On the Interpretability of Attention Networks (Supplementary Material)

Appendix A. Codes for Reproducing Results

All the datasets and codes are available here.

Appendix B. Selective Dependence Classification

Figure 1 illustrates data sampled from an example 1-dimensional base distribution with two foreground classes, and the resulting 2-dimensional mosaic distribution obtained as a result of having $m = 2$ parts per instance. Note the symmetric structure in the scatter plot for the mosaic data, is due to the swap symmetry, i.e. the foreground segment can be either the first or the second segment. This also illustrates that even if the foreground and background are well separated, and the foreground classes are also easily separated, the mosaic data can be significantly more complex. Algorithm 1 in section 3.1 gives the generative model for an instance-label pair in SDC problem.

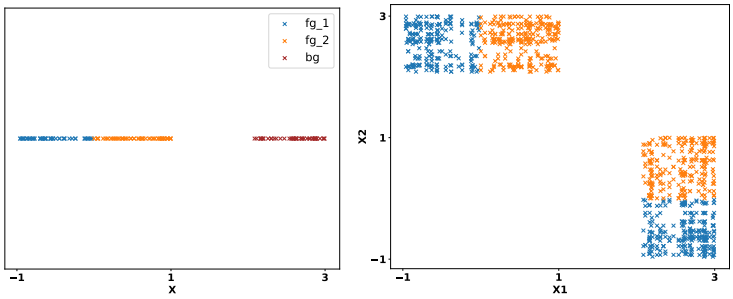

Figure 1: (**left**) Sampled data from D_0(brown), D_1(blue), D_2(orange). (**right**) Mosaic instances.

Remark: There have been links made between attention models and multiple instance learning (MIL) Ilse et al. (2018) and attention models were shown to be a good tool to solve such problems. However, MIL is not an apt problem to study the intricacies of attention models. MIL can effectively be viewed as distinguishing between mosaic instances containing no foreground segment and mosaic instances containing at least one foreground segment. This is distinct from the SDC task where we know the existence of a foreground segment, but are interested in finding the class label of the foreground segment.

Appendix C. Experimental Setup

C.1. Illustration for Interpretability in Image Captioning: A case study

As mentioned in Section 2, We have used a standard method of up-sampling to get a 224×224 image from 14×14 image. Each co-ordinate in the 14×14 $\boldsymbol{\alpha}$ vector corresponds to a square patch in the 224×224 image.

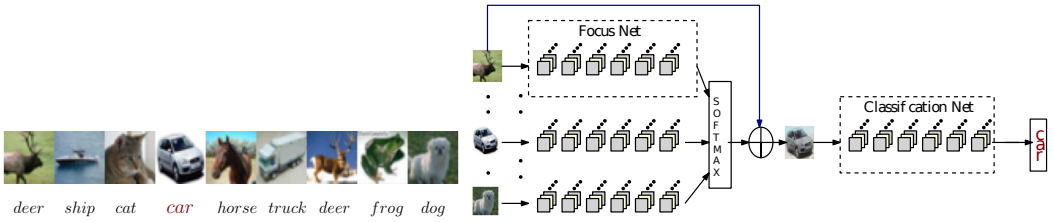

Figure 2: (**Left**) A mosaic instance from CIFAR-SDC Dataset. (**Right**) FCAM Architecture for CIFAR-SDC with averaging at zeroth Layer

Here is a toy example illustrating the interpretability measure in Table 1.

Consider a 4×4 image, where say the word "woman" is predicted as the first word, and the object category of "person" is present in the image according to metadata with the bounding box of this object being the top left quarter of the image. The $\mathbf{v}$ vector here would be $\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. Let the image patches/parts be disjoint 2×2 sub-images of the 4×4 image

A perfect attention model would have $\boldsymbol{\alpha} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. The normalised inner product of an upsampled version of $\boldsymbol{\alpha}$, which would be $\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, and $\mathbf{v}$ would be 1.

A bad attention model might have $\boldsymbol{\alpha} = \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix}$. The normalised inner product of an upsampled version of $\boldsymbol{\alpha}$, which would be $\begin{bmatrix} .3 & .3 & .3 & .3 \\ .3 & .3 & .3 & .3 \\ .2 & .2 & .2 & .2 \\ .2 & .2 & .2 & .2 \end{bmatrix}$, and $\mathbf{v}$ would be approximately 0.6.

A random baseline that randomly chooses one of the four $\boldsymbol{\alpha}$ components to have one and zero elsewhere would have a normalised inner product of 0.25 on average. (corresponding to an inner product of 1 with chance of 25% and 0 with a chance of 75%.)

Also we have categorized the words for each class manually. These words were chosen from vocabulary of the captions. Tables 2, 3 and 4 in the appendix shows the associated words with each object category.

Appendix D. A Synthetic SDC Dataset

We create a 2-dimensional base data, with $k = 3$, foreground classes drawn from distributions D_1, D_2, D_3 which are all normally distributed with different means and identity covariance. The background segments are drawn from D_0, which is a mixture of Gaussians.

Algorithm	averaging layer	attention mechanism	accuracy	FT	NNZ($\boldsymbol{\alpha}$)	Dist($\boldsymbol{\alpha}$)	Ent($\boldsymbol{\alpha}$)
SM-0	zeroth	Softmax (SM)	98.89	77.80	2.407	0.242	0.677
ER-0	zeroth*	Entropy reg.	99.07	77.33	2.139	0.159	0.479
SpMax-0	zeroth	Sparsemax	98.57	77.076	1.612	0.174	0.394
SSM-0	zeroth	Spherical SM	96.16	71.83	3.294	0.338	1.038
HA-0	zeroth	Hard attention	97.264	12.07	1.209	0.037	0.100
SM-2	second	Softmax (SM)	99.67	86.95	4.766	0.422	1.469
ER-2	second	Entropy reg.	99.85	87.89	3.720	0.325	1.099
SpMax-2	second	Sparsemax	99.76	87.17	2.722	0.370	0.979
SSM-2	second	Spherical SM	99.79	89.12	4.962	0.393	1.380
HA-2	second	Hard attention	84.91	10.64	1.247	0.0474	0.121

Table 1: Performance on Synthetic SDC Dataset: Standard FCAM and variants.

We have $m = 9$ segments in each mosaic instance. Each instance $\mathbf{x} \in \mathbb{R}^{2\times 9}$ in mosaic data is associated with a label $\mathbf{y} \in [3]$. We sample 6000 such mosaic instances and set aside 3000 points for testing and use the rest for training the FCAM. Algorithm 1 in section 3.1 is used to generate mosaic instances.

The Focus model f is a multilayer perceptron (MLP) architecture with 2−hidden layers each having 50 units. Classification model $\mathbf{g}$ is also a MLP architecture with single hidden layer having 50 units. The 3 layers of the focus network allow for averaging to be done at either the input level (2-dimensional) or at the first or second hidden layer (50-dimensional). An illustration of the dataset and the architecture is given in the appendix.

We generate synthetic data with D_0 (background) as mixture of Gaussian and D_1 (foreground 1), D_2 (foreground 2),D_3 (foreground 3) as Gaussian distribution with their mean and standard deviation (0.01) as illustrated in the figure 3(a). An illustration of mosaic data (segments $m = 9$) created using base synthetic data is shown in the figure 3 (b).

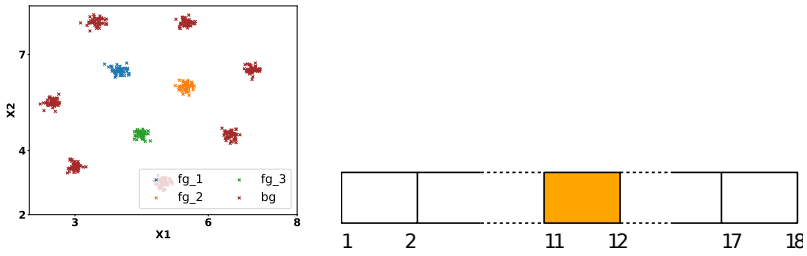

Figure 3: (a) Synthetic Dataset, (b) Mosaic Instance from Synthetic Dataset having one patch from fg2

D.0.1. Experiments on Synthetic SDC Dataset

Figure 4 shows the MLP architecture we employed, with two and one hidden layers in focus and classification modules respectively, each of 50 hidden dimension. We used Adam optimizer with learning rate of 0.0005 and tuned learning rate over search space of $0.001, 0.003, 0.0005$. For the entropy experiments, we considered the λ values in the set $\{0.001, 0.003, 0.005\}$ and trained our models for 5 different random seeds among $\{0, 1, 2, 3, 4\}$. Figure 5 shows the fraction of instances for which the attention vector $\boldsymbol{\alpha}$ scores the true foreground index above a threshold. In table 1, Zeroth Layer averaging with entropy regularisation is average over 4 runs.

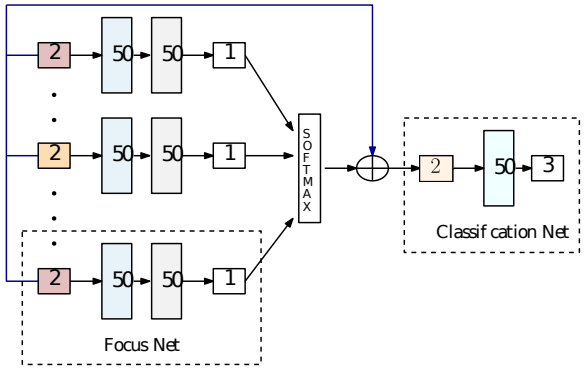

Figure 4: Architecture for Synthetic Dataset with averaging at zeroth Layer

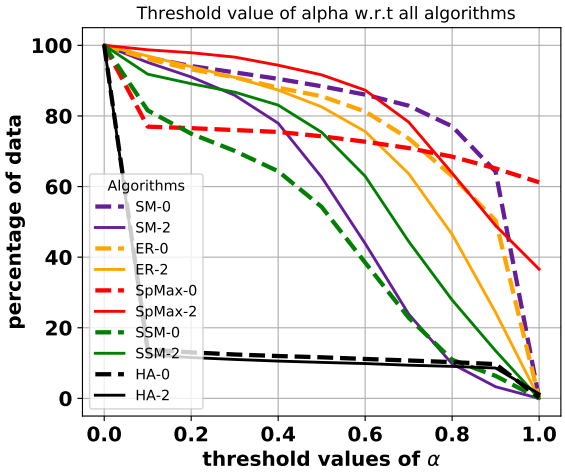

Figure 5: Fraction of test data for which the focus score α_{j^*} for the true foreground index j^* is above a threshold, plotted as function of the threshold for the algorithms mentioned in Table 1

Appendix E. Further Details on Experiments with Synthetic CIFAR10 Dataset

For CIFAR data we use the CNN architecture with 6 CNN along with 3 linear layers in focus and classification modules. We used Adam optimizer with learning rate of 0.0005 and tuned learning rate over search space of $0.001, 0.003, 0.0005$. For the entropy experiments, we considered the λ values in the set $\{0.001, 0.003, 0.005\}$ and trained our models for 3 different random seeds among $\{0, 1, 2\}$. In table 2, Zeroth Layer averaging with entropy regularisation is average over 2 runs.

Category ID	Label	Words associated
1	man	man, men, woman, women, child, children, kid, kids, girl, girls, boy, boys, male, female, person
2	bicycles	bicycles, bicycle, cycles, bike
3	car	car, cars, van, volkswagon, vehicles, bmw, automobile, suv
4	motorcycle	motorcycle, motorcycles, bike, bikes, motorcyclist, motorized, motor, scooters, motorbikes,
5	airplane	airplane, plane, bomber, airplanes, air, crafts, jets, glider, bi-plane aircraft, jet, cargo, airliner,
6	bus	bus, school bus, double decker, busses, vehicles
7	train	train, train engine, cargo train, rails, locomotive, steam engine, train car, diesel train engine, engine
8	truck	truck, fire trucks, tow truck, trucks, pickup truck, trailer, vehicles
9	boat	boat, canoe, cargo boat, ship, trawler, sailboats, rafts
10	traffic light	traffic light, stop light, red light, green light, traffic sign
11	fire hydrant	fire hydrant, firehydrant, hydrant
13	stop sign	stop sign, street sign, sign, signs
14	parking meter	parking meter, meter
15	bench	bench, seat, chairs, lounge
16	bird	bird, red robin, parrot, ostrich, swans, ducks, geese, owl, birds, swan, duck, seagull, duckling, flamingos, pigeons, toucan, seagulls
17	cat	cat, cats, kitten, kittens, animal, animals
18	dog	dog, dogs, bulldog, puppy, pup, animal, animals
19	horse	horse, carriage, animal, animals, horses
20	sheep	sheep, cattle, animal, animals, lamb, lambs
21	cow	cow, cows, calf, calfs, calves, animal, animals, cattle, oxen, ox
22	elephant	elephant, elephants, animal, animals
23	bear	bear, bears, cub, cubs, animal, animals
24	zebra	zebra, zebras, animal, animals

Table 2: Word association table for the case study in Section 2

Category ID	Label	Words associated
25	giraffe	giraffe, giraffes, animal, animals
27	backpack	backpack, bag, bags, backpacks, luggage, back pack
28	umbrella	umbrealla, umbrellas
31	handbag	handbag, handbags, bag, bags, luggage
32	tie	tie, ties
33	suitcase	suitcase, suitcases, luggage, suit case
34	frisbee	frisbee, frisbees, frizbee, frizbees, frisk bee
35	skis	skis, skiing, skier, skiers, ski, spikes, ski
36	snowboard	snowboard, snowboarding, snowboarder, snow board, ski boarder
37	sports ball	sports ball, ball, soccer, baseball, tennis ball, football, volleyball, basketball, soccer ball, soccer balls
38	kite	kite, object, kites
39	baseball bat	baseball bat, bat, bats
40	baseball glove	baseball glove, baseball gloves, gloves, glove, catcher, catch, mitt
41	skateboard	skateboard, skateboarders, skateboarder, skate board, skateboarding, skate boarding
42	surfboard	surfboard, surf board, surfer, boogie, board, wakeboard, surfing
43	tennis racket	tennis racket, tennis racket, tennis rackets, rackets
44	bottle	bottle, bottles, soda, soda, can, drinks, water bottle, water jars
46	wine glass	wine glass, wine glasses, glass, glasses, drink, drinking, drinks
47	cup	cup, cups, mug, mugs, drink, coffee, tea
48	fork	fork, forks, silverware
49	knife	knife, knives, silverware
50	spoon	spoon, spoons, silverware
51	bowl	bowl, bowls, dishes, dish, cup, cups
52	banana	banana, bananas, fruit, fruits
53	apple	apple, apples, fruit, fruits
54	sandwich	sandwich, hamburger, hamburgers, burgers, sandwiches, burger, bun
55	orange	orange, oranges, fruit, fruits
56	broccoli	broccoli, vegetables, vegetable, food, meal
57	carrot	carrot, carrots, vegetable, vegetables, food, meal
58	hot dog	hot dog, hot dogs, hotdog, hotdogs, sandwich, sandwiches, bun
59	pizza	pizza, bread, baked, pizzas, food
60	donut	donut, donuts, cookies, baked, pastries, doughnuts, doughnut, food, dessert, pie
61	cake	cake, cakes, pastries, pastry, dessert, pie

Table 3: Word association table for the case study in Section 2

Category ID	Label	Words associated
62	chair	chair, chairs, furniture, furnitures
63	couch	couch, couches, furniture, furnitures, recliner, recliners
64	potted plant	potted plant, potted plants, pot, pots, plants, plant, vases, flowers, flower, leaves, leaf
65	bed	bed, beds, furniture, furnitures
67	dining table	dinning table, table, tables, furniture, furnitures, dinner table
70	toilet	toilet, bathroom, restroom, toilette seat
72	tv	tv, screen, t.v., television, monitor, monitors, televisions
73	laptop	laptop, computer, monitor, computers, monitors, laptops
74	mouse	mouse
75	remote	remote, remotes, controller
76	keyboard	keyboard, key board, keyboards
77	cell phone	cell phone, cell, phone, phones, mobiles, mobile
78	microwave	microwave, appliances, appliance
79	oven	oven, appliances, appliance
80	toaster	toaster, appliances, appliance
81	sink	sink
82	refrigerator	refrigerator, fridge, refrigerators, fridges
84	book	book, books
85	clock	clock, clocks
86	vase	vase, vases, bouquet, pot
87	scissors	scissors
88	teddy bear	teddy bear, toy, soft toy, stuffed animal, teddy, panda bear, teddy bears, stuffed animals,stuff bears, stuffed panda, bear, doll, dolls, stuffed bear, stuffed bears
89	hair drier	hair drier, hair dryer, hairdryer, hair products, hair product, blow dryer
90	toothbrush	toothbrush, brush, object, tooth, brush

Table 4: Word association table for the case study in Section 2