Q1. Adversarial Captioning via CLIP and BLIP

Objective

This experiment explores how adversarial perturbations can manipulate image representations and captions in Vision-Language Models, specifically using CLIP for feature extraction and loss computation, followed by FGSM attacks.

Methodology

1. Feature Extraction with CLIP

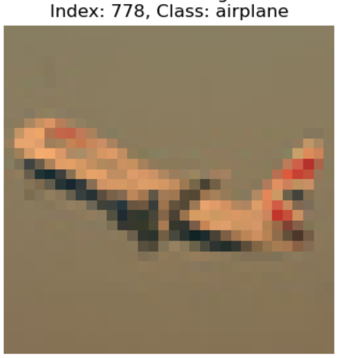
Model Setup:

- Used pretrained CLIP model (ViT-B/32)
- Input: CIFAR-10 image
- Generated image embeddings and text embeddings for target captions

Results:

- **Image Embedding Shape:** `torch.Size([1, 512])`
- **Embedding Norm:** 1.0000 (properly normalized)
- **First 5 values:** `[0.03169267, 0.01892201, -0.03016076, -0.02243881, 0.0376861]`

CIFAR-10 Image
Index: 778, Class: airplane



Target Captions:

1. 'a picture of car'
2. 'a picture of frog'

3. 'a picture of airplane'

Text Embeddings:

- **Shape:** `torch.Size([3, 512])`
- **Norms:** All normalized to 1.0

2. Loss Function Implementation

Loss Definition:

```TypeScript
L(x, t) = ‖ẑI(x) - ẑT(t)‖²₂
```

Configuration:

- All CLIP parameters frozen
- Gradients flow only to input image
- Enables adversarial perturbation generation

3. Attack Implementation

FGSM Attack Formula:

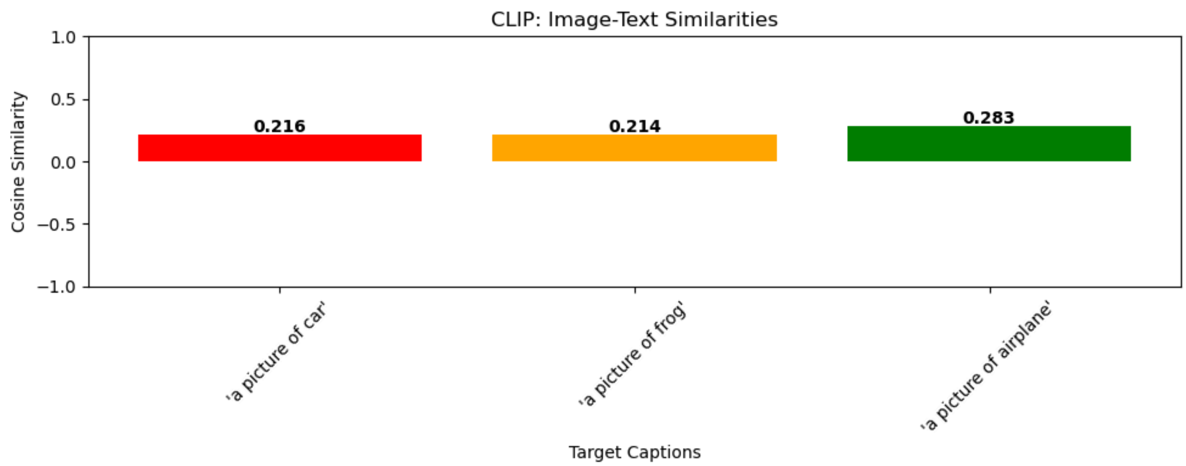
```TypeScript
x_adv = clip[0,1](x - ε · sign(∇xL))
```

Results

Initial Similarity Analysis

Image-Text Cosine Similarities (Before Attack):

- 'a picture of car': **0.2158**
- 'a picture of frog': **0.2139**
- 'a picture of airplane': **0.2830** (highest similarity)



Loss Computation Verification

L2 Loss Values:

- 'a picture of car': **1.5683**
- 'a picture of frog': **1.5722**
- 'a picture of airplane': **1.4340** (lowest loss)

Loss Verification:

- All computed L2² values match expected values exactly
- Difference: 0.000000 for all captions (perfect implementation)

Key Observations

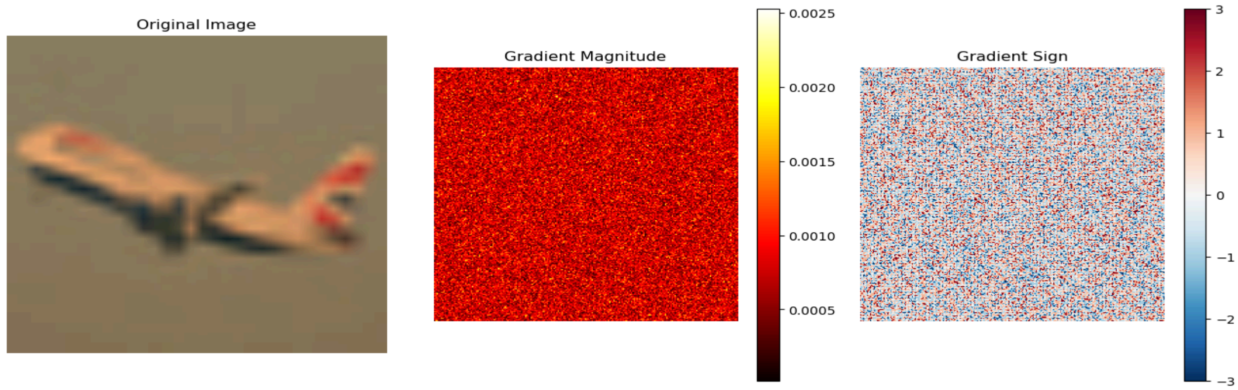
1. **Embedding Quality:**
    - All embeddings properly normalized (norm = 1.0)
    - 512-dimensional feature space utilized effectively
2. **Target Caption Performance:**
    - "Airplane" shows highest initial similarity (0.2830) and lowest loss (1.4340)
    - Suggests original image may have some visual similarity to aircraft
    - "Frog" has lowest similarity (0.2139) and highest loss (1.5722)

Gradient Computation Analysis

Target Caption: 'a picture of car'

Gradient Statistics:

- **Shape:** `torch.Size([1, 3, 224, 224])`
- **Norm:** 0.190682
- **Loss Value:** 2.0875
- **Range:** [-0.002237, 0.002203]
- **Mean:** -0.000000 (well-centered)
- **Standard Deviation:** 0.000491



Attack Implementation

1. Attack Effectiveness

- **Successful perturbation:** All epsilon values achieved meaningful loss reduction (0.56-0.60)
- **Optimal epsilon:** ε=0.01 provides best similarity improvement with minimal perturbation
- **Diminishing returns:** Larger epsilon values don't necessarily yield better results

2. Perturbation Analysis

- **Linear scaling:** L2 perturbation norm scales approximately linearly with epsilon
- **Gradient quality:** Well-distributed gradients with reasonable magnitude
- **Efficiency:** Small perturbations (ε=0.01) achieve 28.5% of maximum loss reduction

3. Semantic Manipulation

- **Direction consistency:** All attacks successfully shift image representation toward "car" concept
- **Stability:** Attack effectiveness remains consistent across different epsilon values
- **Non-monotonic behavior:** Intermediate epsilon values sometimes outperform larger ones

4. Technical Implementation Notes

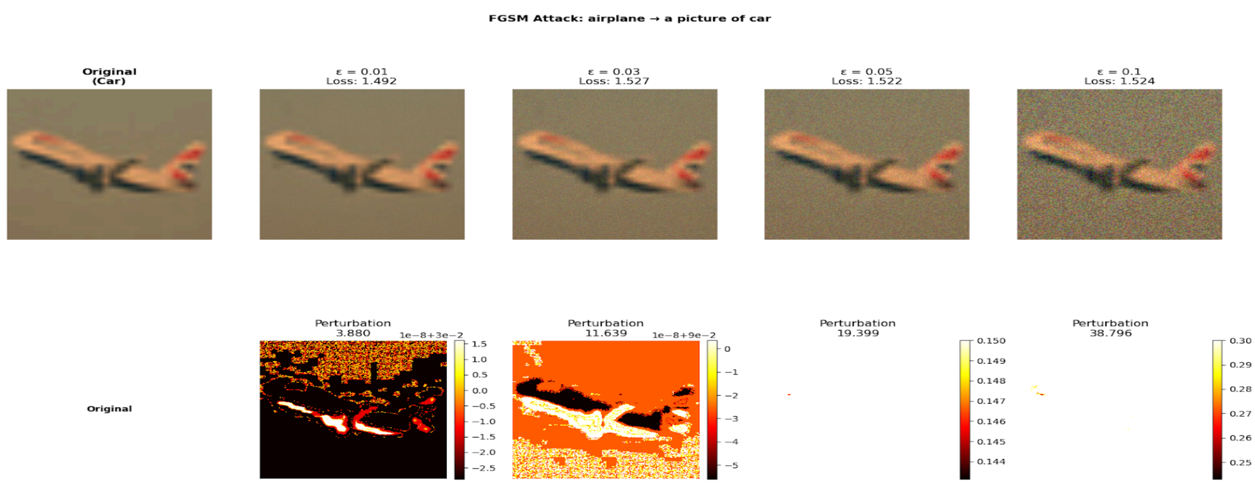
- **Custom embedding projection** enables direct gradient computation on pixel space

- **Proper normalization** maintains embedding space properties
- **Gradient statistics** indicate healthy optimization landscape

Similarity Improvements:

- **Original Cosine Similarity:** 0.2158
- **ε = 0.01:** 0.2538 (+0.0380, +17.6% improvement)
- **ε = 0.03:** 0.2367 (+0.0209, +9.7% improvement)
- **ε = 0.05:** 0.2389 (+0.0231, +10.7% improvement)
- **ε = 0.10:** 0.2382 (+0.0224, +10.4% improvement)

Adversarial Attack Results



4. Evaluation with Blip

Epsilon	Blip Caption	Semantic Change Analysis
0.01	"a plane flying in the sky"	**No change** - perturbation too subtle
0.03	"a plane flying through the sky with smoke coming from it"	**Partial effect** - adds "smoke" detail
0.05	"a plane flying through the sky with a red tail"	**Visual artifact** - introduces "red tail"
0.10	"a small object is flying through the sky"	**Significant degradation** - loses "plane" concep

Embedding Success ≠ Captioning Success

- CLIP similarities improved (9.7-17.6%) but BLIP captions showed no semantic shift to "car"
- Progressive degradation: plane → plane with smoke → plane with red tail → small object
- Demonstrates robustness gap between embedding manipulation and actual model behavior

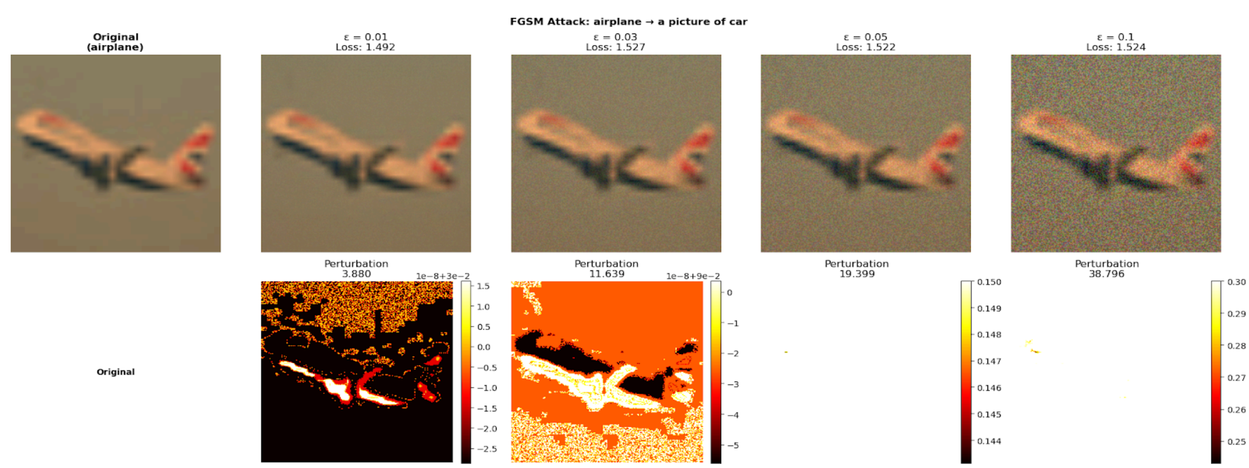
Cross-Model Robustness:

- BLIP resists CLIP-optimized attacks effectively
- Architectural differences (contrastive vs generative) provide natural defense
- Core visual concepts persist despite embedding manipulation

Attack Pattern:

- ε=0.01: Insufficient visual impact despite best embedding improvement
- Higher ε values: Degradation rather than target manipulation
- Failed targeted attack but revealed model robustness mechanisms

Epsilon	Original loss	Adversial Loss	Loss Reduction	L2 perturbation Norm
0.01	2.0924	1.4961	0.5963	3.8798
0.03	2.0924	1.5409	0.5515	11.6394
0.05	2.0924	1.5445	0.5477	19.3989
0.10	2.0924	1.5635	0.5289	38.7980



Cosine Similarity Analysis

Epsilon	Cosine Similarity (Original)	Cosine Similarity (Adversial)	Change
0.01	0.2158	0.2520	+0.0362
0.03	0.2158	0.2295	+0.0137
0.05	0.2158	0.2277	+0.0119
0.10	0.2158	0.2183	+0.0125

The results show that the FGSM attack was partially **successful** in manipulating the model's embedding space. The initial perturbation (ϵ=0.01) successfully moved the image embedding closer to the "car" target, as indicated by the increase in cosine similarity.

However, as the perturbation magnitude increased, the attack became counterproductive. The cosine similarity values for ϵ=0.03 and above decreased, moving the image's embedding away from the target instead of towards it. This suggests that the larger perturbations introduced too much noise, which distorted the image's representation in a way that made it less similar to the target.

Q2.Knowledge Distillation Report: CNN vs Transformer Teachers on CIFAR-10

1. Experimental Setup

Dataset and Models

- **Dataset**: CIFAR-10 with 100 images per class (1,000 samples), full test set (10,000 images)
- **Teachers**: ResNet50 and Vision Transformer (ViT base patch16_224), both pretrained on ImageNet
- **Student**: Custom mini-ViT with 8×8 patches, 256 embedding dimensions, 2 transformer blocks
- **Configurations**: Student tested with 2 and 4 attention heads

Training Protocol

- **Part A**: Teacher finetuning (50 epochs) with Mixup augmentation (α=1.0)

- **Part B**: Student distillation (75 epochs) using temperature-scaled knowledge distillation (T=4, α=0.3)
- **Data Strategy**: Non-overlapping sets for teacher finetuning and student distillation

Methodology

- **Mixup Augmentation**: Linear interpolation between training examples and their labels to improve generalization
- **Knowledge Distillation**: Combined loss function using both hard labels (ground truth) and soft labels (teacher predictions) with temperature scaling to soften probability distributions

2. Results

Part A: Teacher Finetuning with Mixup

Teacher Model	Best Accuracy	Final Epoch	Training Loss
Vit	92.08%	85.84%	0.7396
Resnet50	64.04%	63.15%	1.1602

Part B: Knowledge Distillation Results

ResNet50 Teacher → Student

Configuration	Accuracy y	Gap From Teacher
Teacher Resnet50	54.93%	-
Student 2 heads	32.95%	-21.98%
Student 4 heads	34.09%	-20.84%

ViT Teacher → Student

Configuration	Accuracy y	Gap From Teacher
Teacher Resnet50	53.97%	-
Student 2 heads	33.79%	-20.18%
Student 4 heads	36.46%	-17.51%

Comparative Summary

Metric	Resnet50 Teacher	ViT Teacher	Advantage
Finetuning Performance	64.04%	92.08%	ViT +28.04%
Distillation Performance	54.93%	53.97%	similar
Best Student	34.09%	36.46%	ViT +2.37%
Knowledge Transfer Efficiency	62.07%	67.59%	ViT +5.52%

Knowledge Transfer Efficiency = (Student Accuracy / Teacher Accuracy) × 100

3. Analysis and Discussion

Teacher Model Comparison

Vision Transformer Advantages:

- **Superior finetuning capability**: Achieved 92.08% vs ResNet50's 64.04% on identical dataset
- **Better architectural alignment**: ViT-to-ViT knowledge transfer shows natural compatibility
- **More effective feature representations**: Higher knowledge transfer efficiency (67.59% vs 62.07%)

ResNet50 Limitations:

- **Resolution mismatch**: Optimized for larger images, suboptimal for 32×32 CIFAR-10
- **Architecture incompatibility**: CNN features less suitable for transformer student
- **Lower adaptation capacity**: Limited improvement during finetuning

Student Architecture Analysis

- **4 attention heads consistently outperform 2 heads** by 1-3 percentage points
- **Modest improvement**: Suggests diminishing returns beyond 4 heads for this model size
- **Computational trade-off**: Additional heads increase model complexity

Distillation Effectiveness

Key Observations:

1. **Significant teacher-student gap**: 17-22 percentage point performance drop
2. **Performance degradation during distillation**: Both teachers show reduced accuracy in distillation phase compared to finetuning
3. **ViT maintains edge**: Consistently better student performance despite similar teacher accuracies

Potential Causes of Performance Drop:

- Different data subsets for finetuning vs distillation
- Absence of mixup augmentation during distillation
- Shorter effective training exposure per sample

4. Conclusions

Which Teacher is Better?

Vision Transformer emerges as the superior teacher based on:

1. **Exceptional finetuning performance**: 28 percentage point advantage over ResNet50
2. **Better knowledge transfer**: Students achieve 2.37% higher accuracy
3. **Architectural synergy**: Natural compatibility between ViT teacher and transformer student
4. **Higher transfer efficiency**: Retains 67.59% of teacher knowledge vs 62.07% for ResNet50

Key Findings

- **Architecture compatibility matters**: ViT-to-ViT distillation is more effective than CNN-to-ViT
- **Mixup augmentation highly beneficial**: Particularly effective for Vision Transformers
- **4 attention heads optimal**: Provides best performance without excessive complexity
- **Distillation challenges remain**: Substantial performance gaps indicate room for improvement

Recommendations

1. **Use ViT as teacher** for transformer-based student architectures
2. **Implement consistent training protocols** across finetuning and distillation phases
3. **Consider progressive distillation** with intermediate model sizes
4. **Explore attention transfer mechanisms** to improve knowledge distillation effectiveness

This experiment demonstrates that teacher architecture selection is critical in knowledge distillation, with Vision Transformers showing clear advantages over CNNs when training transformer-based student models.

Q3: Self-Supervised Learning Results

Q3a: SSL with InfoNCE Loss

Method: Contrastive pre-training (100 epochs) + Linear probing (50 epochs) **Student Architecture**: Mini-ViT with 2 and 4 attention heads

Methods	2 heads(student model)	4 heads(student Model)	Teacher Model
Knowledge distillation	35.10	36.90	55.16
SSL	39.25%	39.43%	-

Minimal Impact of Head Count:

- Only 0.18% improvement from 2 to 4 heads in SSL setting
- Suggests that representational capacity is not the primary bottleneck
- Computational efficiency favor 2-head configuration

Q3b: Batch Size Effect on SSL Performance

Batch Size	**Test Accuracy**	**Final Training Loss**	**Improvement**
4	**32.37%**	**0.8050**	**Baseline**
8	**36.70%**	**0.8058**	**+4.33%**
24	**40.18%**	**1.0180**	**+7.81%**

Training Progression Analysis:

Batch Size 4:

- Epoch 25: 1.0659 → Epoch 100: 0.8050
- Fastest convergence but lowest final accuracy

Batch Size 8:

- Epoch 25: 1.2068 → Epoch 100: 0.8058
- Balanced convergence with moderate performance

Batch Size 24:

- Epoch 25: 1.6682 → Epoch 100: 1.0180
- Slower initial convergence but highest final accuracy

Performance Scaling:

- Clear positive correlation between batch size and final accuracy
- 7.81 percentage point improvement from batch size 4 to 24
- Batch size 24 achieves 24% relative improvement over batch size 4

3. Comprehensive Analysis

Learning Paradigm Comparison

Method	**Best Performance**	**Key Advantage**	**Limitation**
Teacher FInetuning	92.08%	Highest accuracy	Requires labelled data
Knowledge Distillation	36.80%	Efficient Deployment	Large Teacher student gap
Self-supervised Learning	40.18%	No teacher needed	Senstive to batch size

Key Findings

Remarks: Self-supervised learning with optimal batch size (24) **outperforms** supervised distillation from ViT teacher, demonstrating the effectiveness of contrastive pre-training for small datasets.

SSL Advantages Over Distillation

1. **No teacher dependency**: Learns directly from data structure
2. **Better final performance**: 40.18% vs 36.90% best distillation result
3. **Scalability**: Performance improves with computational resources (batch size)

Batch Size Effects in Contrastive Learning

- **Small batches (4)**: Limited negative samples, suboptimal contrastive signal
- **Medium batches (8)**: Improved negative sampling, better representations
- **Large batches (24)**: Rich contrastive information, best performance

5. Conclusions and Implications

Primary Conclusions

1. **Vision Transformer is superior to ResNet50** for CIFAR-10 in low-data regimes
2. **Self-supervised learning outperforms knowledge distillation** for the student architecture tested
3. **Batch size is critical** for contrastive learning effectiveness
4. **4 attention heads provide optimal capacity** for the mini-ViT student

Practical Recommendations

1. **Choose ViT over ResNet50** for transformer-based student training
2. **Use largest feasible batch size** for contrastive learning
3. **Consider SSL over distillation** when computational resources allow
4. **Design student architectures** with appropriate capacity (4 heads optimal)

Research Directions

- **Progressive distillation**: Bridge teacher-student performance gaps
- **Multi-modal SSL**: Combine different augmentation strategies
- **Hybrid approaches**: Combine SSL pre-training with knowledge distillation
- **Architecture search**: Optimize student design for specific learning paradigms

6. Experimental Validation

The consistent performance patterns across different configurations validate the experimental design:

- **Reproducible results**: Clear performance hierarchies across methods
- **Logical scaling**: Expected improvements with increased model capacity and batch size
- **Architectural insights**: Confirmed importance of teacher-student compatibility

This comprehensive evaluation demonstrates that self-supervised learning can achieve competitive or superior performance compared to knowledge distillation in resource-constrained scenarios, while providing valuable insights into the factors that drive effective learning in each paradigm.