Um modelo de rede neural siamesa para re-identificação de pessoas em imagens, utilizando rede neural convolucional e *autoencoder*

Fábia Isabella Pires Enembreck
Erikson Freitas de Morais

19 de setembro de 2019

Introdução

- Sistemas de segurança monitorados por câmeras;
  - Manual;
  - Inteligente;

- Re-identificação de pessoas;
  - Verificar se a pessoa já esteve presente em um ambiente;
  - Correspondência entre imagens;
  - Problema sem solução definitiva.

Introdução

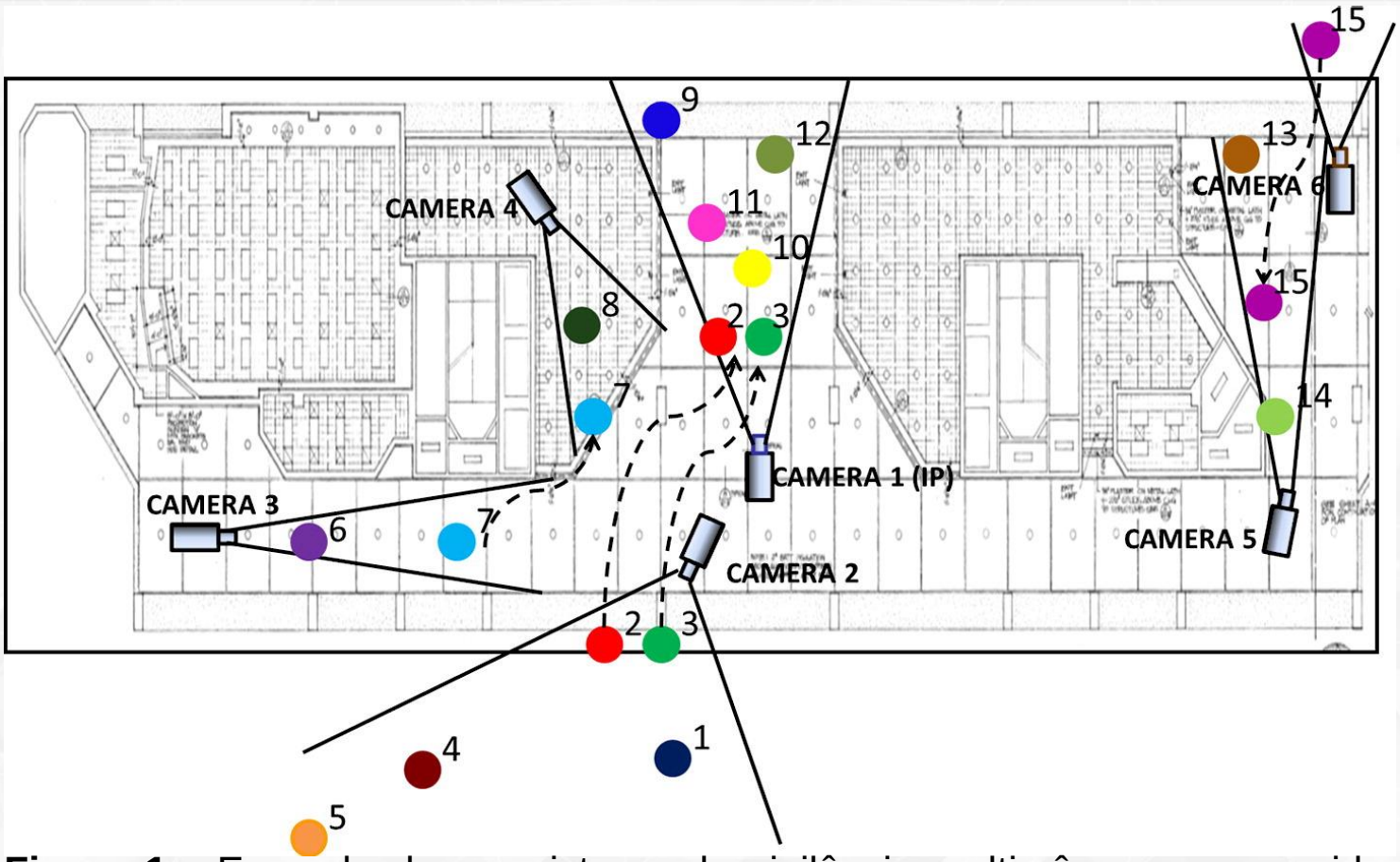


Figura 1: Exemplo de um sistema de vigilância multi-câmera para re-identificação. Adaptado de (BEDAGKAR-GALA; SHAH, 2014) .

Objetivo Principal

- Desenvolvimento de um método para re-identificar pessoas em imagens utilizando técnicas de aprendizagem profunda.

Objetivos Específicos

1) Identificar possíveis técnicas de aprendizagem profunda que possam ser aplicáveis ao problema;
2) Propor e implementar um modelo para re-identificação de pessoas;
3) Selecionar *datasets* públicos para testar o modelo proposto;
4) Realizar experimentos para validação do modelo.

Rede Proposta

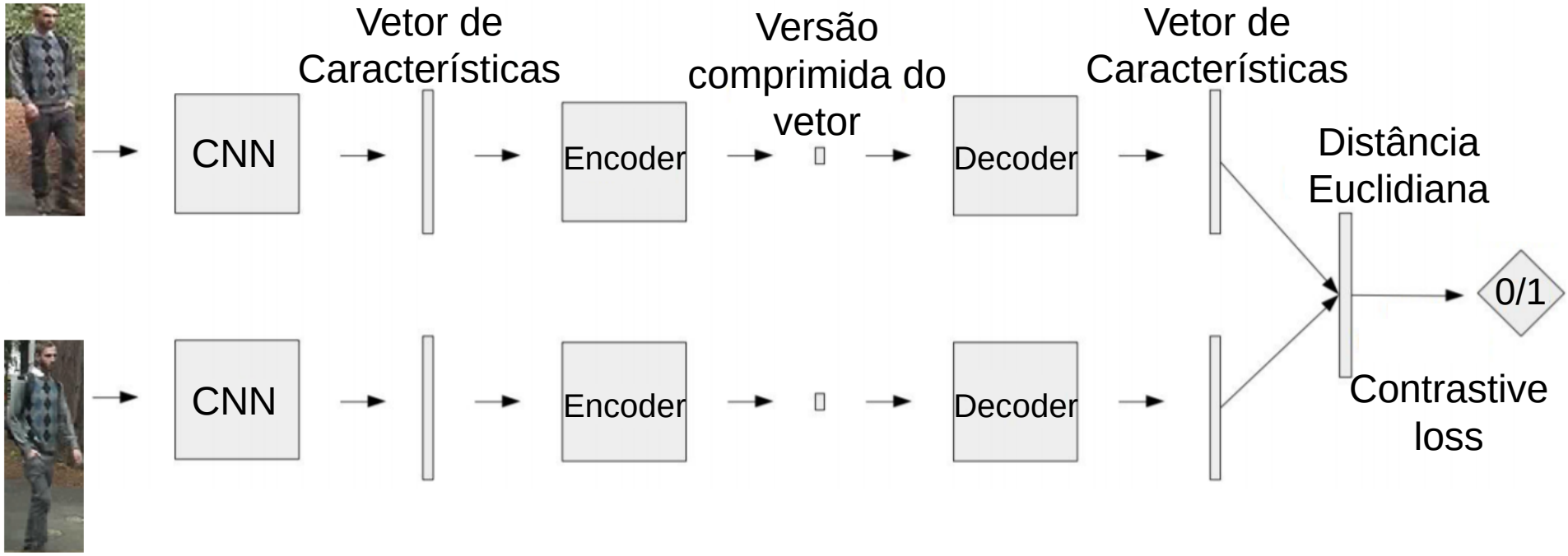


Figura 2: Modelo de Rede Neural proposta.

Rede Neural Convolucional

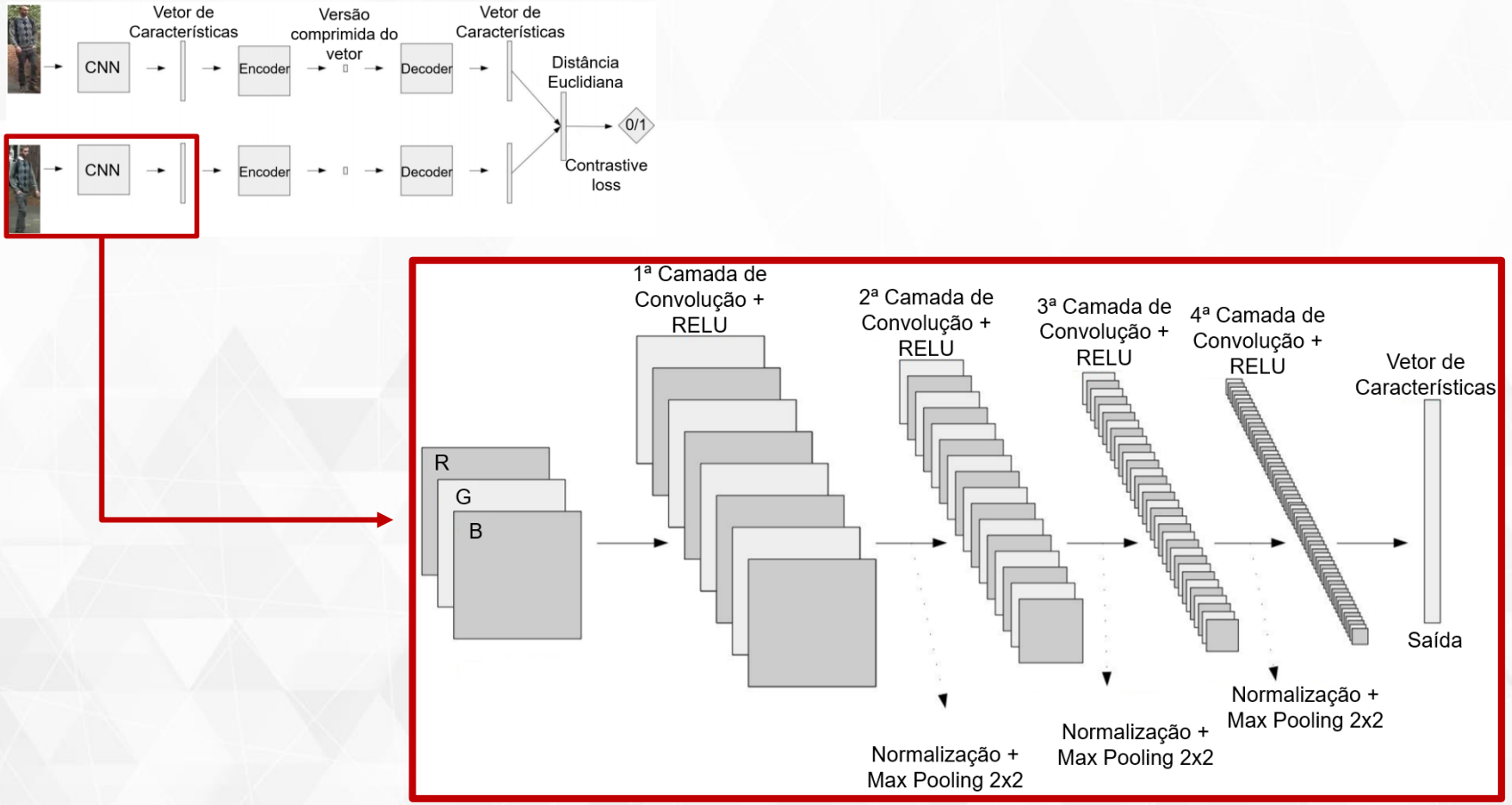


Figura 3: Modelo de Rede Neural proposta: CNN.

Rede Neural Convolucional

- Camada de Convolução;
  - Filtros;
  - Mapas de Características;
  - Aprender Padrões;
- Agrupamento;
  - Max Pooling.

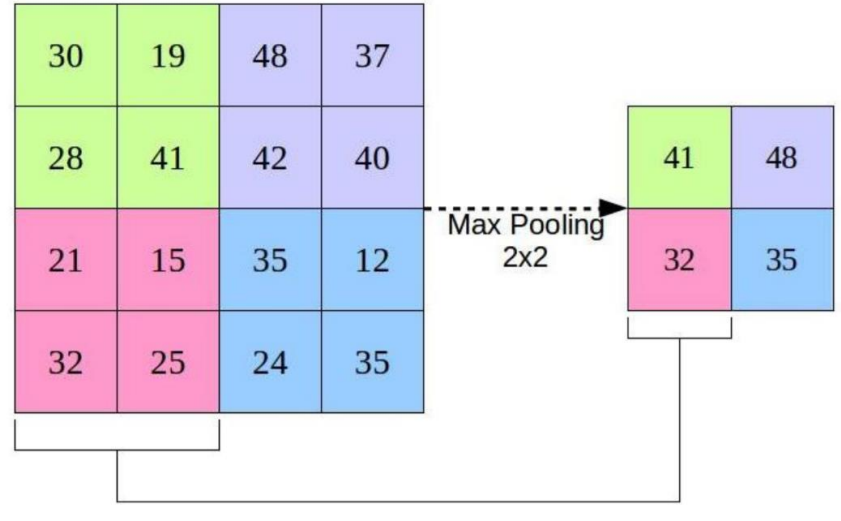


Figura 4: Operação de Max Pooling.

Rede Neural Convolucional

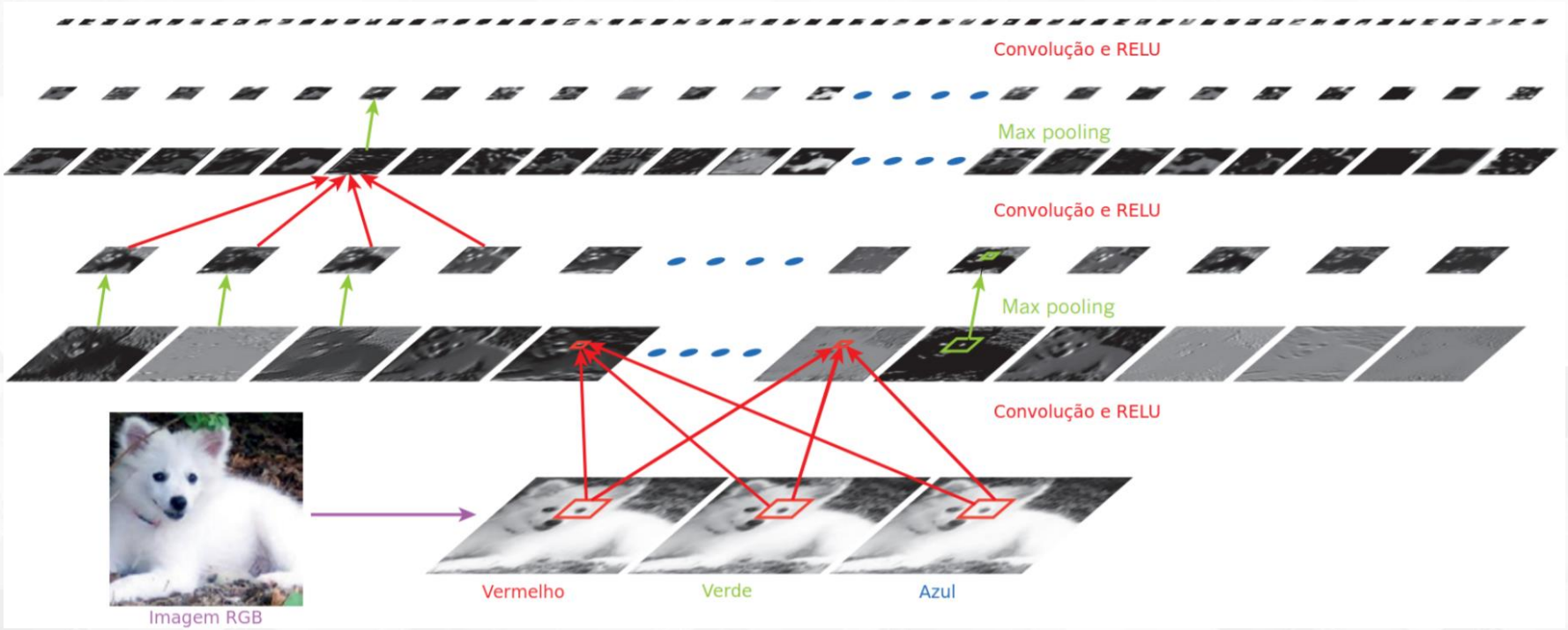
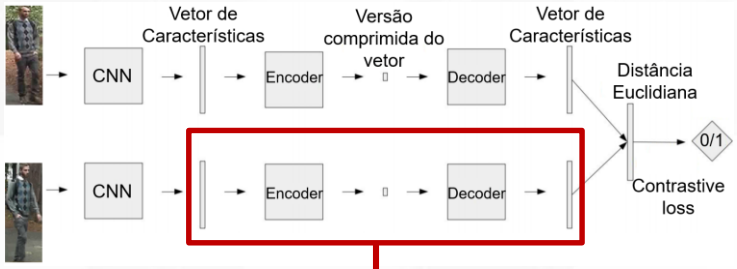


Figura 5: Camadas de uma CNN. Adaptado de (LECUN; BENGIO; HINTON, 2015).

Autoencoder



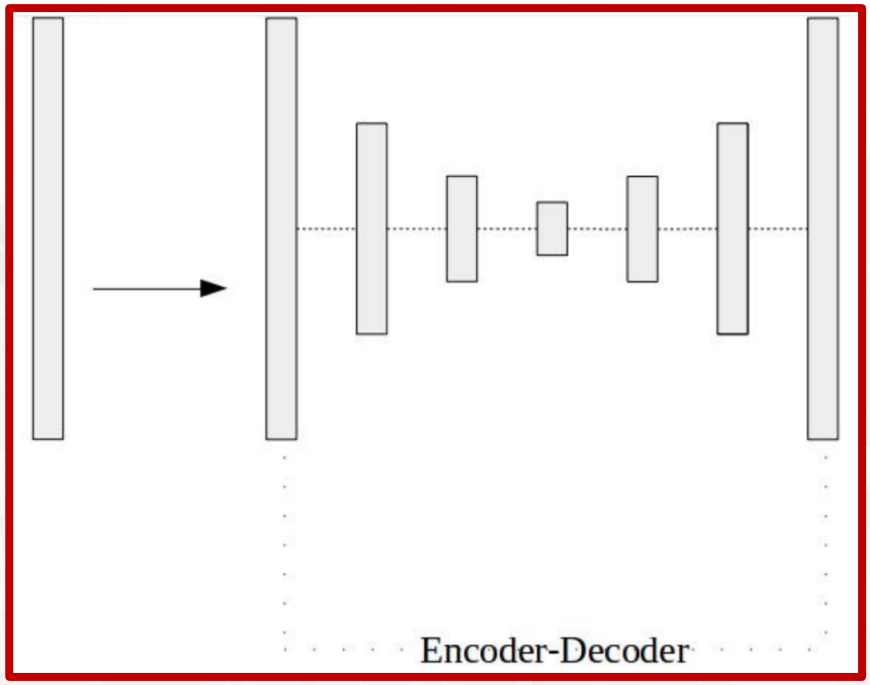


Figura 6: Modelo de Rede Neural proposta: *Autoencoder.*

Autoencoder

- Reproduzir os dados da entrada na saída da rede;
  - Função *encoder*;
  - Função *decoder*;
- Saída aproximada;
  - Priorizar as informações mais importantes;
- *Denoising Autoencoder*;
  - Entrada: dados corrompidos;
  - Saída: dados não corrompidos.

Rede Neural Siamesa

- Duas sub-redes idênticas unidas em suas saídas;
  - Encontrar correspondências;
- Entrada: um par de imagens e um rótulo binário;
- *Contrastive Loss.*

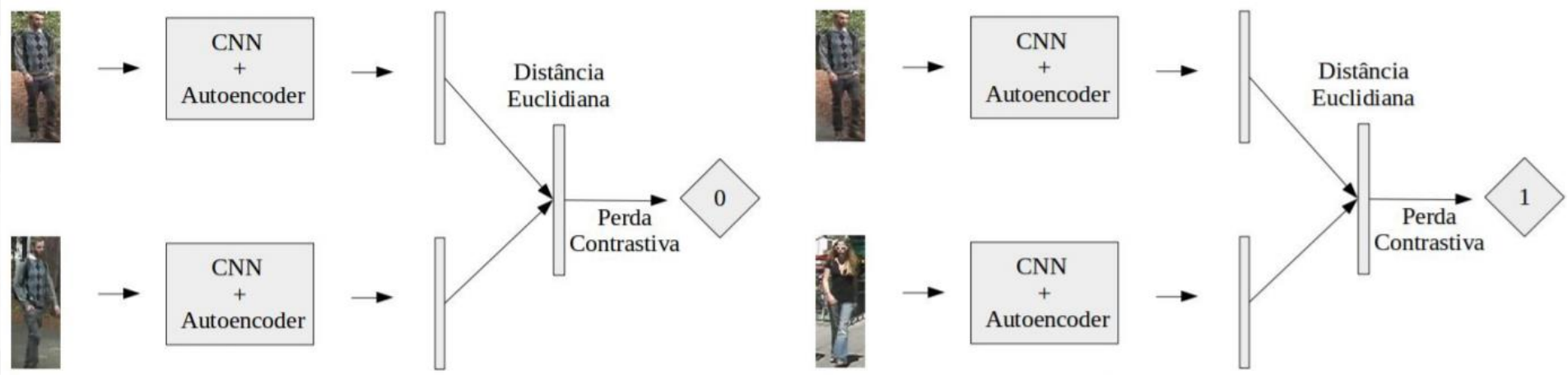
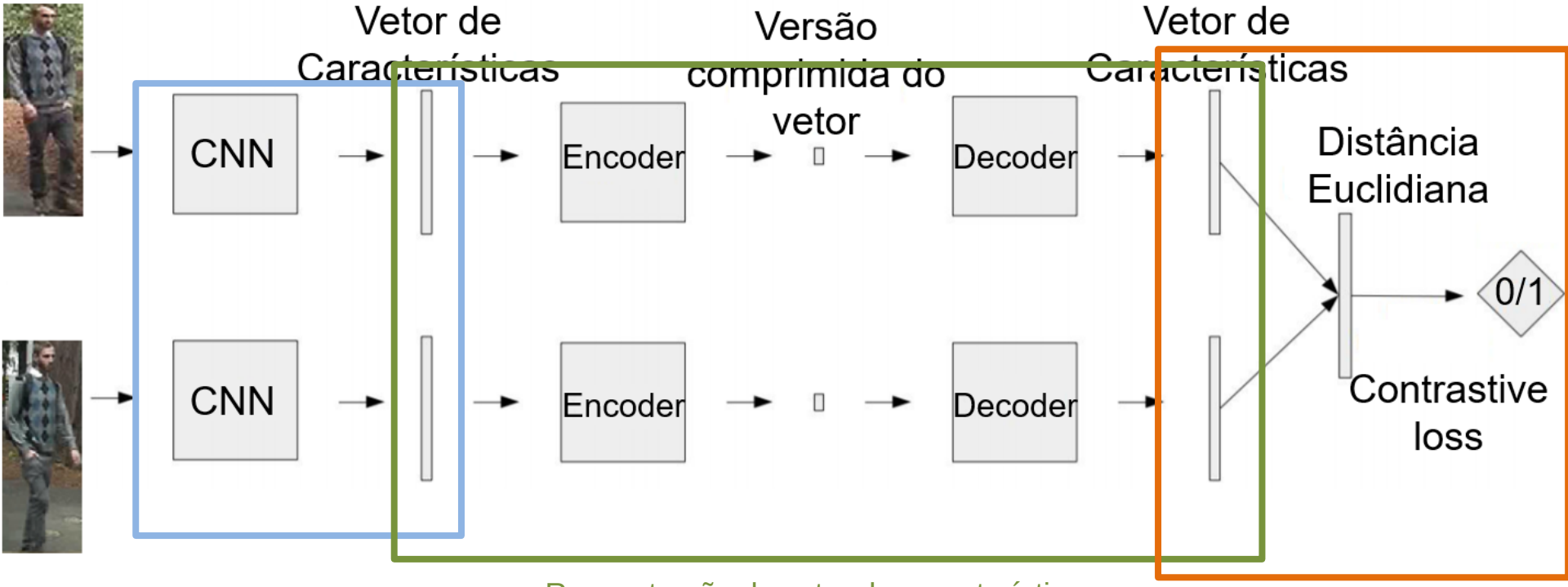


Figura 7: Associação entre pares positivos e pares negativos.

Rede proposta: resumo



Extração de características da imagem de entrada

Reconstrução do vetor de características, de forma que sejam amenizadas variações entre as imagens

Associação entre os dois vetores de características produzidos

Figura 8: Rede neural proposta.

Experimentos e Resultados

- ***Dataset:*** ViPER;
- 2 câmeras;
- 1264 imagens;
- 632 pessoas.

Figura 9: ViPER: exemplos de pares de imagens de 10 pedestres diferentes. Fonte: (GRAY; BRENNAN; TAO, 2007).

Experimentos e Resultados

Nº de Épocas	Acurácia(%)
100	84,47
200	82,92
400	82,92
600	65,2
800	80,23
1000	86,59
1200	87,93

Tabela 1: Acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* VIPeR.

Experimentos e Resultados

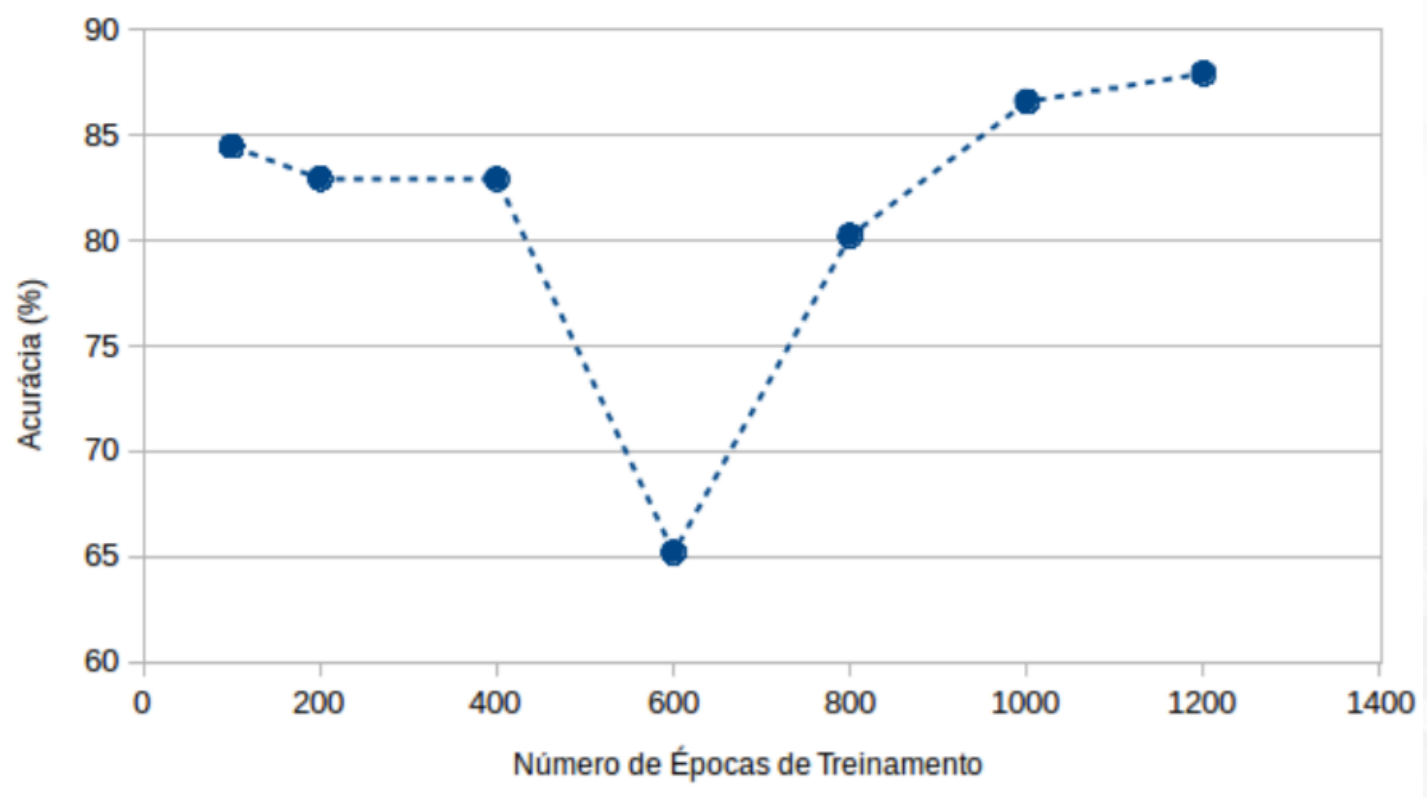


Figura 10: Acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* VIPeR.

Experimentos e Resultados

- ***Dataset:*** iLIDS-VID;
- 2 câmeras;
- 300 pedestres;
- *Single shot* e *Multi shot*.

Figura 11: iLIDS-VID: Imagens de um mesmo pedestre, obtidas por meio de duas câmeras não sobrepostas. Fonte: (WANG et al., 2014).

Experimentos e Resultados

Nº de Épocas	Acurácia(%)	Nº de Épocas	Acurácia(%)	Nº de Épocas	Acurácia(%)
100	93,04	1000	92,4	2000	93,22
200	93,49	1200	89,82	2200	92,06
400	91,2	1400	89,37	2400	92,46
600	94,26	1600	94,09	2600	93,46
800	91,39	1800	90,93	2800	93,82

Tabela 2: Acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* iLIDS-VID.

Experimentos e Resultados

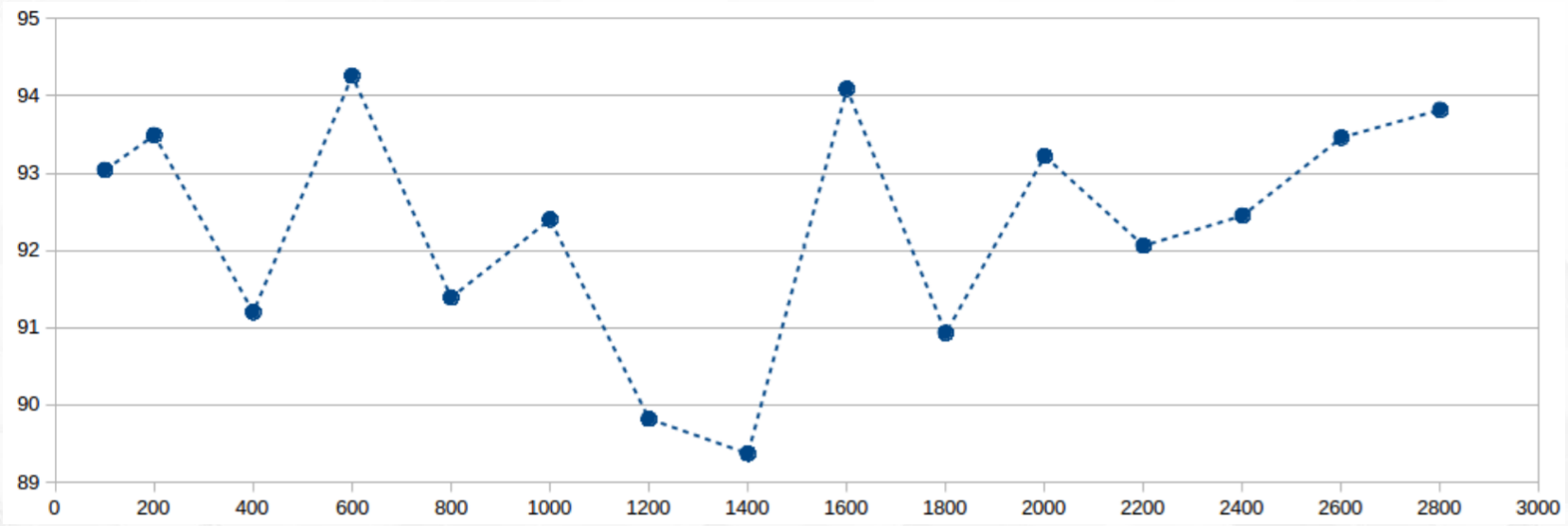


Figura 12: Acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* iLIDS-VID.

Experimentos e Resultados

- E se o *autoencoder* não fosse utilizado?

Nº de Épocas	Acurácia(%)	Nº de Épocas	Acurácia(%)	Nº de Épocas	Acurácia(%)
100	87,42	1000	87,30	2000	87,78
200	87,48	1200	87,43	2200	87,54
400	87,42	1400	87,25	2400	87,27
600	87,73	1600	87,33	2600	87,19
800	87,56	1800	87,51	2800	87,54

Tabela 3: Acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* iLIDS-VID.

Experimentos e Resultados

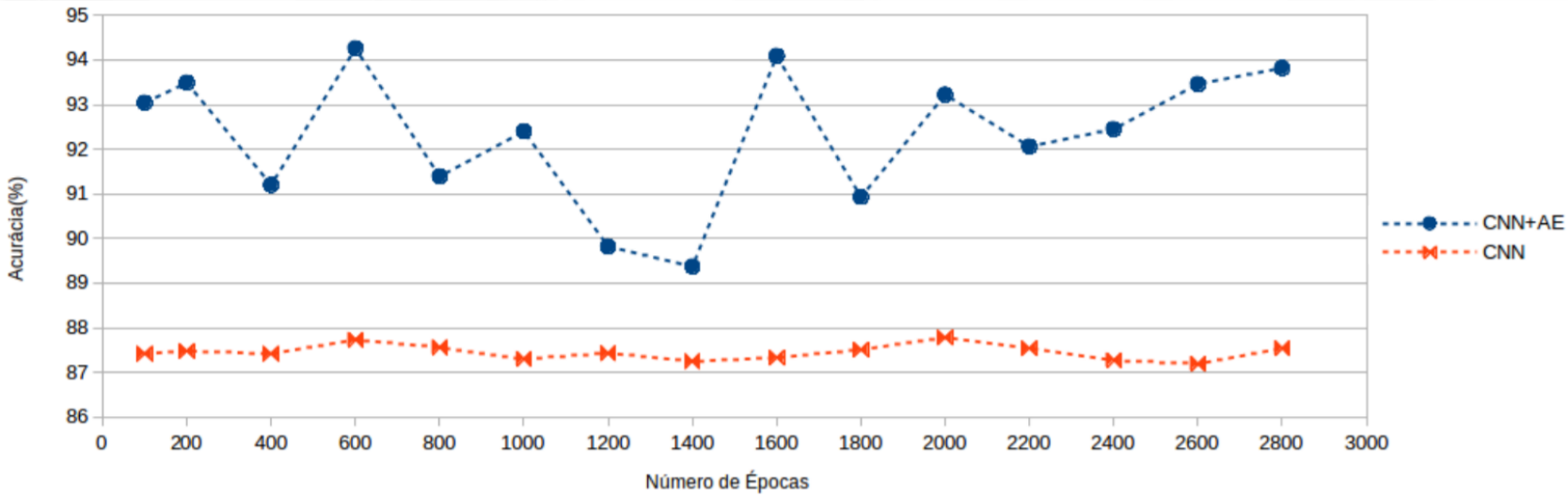


Figura 13: Comparação da acurácia em relação ao Nº de Épocas de treinamento, utilizando o *dataset* iLIDS-VID, para as duas redes implementadas.

Experimentos e Resultados

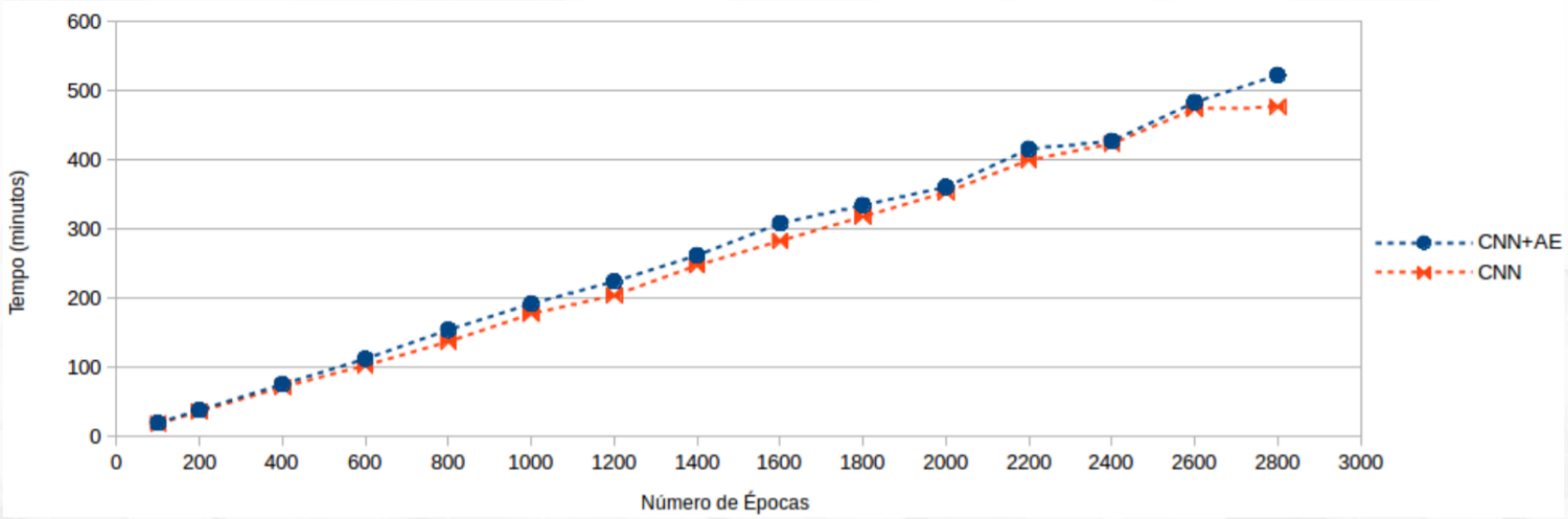


Figura 14: Tempo de execução de cada teste em minutos em relação ao N◦ de Épocas de treinamento, utilizando o dataset iLIDS-VID.

Conclusão e Trabalhos Futuros

- Potencial para re-identificar pessoas;
- CNN + AE;
  - Acurácia;
  - Tempo de execução;

- Realização de mais testes na rede, com outros *datasets;*
- Melhoria da rede;
- *Triplet loss.*

Referências

- BEDAGKAR-GALA, A.; SHAH, S. K. *A survey of approaches and trends in person re-identification*. Image and Vision Computing, Elsevier, v. 32, n. 4, p. 270–286, 2014.
- GRAY, D.; TAO, H. *Viewpoint invariant pedestrian recognition with an ensemble of localized features.* In: EUROPEAN CONFERENCE ON COMPUTER VISION. Proceedings... [S.l.], 2008. p. 262–275.
- LECUN, Y.; BENGIO, Y.; HINTON, G. *Deep learning*. nature, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.
- WANG, T. et al. *Person re-identification by video ranking.* In: EUROPEAN CONFERENCE ON COMPUTER VISION. Proceedings... [S.l.], 2014. p. 688–703.