ML Project - Cross Modal Representation Learning

Group 15
Kunj Mehta - kcm161
Linqi Xiao - lx130
Aishwarya Harpale - ach149
Neil Pillai - ncp67

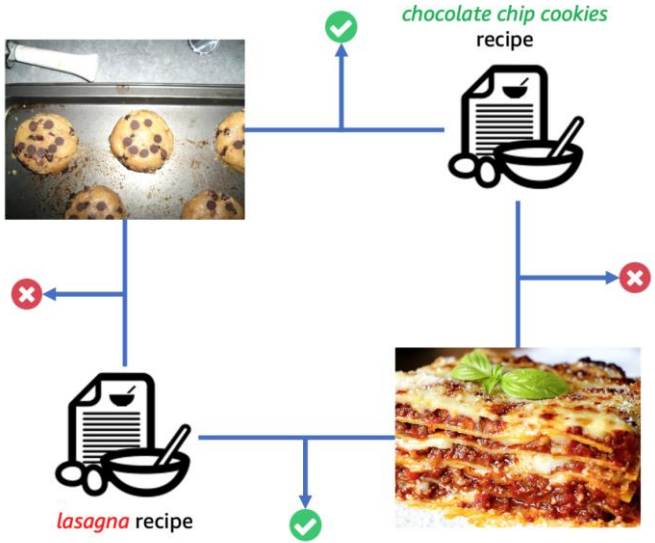
Introduction

In machine learning, multi-view representation learning refers to the setting where the model learns the multiple views the data comes in.

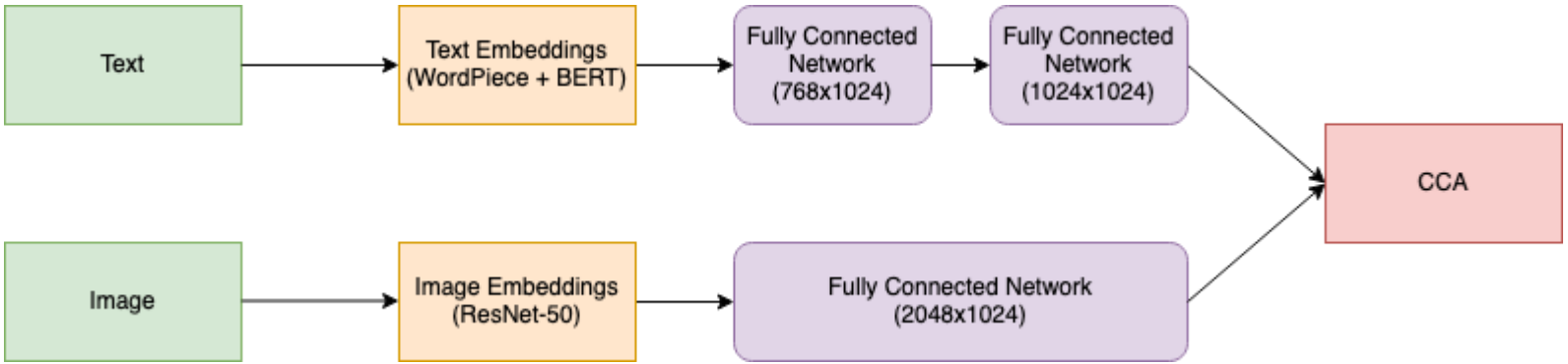
In this project, we apply this concept to cross-modal retrieval for Food AI using Recipe 1 Million dataset.

We tackle the im2recipe and recipe2im problems.

We train CCA models, non-linear feedforward neural networks with parameter sharing using MSE loss and Triplet Loss, and cross-modal transformers to enhance cross-modal representations, and show the effectiveness of our models via visualizations.

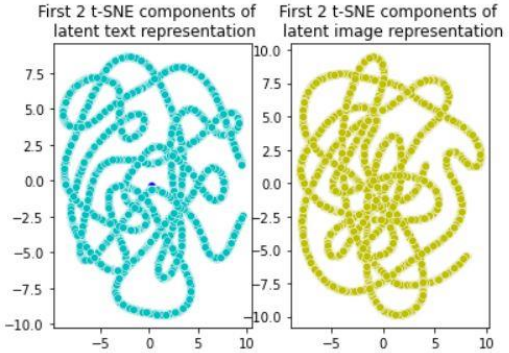


Stage 1 Architecture



[1] Guerrero, Ricardo, Hai Xuan Pham, and Vladimir Pavlovic. "Cross-Modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Representation Learning." arXiv preprint arXiv:2012.01345 (2020).
[2] Marin, Javier, et al. "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images." IEEE transactions on pattern analysis and machine intelligence 43.1 (2019): 187-203.
[3] A. Salvador et al., "Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3068-3076, doi: 10.1109/CVPR.2017.327.

Stage 1 Results and Visualizations



Test Sample Size / Dims	MedR	RK@1	RK@5	RK@10
1k / dim = 2	206.7	0.0035	0.0169	0.0329
1k / dim = 10	22.45	0.0506	0.1991	0.3191
1k / dim = 25	5	0.2229	0.5223	0.6703
1k / dim = 50	2.6	0.3540	0.6824	0.8046
1k / dim = 100	2.	0.4122	0.7384	0.8432
1k / dim = 200	1.9	0.4738	0.7702	0.8508
1k / dim = 500	1	0.5563	0.7981	0.8514
1k / dim = 1000	1	0.5508	0.7712	0.8161
10k / dim = 2	209.65	0.0035	0.015	0.0325
10k / dim = 10	22.85	0.0557	0.2036	0.3212
10k / dim = 25	4.9	0.2208	0.5244	0.6756
10k / dim = 50	2.7	0.3515	0.6797	0.7995
10k / dim = 100	2	0.4133	0.7431	0.8404
10k / dim = 200	1.9	0.4852	0.7773	0.8526
10k / dim = 500	1	0.5543	0.7939	0.8513
10k / dim = 1000	1	0.5567	0.7662	0.8083

Table 3. Dimensional Analysis for recipe2im

Test Sample Size / Dims	MedR	RK@1	RK@5	RK@10
1k / dim = 2	205.6	0.0034	0.0161	0.0321
1k / dim = 10	22.75	0.054	0.1999	0.3222
1k / dim = 25	5.05	0.209	0.5248	0.6799
1k / dim = 50	2.3	0.3559	0.6919	0.8116
1k / dim = 100	2	0.4083	0.7403	0.8433
1k / dim = 200	1.9	0.4769	0.7789	0.8557
1k / dim = 500	1	0.5501	0.7937	0.8451
1k / dim = 1000	1	0.5532	0.7712	0.8158
10k / dim = 2	201.6	0.0033	0.0157	0.0311
10k / dim = 10	22.8	0.0534	0.2	0.3192
10k / dim = 25	4.85	0.2107	0.5294	0.6819
10k / dim = 50	2.6	0.3436	0.6906	0.8056
10k / dim = 100	2	0.4053	0.7384	0.8400
10k / dim = 200	1.95	0.4876	0.7846	0.8592
10k / dim = 500	1	0.5571	0.8049	0.8551
10k / dim = 1000	1	0.5584	0.7686	0.8164

Table 1. Dimensional Analysis for im2recipe

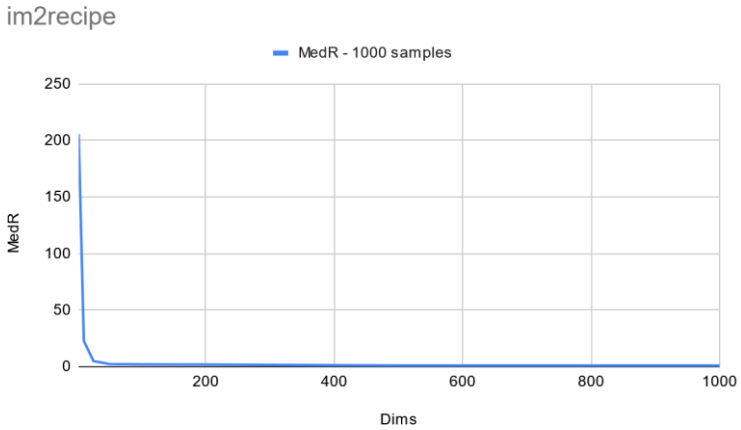
Test Sample / Recipe Comp.	MedR	RK@1	RK@5	RK@10
1k / all	1	0.5516	0.7968	0.85
1k / instructions	2	0.3527	0.6108	0.6873
1k / ingredients	3	0.3555	0.6089	0.6807
1k / title	9.6	0.2215	0.4395	0.5117
10k / all	1	0.5428	0.7882	0.8426
10k / instructions	3	0.3538	0.6054	0.6821
10k / ingredients	3	0.3631	0.6123	0.6843
10k / title	10.9	0.2109	0.4307	0.4992

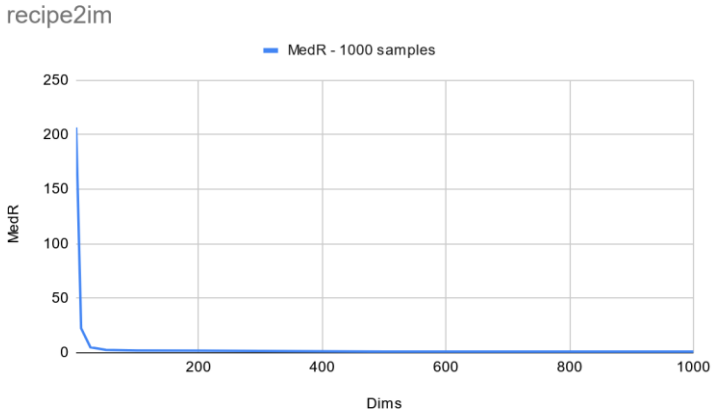
Table 2. Ablation Studies for im2recipe

Test Sample / Recipe Comp.	MedR	RK@1	RK@5	RK@10
1k / all	1	0.5437	0.7927	0.8454
1k / instructions	2.9	0.3605	0.6161	0.6944
1k / ingredients	2.8	0.371	0.604	0.6749
1k / title	9.8	0.2188	0.4379	0.5057
10k / all	1	0.5551	0.7871	0.842
10k / instructions	3	0.3602	0.6132	0.6824
10k / ingredients	3	0.3642	0.6056	0.6761
10k / title	10.65	0.2148	0.4299	0.5042

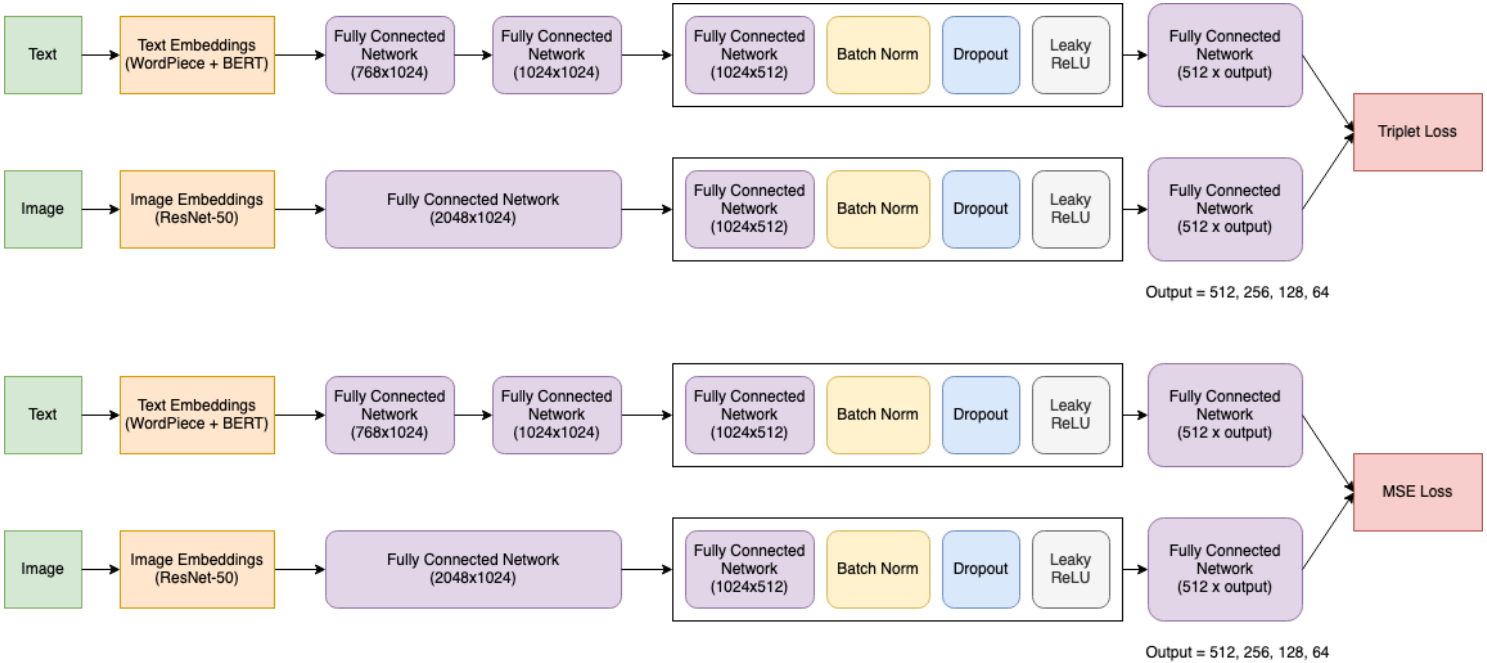
Table 4. Ablation Studies for recipe2im

Stage 1 Results and Visualizations





Stage 2 Architecture



[1] Guerrero, Ricardo, Hai Xuan Pham, and Vladimir Pavlovic. "Cross-Modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Representation Learning." arXiv preprint arXiv:2012.01345 (2020).

Stage 2 MSE Results

Figure 5. Dimensional Analysis and Ablation Studies - MSE Loss

	Dims	Sample	Mean median
	Dimensional Analysis		
im2recipe	64	1000	51.3
	128	1000	12.6
	256	1000	6.95
	512	1000	5.1
	64	10000	510.75
	128	10000	122
	256	10000	59.2
	512	10000	40.7
im2title	64	1000	4.9
	128	1000	2
	256	1000	2
	512	1000	1.4
	64	10000	38.7
	128	10000	13.6
	256	10000	9.1
	512	10000	7
im2ingredients	64	1000	84.6
	128	1000	33.9
	256	1000	13.3
	512	1000	12.25
	64	10000	828.8
	128	10000	334.9
	256	10000	127.05
	512	10000	112.3
im2instructions	64	1000	84.05
	128	1000	47.8
	256	1000	13.7
	512	1000	9
	64	10000	835.5
	128	10000	449.25
	256	10000	129
	512	10000	84.2
	Evaluation and Ablation Studies		
im2recipe	512	1000	5.2
	512	10000	43.1
im2title	512	1000	24.6
	512	10000	247.35
im2ingredients	512	1000	15.2
	512	10000	141.25
im2instructions	512	1000	21.85
	512	10000	204.25

Figure 1. tSNE Plot of shared space for dimensions = 512 using MSE Loss

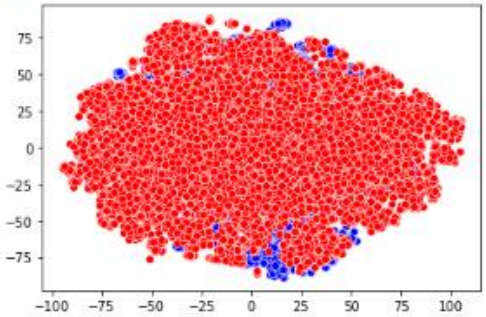


Figure 4. Vector Arithmetic with Shared Space Embeddings MSE Loss / dims = 512: chocolate cake - cake + cupcake

[1] A. Salvador et al., "Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3068-3076, doi: 10.1109/CVPR.2017.327.
[2] Marin, Javier, et al. "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images." IEEE transactions on pattern analysis and machine intelligence 43.1 (2019): 187-203.

Stage 2 Triplet Loss Results

Figure 6. Dimensional Analysis and Ablation Studies - Triplet Loss

	Dims	Sample	Mean median
	Dimensional Analysis		
im2recipe	64	1000	2.4
	128	1000	2
	256	1000	2.1
	512	1000	2.4
	64	10000	14.9
	128	10000	13.9
	256	10000	15
	512	10000	15.9
im2title	64	1000	4.9
	128	1000	5.1
	256	1000	4.9
	512	1000	5.1
	64	10000	39.25
	128	10000	41.6
	256	10000	39.5
	512	10000	42.3
im2ingredients	64	1000	4
	128	1000	3.85
	256	1000	3.9
	512	1000	4.1
	64	10000	28.6
	128	10000	29
	256	10000	28.5
	512	10000	31
im2instructions	64	1000	3
	128	1000	3
	256	1000	3.2
	512	1000	3
	64	10000	20
	128	10000	21.8
	256	10000	24.7
	512	10000	22.1
	Evaluation and Ablation Studies		
im2recipe	256	1000	2
	256	10000	15.1
im2title	256	1000	4.85
	256	10000	40.25
im2ingredients	256	1000	3.8
	256	10000	28.4
im2instructions	256	1000	3.4
	256	10000	24.7

Figure 2. tSNE Plot of shared space for dimensions = 512 using Triplet Loss

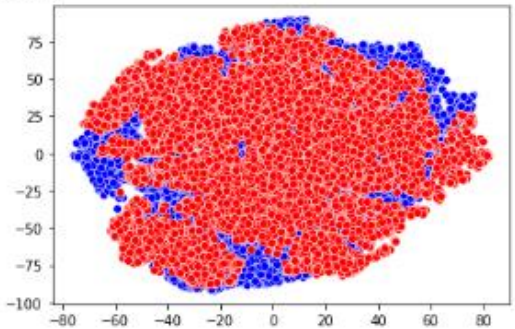


Figure 3. Vector Arithmetic with Shared Space Embeddings Triplet Loss / dims = 512: chocolate cake - cake + cupcake

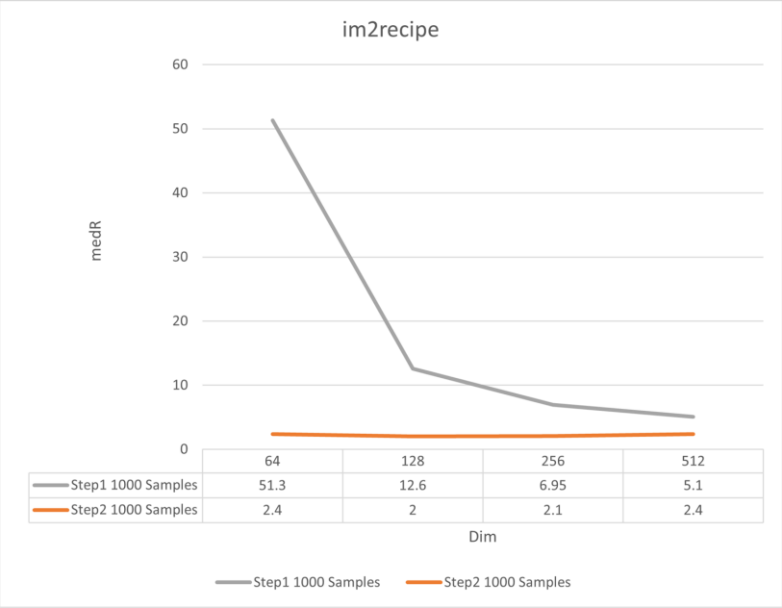
[1] A. Salvador et al., "Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3068-3076, doi: 10.1109/CVPR.2017.327.
[2] Marin, Javier, et al. "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images." IEEE transactions on pattern analysis and machine intelligence 43.1 (2019): 187-203.

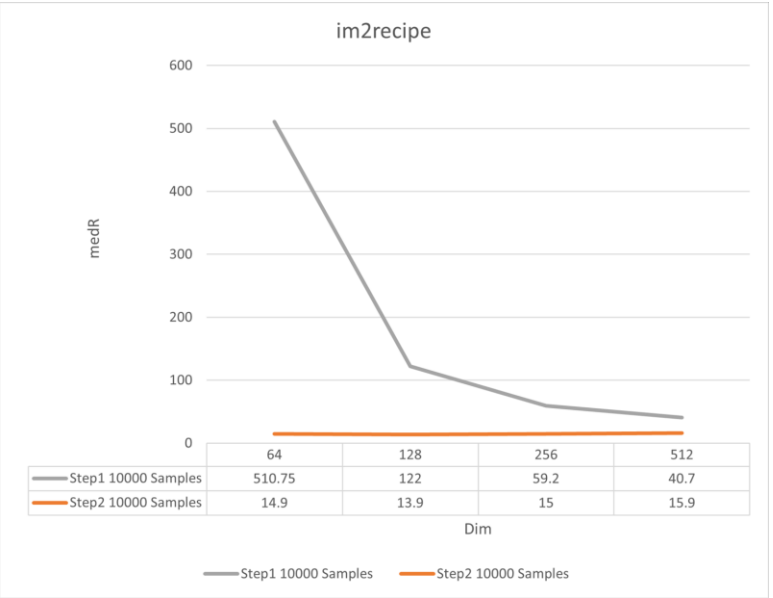
Comparison between MSE Loss and Triplet Loss

im2recipe results	MSE/ 1k samples	Triplet/ 1k samples	MSE/ 10k samples	Triplet/10k samples
MedR	5.2	**2**	43.1	**15.1**
RK@1	0.2348	**0.3794**	0.06055	**0.11216**
RK@5	0.5183	**0.6984**	0.18576	**0.31331**
RK@10	0.6353	**0.7993**	0.27418	**0.43186**

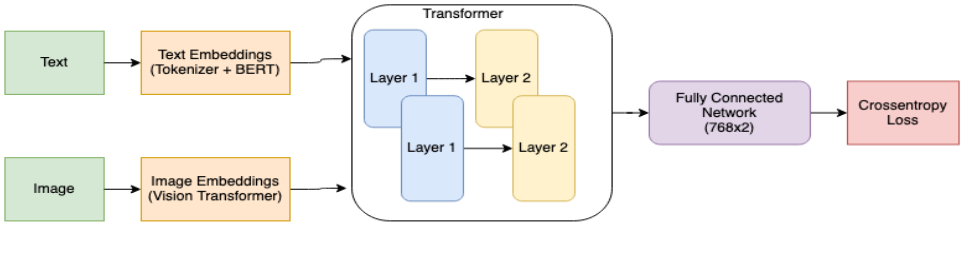
- Comparing results for the best architectures for MSE loss (output layer = 512) and triplet loss (output layer = 256) on full recipe (averaged) embeddings
- The model employing triplet loss outperforms MSE on all metrics

Comparison between MSE Loss and Triplet Loss

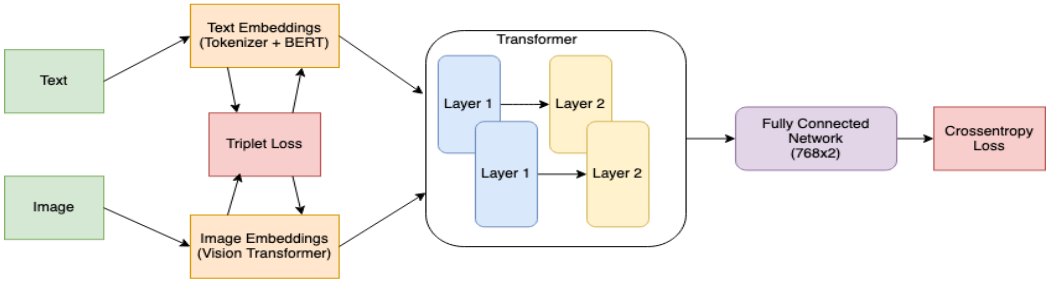




Stage 3 Architecture



Architecture 1



Architecture 2

[1] Guerrero, Ricardo, Hai Xuan Pham, and Vladimir Pavlovic. "Cross-Modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Representation Learning." arXiv preprint arXiv:2012.01345 (2020).
[2] Fu, Han, et al. "Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[3] Shukor, Mustafa, et al. "Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval." *arXiv preprint arXiv:2204.09730* (2022).

Preliminary Results and Future Work

- Retrieval results for ingredients on just the Triplet Loss trained embeddings for 1k samples: medR: 15.7, Recall @1: 0.1, Recall@5: 0.29, Recall@10: 0.42
- Retrieval results after cross-modal attention on ingredients 1k samples: medR = 496.7, Recall@1: 0.0008, Recall@5: 0.0039, Recall@10: 0.007
- Finish optimizing the Stage 3 transformer architecture
- Visualize the attention maps to show cross-modal attention

THANK YOU!