Image-and-Text Fusion for improved Image Classification

Team 4
DATCHANAMOURTY Rohitkumar, SOUTHIRATN Thibaud, Kim Taegyu, LANG Yohan

Table of Contents

1. Context & Motivation
2. Idea & Project Overview
3. Methods
4. Results
5. Future Improvements & Conclusion

Context & Motivation

Image and Text fusion for UPMC Food-101 using BERT and CNNs

Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa

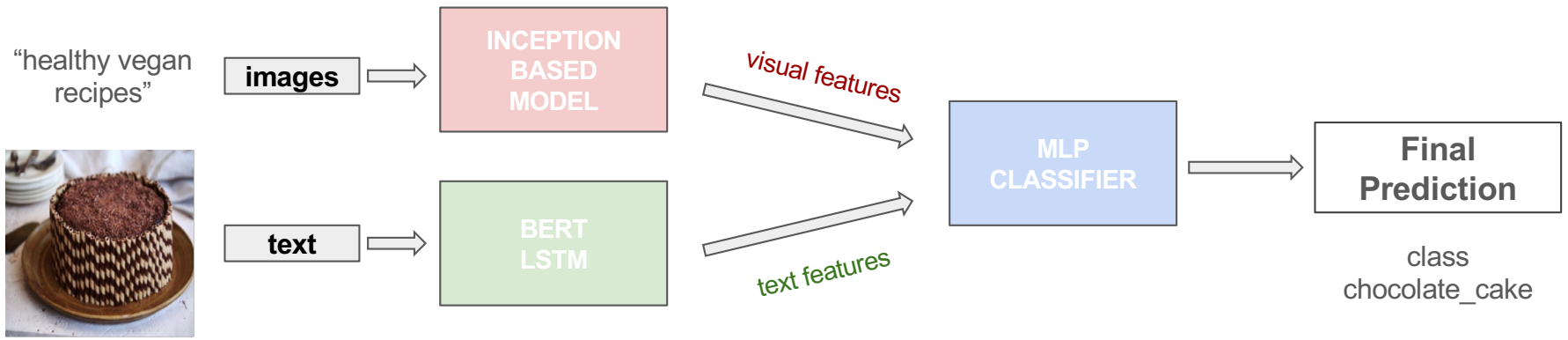
Department of Theoretical and Applied Science, University of Insubria, Varese, Italy

UPMC Food-101

- Multimodal data: image and text
- 101 classes
- Highly noisy: 5% of irrelevant images

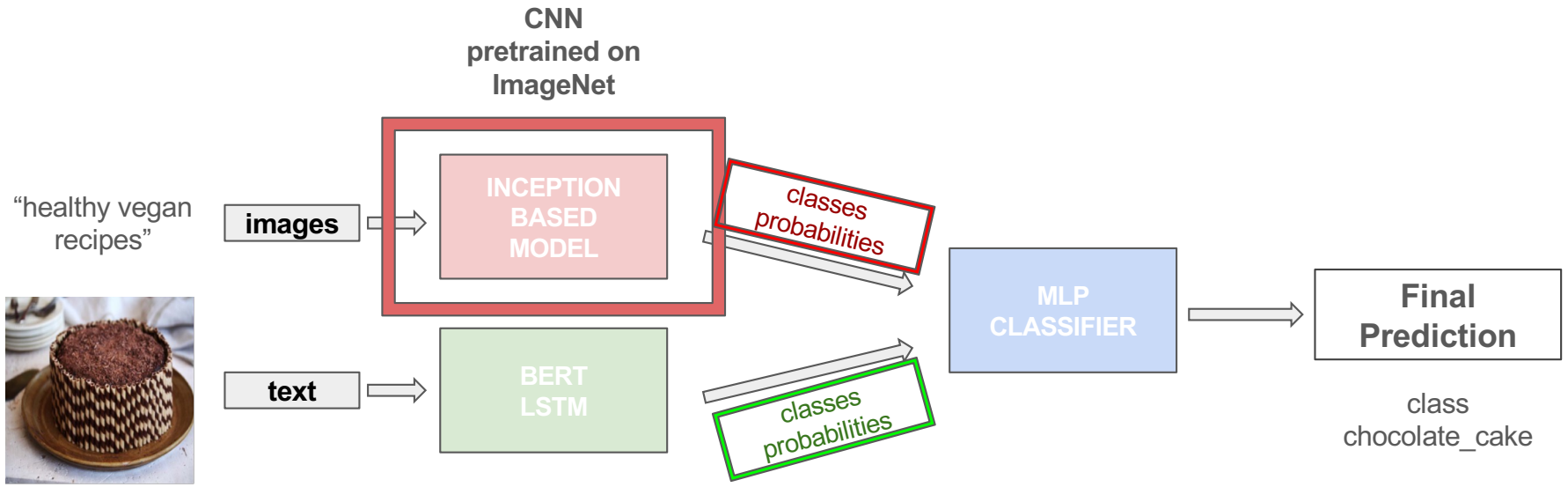
Baseline - Gallo et a.l. (2020): Early Fusion

Use multimodal information for improved classification



Feature Level Fusion

Baseline - Gallo et a.l. (2020): Late Fusion



Decision Level Fusion

Baseline - Gallo et a.l. (2020): Results

Test Accuracy on UPMC Food-101

Images Model	Text Model	Late Fusion	Early Fusion
71.67%	84.41%	84.59%	92.50%

→ **Previous SOTA *(Narayana et al.)* : 92.3%**

Idea & Project Overview

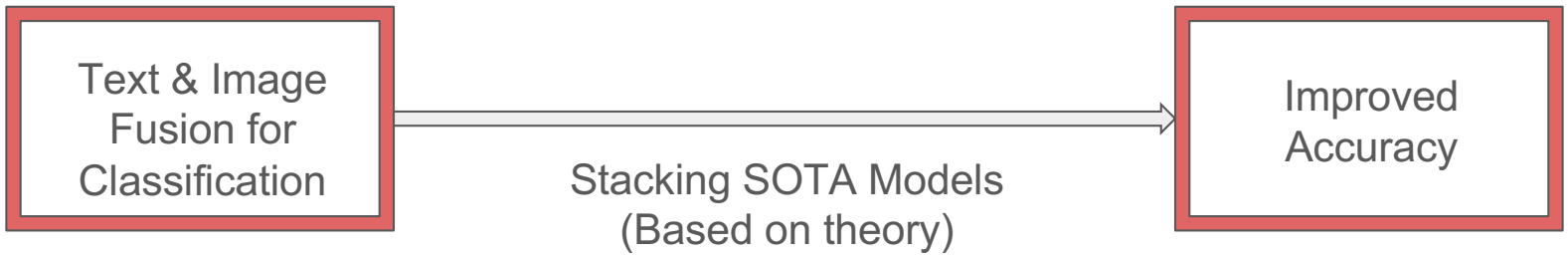
Multimodality in Image Classification

Gallo et a.l. (2020) → Decent Performance

Today: Multimodal I/O (e.g. ***MiniGPT-5*** - Zheng et a.l, 2023…)

Our project:

Improve performance (*Training & Inference Efficiency*)



Improving accuracy - Stacking Text+Image

Gallo et a.l. (2020) → ***Stacking 2 "good but not SOTA models" can perform like SOTA.***

Time constraints: only focus on Early Fusion.

	Vision Model	*Text* Model
Gallo et a.l. (2020)	InceptionV3 (2019)	BERT + LSTM (2018)
Our Project	**EfficientNetV2 (2021)**	**AnglE (2023) + Encoder***

*Encoder**: {BERT, RoBERTa, LLaMa…} or native ***UAE*** *(Universal AnglE Embedding)*

Project Pipeline

Replicate Baseline

Appropriate:
- Dataset
- Text Architecture
- Vision Architecture
- Stacking Method

Early Stack EfficientNetV2 & AnglE

- Evaluate our implementation
- Attempt to outperform baseline

Optimize hyperparameter choice

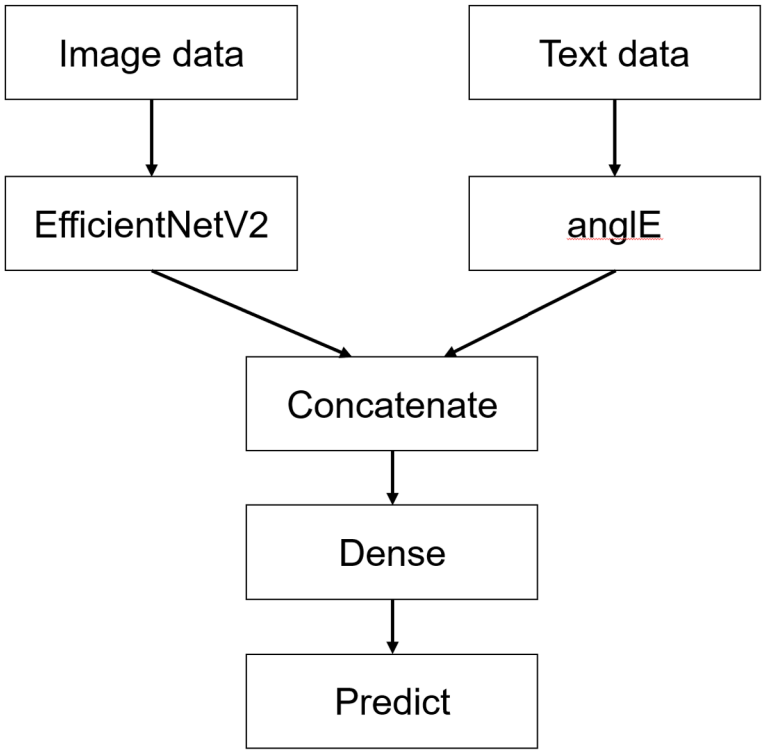
- Learning Rate for better accuracy

Extend to Image Captioning

- Fine-Tune text model on custom dataset
- Attempt to generate captions using a GIT model

Methods

Model Architecture



EfficientNetV2 over InceptionV3

EfficientNetV2 (2021) = Improved Version of EfficientNet (2019).

Better than InceptionV3 on:

- Computational Cost
- Smaller Model Size
- Scaling
- Faster Inference
- Better Accuracy…

	EfficientNetV2	**InceptionV3**
Architecture	Compound Scaling	Inception Modules
Model Size (↓)	+	-
Parameters # (↓)	+	-
Training Speed (↑)	+	-
Inference Speed (↑)	+	-

EfficientNetV2 - Training-aware NAS

Jointly optimize *Accuracy*, *Param.* & *Training Efficiency* on **modern accelerators**

Stage-based design choices:

- Convolutional Operation Types {*MBConv*, *Fused-MBConv*}
- # of Layers
- Kernel Size {3x3, 5x5}
- Expansion Ratio {1, 4, 6}

Reinforcement Learning: Optimize $A \cdot S^w \cdot P^v$

We need better Accuracy (A), less training step time (S) at the same time, a network with less parameters (P)

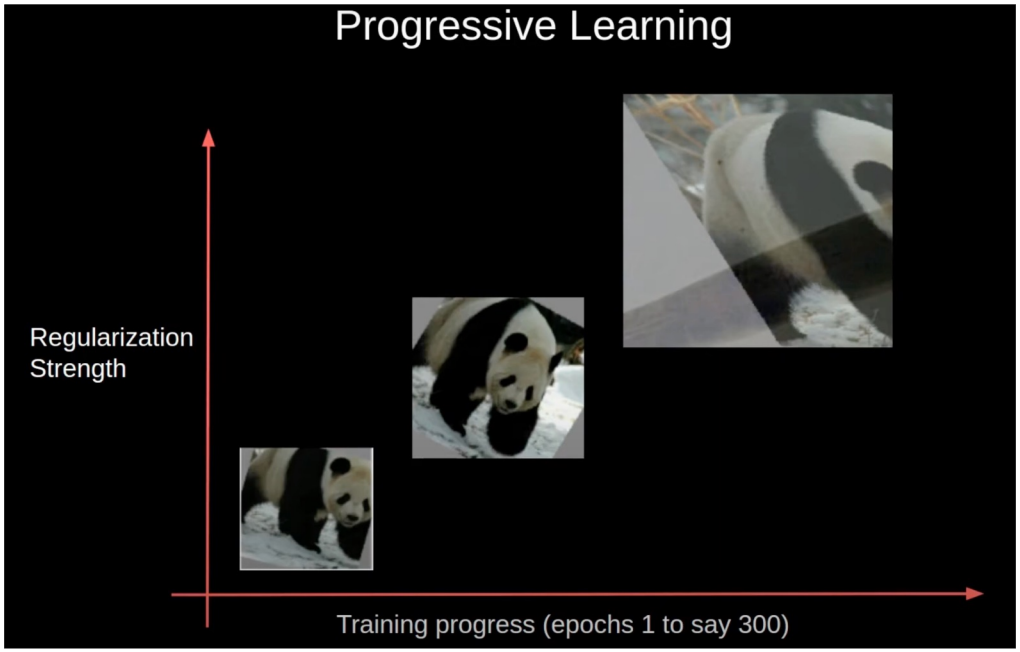
w and v are constants experimentally determined

EfficientNetV2 - Progressive Learning

Idea: Resolution is proportional to computation costs.

→ Adaptive Regularization (e.g. data augmentation) along with image size.

	Early Training	**After**
Image Size	Small	Bigger
Regularization	Weak	Strong



AnglE over base BERT+LSTM

AnglE = Input Layer + 3 Objectives

	AnglE + Encoder	**Base BERT+LSTM**
Embedded Input	Whole Sentence	Tokenized Sentence (Sequence)
Embedding Dimension	Consistent (1D): [1*1028]	[Sequence_length*768]
Prediction Task & Method	Semantic Textual Similarity (STS) based on new ***"Angle Objective"***	Next Sentence Prediction (NSP) based on **Masked Language Modeling** (MLM)

Quality & Efficiency*: AnglE > Traditional STS Cosine Similarity > BERT (word2vec)*

AnglE - Cosine Saturation Zone

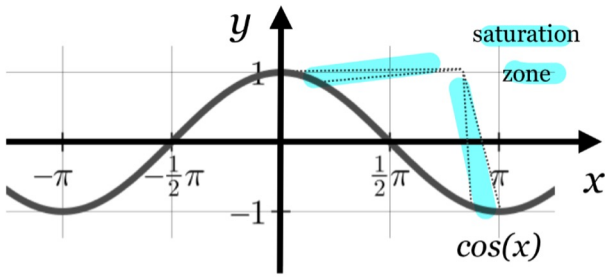
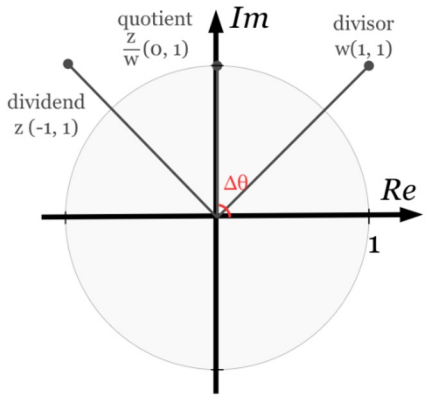


Figure 1: The saturation zones of the cosine function. The gradient at saturation zones is close to zero. During backpropagation, if the gradient is very small, it could kill the gradient and make the network difficult to learn.

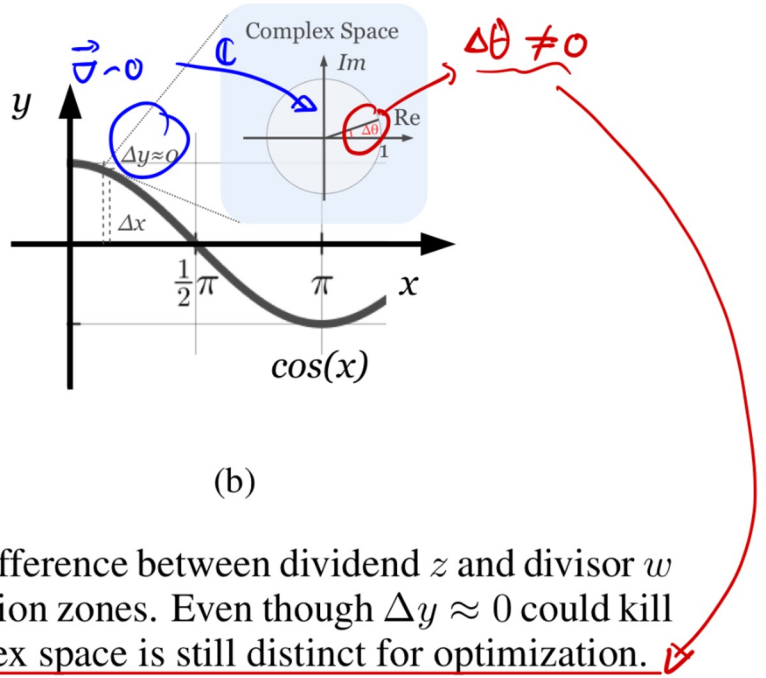
Work in $\mathbb{C}$!

AngIE - Angle Difference

$$\arg\left(\frac{z_1}{z_2}\right) = \arg(z_1) - \arg(z_2) = \theta_1 - \theta_2$$



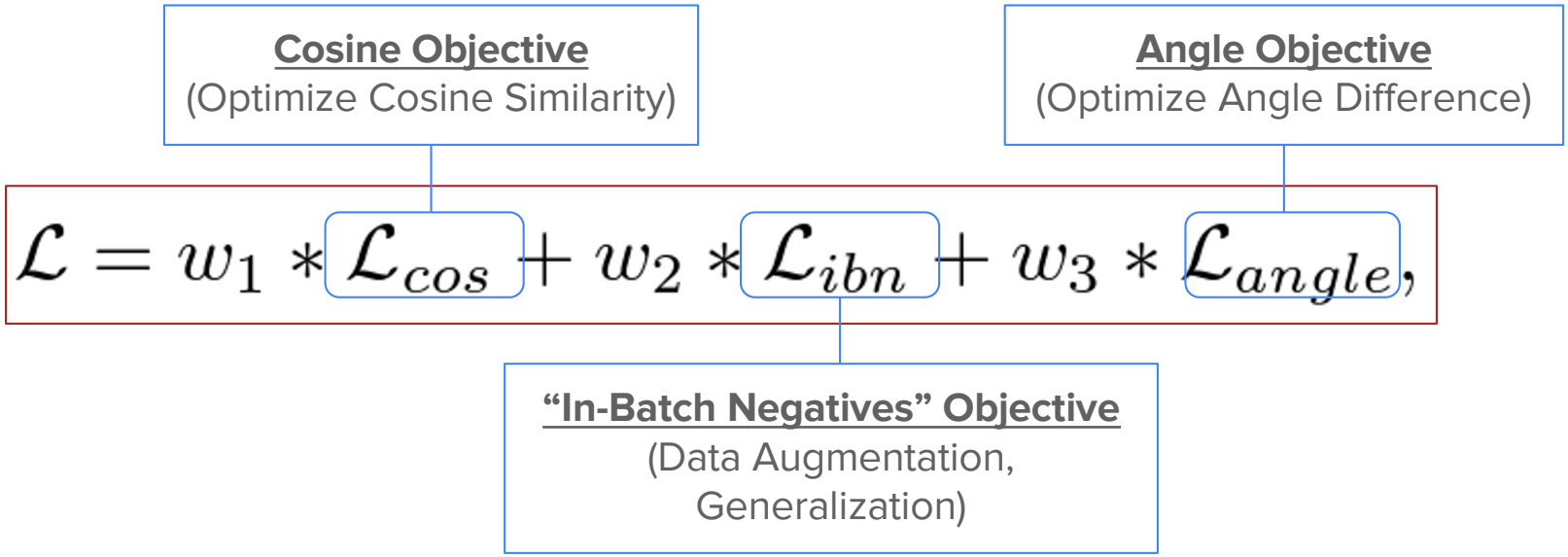
(a)



(b)

Figure 2: (a) Division in complex space. $\Delta\theta$ is the angle difference between dividend z and divisor w in complex space. (b) Angle optimization in cosine saturation zones. Even though $\Delta y \approx 0$ could kill the gradient, the corresponding angle difference in complex space is still distinct for optimization.

AnglE - Angle Objective



(w1, w2, w3 are constants)

Results

Results - Best Performance

For now…

(%)	*Image Accuracy*	*Text Accuracy*	*Stacked Accuracy*
Baseline	84.4	71.7	**92.5**
Our Project	**85.0**	**72.3**	89.2

Still improving !
Expected to outperform baseline

Future Improvements & Conclusion

More Experiments

Given more time:

Outperform baseline in Stacked Accuracy

Extension to Image Captioning

Attempt: Fine-Tune a GIT Model on our own captions.

→ Overfit (scarce data) + Time Constraints

Possible applications: Add LLM for recipe generation !

(Zhu et a.l., 2023 - Stanford University “ChefNet” for Image Captioning & Recipe Matching)

References & GitHub Repository

1) Image and Text fusion for UPMC Food-101 using BERT and CNNs | IEEE Conference Publication
2) [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
3) Rethinking the Inception Architecture for Computer Vision (CVPR), 2016, pp. 2818-2826
4) [2309.12871] AnglE-optimized Text Embeddings
5) [2104.00298] EfficientNetV2: Smaller Models and Faster Training
6) ChefNet: Image Captioning and Recipe Matching on Food Image

GitHub Repository: https://github.com/dat-rohit/multimodal-model-Food101