{P} \leftarrow \mathbf{0}_{dim(\mathcal{X})}$
while $f(\mathbf{x}')_t \leq \gamma$ **and** $k <$ max_iter **do**
    compute gradient $\mathbf{g}$ in input space
    select m most salient pixels: $I =$ `select_q_largest_masked`$(|\mathbf{g}|, \mathbf{P} < p)$
    update these pixels: $\forall i \in I : \mathbf{x}'[i] \leftarrow \mathbf{x}'[i] +$ `sign`$(\mathbf{g}[i]) \cdot \delta$
    clip to input domain: $\mathbf{x}' \leftarrow$ `clip`$(\mathbf{x}')$
    $\forall i \in I : \mathbf{P}[i] \leftarrow \mathbf{P}[i] + 1$
    $k \leftarrow k + 1$
return $\mathbf{x}'$

- Designing Counterfactual Generators using Deep Model Inversion
  - ▶ Maheep Notes

```
argmin(lambda_1*sigma_on_l(layer_l(x'), layer_l(x)) +
lambda_2*L_mc(x';F) + lambda_3*L_cf(F(x'), y'))
```

where,

`layer_l` :The differentiable layer "l" of the neural network, it is basically used for semantic preservation.

`L_mc`: It penlaizes x' whcih do not lie near the manifold. L_mc can be Deterministic Uncertainty

Quantification (DUQ).
`L_fc`: It ensures that the prediction for the counterfactual matches the desired target

- ECINN: Efficient Counterfactuals from Invertible Neural Networks
  - ▶ Maheep's Notes

    `x' = f_inv(f(x) + alpha*delta_x)`
    where,
    `x'` :Counterfactual image.
    `f`: INN and therfore `f_inv` is the inverse of `f`.
    `delta_x`: the infoprmation to be added to convert the latent space of image to that of counterfactual image.
    `||z + alpha_0*delta_x- μ_p || = ||z + alpha_0*delta_x - μ_q ||` where the z + alpha_0*delta_x is the line separating the two classes and μ_q and μ_q are the mean distance from line. Therefore
    `alpha = alpha_0 + 4/5*(1-alpha_0)`

- EXPLAINABLE IMAGE CLASSIFICATION WITH EVIDENCE COUNTERFACTUAL
  - ▶ Maheep's Notes

Algorithm 1 SEDC

Inputs:
I % Image to classify
$C_M : I \rightarrow \{1,2,...,k\}$ % Trained classifier with k classes
$S = \{s_i, i = 1,2,...,l\}$ % Segmentation of the image with l segments

Procedure:
$E = \{\}$ % List of explanations
for s_i in S **do**
  if class change after removing s_i from I **then**
    $E = E \cup \{s_i\}$
  end if
end for
while $E = \emptyset$ **do**
  Select $best$ % Best-first: segment set with highest reduction in predicted class score
  $best\_set$ = expansions of $best$ with one segment
  for C_0 in $best\_set$ **do**
    if class change after removing C_0 from I **then**
      $E = E \cup \{C_0\}$
    end if
  end for
end while

Output:
Explanations in E

- Explaining Visual Models by Causal Attribution
  - ▶ Maheep Notes

- Explaining the Black-box Smoothly-A Counterfactual Approach
  - ▶ Maheep's Notes

    `L_cgan = log(P_data(x)/q(x)) + log(P_data(c|x)/q(c|x))`

    where P_data(x) is the data distribtion and learned distribution q(x), whreas P_data(c|x)/q(c|x) = r(c|x) is the ratio of the generated image and the condition.

2.) **Classification model consistency**: The generated image should give desired output. Therefore the condition-aware loss is introduced, i.e. `L := r(c|x) + D_KL (f(x')||f(x) + delta),`, where f(x') is the output of classifier of the counterfactual image is varied only by delta amount when added to original image logit. They take delta as a knob to regularize the genration of counterfactual image.
3.) **Context-aware self-consistency**: To be self-consistent, the explanation function should satisfy three criteria

> (a) Reconstructing the input image by setting = 0 should return the input image, i.e., G(x, 0) = x.
> (b) Applying a reverse perturbation on the explanation image x should recover x.

To mitigate this conditions the author propose an identity loss. The author argues that there is a chance that the GAN may ignore small or uncommon details therfore the images are compared using semantic segemntation with object detection combined in identity loss. The identity loss is : L_identity = L_rec(x, G(x, 0))+ L_rec(x, G(G(x,delta), -delta))

- Explaining the Behavior of Black-Box Prediction Algorithms with Causal Learning
  - ▶ Maheep's Notes

    `V = (Z,Y')`
    `Y' = g(z1,.....zs, epsilon)` On the basis of possible edge types, they find out which high level causes, possible causes or non-causes of the balck-box output Y'.

- Explaining Classifiers with Causal Concept Effect (CaCE)
  - ▶ Maheep's Notes

    `Effect = E(F(I)|do(C = 1)) - E(F(I)|do(C = 0))` where F gives output on image I and C is the concept. This can be done at scale by intervening for a lot of values in a concept and find the spurious corrlation. But due to the insufficient knowlegde of the Causal Graph teh author porposes a VAE which can calculate the precise CaCE by by generating counterfactual image by just changing a concept and hence computing the difference between the prediction score.

- Fast Real-time Counterfactual Explanations
  - ▶ Maheep's Notes

    1.) **Adverserial loss**: It measures whether the generated image is indistinguishable from the real world images
    2.) **Domain classification loss**: It is used to render the generate image x + G(x,y') conditional on y'. `L = E[-log(D(y'|x + G(x,y')))]` where G(x, y') is the perterbuation introduced by generator to convert image from x to x'
    3.) **Reconstruction loss**: The Loss focuses to have generator work propoerly so as to produce the image need to be produced as defined by the loss. `L = E[x - (x + G(x,y') + G(x + G(x,y'), y))]` 4.) **Explanation loss**: This is to gurantee that the generated fake image produced belongs to the distribution of H. `L = E[-logH(y'|x + G(x,y'))]`
    5.) **Perturbation loss**: To have the perturbation as small as possible it is introduced. `L =`

`E[G(x,y') + G(x + G(x,y'),y)]`
All these 5 losses are added to make the final loss with different weights.

- GENERATIVE_COUNTERFACTUAL_INTROSPECTION_FOR_EXPLAINABLE_DEEP_LEARNING
  - ▶ Maheep's Notes

    `min(lambda*loss(I(A')) + ||I - I(A'))`, where loss is cross-entropy for predicting image I(A') to label c'.

- Generative_Counterfactuals_for_Neural_Networks_via_Attribute_Informed_Perturbations
  - ▶ Maheep's Notes

    `min(E[sigma_for_diff_attributes*(-a*log(D(x')) - (1-a)*(1-D(x)))]) + E[||x - x'||]` where D(x') generates attributes for counterfactual image.
    To generate the counterfactual 2 losses are produced,one ensures that the perturbed image has the desired label and the second one ensures that the perturbation is minimal as possible, i.e.
    `L_gen = Cross_entropy(F(G(z, a)), y) + alpha*L(z,a,z_0, a_0)`
    The L(z,a,z0,a0) is the l2 norm b/w the attribute and the latent space.

- Question-Conditioned Counterfactual Image Generation for VQA
  - ▶ Maheep's Notes

- FINDING AND FIXING SPURIOUS PATTERNS WITH EXPLANATIONS
  - ▶ Maheep's Notes

    `P(Spurious | Main) = P(Spurious | not Main) = 0.5`
    The second step consist of minimizing the potential for new SPs by setting the `P(Main|Artifact) = 0.5)`.
    SPIRE moves images from {Both, Neither} to {Just Main, Just Spurious} if p > 0.5, i.e. p = P(Main|Spurious) but if p < 0.5 then SPIRE moves images from {Just Main, Just Spurious} to {Both, Neither}.

- Contrastive_Counterfactual_Visual_Explanations_With_Overdetermination
  - ▶ Maheep's Notes

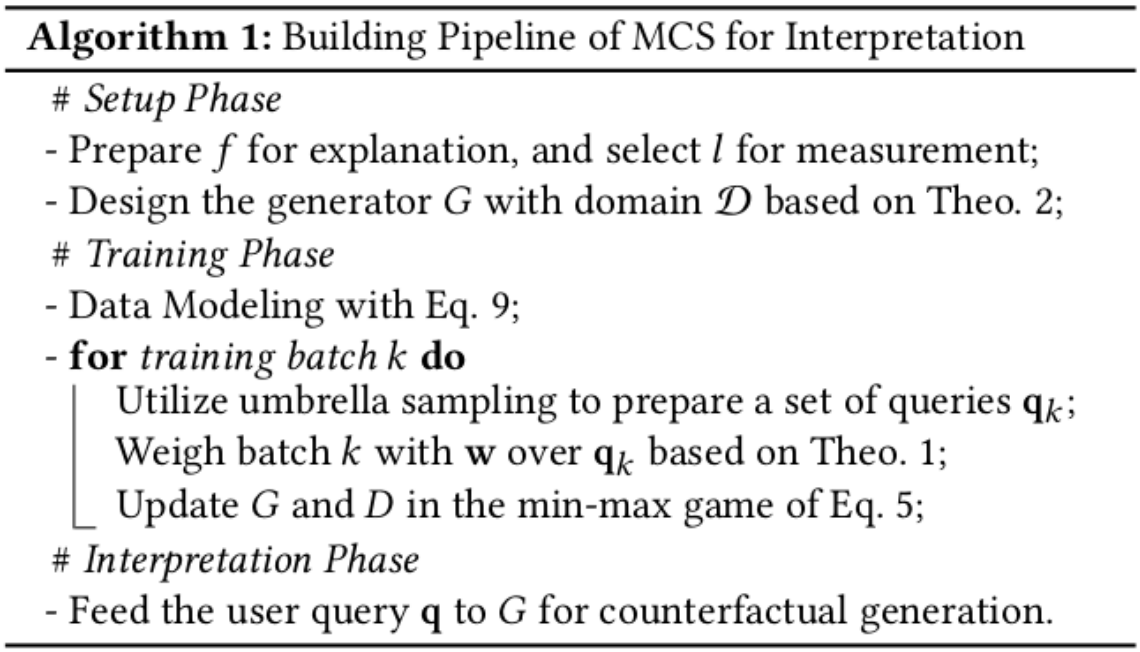
- Training_calibration-based_counterfactual_explainers_for_deep_learning
  - ▶ Maheep's Notes

    Validity: ratio of the counterfactuals that actually have the desired target attribute to the total number of counterfactuals
    The confidence of the **image** and **sparsity**, i.e. ratio of number of pixels altered to total no of pixels. Th eother 2 metrcs are **proximity**, i.e. average l2 distance of each counterfactual to the K-nearest training samples in the latent space and **Realism score** so as to have the generated image is close to the true data manifold.
    TraCE reveals attribute relationships by generating counterfactual image using the different

attribute like age "A" and diagnosis predictor "D".
`delta_A_x = x - x_a'` ; `delta_D_x = x - x_d'`
The x_a' is the counterfactual image on the basis for age and same for x_d'.
`x' = x + delta_A_x + delta_D_x` and hence atlast we evaluate the sensitivity of a feature by `F_d(x') - F_d(x_d')`, i.e. F_d is the classifier of diagnosis.

- Generating Natural Counterfactual Visual Explanations
  - ▶ Maheep's Notes
- On Causally Disentangled Representations
  - ▶ Maheep's Notes
- INTERPRETABILITY_THROUGH_INVERTIBILITY_A_DEEP_CONVOLUTIONAL_NETWORK
  - ▶ Maheep's Notes

    `z' = z + alpha*w` where `x' = phi_inverse(z + alpha*w)`. Any change orthogonal to w will create an “isofactual. To show that their counterfactuals are ideal, therfore they verify that no property unrelated to the prediction is changed. Unrealted properties = e(x), `e(x) = vT*z`, where v is orthogonal to w. `e(x') = vT*(z + alpha* w) = vT*z = e(x)`. To measure the difference between the counterfactual and image intermediate feature map h, i.e. `m = |delta_h|*cos(angle(delta_h, h))` for every location of intermediate feature map.
- Model-Based Counterfactual Synthesizer for Interpretation
  - ▶ Maheep's Notes



Algorithm 1: Building Pipeline of MCS for Interpretation

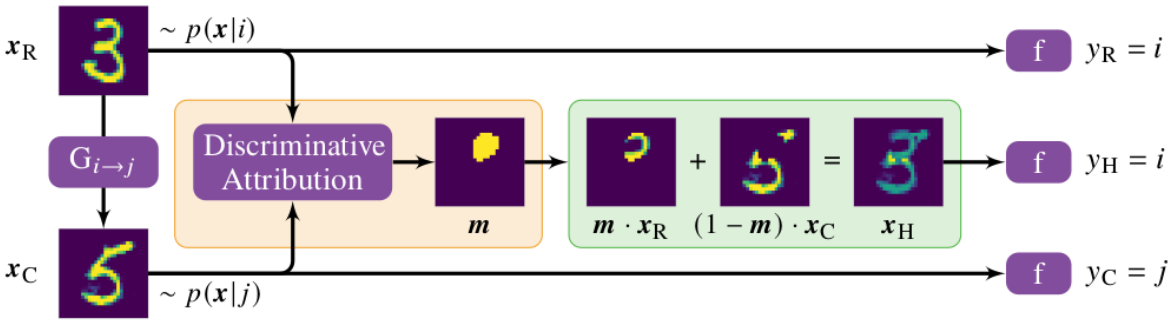
 # Setup Phase
- Prepare f for explanation, and select l for measurement;
- Design the generator G with domain $\mathcal{D}$ based on Theo. 2;

 # Training Phase
- Data Modeling with Eq. 9;
- **for** *training batch* k **do**
  - Utilize umbrella sampling to prepare a set of queries $\mathbf{q}_k$;
  - Weigh batch k with $\mathbf{w}$ over $\mathbf{q}_k$ based on Theo. 1;
  - Update G and D in the min-max game of Eq. 5;

 # Interpretation Phase
- Feed the user query $\mathbf{q}$ to G for counterfactual generation.

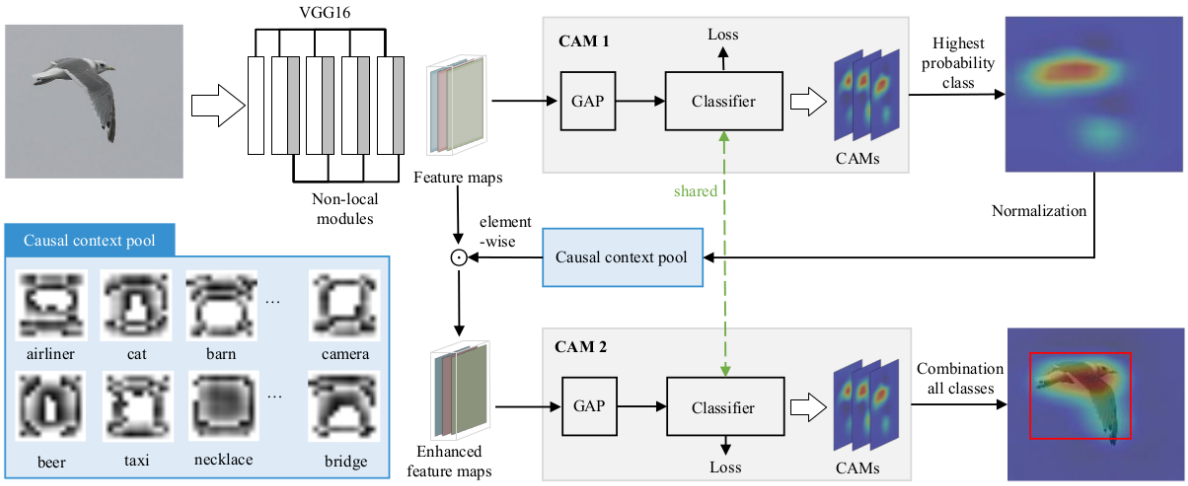
- The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples
  - ▶ Maheep's Notes

- Discriminative Attribution from Counterfactuals
  - ▶ Maheep's Notes

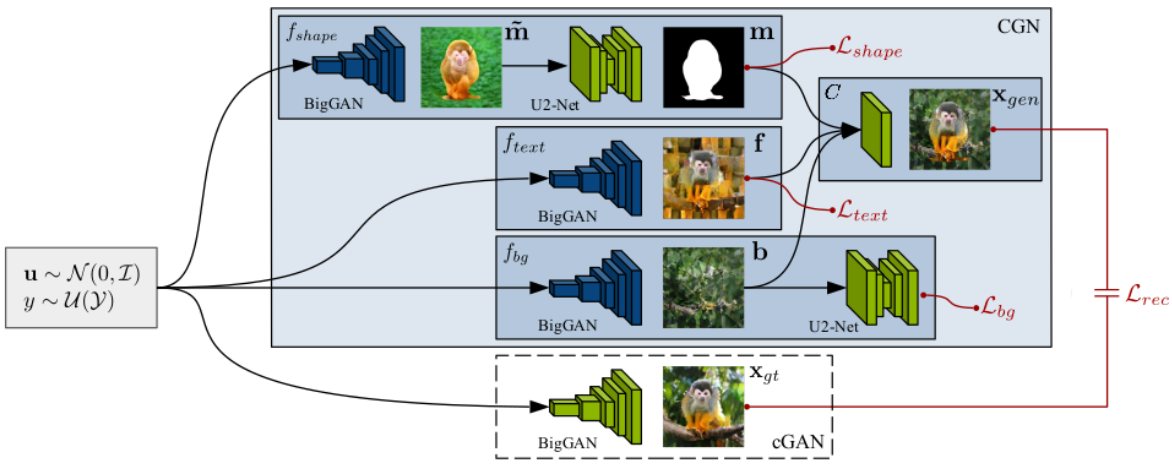


- Causal Interventional Training for Image Recognition
  - ▶ Maheep's Notes

- Improving_Weakly_supervised_Object_Localization_via_Causal_Intervention
  - ▶ Maheep's Notes



- COUNTERFACTUAL GENERATIVE NETWORKS
  - ▶ Maheep's Notes

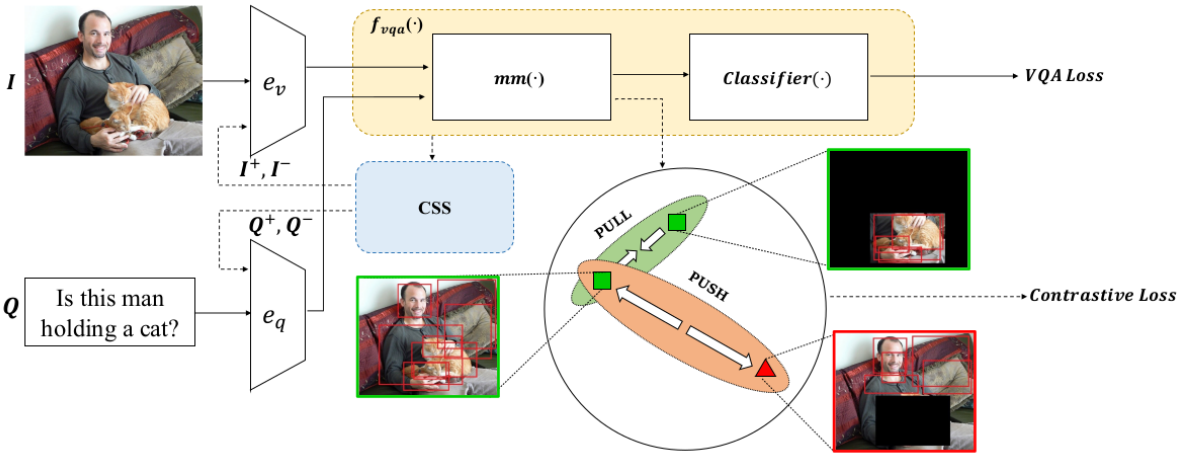


- Discovering Causal Signals in Images
  - ▶ Maheep's Notes

    `X causes Y`, the cause, noise and mechanism are independent but we can identify the footprints of causality when we try to `Y causes X` as the noise and Y will not be independent.

- Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering
  - ▶ Maheep's Notes

    `Q` and `V(mm(Q,V) = a)`, and joint embedding of `Q` and `V_+ (factual)(mm(Q,V+) = p)` by taking a cosine similarity b/w them. They also aim to decrease mutual information b/w `mm(Q,V-) = n` and `a` by taking cosine similarity`(s(a,n))`. The final formula becomes: `L_c = E[-log(e^s(a,p)/e^s(a,p)+e^s(a,n))]`
    The total loss becomes `L = lambda_1*L_c + lambda_2*L_vqa`



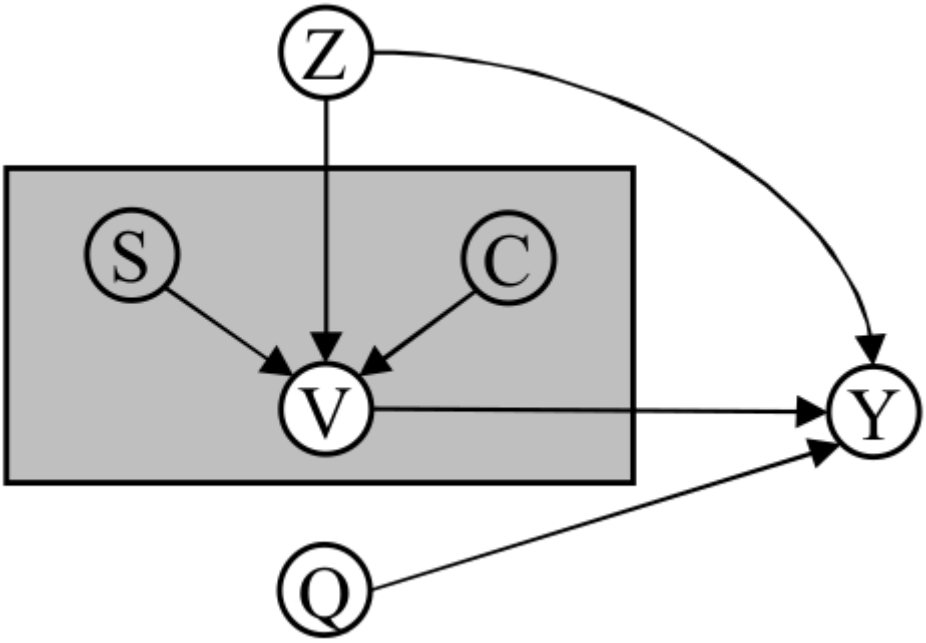
- Latent Causal Invariant Model
  - ▶ Maheep's Notes

    `S` and (b) others that are spuriously correlated `Z` from V(latent variable).
    There exists a spurious correlation b/w `S` and `Z`. The author argues that we will get a `p(y|do(s*)) = p(y|s*)`

- Two Causal Principles for Improving Visual Dialog
  - ▶ Maheep's Notes

    `a_i`(answer) is a sentence observed from the “mind” of user u during dataset collection. Then, `sigma(P(A)*P(u|H))`, `H` is history and `A` is answer can be approximated as `sigma(P(A)P(a_i|H))`.They further use `p(a_i|QT)`, where `QT` is Question Type to approximate `P(a_i|H)` because of two reasons: First, `P (a_i|H)` essentially describes a prior knowledge about `a_i` without comprehending the whole `{Q, H, I} triplet`.

- Weakly-Supervised Video Object Grounding via Causal Intervention
  - ▶ Maheep's Notes



`Z` that occurs due to some specific objects occuring frequently. The style confounder is replaced by using the contrastive learning, where the counterfactual examples are created by taking the vectors from a memory bank by taking the top sleected top regions for described object and then the selected regions and frames are grouped together into frame-level content(H_c) and region-level content(U_c), and the rest of the regions are grouped as U_s and H_s. These regions are the converted to counterfactual using these memory vectors which were created by taking the randomly selected regions in training set. The most similar one and replaces the original one, to generate examples to have them hard to distinguish from real ones contrastive learning is used. The equation looks like:

```
IE(p|do(U_s = U_s_generated)) < IE(p|do(U_c = U_c_generated))
IE(p|do(H_s = H_s_generated)) < IE(p|do(H_c = H_c_generated))
```

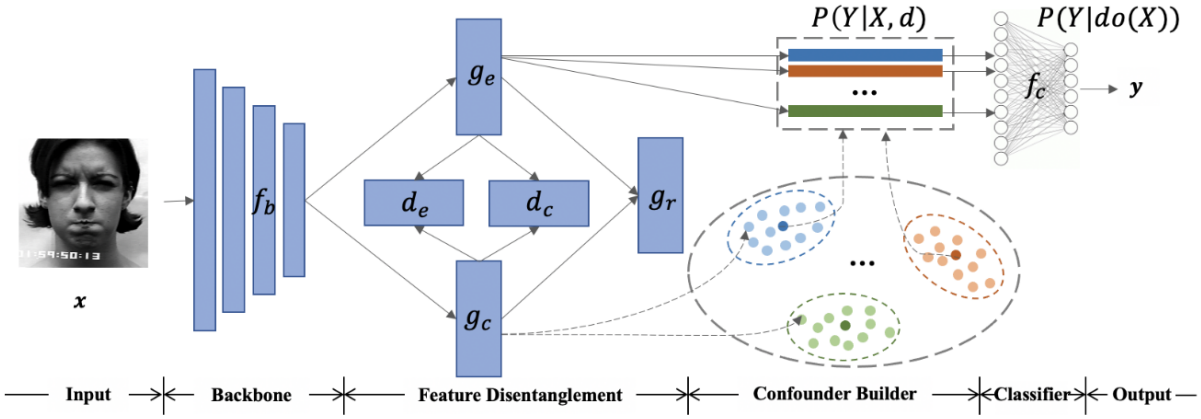
where the `IE` is Interventional Effect. As for the next confounder they uses the textual embedding of o_k(object) essentially provides the stable cluster center in common embedding space for its vague and diverse visual region embeddings in different videos. Therefore, by taking the textual embedding of the object as the substitute of every possible object z and apply backdoor adjustment.

- Towards Unbiased Visual Emotion Recognition via Causal Intervention
  - ▶ Maheep's Notes

    IERN, which is composed of four parts:
    1.) **Backbone**
    > It extracts the feature embedding of the image.
    2.) **Feature Disentanglement**
    > It disentangles the emotions and context from the image, having emotion dicriminator(d_e) and context discriminator(d_c) which ensures that the extracted feature are separated and has the desired feature. The loss comprises as :
    `L = CE(d_e(g_e(f_b(x))), y_e) + MSE(d_c(g_e(f_b(x))), 1/n)` where g_e is emotion generator and y_e is the emotion label and n is the number of counfounder and the same loss is for context replacing d_e, g_e and d_c by d_c, g_c and d_e, here n represents

number of emotions. To ensure that the separated features fall within reason-able domains, IERN should be capable of reconstructing the base feature f_b(x), i.e. `L =MSE(g_r(g_e(f_b(x)), g_c(f_b(x))), f_b(x))`

3.) **Confounder Builder**
> The purpose of the confounder builder is to combine each emotion feature with different context features so as to avoid the bias towards the observed context strata.
4.) **Classifier**
> It is simply used for prediciton.

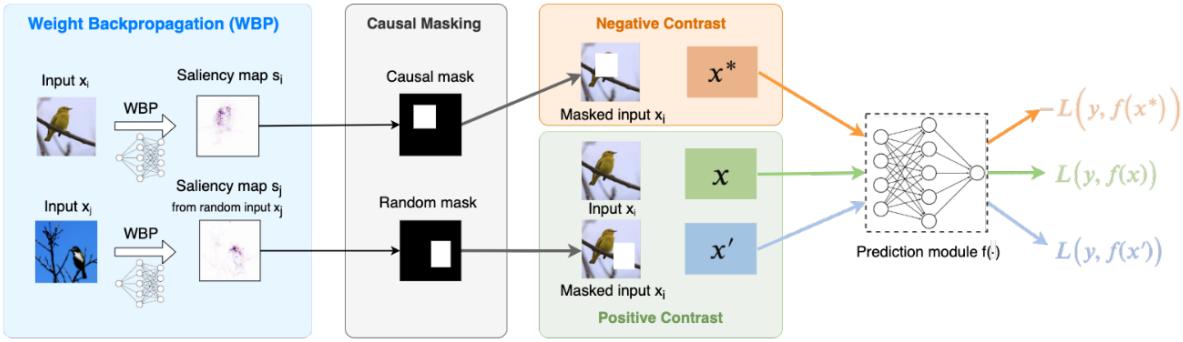


- Human Trajectory Prediction via Counterfactual Analysis
  - ▶ Maheep's Notes

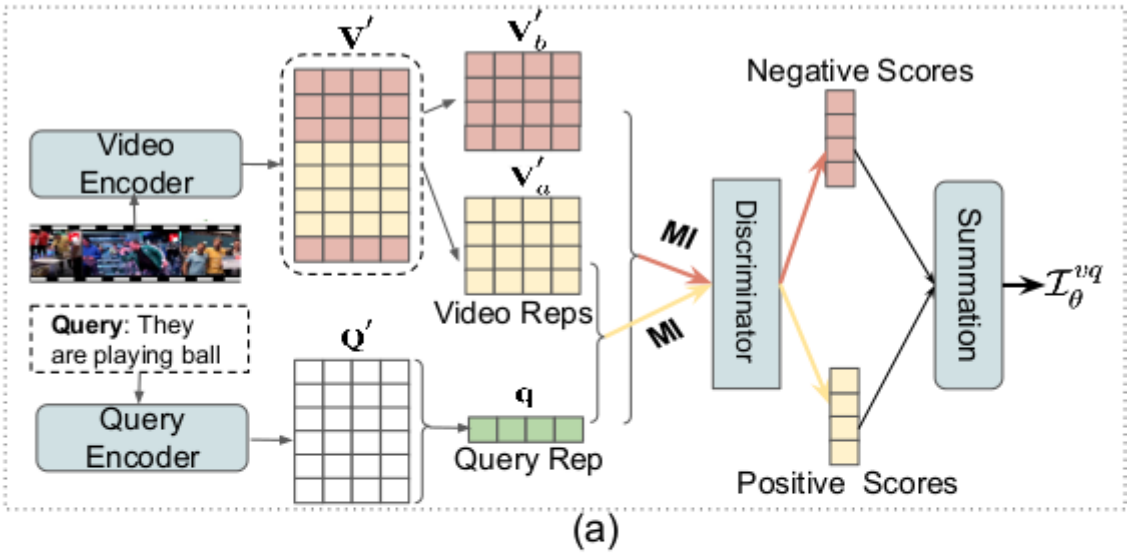
    They Y_causal is defined as `Y_causal = Y_i - Y-i(do(X_i = x_i))`
    They define a generative model which generates trajectory by a noise latent variable Z indicated by `Y*_i`. Finally the loss is defined as:
    `Y_causal = Y*_i - Y*_i(do(X_i = x_i))`
    `L_causalGAN = L2(Y_i, Y_causal) + log(D(Y_i)) + log(1-D(Y_causal))`, where D is the discriminator.

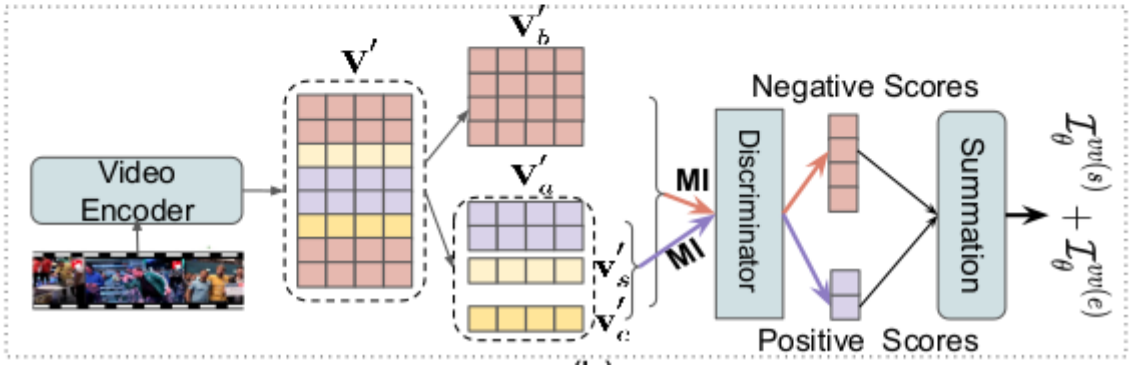
- Proactive Pseudo-Intervention: Contrastive Learning For Interpretable Vision Models
  - ▶ Maheep's Notes

    `f(theta)` are produced and the main features are masked out of the image giving us `x*`. Now the loss becomes `L = sigma(l(x*,not_y;f(theta)))`
    A trivbial solution can be the saliency maps copvers the whole image therefore L1-norm of saliency map is used to encourage succinct (sparse) representations. The another problem that now arises is that the model can learn a shortcut that when it get a masked image then it has to always give `not_y` as prediction, so as to counter it the author proposes to send images with random masks on them, making the loss `L = sigma(l(x',y;f(theta)))`

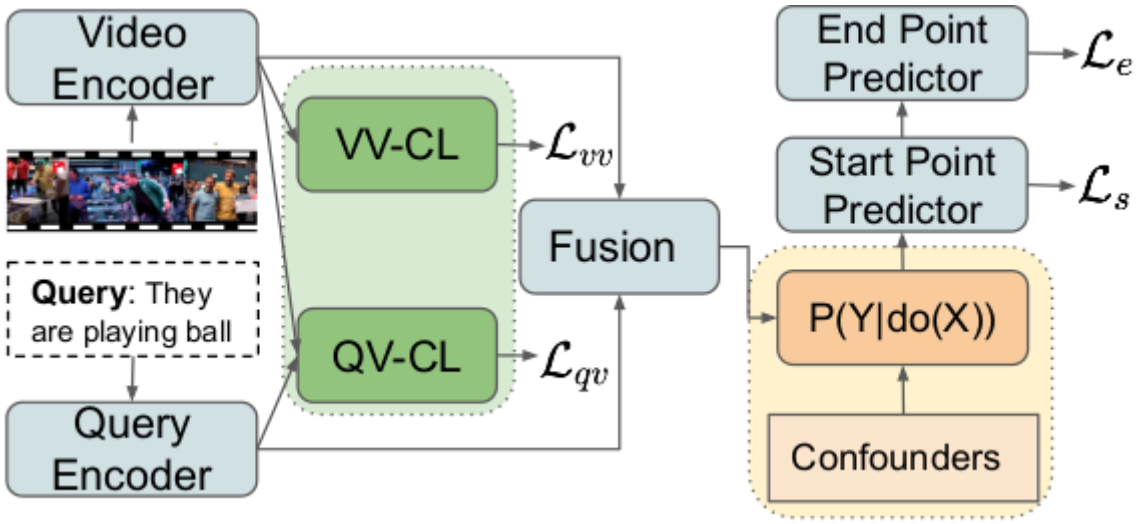


- Interventional Video Grounding with Dual Contrastive Learning
  - ▶ Maheep's Notes



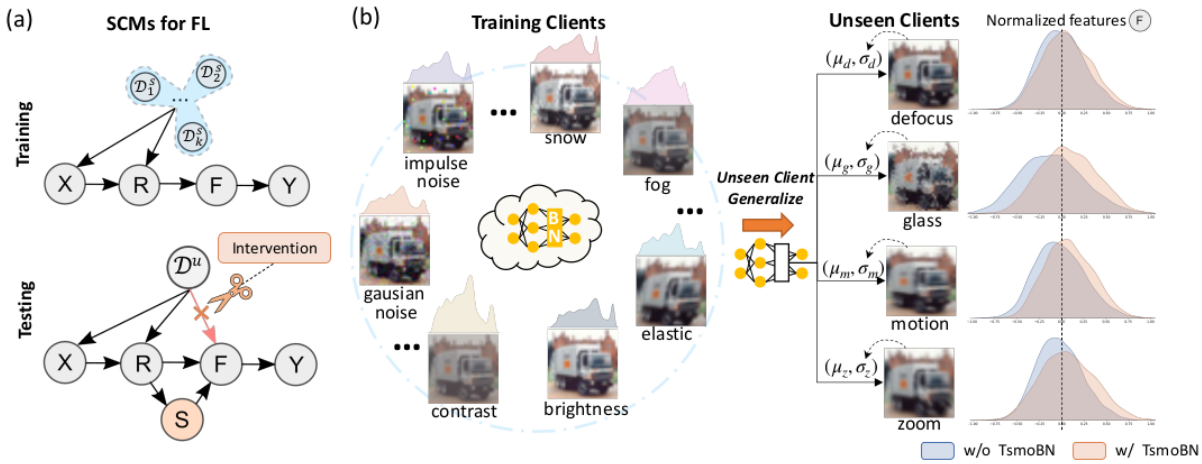


3. The output of two feature encoders are fed to a fusion module with a context-query attention mechanism to capture the cross-modal interactions between visual and textual features.
4. Next, to mitigate the spurious correlations between textual and visual features, they use causal interventions P (Y |do(X)) with event as surrogate confounders to learn representations.
5. Finally, two losses L_s and L_e for the start and end boundaries are introduced.

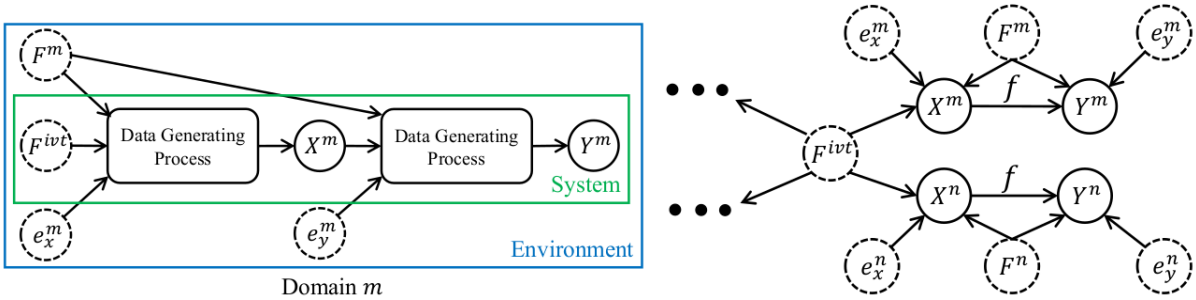


- Causality matters in medical imaging
  - ▶ Maheep's Notes
- TSMOBN GENERALIZATION FOR UNSEEN CLIENTS IN FEDERATED LEARNING
  - ▶ Maheep's Notes

    `D_s_i` for datasets of different domain, i.e. coming from different users but used in training, `X` are the samples, `R` are the raw extracted features of `X`, `F` is the normalized feature representaiton of `R` and `Y` is the classifier. To remove the confounding effects brought by `D_u` , a direct way is using causal intervention on normalized features (i.e., do(F)) to let the feature distribution similar to training distributions. This intervention by introducing the surrogate variable `S`, which is test-specific statistics of raw features `R` during testing by obtaining the test normalized features that have similar distributions as the training normalized features. More specifically by calculating the mean and variance pair at test time in BN to normalize features. Additionally they further propose to use momentum to integrate relations among different batches, thus reducing the variances. Precisely by giving the unseen client with M batches of data to be tested in sequential manner.



- Learning Domain Invariant Relationship with Instrumental Variable for Domain Generalization
  - ▶ Maheep's Notes



- Latent Space Explanation by Intervention
  - ▶ Maheep's Notes

    `Z = {z1, z2,....., zn}` which is intervened to flip the output of the model and are finally vsualized from the hidden representation using the loss `L = l(g(phi'(x), x))`, where `phi'(x)` is the counterfactual model. The next goal is to ensure that the generated concepts follow the same concepts the discriminator employ. They achieve this by maximizing the amount of information that the explanatory learner (i.e., g) extracted from the latent representation with respect to the discriminative learner's (i.e., f_K) information.

- The Blessings of Unlabeled Background in Untrimmed Videos
  - ▶ Maheep's Notes

    `P(x_1, x_2,....., x_n | Z = z) = TT P(x_t | Z = z)`, i.e. the features will become independent if we are able to obeserve `Z` but if there exists an unobserved confounder c, which affects multiple input video features within x and segment-level labels A. Then, x would be dependent, even conditional on z, due to the impact of c, in this case with the blessings of weak ignorability we can replace the expectation over C with a single z in `E[E[A|X = x, C = c]] = A`

- Selecting Data Augmentation for Simulating Interventions
  - ▶ Maheep's Notes

- Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification
  - ▶ Maheep's Notes

    `Y_effect = E[Y(A = A, X = X)] - E[Y(A = A', X = X)]`
    The loss comprises as:
    `L = L_crossentropy(Y_effect, y) + L_others`, where L_others represents the original objective such as standard classification loss.

- Meaningful Explanations of Black Box AI Decision Systems
  - ▶ Maheep's Notes

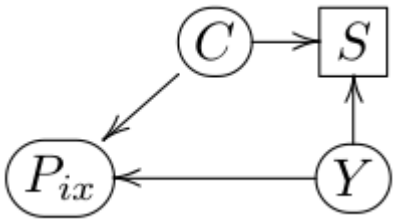
- Are VQA Systems RAD? Measuring Robustness to Augmented Data with Focused Interventions
  - ▶ Maheep's Notes

`RAD =|J(D;F) and J(D';F)|/|J(D;F)|`
, where `J(D;F)` as the set of example indices for which a model `f` correctly predicts `y`. `D'` represents the augmented example which is prepared as *VQA dataset there are three answer types: "yes/no", "number" and "other", and 65 question types. In augmentations, they generate "yes/no" questions from "number" and "other" questions, i.e. What color is the ? is changed to Is the color of is ?*
RAD is in [0, 1] and the higher the RAD of f is, the more robust f is.

- Adversarial Robustness through the Lens of Causality
  - ▶ Maheep's Notes

    `P_theta(X, Y) = sigma(P_theta(Y,s|X)*P_theta(X))`,
    where s is the spurious correlation. As we know that the distribtuion of `X` can be hardly changed therefore `P_theta(X) = P(X)`. Therefore it can be assumed that the difference b/w `P_theta(Y,s|X)` and `P(Y,s|X)` is the main reason of the adverserial inrobustness.
    Therefore they define the loss as:
    `min CE(h(X + E_adv ; theta), Y) + CE(h(X; theta), Y) + CE[P(Y| g(X, s)), P(Y|g(X + E_edv, s))]`
    where `E_adv` adverserail perturbation, theta are parameters of the model, and g represents the paramter optimized to minimize the `CE`, i.e. Cross Entropy loss.

- Causality-aware counterfactual confounding adjustment for feature representations learned by deep models
  - ▶ Maheep's Notes

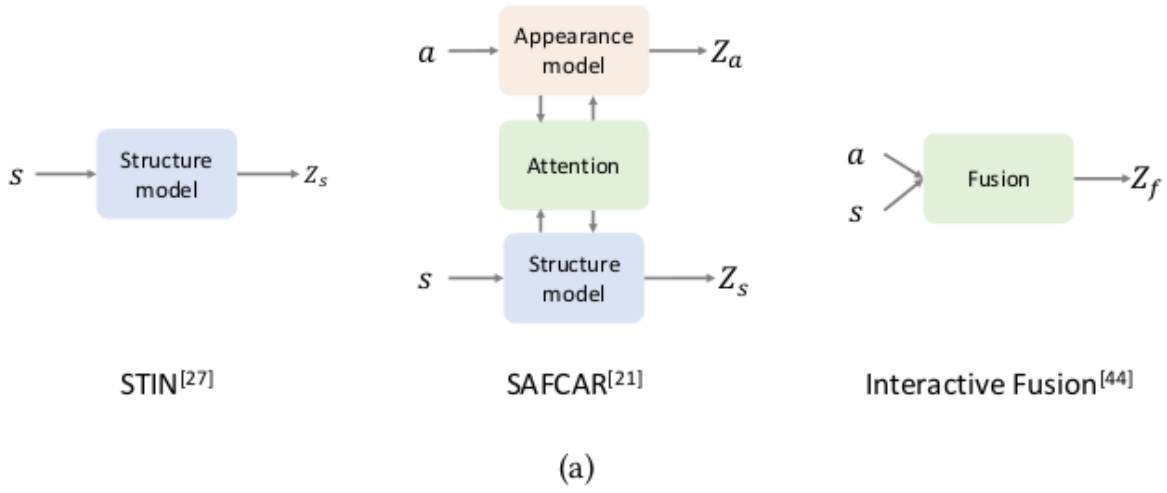
    In order to remove/reduce the influence of C on the predictive performance of the classifier, they apply the causality-aware adjustment proposed to generate counterfactual features, X'. These counterfactual examples are used to train a logistic regression classifier, and then use the same algorithm to generate counterfactual in test set X_test' to generate predictions that are no longer biased by the confounder.

- Domain Generalization using Causal Matching
  - ▶ Maheep's Notes

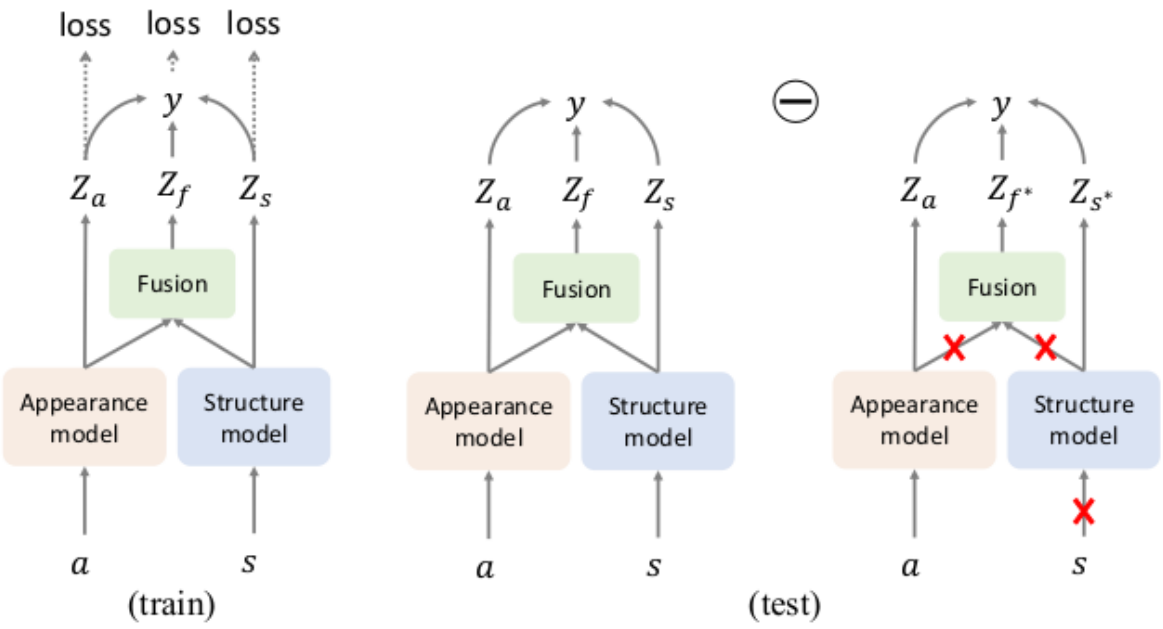
Algorithm 1 MatchDG

In: Dataset $(d_i, x_i, y_i)_{i=1}^{n}$ from m domains, τ, t
Out: Function $f : \mathcal{X} \rightarrow \mathcal{Y}$
Create random match pairs Ω_Y.
Build a $p * q$ data matrix $\mathcal{M}$.
Phase I
while notconverged **do**
  for $batch \sim \mathcal{M}$ **do**
    Minimize contrastive loss (4).
  end for
  if epoch % t == 0 **then**
    Update match pairs using Φ_{epoch}.
  end if
end while
Phase II
Compute matching based on Φ.
Minimize the loss (3) with learnt match function Φ to obtain f.

- Counterfactual Debiasing Inference for Compositional Action Recognition
  - ▶ Maheep's Notes

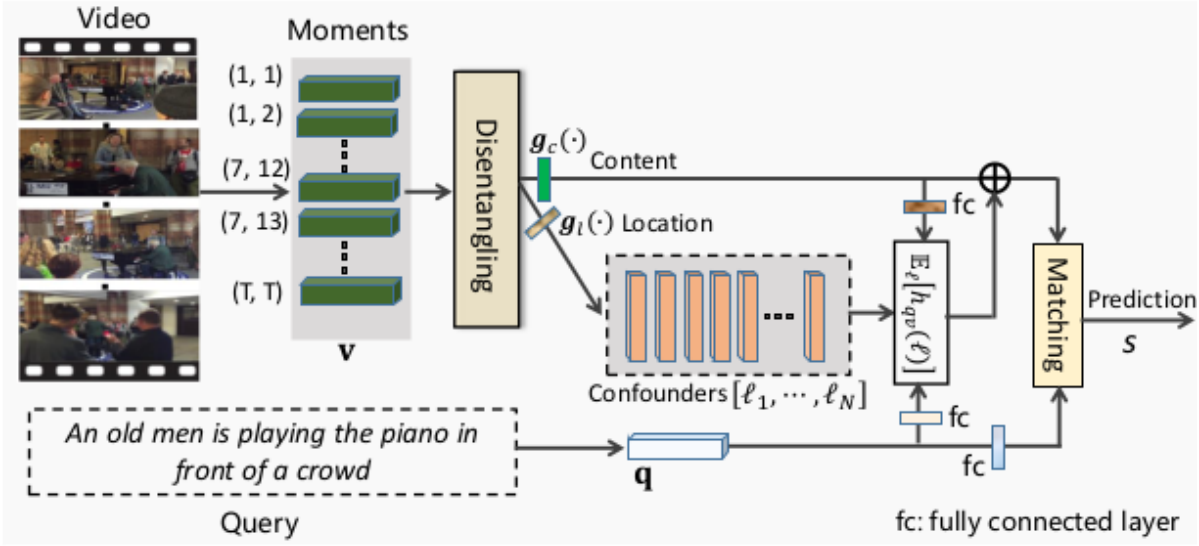


(a)



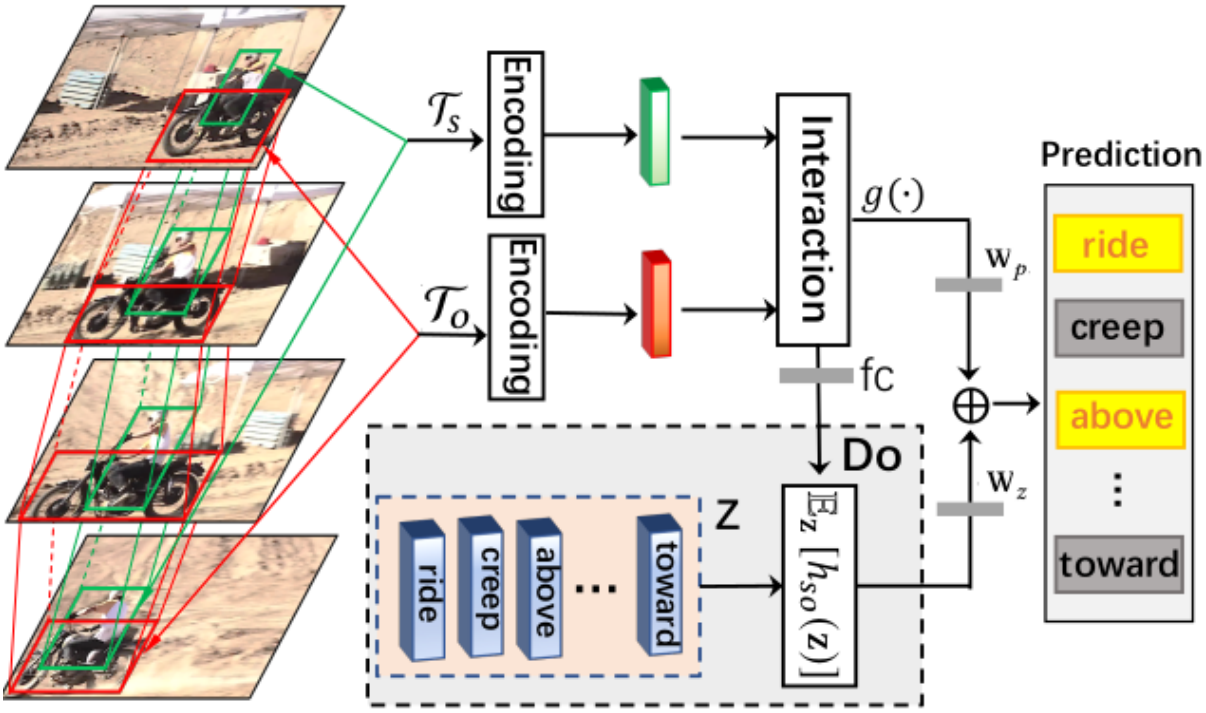
(b)

- Deconfounded Video Moment Retrieval with Causal Intervention
  - ▶ Maheep's Notes



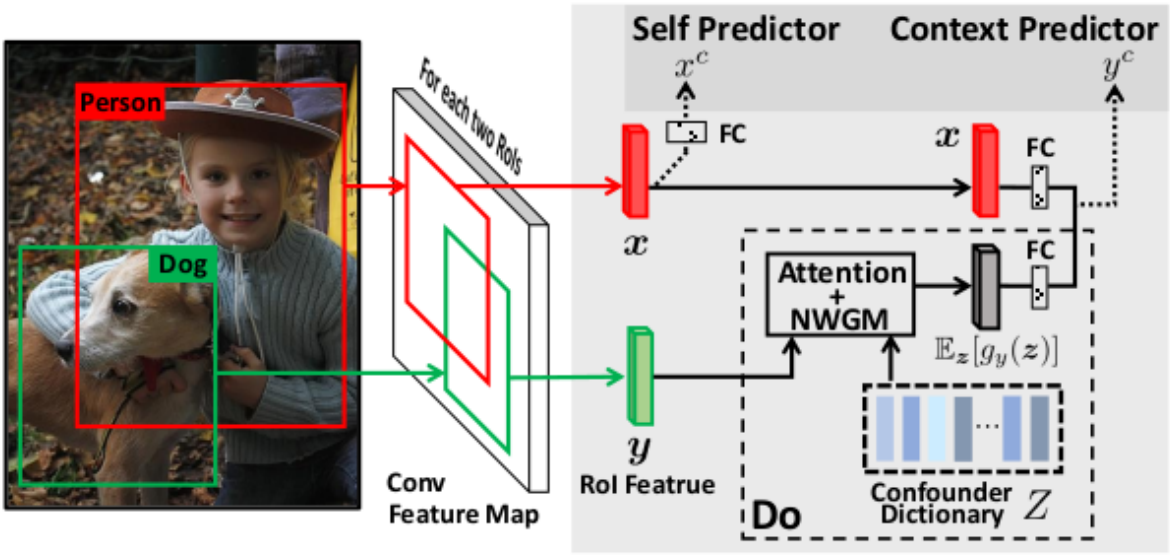
- Intervention Video Relation Detection
  - ▶ Maheep's Notes

    `L = L_obj + lambda*L_pred`, where `L_obj` is the cross entropy loss function to calculate the loss of classifying video object trajectories and `L_pred` is binary cross entropy loss used for predicate prediciton.



- Visual Commonsense R-CNN
  - ▶ Maheep's Notes

    `E[g(z)]` to get the top confounders from the dictionary. Now there is a complexity arise where the confounder is the doctionary act as the colliders therefore they are eridacted through the use of Neural Causation coefficient(NCC).


Person
Dog
For each two RoIs
Conv
Feature Map
Self Predictor
Context Predictor
x^c
y^c
FC
x
FC
x
Attention
+
NWGM
FC
$\mathbb{E}_z[g_y(z)]$
y
RoI Featrue
Confounder
Dictionary Z
Do