# Encoders visuais e textuais
## *Encoders e espaço latente multimodal*

Alysson Machado de Oliveira Barbosa

alysson.barbosa@ee.ufcg.edu.br

Universidade Federal de Campina Grande – UFCG
Programa de Pós-Graduação em Engenharia Elétrica – PPgEE

**Orientadores:**

Prof. Dr. Wamberto J. L. de Queiroz

wamberto@dee.ufcg.edu.br

Profa. Dra. Luciana Ribeiro Veloso

luciana.veloso@dee.ufcg.edu.br

2026

# Sumário

1. Introdução
2. *Encoders* e representações
3. Alinhamento multimodal
4. Arquiteturas multimodais
5. Considerações finais
6. Referências

# Problema e objetivo

- Imagens, textos e outros sinais têm estruturas diferentes.
- Um *encoder* transforma cada entrada em representações numéricas que podem ser usadas em tarefas posteriores.
- A questão central é como aprender vetores comparáveis sem ignorar a natureza de cada modalidade.

$$x \xrightarrow{f_v} h_v \xrightarrow{g_v} z_v \qquad\qquad t \xrightarrow{f_t} h_t \xrightarrow{g_t} z_t$$

$f_v$ e $f_t$: *encoders*.
$g_v$ e $g_t$: pooling e projeção para o espaço de comparação.

# *Encoder* textual

- O tokenizador converte a frase em unidades discretas, como palavras ou subpalavras.
- *Embeddings* de token e posição formam a entrada do Transformer, a autoatenção contextualiza cada unidade.
- A saída pode ser uma sequência de vetores. Um *token* de resumo ou uma operação de *pooling* produz o vetor usado pela tarefa.

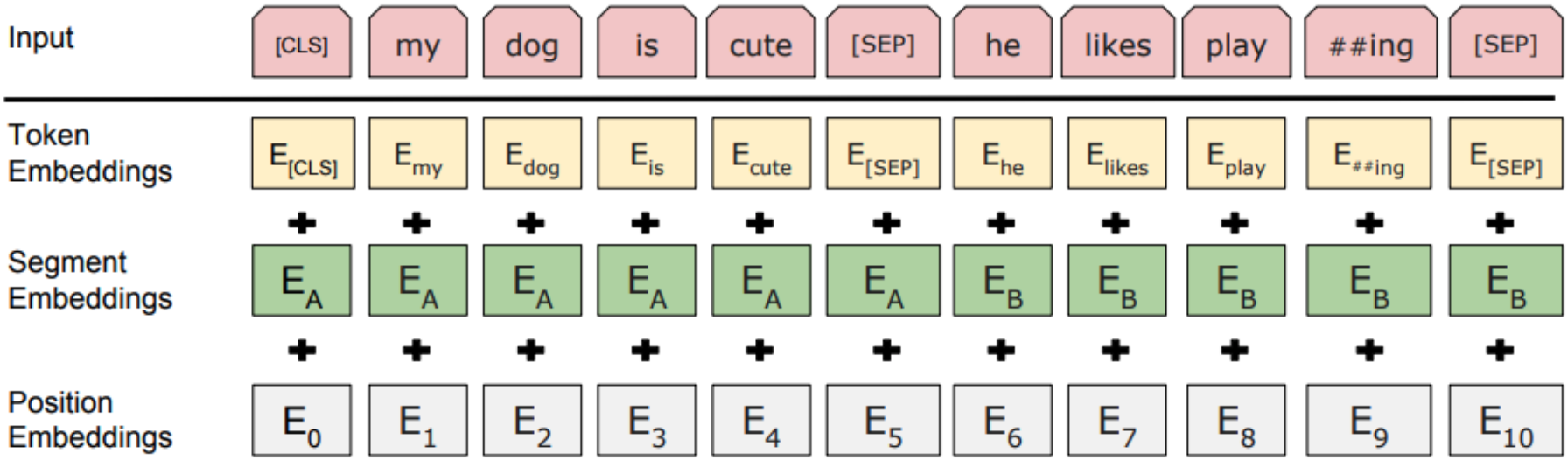



Figura 1: Representação de entrada no BERT.
Fonte: Devlin et al. (2019, fig. 2).

# *Encoder* visual

- Redes convolucionais extraem características de regiões, o ViT oferece outro caminho, tratando recortes da imagem como tokens.
- No ViT, cada *patch* é projetado em um vetor, recebe informação de posição e passa por camadas de Transformer.
- O vetor de saída é resumido ou projetado para a representação exigida pela tarefa.

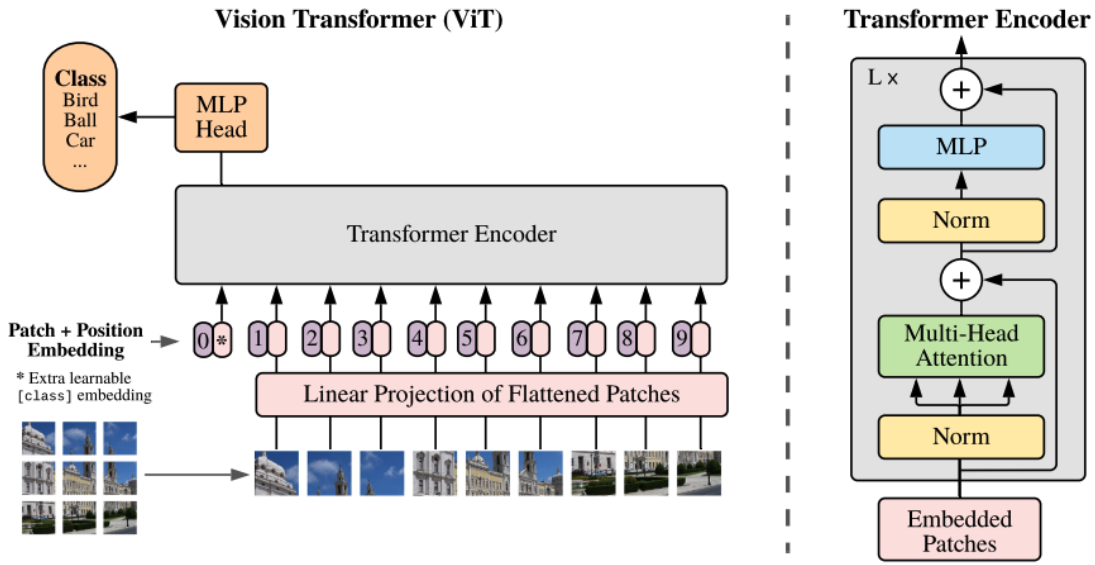



Figura 2: Divisão da imagem em *patches* e processamento pelo ViT.
Fonte: Dosovitskiy et al. (2021, fig. 1).

# Da representação ao *embedding*

- O *encoder* pode produzir vários estados: um por *token* textual ou por região da imagem.
- *Pooling* ou um *token* especial resume esses estados, uma projeção leva cada modalidade à dimensão compartilhada $d_e$.
- A normalização permite comparar os vetores pela similaridade do cosseno.

$$z_v = \frac{W_v h_v}{\|W_v h_v\|_2}, \qquad z_t = \frac{W_t h_t}{\|W_t h_t\|_2}$$

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figura 3: Projeção, normalização e similaridades.
Fonte: Radford et al. (2021, fig. 3).

# Dois *encoders* para imagem e texto

- Um *encoder* visual e um *encoder* textual processam as entradas separadamente.
- Durante o treino, o modelo aprende a atribuir alta similaridade a imagens e textos correspondentes.
- Em inferência, *embeddings* já calculados podem ser comparados sem executar os dois *encoders* em conjunto.

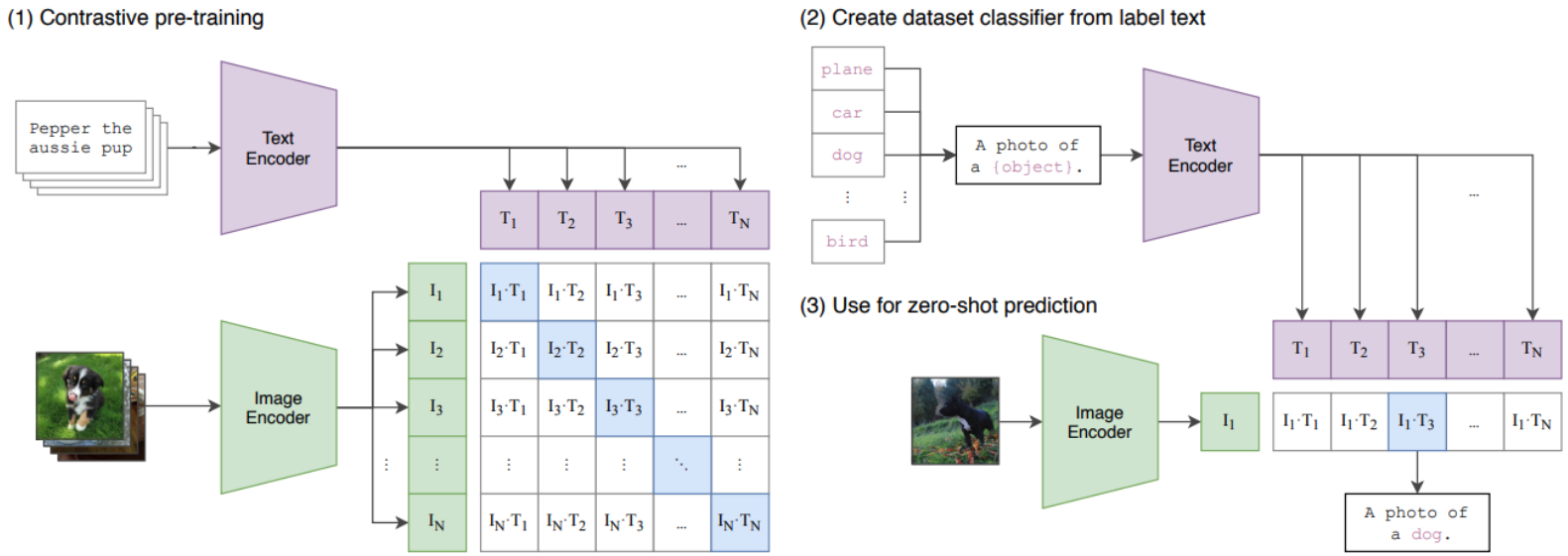



Figura 4: Treino contrastivo e classificação zero-shot no CLIP.
Fonte: Radford et al. (2021, fig. 1).

# Aprendizagem contrastiva

Considere $B$ pares de treino e a similaridade entre a imagem $i$ e o texto $j$:

$$s_{ij} = \frac{z_i^v \cdot z_j^t}{\tau}$$

- Os pares corretos ocupam a diagonal da matriz $B \times B$, os demais pares do lote servem como contrastes.
- A perda imagem $\rightarrow$ texto classifica o texto correto para cada imagem.

$$\mathcal{L}_{v \rightarrow t} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})}$$

- Uma perda simétrica faz também texto $\rightarrow$ imagem, a média das duas atualiza ambos os *encoders*.

(Radford et al., 2021)

# Geometria do espaço compartilhado

- Depois da normalização, o produto interno coincide com a similaridade do cosseno:

$$\text{sim}(z_v, z_t) = z_v^\top z_t = \cos(\theta)$$

- O treino aproxima os pares observados e separa alternativas dentro do lote.
- Assim, a vizinhança do espaço reflete os padrões úteis para o objetivo e presentes nos dados.
- Proximidade não garante que cada palavra esteja alinhada a uma região específica, nem que o vetor tenha um significado universal.

# Busca e classificação zero-shot

- Para busca, codifica-se a consulta e ordenam-se as imagens ou textos pelo cosseno entre *embeddings*.
- Para classificação zero-shot, cada classe é convertida em uma descrição textual candidata.

$$\hat{c} = \arg\max_{c \in \mathcal{C}} \ \mathrm{sim}\big(z_{\text{imagem}}, z_{\text{texto}(c)}\big)$$

- O classificador é formado pelas descrições das classes, não é necessário treinar uma nova cabeça para cada conjunto de rótulos.
- O resultado depende das descrições usadas e da cobertura dos conceitos nos dados de pré-treino.

(Radford et al., 2021)

# Embeddings globais e interação entre modalidades

**_Encoders_ separados**

- Cada modalidade é codificada pelo seu próprio *encoder*.
- A comparação ocorre entre vetores globais.
- É adequada para busca em coleções grandes.

**Fusão com atenção cruzada**

- *Tokens* visuais e textuais trocam informação dentro do modelo.
- A representação pode depender do par específico de entradas.
- É útil para perguntas e relações locais, com maior custo de inferência.

(Radford et al., 2021; Tan e Bansal (2019))

# Interação visual e textual no LXMERT

- A imagem é representada por regiões de objetos e a frase por *tokens*.
- *Encoders* específicos modelam cada modalidade, camadas de atenção cruzada conectam palavras e regiões.
- Esse desenho foi proposto para tarefas que exigem relações entre partes da pergunta e da imagem.

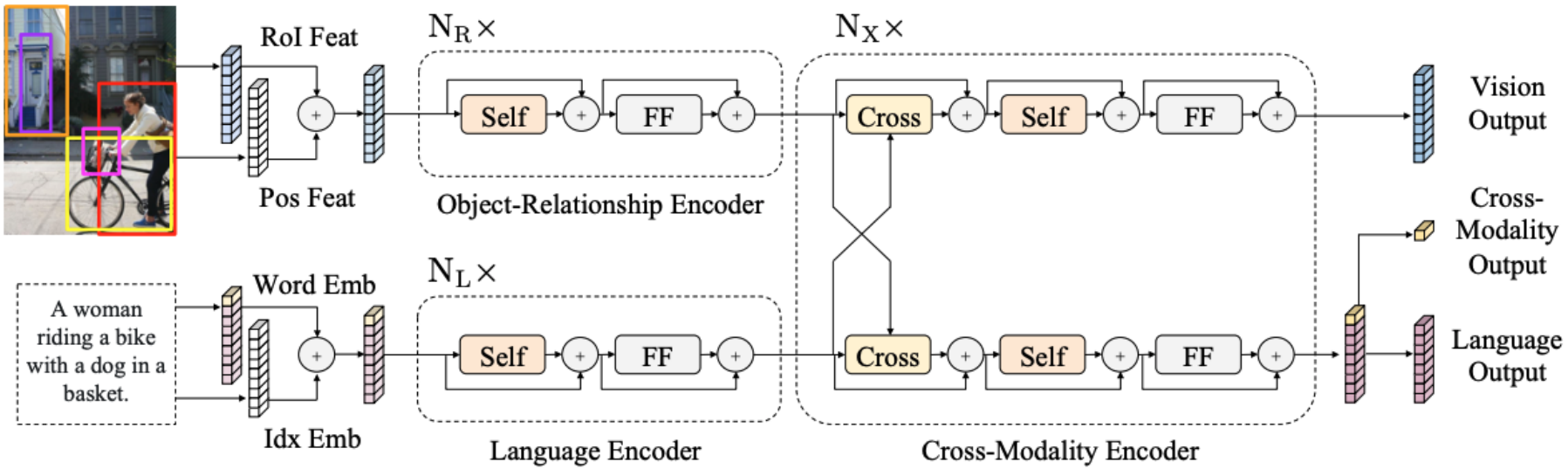



Figura 5: *Encoders* de visão, linguagem e *cross-modality*.
Fonte: Tan e Bansal (2019, fig. 1).

# ImageBind: modalidades ligadas por dados pareados

- O ImageBind reúne representações de imagem, texto, áudio, profundidade, dados térmicos e sinais inerciais (IMU).
- Em vez de observar todos os pares possíveis, o método usa associações naturais, como vídeo com áudio e imagem com profundidade.
- O alinhamento por uma modalidade comum pode produzir correspondências emergentes entre modalidades não pareadas diretamente.

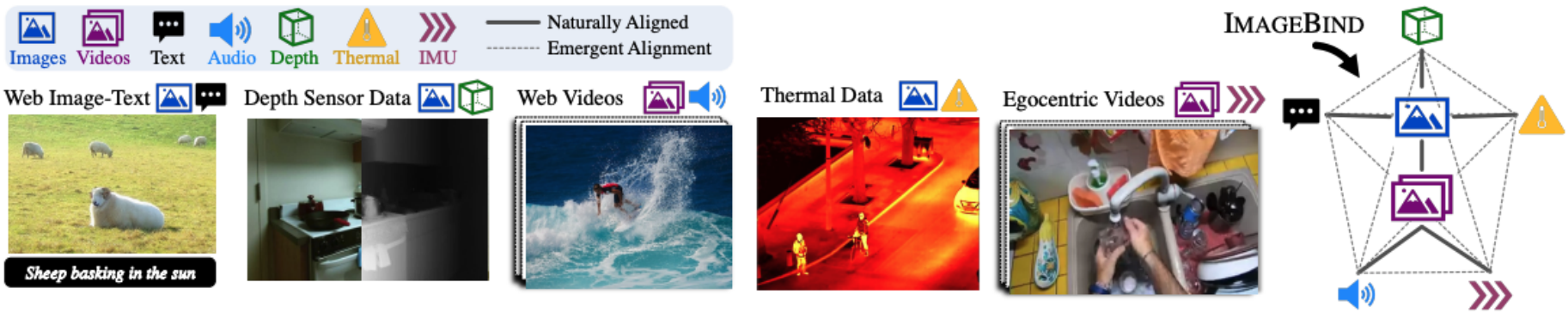



Figura 6: Dados pareados e alinhamento emergente no ImageBind.
Fonte: Girdhar et al. (2023, fig. 2).

# O que um espaço multimodal permite

- Recuperar itens de uma modalidade usando uma consulta de outra.
- Combinar *embeddings* para compor mais de um conceito.
- Usar uma modalidade como consulta em sistemas preparados para receber outra.

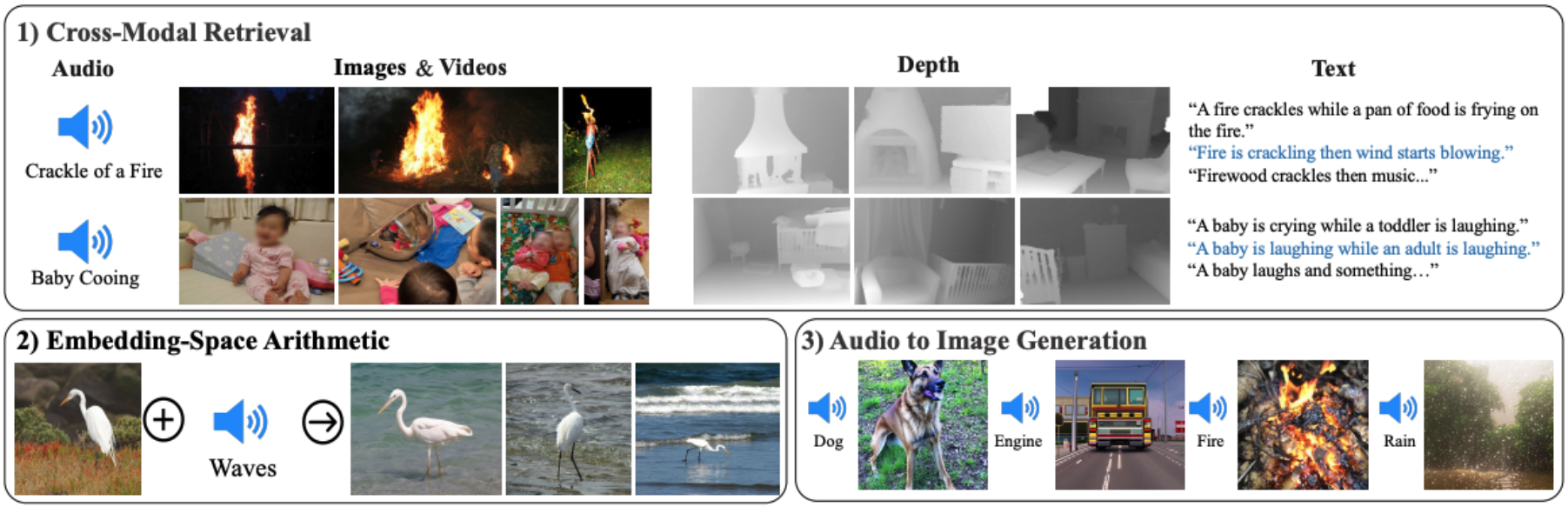



Figura 7: Recuperação, composição de *embeddings* e geração condicionada por áudio no ImageBind.
Fonte: Girdhar et al. (2023, fig. 1).

# O espaço reflete dados e objetivo

- Pares imagem-texto da web podem conter legendas incompletas, genéricas ou ruidosas.
- O modelo aproxima as associações vistas no treino, isso não equivale a uma descrição completa do conteúdo.
- Similaridade global não comprova localização correta, compreensão de relações ou raciocínio.
- Avalie cada capacidade com métricas próprias: recuperação, classificação zero-shot ou tarefas de interação visual e textual.

(Radford et al., 2021; Jia et al., 2021; Tan e Bansal, 2019)

# Síntese

- *Encoders* convertem entradas de modalidades distintas em representações numéricas.
- Dados pareados e objetivos contrastivos podem alinhar essas representações num espaço compartilhado.
- A similaridade facilita busca e classificação sem exemplos rotulados da tarefa, a fusão por atenção permite interações mais locais.
- O espaço aprendido depende das arquiteturas, dos dados e do objetivo de treinamento.

# Referências


Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). *Bert: Pre-training of deep bidirectional transformers for language understanding.* In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Dosovitskiy, A. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). *Learning transferable visual models from natural language supervision* In International conference on machine learning (pp. 8748-8763). PmLR.

Tan, H., & Bansal, M. (2019, November). *Lxmert: Learning cross-modality encoder representations from transformers.* In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 5100-5111).

# Referências




Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023, June). *ImageBind one embedding space to bind them all.* In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15180-15190). IEEE.



Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). *Scaling up visual and vision-language representation learning with noisy text supervision.* In International conference on machine learning (pp. 4904-4916). PMLR.

Obrigado!

alysson.barbosa@ee.ufcg.edu.br




Universidade Federal
de Campina Grande