Vision-Language Multimodal Fusion in Dermatological Disease Classification

Moreno La Quatra, Nicole Dalia Cilia, Vincenzo Conti,
Salvatore Sorce, Giovanni Garraffa, and Valerio Mario Salerno

Kore University of Enna, Italy

The Challenge in Dermatological Diagnosis

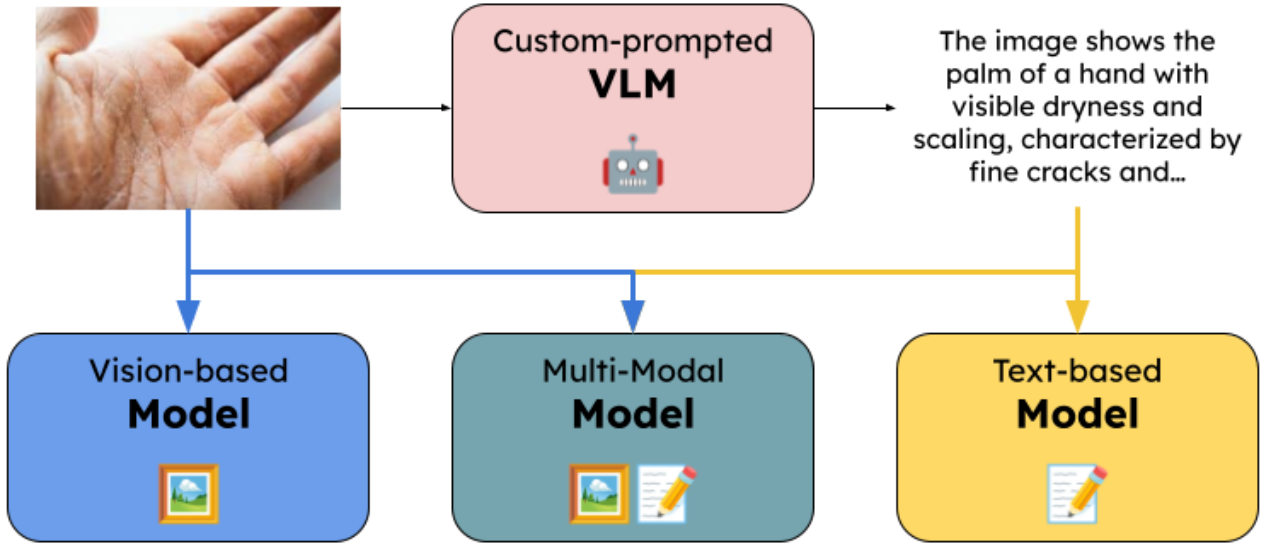
Current challenges in dermatological diagnosis:

- 🏥 **Expert Dependency**: Heavy reliance on specialist dermatologists
- ⏳ **Resource Intensive**: Time-consuming manual analysis
- 🔄 **Inconsistency**: Variation in expert interpretations
- 📊 **Scalability**: Limited ability to handle increasing cases

Our Solution

We propose a **multimodal approach** that:

1. Leverages advanced Vision-Language Models for expert-like textual descriptions
2. Combines visual and textual information
3. Enhances classification accuracy



Key Contributions

1. 🔄 **Novel Multimodal Framework**
    - Integration of visual data with AI-generated descriptions
    - Specialized for dermatological diagnosis
2. 🔍 **Advanced Fusion Strategies**
    - Attention Pooling
    - Gating Mechanism
    - Dual Feature-wise Linear Modulation
3. 📊 **Open Resource**
    - Dataset with AI-generated annotations

System Architecture

Key components:

- Vision Backbone (Vision Transformer)
- Text Backbone (Text-based Transformer Encoder)
- Fusion Network
- Classification Head

System Architecture

Vision Backbone:

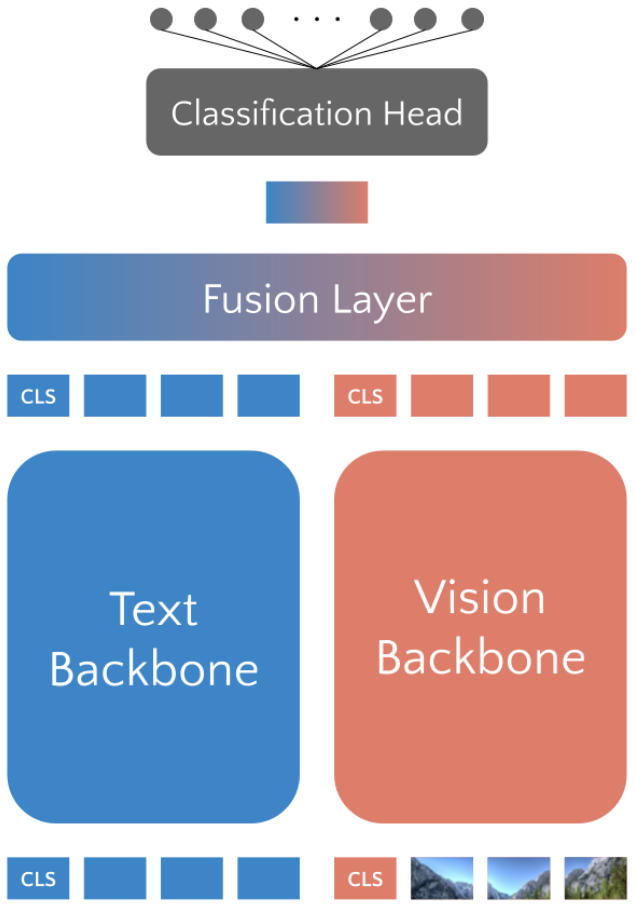
- Analyzes visual patches
- Extracts image embeddings

Text Backbone:

- Analyzes synthetic medical reports
- Generates text embeddings

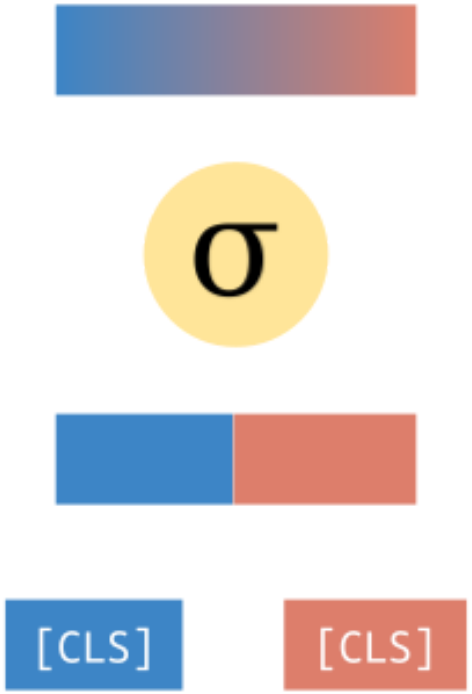
Fusion Network & Classification Head:

- Combines visual and textual features
- Predicts skin condition class



Gating Mechanism

- Controls **information flow** between modalities
- Uses sigmoid activation for feature selection
- Filters out irrelevant information



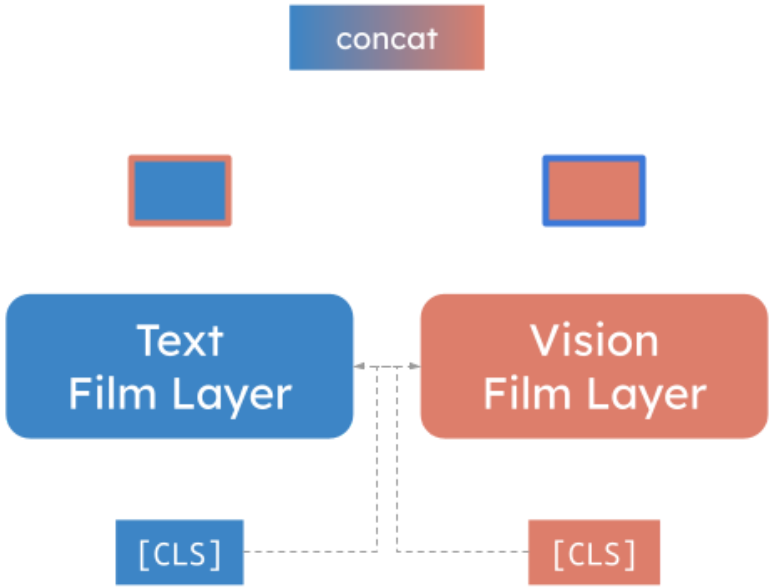
Dual FiLM Fusion

- **Bidirectional** feature modulation
- Each modality *influences* the other
- Applies scale and shift operations

$$T_{mod} = T_{CLS} \odot (1 + \gamma_{v \to t}) + \beta_{v \to t}$$

- γ and β are scale and shift parameters, one for each direction (visual to text and text to visual)

It aims at modeling how visual information influences textual information and vice versa.



Attention Pooling Fusion

- Dynamically weighs importance of input
- Processes visual and textual sequences separately
- Applies weighted attention to both modalities

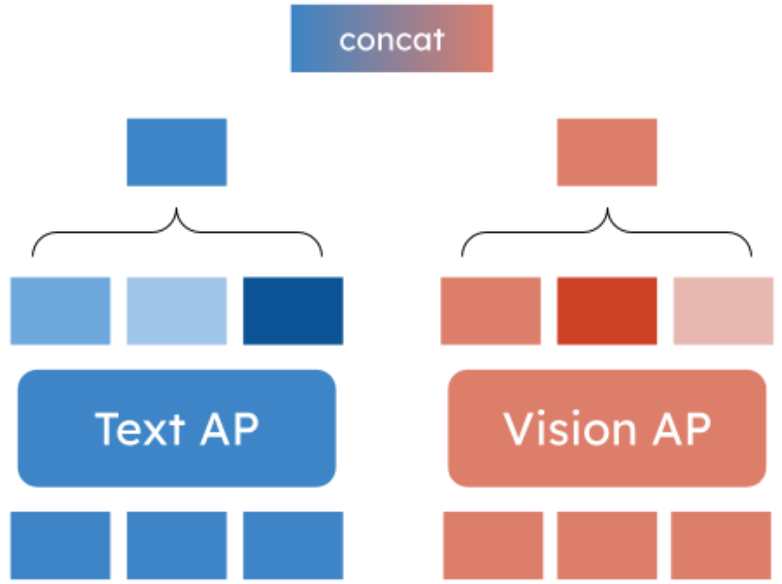
$$W_v = \text{softmax}(VW_v^a)$$

$$V_p = W_v^T V$$

$$W_t = \text{softmax}(TW_t^a)$$

$$T_p = W_t^T T$$

$$F = [V_p; T_p]$$



Dataset: DermNet Enhanced

Dataset Statistics:

```
Training:   15,557 images
Testing:    4,002 images
Classes:    23 skin conditions
```

Enhancement:

- AI-generated medical descriptions
- Standardized annotation format
- Balanced class distribution

Text Generation Process

Using InternVL Model:

Input prompt:

```
"You are a doctor. Please describe the image from
a medical perspective in an objective manner..."
```

Generated description example:

```
"The image shows a well-circumscribed lesion with
irregular borders and varying pigmentation..."
```

Experimental Results

Key Findings:

- Attention Pooling consistently performs best
- Visual information dominates classification
- Text can provides complementary context when effectively combined

Model	Training	Accuracy	F1-Score
BERT 📝	Fine-tuned	35.66%	0.2992
ViT 🖼️	Fine-tuned	70.59%	0.6648
Concat 🖼️ + 📝	Fine-tuned	70.44%	0.6696
Gating 🖼️ + 📝	Fine-tuned	70.39%	0.6671
DualFiLM 🖼️ + 📝	Fine-tuned	70.41%	0.6702
Att. Pool. 🖼️ + 📝	Fine-tuned	**71.31%**	**0.6794**

Impact Analysis

1. **Clinical Applications**
    - Enhanced diagnostic support
    - Reduced dependency on specialists
    - Standardized assessment process
2. **Research Contributions**
    - Open-source dataset
    - Reproducible methodology
    - Framework for future studies

Future Directions

Short-term Goals:

- Larger scale validation
- Clinical environment testing
- Enhanced VLM integration

Long-term Vision:

- Extension to other medical domains
- Real-time diagnostic support
- Integration with healthcare systems

Conclusions

Our research demonstrates:

1. **Effectiveness** of multimodal fusion in medical imaging
2. **Value** of AI-generated medical descriptions
3. **Superiority** of Attention Pooling for feature fusion
4. **Potential** for practical clinical applications

Thank You!

Contact Information:

✉️ Email: {name.surname}@unikore.it

💻 Project repository: github.com/MorenoLaQuatra/vl-dermnet-annotations

Acknowledgments:

D.A.R.E. "DigitAI lifelong pRevEntion" project

LifeMap project